Peer-Assessment: R Data Analysis
Do traffic signs reduce accidents?

Question 1.

The first step is to create a scatter plot to observe the general trend between the rate of car accidents and the number of signs. By typing in R:

```
plot(signs, rate) #creates scatterplot
cor(signs, rate)  #computes correlation
```

We get the following scatter plot:

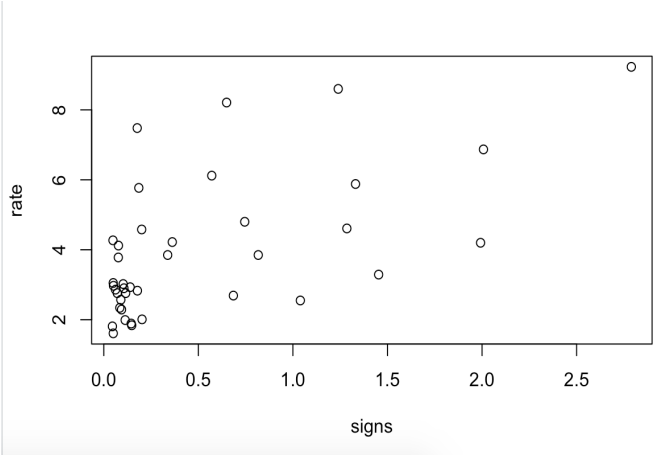

Figure 1: Rate of accidents vs Number of Signals

From the scatter plot we can see that there is a somewhat positive (maybe linear) relationship between the two variables. Moreover, the correlation coefficient between the two variables is 0.6031906, which supports our previous hypothesis that the variables are positively correlated. Thus, we can assume a linear relationship between the two variables. Therefore we will fit a simple linear regression model without transforming the data.

Response quality	Description	Points (out of 10)
Poor	The student does not answer any parts of the question correctly	0
OK	The student does 1 of the following correctly: produces the scatterplot, calculates the correlation coefficient, or concludes there is a linear relationship between the variables	4
Good	The student does 2 of the following correctly: produces the scatter plot, calculates the correlation coefficient, or concludes there is a linear relationship between the variables	7
Perfect	The student answers all parts of the question correctly	10

Question 2.

The model to be estimated is: $rate = \beta_0 + \beta_1 signs + \epsilon$. By using the `lm` function in R, we get the following output.

```
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

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
```

The estimate for the intercept is: $\hat{\beta}_0 = 3.0129$ and the estimate for the slope is: $\hat{\beta}_1 = 1.8023$ and the estimate for the error term is $\hat{\sigma^2} = 1.605^2$, or 2.5760. Thus, the equation for the least squares line is:

$$rate = 3.0129 + 1.8023 signs$$

The estimate for $\hat{\beta}_1$ means that an increase in one signal per mile of roadway increases the accident rate by 1.8023 units with a standard error of 0.3918.

An easy way to determine if the slope is statistically significant is to compute its 95% confidence interval. By typing in R:

```
confint(model, level=.95)
```

The result from the 95% CI is (1.008440, 2.596106). Since 0 is not included in the CI we can conclude that the explanatory variable $signs$ is statistically significant at the significance level $\alpha = 0.05$.

Response quality	Description	Points (out of 10)
Poor	The student does not answer any parts of the question correctly	0
OK	The student answers 1 of the 4 parts of the question correctly	2
Good	The student answers 2 of the 4 parts of the question correctly	5
Great	The student answers 3 of the 4 parts of the question correctly	7
Perfect	The student answers all parts of the question correctly	10

Question 3. Recall that since we are fitting a linear model and that we are making statistical inferences, we have assumed that the data follows a linear trend, that the errors are normally distributed, and the variance of the errors is constant. To test this assumptions we will use three visual displays that can be generated by typing in R:

```
plot(signs, rate)
plot (fitted(model), resid(model))
qqnorm(resid(model))
qqline(resid(model))
```

From the scatter plot in Figure 1 (Question 1) we can see that the data follows a somewhat linear relationship and that there are no obvious outliers. From the residual plot in Figure 2 we can see that there is a cluster of points just underneath the zero line on the left, and as we move to the right the residuals seem to be further away from 0, suggesting that there is some heteroscedasticity (non-constant variance) in the residuals. Due to the nature of this cluster of points together in the lower left corner, we may also see that there are problems with the independence assumption.

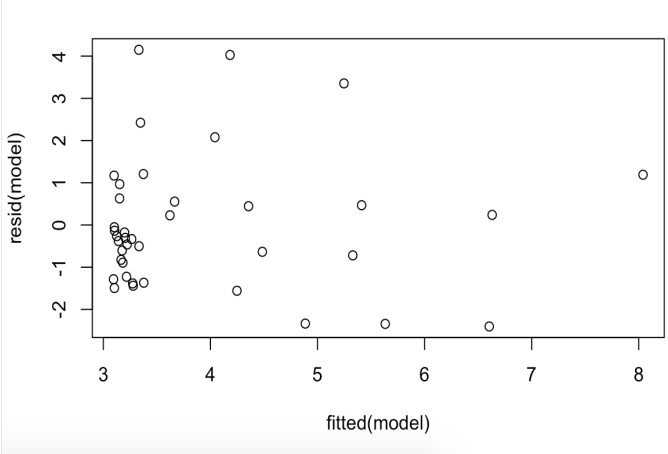

Figure 2: Residual Plot

Finally, we can see that the qq-plot in Figure 3 has an S-shape, especially on the upper tail. This might suggest that the error term is not normally distributed.

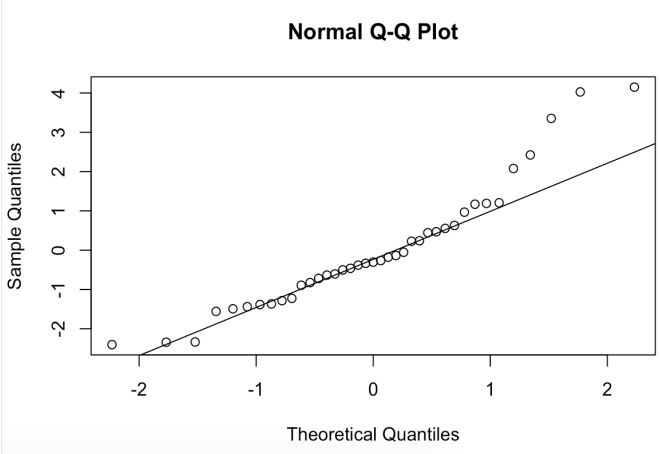

Figure 3: Normal Q-Q Plot

Response quality	Description	Points (out of 10)
Poor	The student does not answer any parts of the question correctly	0
OK	The student properly produces 1 plot and interprets it correctly	4
Good	The student properly produces 2 plots and interprets correctly	7
Perfect	The student answers all parts of the question correctly	10

Question 4

We are interested in making a prediction and obtaining a prediction interval when there are 1.25 signals per mile on the road. To do this in R we can type:

```
predict(model, data.frame(signs=1.25), interval="prediction")
```

The point prediction is 5.26571 and the 95% prediction interval is (1.919705, 8.611715).

Response quality	Description	Points (out of 10)
Poor	The student does not answer any parts of the question correctly	0
OK	The student answers part of the question correctly	5
Perfect	The student correctly calculates the confidence interval	10