Homework 1

Introduction

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.

The data file includes the following columns:

Rate: 1973 accident rate per million vehicle miles.

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

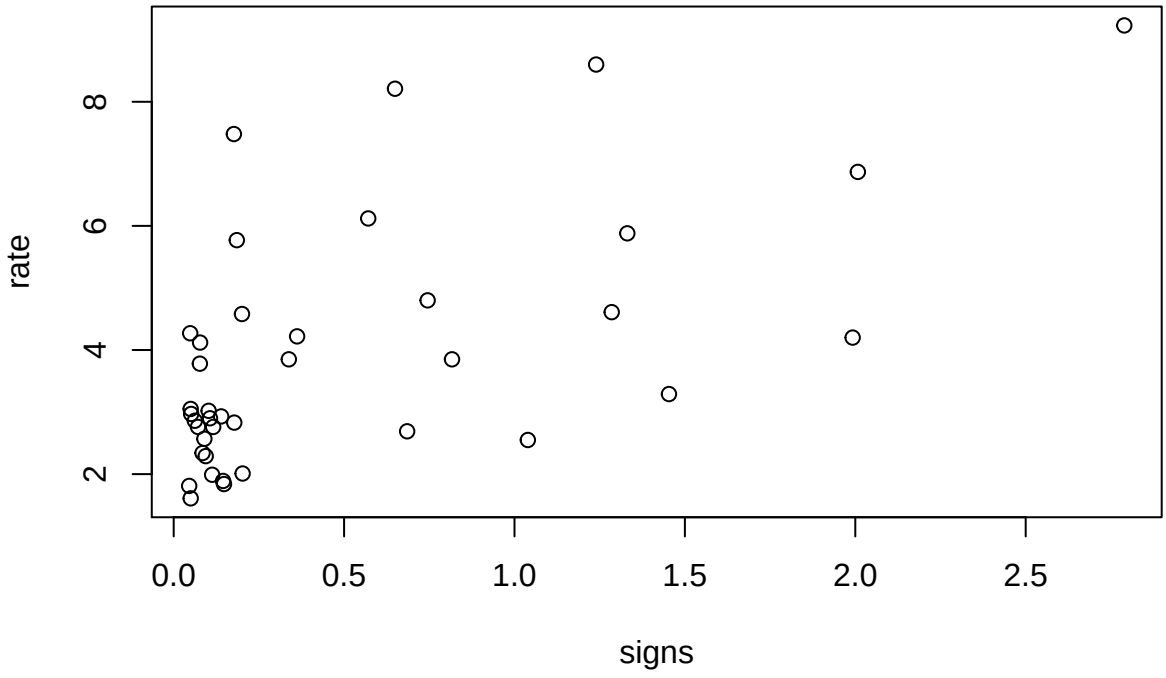
```
data = read.csv("Highway1.csv", head = TRUE, sep = ",")

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

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

cor_coef = round(cor(signs, rate),3)
cor_coef_log = round(cor(log(signs), log(rate)),3)

plot(signs, rate)
```



Question 1

A and B

The scatterplot above shows a weak, positive, linear correlation between signals per mile of roadway and the accident rate per million vehicle miles with a correlation coefficient of 0.603. There doesn't appear to be

any outliers in the data. The positive correlation coefficient suggests that an increase in signage may lead to higher accidents, but the value is not that close to 1, so the correlation is not strong.

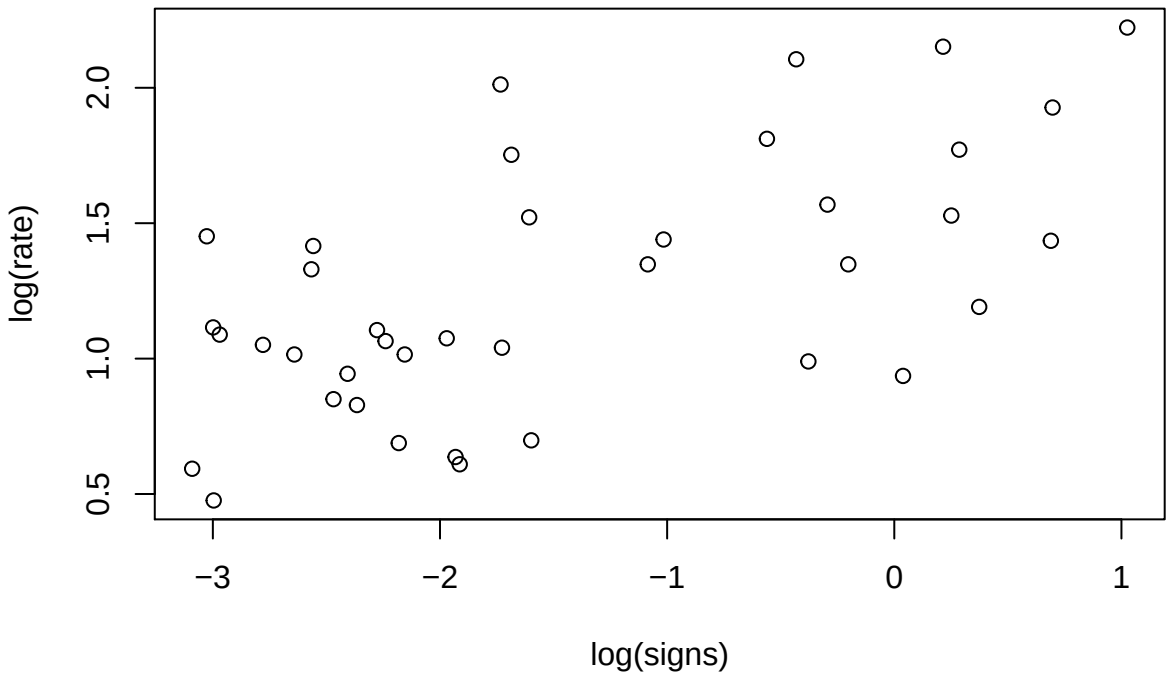
C

Simple linear regression appears to be a reasonable model for the above data based on the scatterplot and correlation coefficient. I did not note anything particularly unusual.

D

I took the logs of the data based on the above scatterplot. There are quite a few data points clustered in the 0.0 - 0.25 range. A log transformation might spread this data out and improve the correlation. It did improve the look of the plot (seen below) and the correlation (0.604), but not dramatically.

```
plot(log(signs), log(rate))
```



Question 2

```
model = lm(rate ~ signs)
model_log = lm(log(rate) ~ log(signs))
coefs = model$coefficients
slope = round(coefs[2],3)
intercept = round(coefs[1], 3)
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
```

```
summary(model_log)
```

```
##
## Call:
## lm(formula = log(rate) ~ log(signs))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6543 -0.2631  0.0080  0.2497  0.8166
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.58183    0.09200  17.194  < 2e-16 ***
## log(signs)   0.22277    0.04839   4.604 4.76e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3741 on 37 degrees of freedom
## Multiple R-squared:  0.3642, Adjusted R-squared:  0.3471
## F-statistic:  21.2 on 1 and 37 DF,  p-value: 4.758e-05
```

```
1-pt(4.6,37)
```

```
## [1] 2.409184e-05
```

A

The model parameters are the slope (1.802) and intercept (3.013) of the linear regression model.

B

The equation to the least squares line is:

$$\begin{aligned} Y =& \beta_0 + \beta_1 X \\ =& 3.013 + 1.802X \end{aligned}$$

C

The interpretation of the slope is that there is a 1.8 increase in the accident rate per million vehicle miles for every 1 signal increase per mile of roadway on average. The null hypothesis, $H_0 : \beta_1 = 0$ is rejected because the p-value is close to 0, 4.82e-05, and using the standard error as an approximation for the standard deviation puts 0 more than 4 standard deviations away from a slope of 0. We can also conclude that β_1 is statistically positive because the probability of a t-distribution is 2.409184e-05, very close to 0,

D

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

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

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

```
##                 2.5 %    97.5 %
## (Intercept) 1.3954191 1.7682388
## log(signs)  0.1247354 0.3208132
```

The confidence interval of the slope is between 1.008 and 2.596. This confirms our earlier assessment that we can reject the null hypothesis, $H_0 : \beta_1 = 0$, because 0 is not within the bounds of this confidence interval.

Question 3

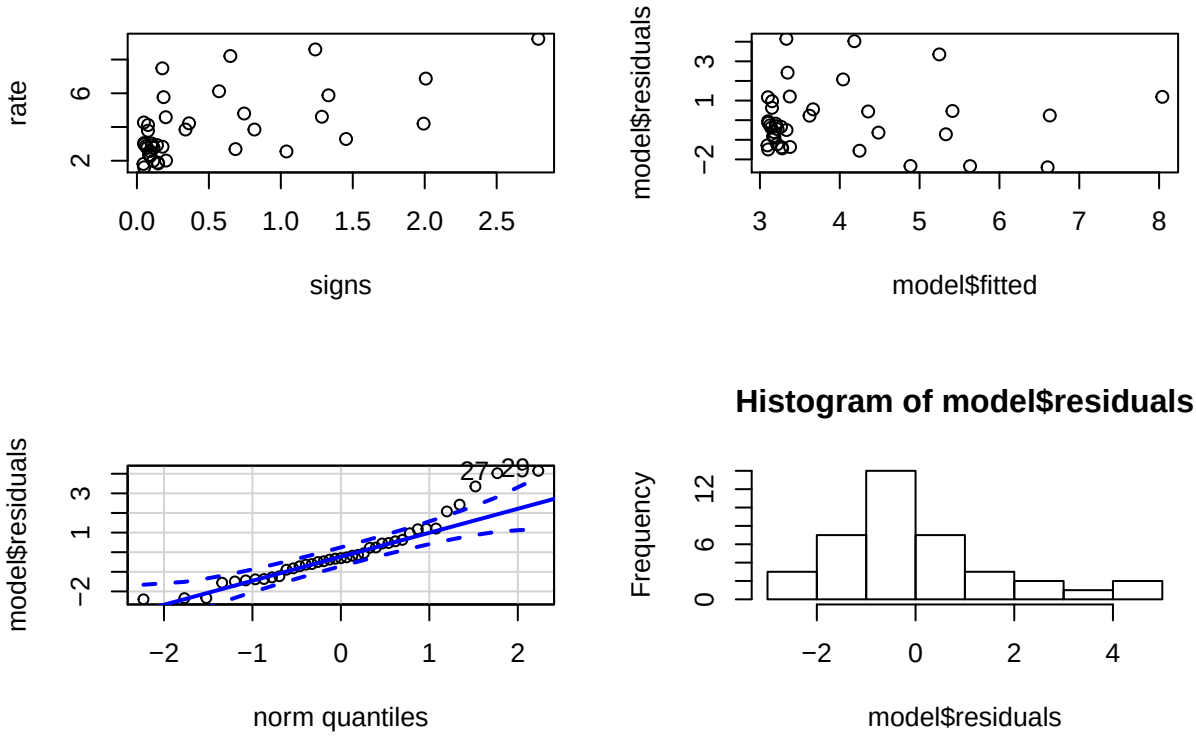
Assumptions

1. Linearity
2. Constant Variance
3. Independence
4. Normality

```
library(car)
```

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

```
par(mfrow=c(2,2))
p1 = plot(signs, rate)
p2 = plot(model$fitted, model$residuals)
p3 = qqPlot(model$residuals)
p4 = hist(model$residuals)
```



Linearity

Looking at the first plot I can see that there is a positive linear correlation between signs and rate so I would conclude that the linearity assumption holds.

Constant Variance

The second plot, which is the fitted values versus the model residuals, we can see that the residuals are scattered around the 0 line. Although there is a cluster at the lower end of the fitted values, I would say the constant variance holds.

Independence

The second plot also indicates that we have uncorrelated errors because the values are scattered with only one cluster. This is not a true test of independence, but can be used as a proxy.

Normality

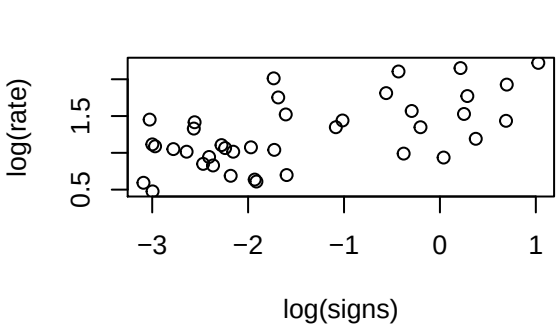
The model deviates from the normality assumption as seen in the q-q plot. THere are values outside the confidence interval on upper tail of the plot. The histogram is right skewed as a result.

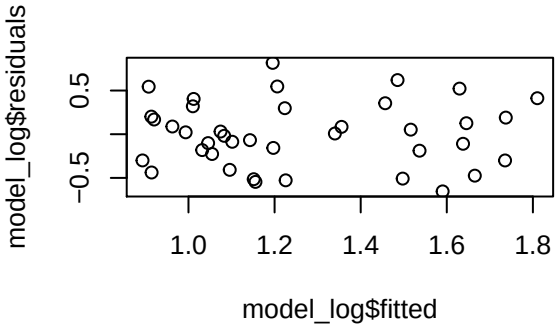
Log model

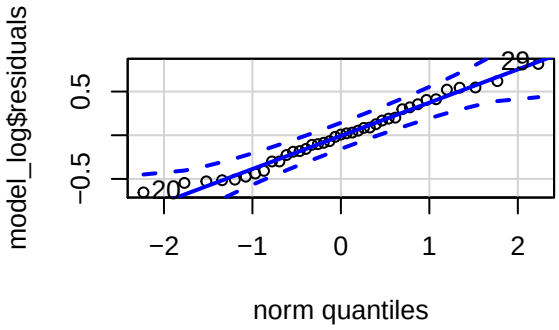
I decided to run these plots again taking the log of the predictor and response variable to see if it gave a better result.

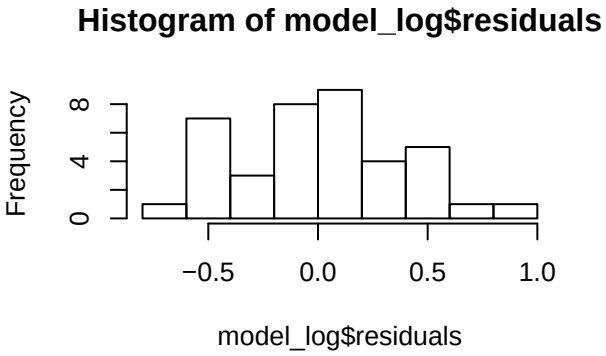
```
par(mfrow=c(2,2))
p1 = plot(log(signs), log(rate))
p2 = plot(model_log$fitted, model_log$residuals)
```

```
p3 = qqPlot(model_log$residuals)
p4 = hist(model_log$residuals)
```









Linearity

Looking at the first plot I can see that there is a positive linear correlation between the log signs and log rate so I would conclude that the linearity assumption holds.

Constant Variance

The second plot, which is the fitted values versus the model residuals, we can see that the residuals are scattered around the 0 line. The new log model does not have the cluster of residuals that the previous model had, so this is a stronger indication of constant variance.

Independence

The second plot also indicates that we have uncorrelated errors because the values are scattered without any clustering. Again, this is not a true test of independence, but can be used as a proxy.

Normality

The model does not deviate too radically as seen in the q-q plot as it did in the previous model. Therefore the normality assumption holds for this model, where it did not for the previous, because all pointss fall within the confidence intervals.

Outliers

I do not see any extreme outliers in the data of the second model, however, the first model's q-q plot indicates some outliers on the upper tail end of the data.

Question 4

```
new = data.frame(signs = 1.25)

predict.lm(model, new, interval="predict", level = .95)
```

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

```
exp(predict.lm(model_log, new, interval="predict", level = .95))
```

```
##        fit   lwr      upr
## 1 5.111739 2.332 11.20492
```

The top prediction is the untransformed model and the bottom is the log model. The predictions are not far off from each other, but the prediction bands are much larger on the log-log model. The results show that there is a predicted rate of 1.92 and 8.61 accidentes per mile with an average of 1.25 signs per mile for the first model, and a predicted rate of between 2.33 and 11.20 for the log-log model.