西南大学 2022: STATS 201 Assignment 2

RUNZE LIAO 22202032102007

22.5.18

Enter your name here

```
# Replace "Model Student" with your name in quotes,
# E.g., myname="Ruoxi Xu"
myname="runze liao"
```

Question 1

Background

A manufacturer of autonomous vehicles wanted to predict the stopping distance as a function of vehicle speed (km/h).

Stopping distance (m) was measured as the distance traveled after emergency braking was activated. You may assume that the observations are independent.

The code below automatically generates the data for a randomly chosen student. The data are in the dataframe **Stop.df**. Variable **Speed** is the explanatory variable, and **Dist** is the response variable.

Questions of interest

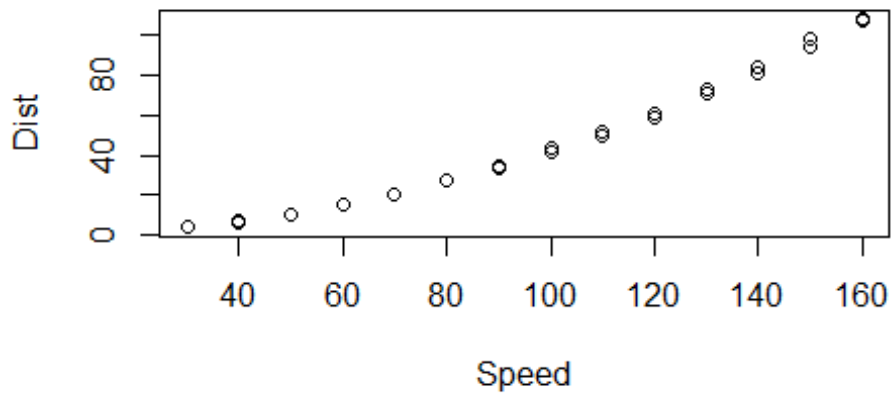
Make inference about the stopping distance for a given speed. In particular,

- What is the effect of a doubling in speed?
- Estimate the typical stopping distance for emergency braking at speeds of 50 km/h and 100 km/h.
- The manufacturer desires that the vehicle stops in less than 12 m at 50 km/h, at least 999 times out of 1000. Has this been achieved? [Hint, calculate the 0.998 level prediction interval.]

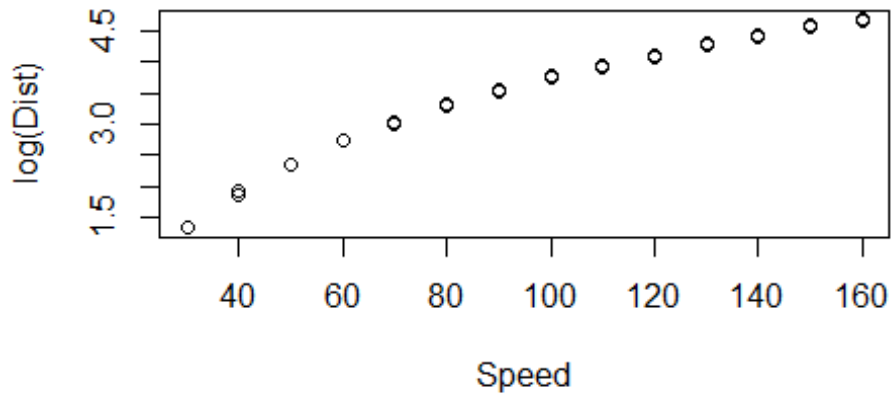
```
## Warning: package 's20x' was built under R version 4.0.5
```

Plot data

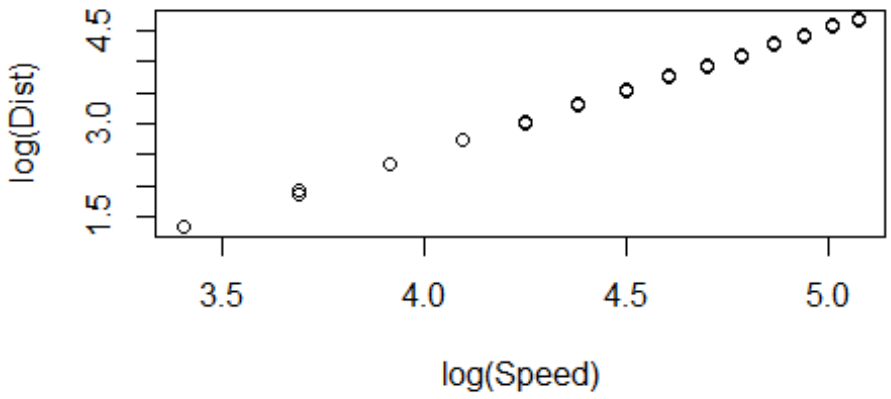
```
# Add R code below to draw appropriate scatter plot(s)
plot(Dist~Speed, data = Stop.df)
```



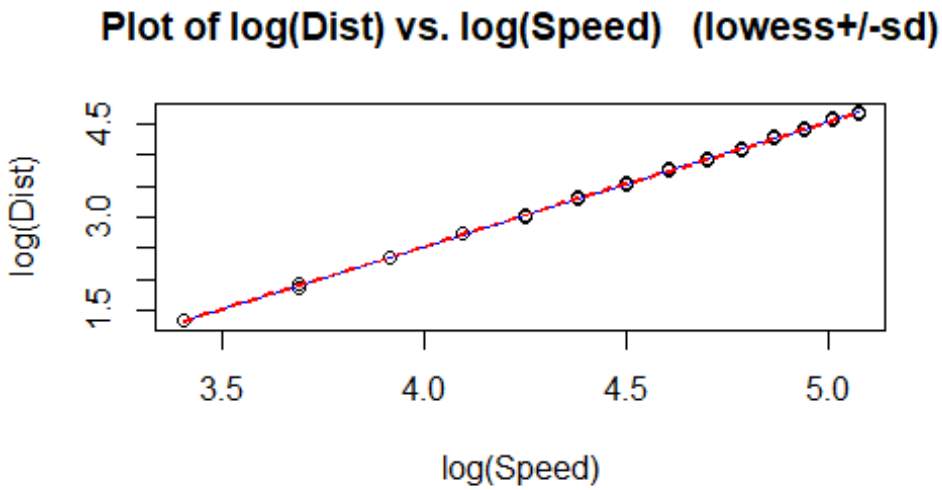

```
plot(log(Dist)~Speed, data = Stop.df)
```




```
plot(log(Dist)~log(Speed), data = Stop.df)
```



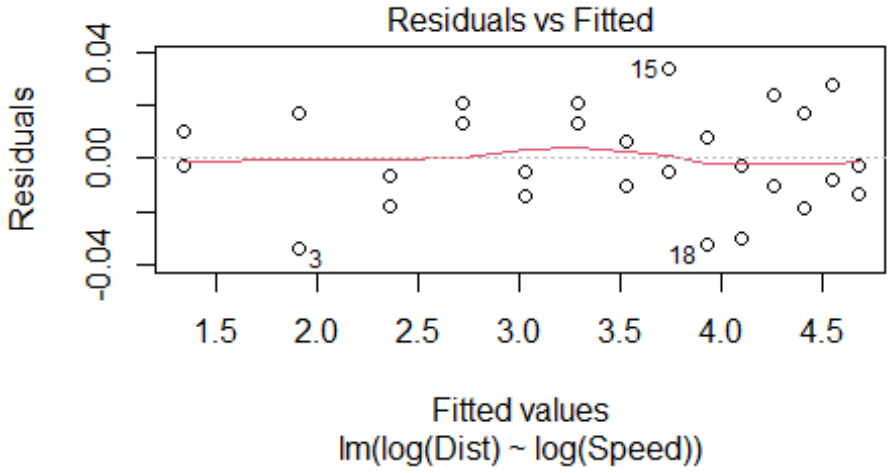

```
trendscatter(log(Dist)~log(Speed),data = Stop.df)
```



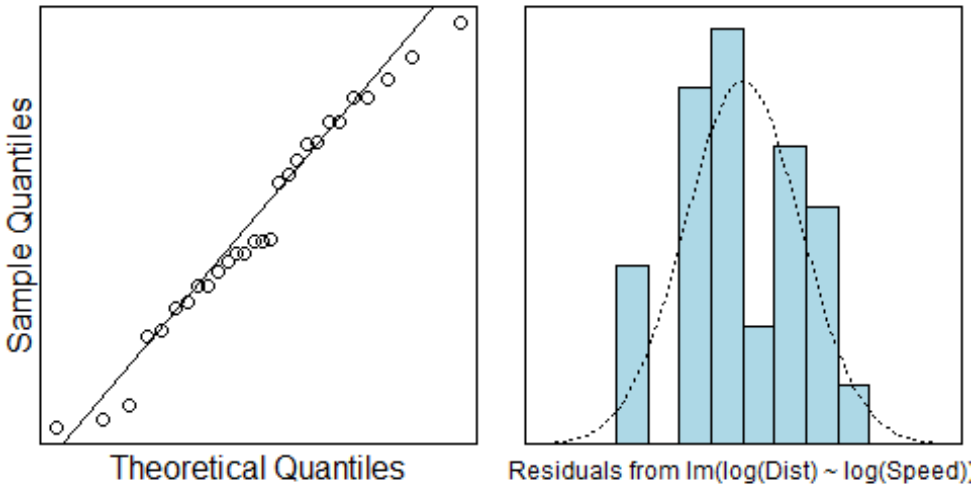
In this graph, we can see a apparent looks reasonably linear on the log scale and after we fit the log(Dist), it seems that it has a Power law relation. So we will fit the model with log(Dist) and log(Speed)

Fit a power-law model and do assumption checks

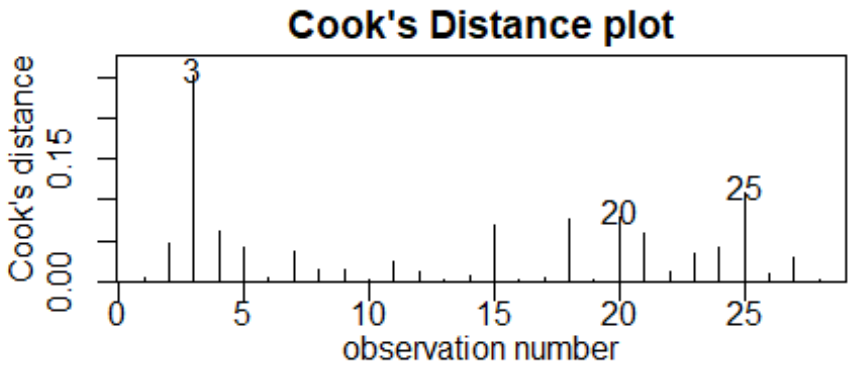
```
Stop.lm = lm(log(Dist)~log(Speed), data = Stop.df)
plot(Stop.lm, which = 1)
```




```
normcheck(Stop.lm)
```




```
cooks20x(Stop.lm)
```



The EOV check is good, the residuals satisfy the normal distributions. The point 3 seems strange, however, we will keep it as it does not > 0.4.

```
summary(Stop.lm)

##
## Call:
## lm(formula = log(Dist) ~ log(Speed), data = Stop.df)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.033919 -0.011019 -0.002776  0.014365  0.033273
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.467523   0.031587  -173.1   <2e-16 ***
## log(Speed)   2.000024   0.007064   283.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01859 on 26 degrees of freedom
## Multiple R-squared:  0.9997, Adjusted R-squared:  0.9997
## F-statistic: 8.016e+04 on 1 and 26 DF,  p-value: < 2.2e-16

exp(confint(Stop.lm))

##                    2.5 %      97.5 %
## (Intercept) 0.003956278 0.004504881
## log(Speed)  7.282719779 7.497311256
```

The p-value of intercept and log(Speed) are all < 0.05, we have strong evidence to believe the assumption.

The estimate of intercept is from 0.0040 to 0.0045 and the estimate of Speed is from 7.28 to 7.50.

Inference, i.e, answer questions of interest

```
# Add R code below
pred.df = data.frame(Speed = c(50,100))
exp(predict(Stop.lm,pred.df,level = 0.998, interval = "confidence"))
```

```
##        fit      lwr      upr
## 1 10.55520 10.37032 10.74338
## 2 42.22152 41.68918 42.76066
```

If the Speed = 50, the Distance will be the 10.37 to 10.74, while when the Speed = 100, then the distance will be 41.69 to 42.76.

Method and Assumption Checks

By plotting the original data, we cannot find a linear regression, and we use the Eov check and Normchenck, however, it does not satisfy by fitting the simple linear regression.

After logging the variables, the scatter plots showed the desired linear relationship between Speed and Distance, so we fitted a power law model explaining log Speed and log Distance. The trendscatter using the power law model seems a straight line.

The underlying model assumptions appear valid, however, we have a strange obseravtion point 3, however, it does not > 0.4, so we will keep it. Our final model is:

$$log(Dist)_i = \beta_0 + \beta_1 * log(SpeWed)_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$

Our model explained 99.97% of the vairability in the Speed & Distance model.

Executive Summary

We want to find the relationship between the Speed and the Distance, and we also are interested in estimate the Speed.

The model was that the Distance would follow a power law relationship with the car's Speed.
More specifically, the Distance will increase with the square of the Speed.

The Distance will follow a power law modelwith respect to the Speed.

- What is the effect of a doubling in speed?
  As the power law parameter is 2, the Distance will increase with the square of the Speed, so the stop distance will be 4 times than before.

- Estimate the typical stopping distance for emergency braking at speeds of 50 km/h and 100 km/h. We can see that If the Speed = 50, stopping distance for emergency braking be the 10.37 to 10.74, while when the Speed = 100, then stopping distance for emergency braking will be 41.69 to 42.76.

- The manufacturer desires that the vehicle stops in less than 12 m at 50 km/h, at least 999 times out of 1000. Has this been achieved? [Hint, calculate the 0.998 level prediction interval.]
  If the Speed = 50, the Distance will be the 10.37 to 10.74, while when the Speed = 100, then the distance will be 41.69 to 42.76.

Question 2

Background and questions of interest

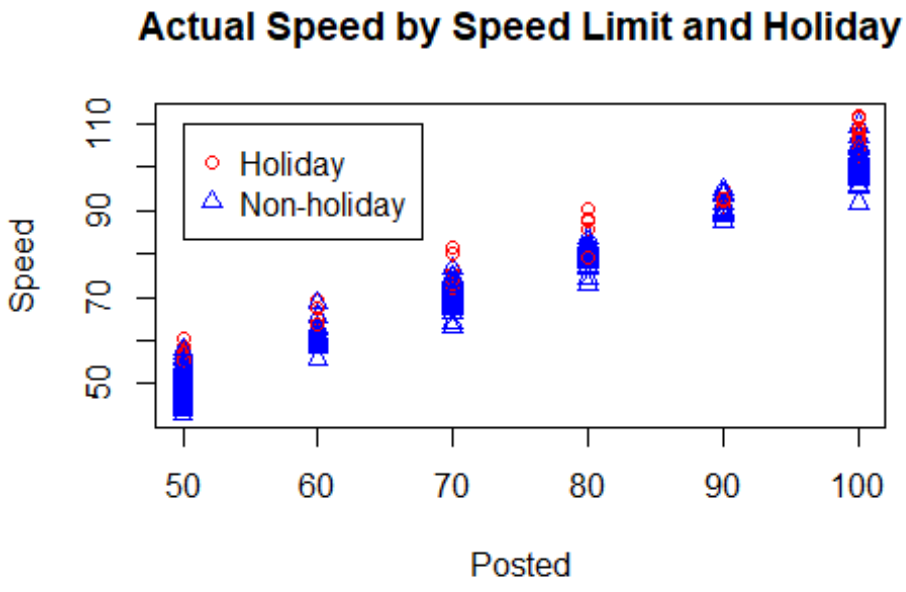
In New Zealand, the police are usually tolerant of speeding up to 10 km/h over the posted speed limit. However this tolerance is reduced to 5 km/h during public holidays.

We investigated the effect of the posted speed limit on the actual speed of vehicles, and also the effect of the reduced police tolerance for speeding that applies on holidays. Also, we wished to see if the effect of the reduced tolerance during holidays depended on the posted speed limit.

A total of 445 independent observations were made for posted speed limits between 50 and 100 km/h. All measurements were made during periods of unrestricted traffic flow on straight stretches of road.

Read in and inspect the data:

```
Speed.df=read.table("Speed.txt", header=T)
plot(Speed~Posted,main="Actual Speed by Speed Limit and Holiday", col=i
felse(Holiday=="N","red","blue"),pch=ifelse(Holiday=="N",1,2),data=Spee
d.df)
legend(50,110,pch=c(1,2),col=c("red","blue"),legend=c("Holiday","Non-ho
liday"))
```



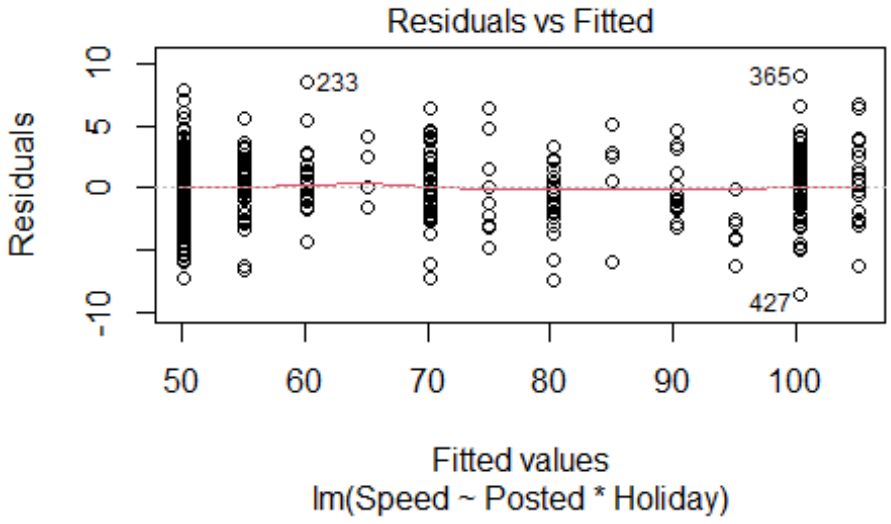
Comment on plot

The number of holiday cars is less than the Non-holidays. And we can see that the average speed in their own posted area, the Holiday's Speed is larger than the Non-

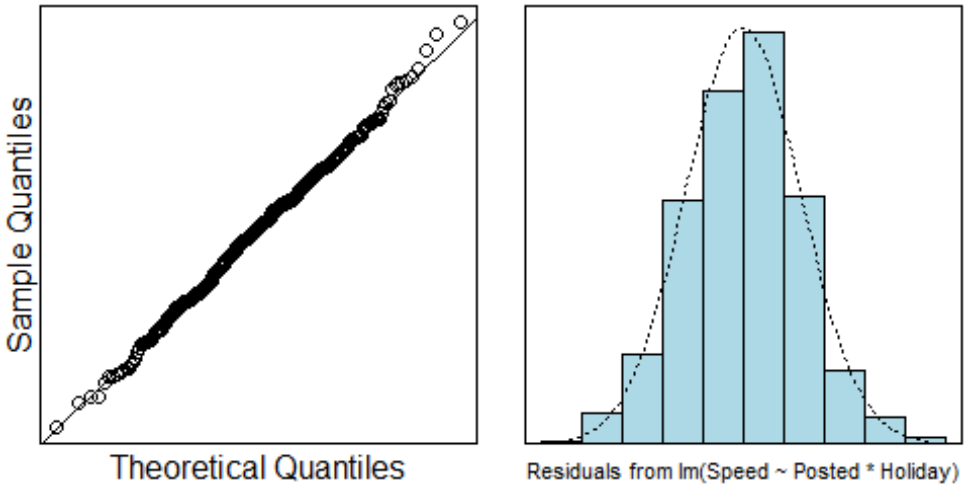
holiday. The fact that appears this problem maybe in holiday, the traffic management is easy.

Fit model and check assumptions

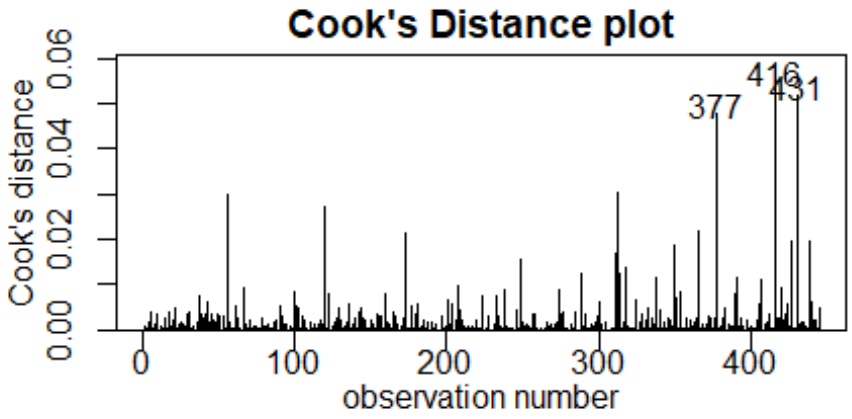
```
# Add R code below
library(s20x)
Speed.fit = lm(Speed~Posted*Holiday, data = Speed.df)
plot(Speed.fit, which = 1)
```




```
normcheck(Speed.fit)
```




```
cooks20x(Speed.fit)
```

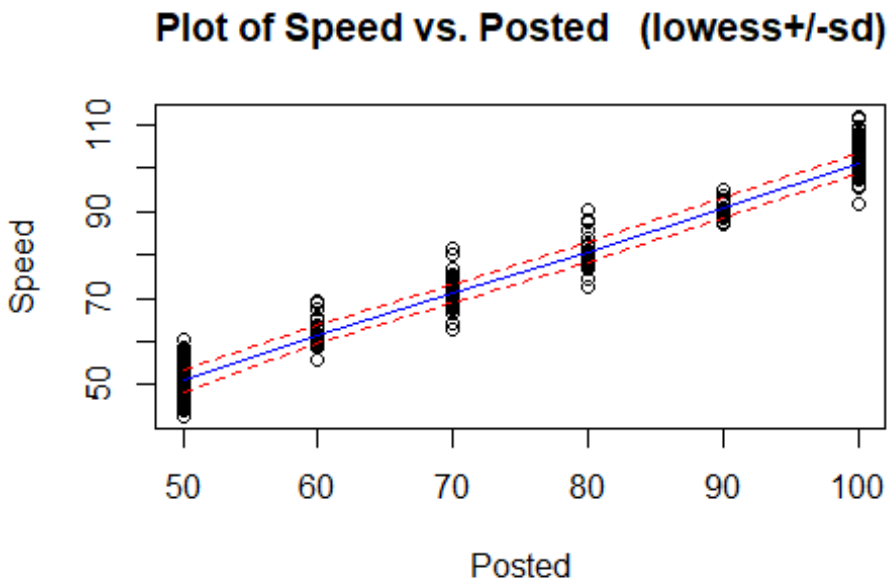



```
summary(Speed.fit)

##
## Call:
## lm(formula = Speed ~ Posted * Holiday, data = Speed.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5510 -1.9235  0.0614  1.8614  9.0490
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.173458   1.061487   4.874 1.53e-06 ***
## Posted           0.998491   0.014727  67.801  < 2e-16 ***
## HolidayY        -5.247256   1.179569  -4.448 1.10e-05 ***
## Posted:HolidayY  0.005757   0.016346   0.352    0.725
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.856 on 441 degrees of freedom
## Multiple R-squared:  0.9825, Adjusted R-squared:  0.9824
## F-statistic:  8271 on 3 and 441 DF,  p-value: < 2.2e-16
```

The residuals plot is good, the normcheck is good which indicates the residuals are normally distributed. The all observations seem no strong influence point. However, through the summary we can see that the p-value of Posted:HolidayY(0.725) is larger than 0.05 so much, so we wonder if we can remove the interaction between the Posted and Holiday.

Reproduce plot with fitted lines superimposed.

```
# Add R code below
trendscatter(Speed~Posted, data = Speed.df)
```

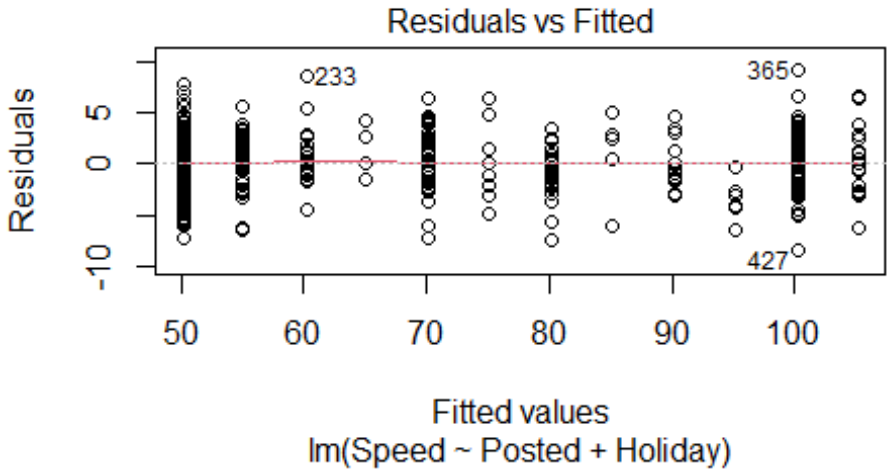


The line between the holiday and Non-holiday is parallel, So we fitted the model without interaction.

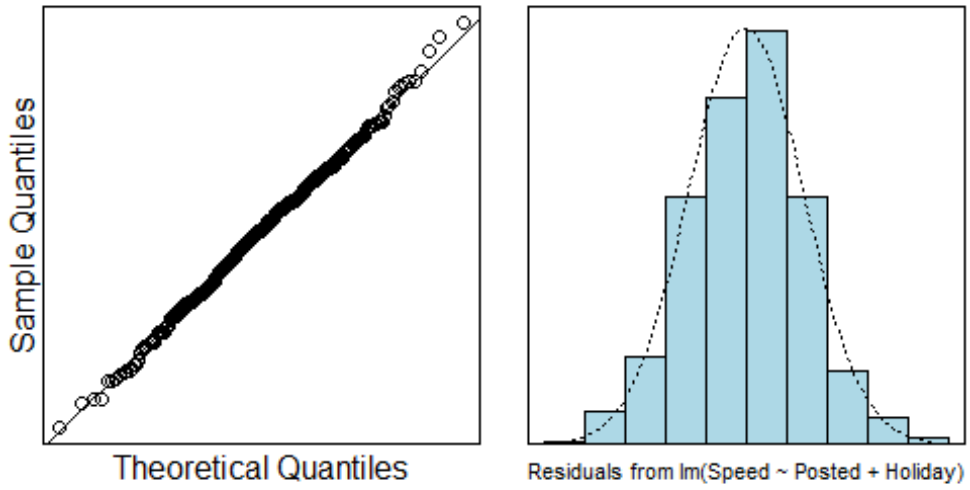
```
Speed.fit2 = lm(formula = Speed ~ Posted + Holiday, data = Speed.df)
```

Then we plot some graphs to see if it satisfy all assumptions.

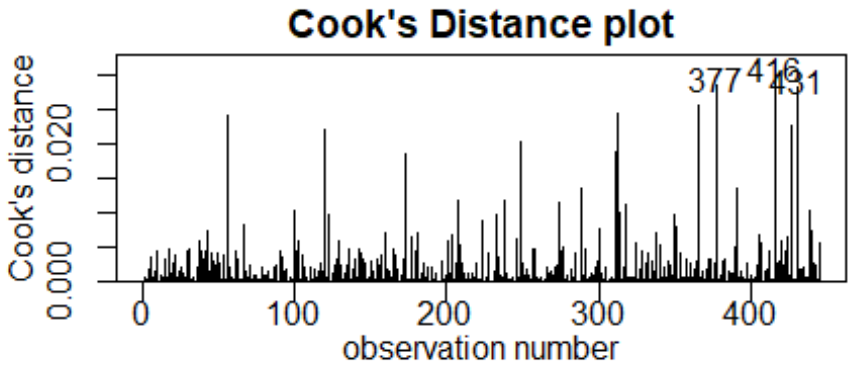
```
plot(Speed.fit2, which = 1)
```



```
normcheck(Speed.fit2)
```



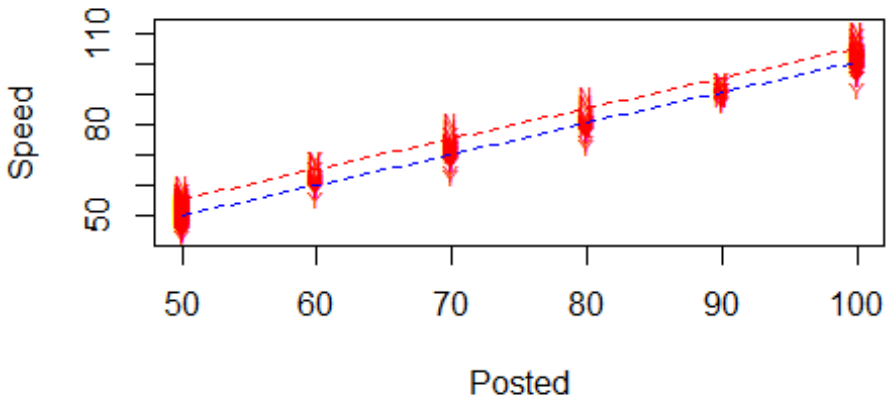

```
cooks20x(Speed.fit2)
```



From the graphs we could see that the Residuals is good, the distribution of residuals is like Normal, the cooks distance is good, no strong influence point. It satisfied all assumption.

```
summary(Speed.fit2)

##
## Call:
## lm(formula = Speed ~ Posted + Holiday, data = Speed.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -8.5178 -1.9229  0.0589  1.8905  9.0822
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.851280   0.537983   9.018   <2e-16 ***
## Posted       1.003164   0.006383 157.153   <2e-16 ***
## HolidayY    -4.849907   0.344058 -14.096   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.853 on 442 degrees of freedom
## Multiple R-squared:  0.9825, Adjusted R-squared:  0.9825
## F-statistic: 1.243e+04 on 2 and 442 DF,  p-value: < 2.2e-16

confint(Speed.fit2)

##                 2.5 %    97.5 %
## (Intercept)  3.7939585  5.908602
## Posted       0.9906186  1.015709
## HolidayY    -5.5261009 -4.173713

pred.df1 = data.frame(Posted = c(50,60,70,80,90,100),Holiday = "Y")
pred.df2 = data.frame(Posted = c(50,60,70,80,90,100),Holiday = "N")
predict(Speed.fit2,pred.df1,interval = "confidence")

##         fit      lwr       upr
## 1  50.15957 49.77704  50.54211
## 2  60.19121 59.87321  60.50922
## 3  70.22285 69.92723  70.51848
## 4  80.25450 79.93023  80.57876
## 5  90.28614 89.89323  90.67904
## 6 100.31778 99.83293 100.80263

predict(Speed.fit2,pred.df2,interval = "confidence")

##         fit       lwr       upr
## 1  55.00948  54.35653  55.66243
## 2  65.04112  64.42269  65.65956
## 3  75.07276  74.46444  75.68108
## 4  85.10440  84.48060  85.72820
## 5  95.13604  94.47296  95.79912
## 6 105.16768 104.44539 105.88997

plot(Speed ~ Posted, data = Speed.df, pch = substr(Holiday,1,1), cex =
0.7, col = ifelse(Holiday == "Yes","blue","red"))
lines(x = c(50,60,70,80,90,100),y = predict(Speed.fit2,pred.df1),col =
"blue",lty = 2)
lines(x = c(50,60,70,80,90,100),y = predict(Speed.fit2,pred.df2),col =
"red",lty = 2)
```



We can see that they are parrael line and have no interaction.

Methods and assumption checks

To explain the effect of the posted speed limit on the actual speed of vehicles, and also the effect of the reduced police tolerance for speeding that applies on holidays. First we build a interaction model by having a numeric and a factor variable, however, when we do the summary we see that the p-value of the interaction part is larger than 0.05, we have no strong evidence to have it, moreover, we plotted the lines between Posted and Speed with whether having the holiday, we saw that they are parallel, which indicated that the two variables have no interaction.

So we have a latest model without interaction, and it satisfied all the assumptions, we hold the belief that the modle is quiet good.
Our final model is:

$$Speed_i = \beta_0 + \beta_1 * Posted_i + \beta_2 * Holiday_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma^2)$ and "Holiday" is 1 if this day is Holiday, otherwise the "Holiday" is 0.

Our model explained 98.25% of the variablity in the Holiday Speed experiment. ## Executive Summary We are interested in building a model to estimate the relation between Speed and Posted with whether it is Holiday.

The relationship between Posted and Holiday has no interaction.

- With the Posted increased 1, the Speed will increase from 0.99 to 1.02.
- When the Posted are the same, if it is in the holiday, the Speed of the vehicals the decrease about 4.17 to 5.53.

We estimate that when Posted speeds are 50, 60, 70, 80, 90, 100 and is a holiday, expected Speeds are between 50.16, 60.19, 70.22, 80.25, 90.29 and 100.32.

We estimate that when Posted speeds are 50, 60, 70, 80, 90, 100 and is no-holiday the expected Speeds are between 55.01, 65.04, 75.07, 85.10 and 105.17 respectively.