STAT400 - Homework 11

Your Name

Due 12/10/2020 by 4pm

Be sure to `set.seed(400)` at the beginning of your homework. Please use 2000 bootstrap samples in the problems below.

```
#reproducibility
set.seed(400)

# bootstrap samples
B <- 2000
```

```
# useful libraries
library(tidyverse)
library(bootstrap)
library(resample)
library(boot)
library(simpleboot)
data(Verizon)
```

1. Nike has hired you to help analyze their data on their customers who run. They want to make sure that you understand how their running gear fits their customers. A sample of 25 randomly selected customers was selected, and the customers were asked to submit their weights. The data:

   ```
   wt <- c(149, 136, 139, 117, 137, 132, 122, 130, 134, 153, 140, 151, 203, 143, 145, 123, 127, 146,
   ```

   a. Calculate the sample standard deviation s for these weights.
   b. To do the following, use the `boot` and `simpleboot` packages as shown in the class handouts.
      i. Compute the bootstrap bias and standard error for s.
      ii. Plot a histogram and qq-plot of the bootstrap distribution.
      iii. Based on these results: (1) Is there evidence of bias and skewness of the bootstrap distribution for s? (2) Is it appropriate to assume that the distribution of s is normally distributed?
      iv. Construct 4 types of intervals that we discussed in class by using `type=c("norm","basic", "perc", "bca")` in the boot.ci command
      v. Plot the four intervals onto a histogram of the sampling distribution using the command `geom_segment`.
   c. Construct a "studentized" bootstrap t CI and also plot it onto your histogram of the sampling distribution.
   d. What final result would you report to Nike? Explain your reasoning.

```
#a)
sdhat <- sd(wt)
```

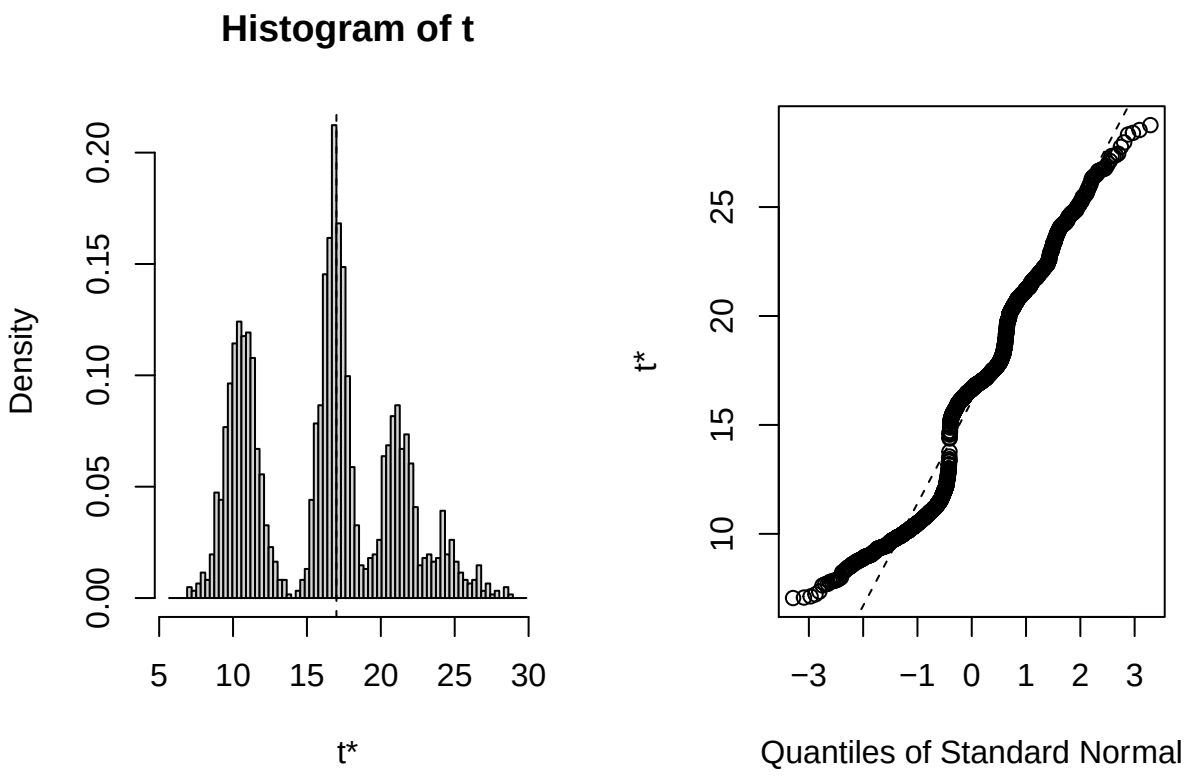
```
samplesd <- function(x,i){
  return(sd(x[i]))
}
```

```
wtb <- boot(data=wt,statistic = samplesd,R=2000)
```

```
wtb
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = wt, statistic = samplesd, R = 2000)
##
##
## Bootstrap Statistics :
##     original     bias    std. error
## t1* 16.99588 -0.9057863    4.694239
```

```
plot(wtb)
```



Yes, it does seems to be skewed and it is not appropriate to assume normality.

```
ci<-boot.ci(wtb,type=c("norm","basic", "perc", "bca"))
```

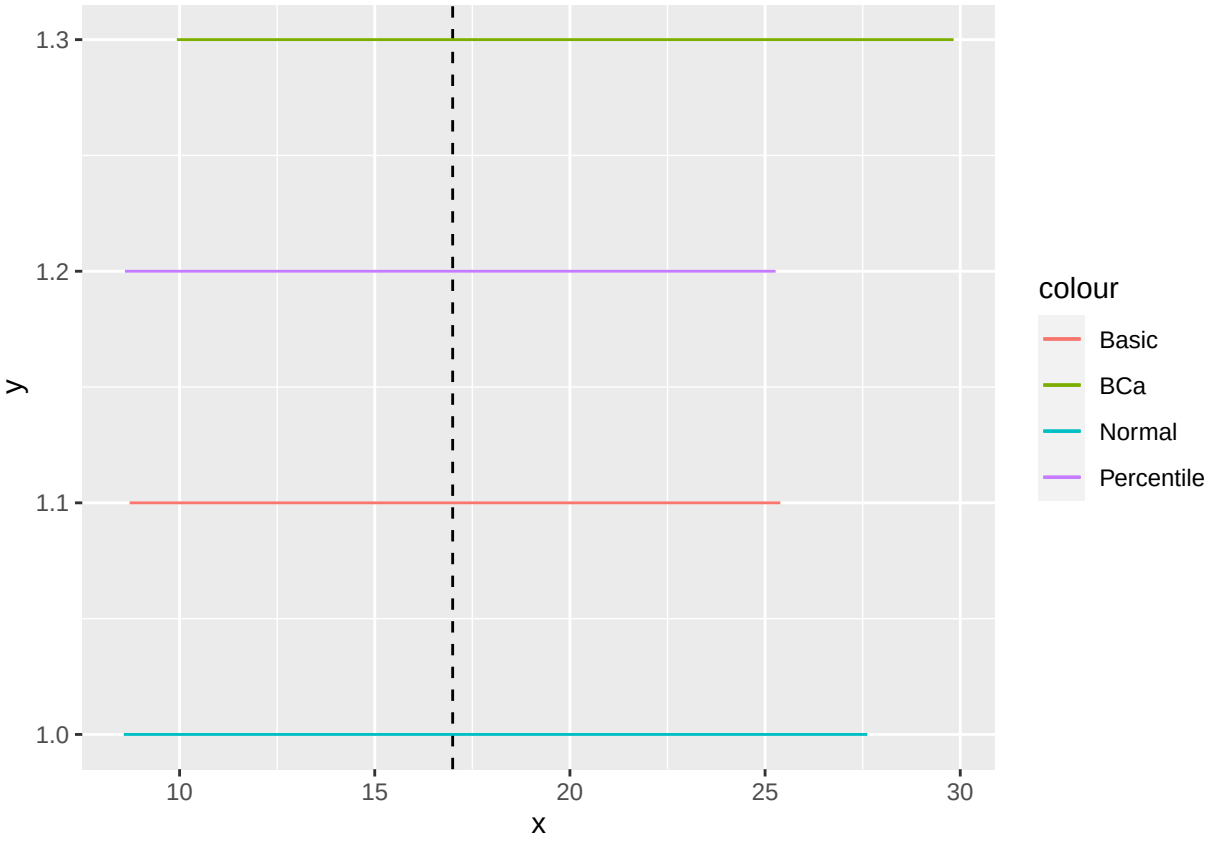
```
## Warning in norm.inter(t, adj.alpha): extreme order statistics used as endpoints
```

```
ci
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
```

```
## CALL :
## boot.ci(boot.out = wtb, type = c("norm", "basic", "perc", "bca"))
##
## Intervals :
## Level      Normal              Basic
## 95%   ( 8.70, 27.10 )   ( 8.92, 25.03 )
##
## Level     Percentile            BCa
## 95%   ( 8.96, 25.07 )   (10.07, 28.76 )
## Calculations and Intervals on Original Scale
## Warning : BCa Intervals used Extreme Quantiles
## Some BCa intervals may be unstable
```

```
ggplot()+geom_vline(aes(xintercept = sd(wt)), lty = 2) +
  geom_segment(aes(x=8.57,y=1,xend=27.62,yend=1,colour="Normal"))+
geom_segment(aes(x=8.72,y=1.1,xend=25.39,yend=1.1,colour="Basic"))+
geom_segment(aes(x=8.60,y=1.2,xend=25.27,yend=1.2,colour="Percentile"))+
  geom_segment(aes(x=9.93,y=1.3,xend=29.83,yend=1.3,colour="BCa"))
```



2. For the `Verizon` dataset from the class handouts construct a 95% CI for the difference of two medians. Construct 4 types of intervals that we discussed in class by using `type=c("norm","basic", "perc", "bca")` in the boot.ci command.

   a. Are the intervals similar for all the methods? Why or why not?
   b. Let $\tilde{\mu}_1$ = the population median repair time for ILEC customers and $\tilde{\mu}_2$ = the population median of repair time for CLEC customers. Based on the results of the BCa interval, would you reject this hypothesis? Explain your answer.

$$H_0 : \tilde{\mu}_1 = \tilde{\mu}_2 H_a : \tilde{\mu}_1 \neq \tilde{\mu}_2$$

```
ilec_times <- Verizon[Verizon$Group == "ILEC",]$Time
clec_times <- Verizon[Verizon$Group == "CLEC",]$Time


median <- function(x){
 median(x)
}
```

A) They are pretty different likely due to skewness and lack of normality.

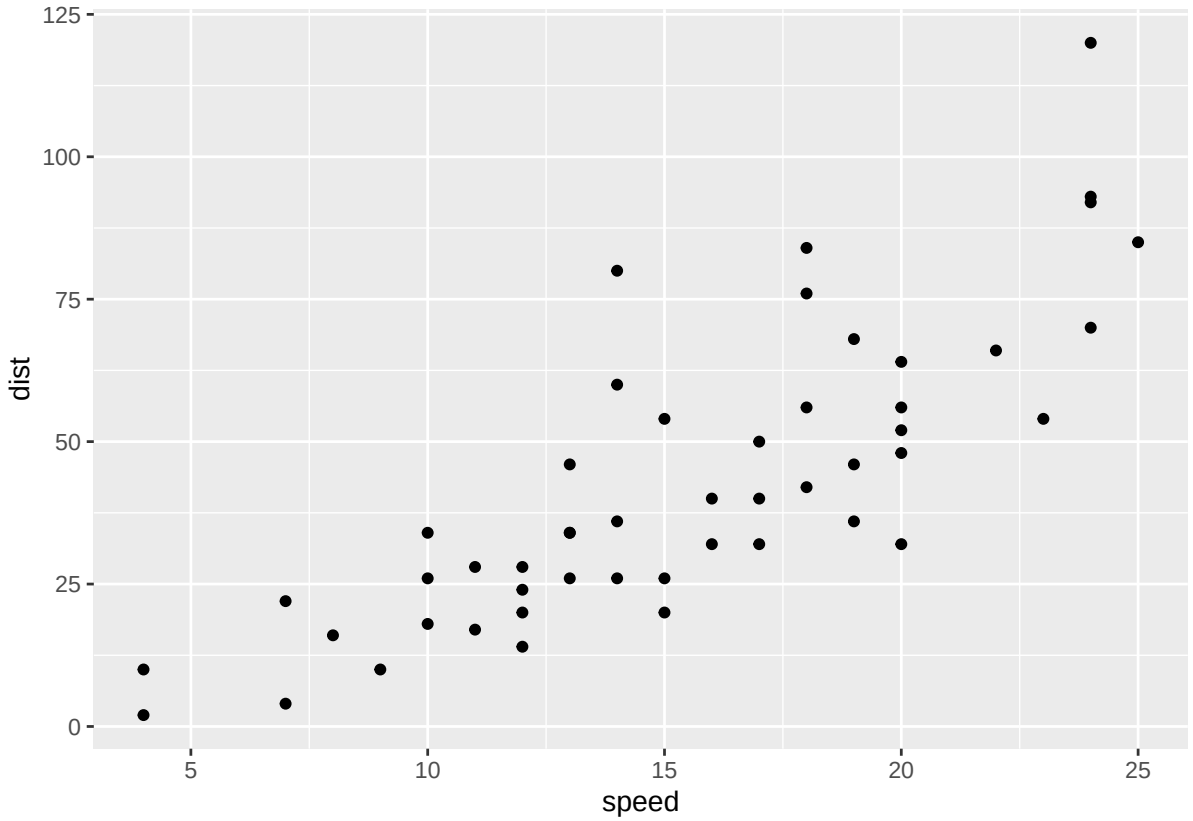
B) I would reject the null since most of the the CI's did not include 0, which means that the medians are different for different costumers.

I had the CI's but when I went to knit it gave me this error. Error: C stack usage 7971476 is too close to the limit

3. This data set is the `cars` data in `R`. The goal is to create a regression model about the relationship between stopping distance (`dist`) and speed (`speed`) in cars.

   a. Create a scatter plot of `speed` vs. `dist` in `mpg` and describe the relationship.

```
ggplot(cars) + geom_point(aes(x=speed,y=dist))
```



there appears to be a linear relationship between speed and distance, where when speed increases the

distance to stop increases.

b. Fit a simple linear model of `dist` on `speed`. Describe the result, including diagnostic plots, and create 95% CIs for the coefficients.

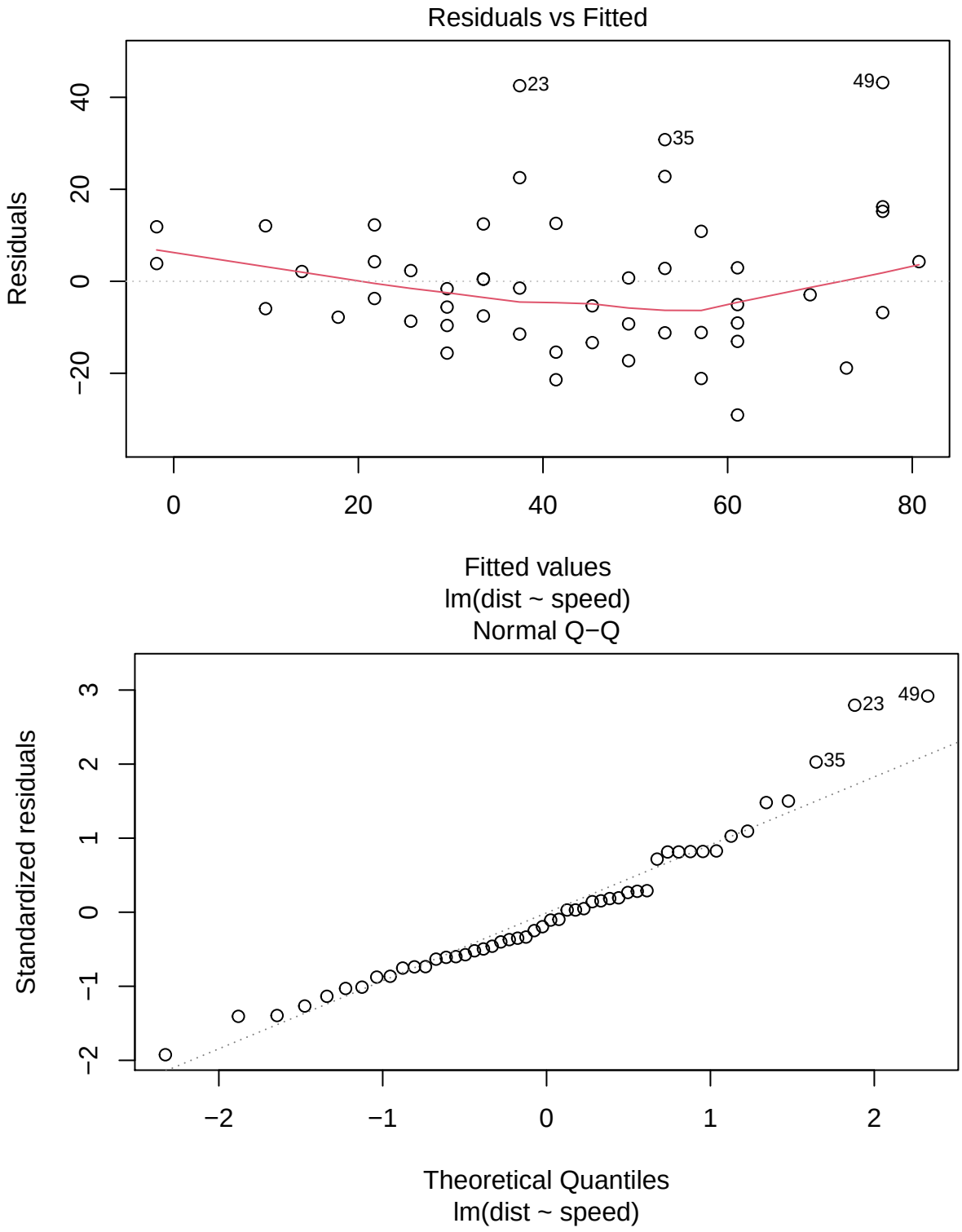
```
lm <- lm(dist~speed,data=cars)
confint(lm, 'speed', level=0.95)
```

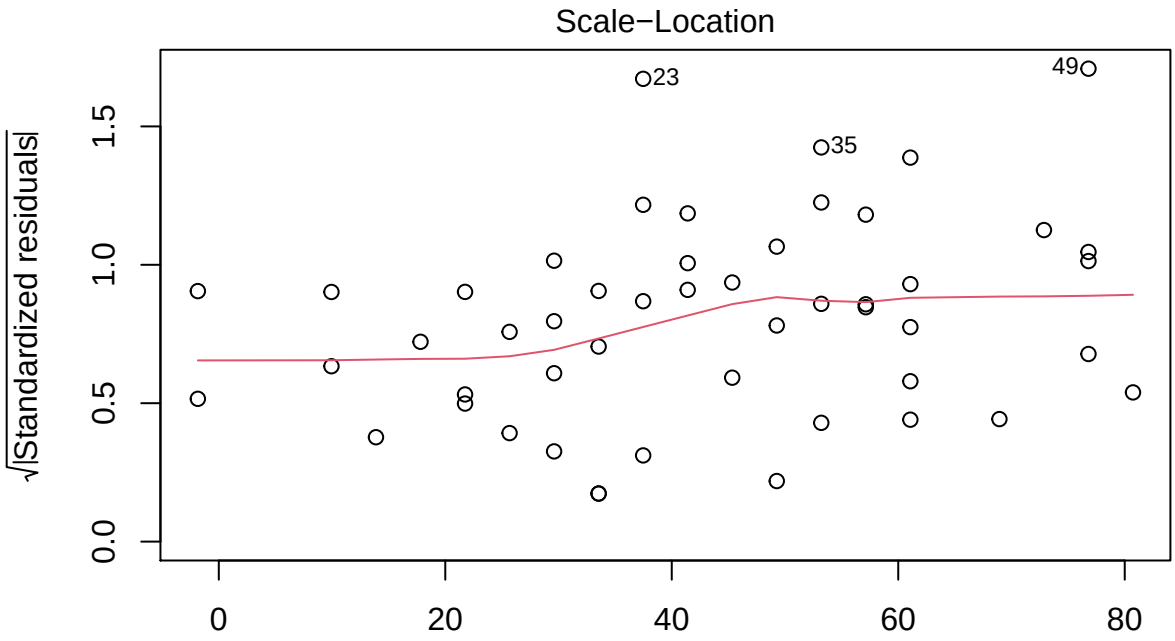
```
##          2.5 %   97.5 %
## speed 3.096964 4.767853
```

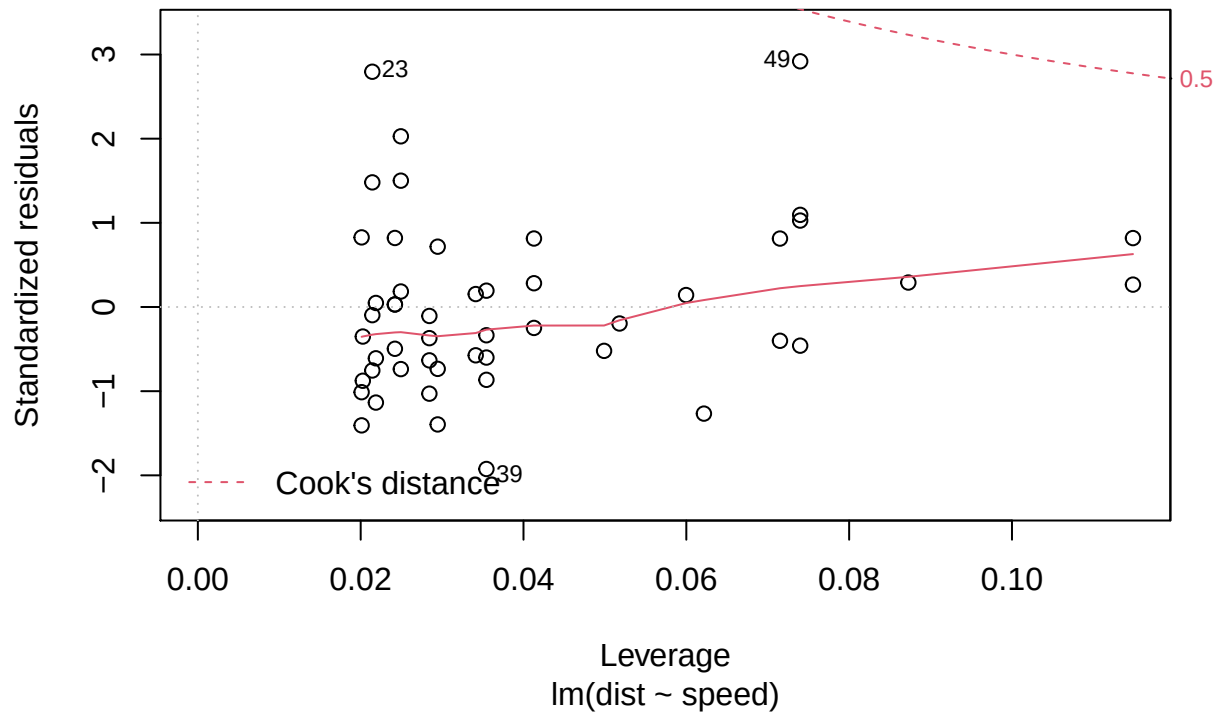
```
summary(lm)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```
plot(lm)
```


Residuals vs Fitted
Residuals
Fitted values
lm(dist ~ speed)
23
35
49
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(dist ~ speed)





For a increase of 1 mph there appears to be a increase of 3.9324 ft in stopping distance(95%CI=(3.096964,4.767853)), in the diagnostic plots there appears to be no violation of the normal assumption nor a violation of the constant variance assumption. (R^2=0.65)

`c. Perform the paired bootstrap and compute the bootstrap bias and standard error for the coefficients.`

```r
reg_func <- function(dat, idx) {
  # write a regression function that returns fitted beta
  df_star <- dat[idx,]
  m1 <- lm(dist ~ speed, data = df_star)
  coef(m1)
}

paired.boot <- boot(cars, reg_func, R = 2000)

boot.ci(paired.boot, conf = 0.95, type =c("norm","basic", "perc", "bca"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 2000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = paired.boot, conf = 0.95, type = c("norm",
##     "basic", "perc", "bca"))
##
## Intervals :
## Level      Normal              Basic
## 95%   (-28.80,  -6.00 )   (-28.45,  -5.16 )
##
## Level     Percentile            BCa
## 95%   (-30.00,  -6.71 )   (-31.34,  -7.33 )
## Calculations and Intervals on Original Scale
```

d. Perform the bootstrap using the residuals and compute the bootstrap bias and standard error for the

```r
reg_func_2 <- function(dat, idx) {
  # write a regression function that returns fitted beta
  # from fitting a y that is created from the residuals
  m1 <- lm(dist ~ speed, data = dat)
  resids <- m1$residuals

  # resample the residuals
  resids_star <- resids[idx]

  # make new response data and fit model
  y_star <- m1$fitted.values + resids_star
  dat_star <- data.frame(dist = y_star, speed = cars$dist)
  m1_star <- lm(dist ~ speed, data = dat_star)

  # get coefs
  coef(m1_star)
}



resid.boot <- boot(cars, reg_func_2, R = 2000)
```

e. Which method (simple linear regression, paired bootstrap, or bootstrap using the residuals) would yo