Assignment1

Yeonsil Choi

Question 1

Suppose that we have the following 9 observations for random variable X: 12.21, 14.37, 17.18, 11.74, 13.84, 14.26, 15.42, 13.52, 17.97. Test

$$H_0: \mu = 16$$

vs

$$H_1: \mu < 16$$

at

$$\alpha = 0.05$$

```
#The 9 observations are stored in x.
x = c(12.21, 14.37, 17.18, 11.74, 13.84, 14.26, 15.42, 13.52, 17.97)

#Test (H0: true mean = 16) vs (H1: true mean < 16) at alpha = 0.05 (one-
sided)
sample_mean = mean(x) #x_bar
sample_mean

## [1] 14.50111

sample_sd = sd(x) #s
sample_sd

## [1] 2.073701

#Obtain a test statistic
t_stat = (sample_mean - 16)/(sample_sd/sqrt(9)) # (x_bar-mu0)/(s/sqrt(n))
t_stat

## [1] -2.168426

#Calculate the p_value
p_val = pt(t_stat,df=8)
p_val

## [1] 0.03098519

p_val < 0.05

## [1] TRUE
```

```
#Since p_val < alpha = 0.05 we reject H0.

#We can use a default function in R for the same test
t.test(x = x, mu = 16, alternative = c("less"), conf.level = 0.95)

##
##  One Sample t-test
##
## data:  x
## t = -2.1684, df = 8, p-value = 0.03099
## alternative hypothesis: true mean is less than 16
## 95 percent confidence interval:
##      -Inf 15.78649
## sample estimates:
## mean of x
##  14.50111
```

Therefore, we reject H0.

Question 2

Consider the SLR model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where

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

. We have obtained the following data results from 10 observations (i.e. n = 10):

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = -2022$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 102$$

$$\bar{x} = 5$$

$$\bar{y} = -90$$

(a) Find the LS estimates of

$$\beta_0$$

and

$$\beta_1$$

.

```
#the summation of (x_i - x_bar)(y_i - y_bar) with i = 1, ... n is -2022 and
#the summation of (x_i - x_bar)^2 with i = 1, ... n is 102
#Since beta_1_hat is the summation of (x_i - x_bar)(y_i - y_bar) with i = 1,
... n over
#the summation of (x_i - x_bar)^2 with i = 1, ... n, we can obtain beta_1_hat
beta_1_hat = -2022/102
beta_1_hat

## [1] -19.82353

#and beta_0_hat is y_bar - beta_1_hat * x_bar
#All the information is already given, so we can simply plug them in to the
equation.
x_bar = 5
y_bar = -90
beta_0_hat = y_bar - beta_1_hat * x_bar
beta_0_hat

## [1] 9.117647

##beta_0_hat and beta_1_hat are LS estimates of beta_0 and beta_1
```

(b) Using the estimates, obtain the fitted value of y at x = 3

```
x = 3
y_hat = beta_0_hat + beta_1_hat * x
y_hat

## [1] -50.35294
```

(c) Suppose that

$$\sum_{i=1}^{n}(y_i - \widehat{\beta_0} - \widehat{\beta_1}x_i)^2 = 47.13$$

(d) Obtain an unbiased estimate of

$$\sigma^2$$

```
#Since the summation of e_i^2/(n-2) is unbiased estimator of sigma^2, we can
simply plug them in to the equation.
n = 10
unbiased_esm = 47.13/(n-2)
unbiased_esm

## [1] 5.89125
```

(d) Construct a 95% confidence interval for E(Y|x = 3)

```
#Since we already have all the values, we can simply plug them in to the
equation.
cv = qt(0.975, n - 2) #critical value of t
cv

## [1] 2.306004
```

```r
#abs(qt(p = 0.05 / 2, df = n - 2))

sigma_hat = sqrt(unbiased_esm)
sigma_hat

## [1] 2.42719

##95% confidence interval for E(Y|x = 3)

c(y_hat - (cv * sigma_hat * sqrt((1/n) + (3 - x_bar)^2/102)), y_hat + (cv *
sigma_hat * sqrt((1/n) + (3 - x_bar)^2/102)))

## [1] -52.44131 -48.26457

#c(y_hat - (cv *  unbiased_esm* (1/n + 4/102)), y_hat + (cv * unbiased_esm*
(1/n + 4/102)))
```

Question 3

Consider the SLR model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where

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

.

(a) Show that the fitted regression line (

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

(b)) passes through the point (

$$\bar{x}, \bar{y}$$

(c)).

When x =

$$\bar{x}$$

,

$$y = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Since

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

, plug it in to the equation.

$$y = (\bar{y} - \widehat{\beta_1}\bar{x}) + \hat{\beta}_1 \bar{x} = \bar{y}$$

Therefore, the fitted regression line passes through the point

$$(\bar{x}, \bar{y})$$

.

(b) Show that SST = SSR + SSE.

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

(c) Let

$$y_i - \hat{y}_i = A$$

(d) and

$$\hat{y}_i - \bar{y} = B$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (A + B)^2 = \sum_{i=1}^{n} A^2 + \sum_{i=1}^{n} B^2 + 2\sum_{i=1}^{n} A\,B$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

$$2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 2\sum_{i=1}^{n}(y_i\hat{y}_i - \hat{y}_i^{\,2} - y_i\bar{y} + \bar{y}\hat{y}_i) = 2\sum_{i=1}^{n}(\hat{y}_i(y_i - \hat{y}_i) - \bar{y}(y_i - \hat{y}_i))$$

(e) Note that

$$\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$$
$$\sum e_i = \sum (y_i - \hat{y}_i) = 0$$
$$\sum x_i e_i = 0$$

(f) Plugging them to the above equation.

$$2\sum_{i=1}^{n}(\hat{y}_i(y_i - \hat{y}_i) - \bar{y}(y_i - \hat{y}_i)) = 2\sum_{i=1}^{n}(\hat{y}_i e_i - \bar{y} e_i) = 2\sum_{i=1}^{n}[\,(\beta_0 + \widehat{\beta_i x_i})e_i - \bar{y} e_i]$$

$$= 2[\widehat{\beta_1}\sum_{i=1}^{n} x_i\, e_i + \widehat{\beta_0}\sum_{i=1}^{n} e_i - \bar{y}\sum_{i=1}^{n} e_i] = 0$$

So,

$$2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

Therefore,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Question 4

For question 4, you will use an imported data. Run the following R code (in blue) and use the hw1 data data set to answer the questions. R codes: hw1 data = read.csv("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw1 data1.csv") The imported data set contains 100 observations with 2 variables: x1 and x2. Include your R codes and output for the following questions.

```
hw1_data =
read.csv("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw1_data1.c
sv")
#(a) Count the number of observations whose x1 are greater than 6
nrow(hw1_data[hw1_data$x1 > 6, ])

## [1] 26

#(b) Count the number of observations whose x1 are greater than 6 and x2
equal to H
nrow(hw1_data[hw1_data$x1 > 6 & hw1_data$x2 == "H", ])

## [1] 23

#(c) Consider a subset A that contains all observations with x2 = H.
#Compute the mean, median and standard deviation of the x1 values in subset
A.
subset.A <- hw1_data[which(hw1_data$x2 == 'H'), ]
subset.A

##           x1 x2
## 2   2.864985  H
## 3   8.158611  H
## 5   9.916887  H
## 6   8.380810  H
## 8   4.568047  H
## 9   4.761942  H
## 10  7.698870  H
## 11  4.737382  H
## 13  7.830424  H
## 16  7.483162  H
## 17  5.694272  H
## 18  5.810317  H
## 21  5.880679  H
## 23  2.177601  H
## 26  9.343799  H
## 27  5.198466  H
## 28  6.277103  H
## 29  4.308971  H
## 32  4.133553  H
## 33  5.612404  H
## 34  3.863486  H
## 37  6.961105  H
```

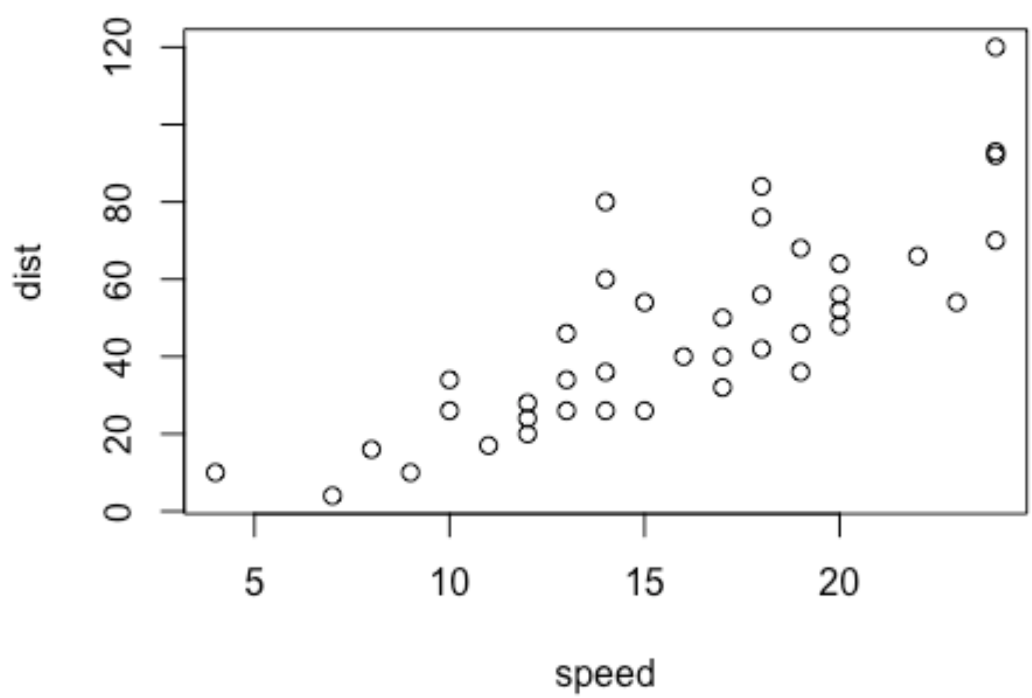
```
## 40  6.810860  H
## 41  5.724392  H
## 42  4.730367  H
## 43  5.018466  H
## 44  5.617432  H
## 45  5.440828  H
## 46  1.563948  H
## 47  3.360676  H
## 48  5.684439  H
## 49  5.877080  H
## 50  5.361127  H
## 51  4.179256  H
## 52  5.364444  H
## 54  3.751353  H
## 55  5.066975  H
## 56  7.742642  H
## 57  4.224449  H
## 61  2.778969  H
## 62  7.263755  H
## 64  7.702123  H
## 66  5.428301  H
## 69  4.683760  H
## 72  3.405323  H
## 75  5.438166  H
## 76  8.170413  H
## 79  7.892661  H
## 80  5.536360  H
## 83  7.552486  H
## 87  6.616907  H
## 88  6.032655  H
## 89  7.921882  H
## 90  7.263723  H
## 93  7.317767  H
## 96  6.217891  H
## 100 8.071647  H

#summary(subset.A$x1)
mean(subset.A$x1)

## [1] 5.832919

median(subset.A$x1)

## [1] 5.684439

sd(subset.A$x1)

## [1] 1.790704

#(d) The sample mean of x1 is 4.435. Can we argue that the true mean of x1
differs from 4?
```

```r
#Conduct a t-test at significance level = 0.05
result = t.test(hw1_data$x1, mu = 4)
result

##
##  One Sample t-test
##
## data:  hw1_data$x1
## t = 1.7192, df = 99, p-value = 0.08871
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  3.932958 4.936674
## sample estimates:
## mean of x
##  4.434816

#names(result)
result$statistic

##        t
## 1.719151

result$p.value

## [1] 0.08871225

result$p.value < 0.05

## [1] FALSE

#Since p-value > 0.05, we do not reject H0.
#We can argue that the true mean of x1 does not differ from 4.

#(e) Consider the statement: "Given that x2 equals to H, the true mean of x1
is larger than 4."
#Is this statement convincing? Use a t-test (alpha = 0.05).
result2 = t.test(subset.A$x1, alternative = "greater", mu = 4)
result2

##
##  One Sample t-test
##
## data:  subset.A$x1
## t = 7.7278, df = 56, p-value = 1.086e-10
## alternative hypothesis: true mean is greater than 4
## 95 percent confidence interval:
##  5.436223      Inf
## sample estimates:
## mean of x
##  5.832919

result2$p.value
```

```
## [1] 1.085591e-10

result2$p.value < 0.05

## [1] TRUE

#Since p-value < 0.05, we reject H0.
```

Question 5

```
set.seed (20)
idx = sample(nrow(cars), 40, replace=FALSE)
cars2 = cars[idx, ]

#(a) Make a scatterplot that shows the relationship between x and Y
plot(dist~speed, data=cars2) #plotting the simulation data
```




```
##It reveals a positive correlation between them

#(b)
x = cars2$speed
x
```

```
##  [1] 19 20  4 18 17 17 20  7 12 10 10 14 13  9 14 20 24 15 18 14  8 24 13
22 12
## [26] 16 18 13 23 20 15 17 11 14 24 19 12 19 24 18
```

```
y = cars2$dist
y
```

```
##  [1]  68  64  10  76  32  40  52   4  24  26  34  36  34  10  26  48  70
26  42
## [20]  60  16  93  46  66  20  40  84  26  54  56  54  50  17  80  92  46
28  36
## [39] 120  56
```

```
Sxy = sum((x - mean(x)) * (y - mean(y)))
Sxy
```

```
## [1] 3992.65
```

```
Sxx = sum((x - mean(x)) ^ 2)
Sxx
```

```
## [1] 978.775
```

```
# beta parameter estimation
beta_1_hat = Sxy / Sxx
beta_0_hat = mean(y) - beta_1_hat * mean(x)
c(beta_0_hat, beta_1_hat) #LS estimates for beta_0 and beta_1
```

```
## [1] -18.411765   4.079232
```

```
unbiased_estimate = sum((y - beta_0_hat - beta_1_hat * x)^2)/(nrow(cars2)-2)
#unbiased estimate for sigma^2
unbiased_estimate
```

```
## [1] 245.2357
```

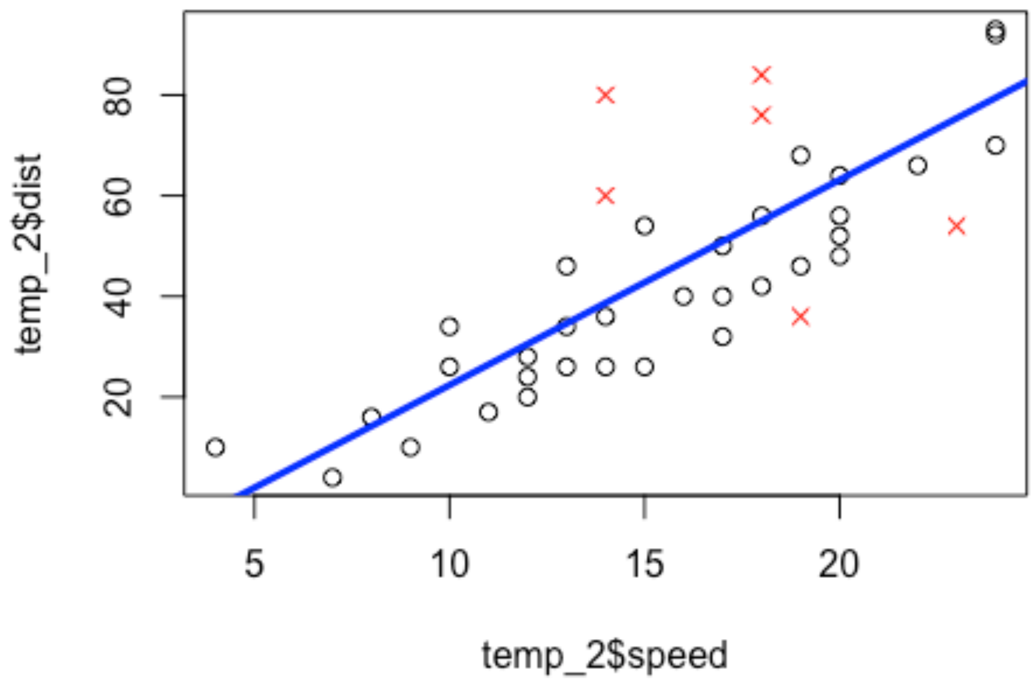
```
#We can also obtain an unbiased estimate using lm function
cars_lm = lm(dist~speed,data=cars2)
cars_lm
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars2)
##
## Coefficients:
## (Intercept)        speed
##     -18.412        4.079
```

```
# The summary function gives summary information of the lm object
# "Residual standard error" represents sigma_hat (not sigma_hat^2)
# sigma_hat^2 = sum(e^2)/(n-2) = (15.66)^2 = 245.2356
summary(cars_lm)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.094 -10.638  -4.014  11.263  41.303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.4118     8.3470  -2.206   0.0335 *
## speed         4.0792     0.5006   8.149 7.26e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.66 on 38 degrees of freedom
## Multiple R-squared:  0.6361, Adjusted R-squared:  0.6265
## F-statistic: 66.41 on 1 and 38 DF,  p-value: 7.26e-10

#(c) Using the estimates, calculate the residuals e_4, e_7 and e_10
cars2$dist[4] - (beta_0_hat + beta_1_hat*cars2$speed[4])

## [1] 20.98559

cars2$dist[7] - (beta_0_hat + beta_1_hat*cars2$speed[7])

## [1] -11.17287

cars2$dist[10] - (beta_0_hat + beta_1_hat*cars2$speed[10])

## [1] 3.619448

#We can also obtain them using lm function
cars_lm$residuals[4]

##       34
## 20.98559

cars_lm$residuals[7]

##        41
## -11.17287

cars_lm$residuals[10]

##        8
## 3.619448

#(d)Find the residuals whose absolute values are greater than 20.
#Indicate those residuals in the scatterplot with different a color and
shape.
residual.set = cars_lm$residuals[which(abs(cars_lm$residuals)>20)]
residual.set
```

```
##        34        22        35        45        23        36        49
##  20.98559  21.30252  28.98559 -21.41056  41.30252 -23.09364  40.51020

#class(residual.set)
#class(cars2)

temp = cbind(cars2, cars_lm$residuals)
temp_2 = cbind(cars2, cars_lm$residuals)
colnames(temp) = c("speed", "dist", "residual")
colnames(temp_2) = c("speed", "dist", "residual")

temp = subset(temp, abs(temp$residual) > 20)
temp_2 = subset(temp_2, abs(temp_2$residual)<=20)

temp = temp[,1:2]
temp_2 = temp_2[,1:2]

plot(temp_2$speed, temp_2$dist)
points(temp$speed, temp$dist, col = "red", pch = 4)


#(e)
# Fitted values
#cars_lm$fitted.values # same as fitted(cars_lm)

# Adds the fitted line to the current plot
abline(cars_lm, lwd = 3, col = "blue")
```




```
# Predict the distance taken to stop when the speed of the car is 17
predict(cars_lm, newdata=data.frame(speed=17))

##        1
## 50.93517

#(f) State the goodness of fit for the fitted model.
#What percentage of the variation in the response variable is explained by
the fitted model?

#compute R^2 of the fitted model for the dataset.
summary(cars_lm)

##
## Call:
## lm(formula = dist ~ speed, data = cars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.094 -10.638  -4.014  11.263  41.303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.4118     8.3470  -2.206   0.0335 *
```

```
## speed         4.0792     0.5006   8.149 7.26e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.66 on 38 degrees of freedom
## Multiple R-squared:  0.6361, Adjusted R-squared:  0.6265
## F-statistic: 66.41 on 1 and 38 DF,  p-value: 7.26e-10

summary(cars_lm)$r.squared #print R^2 from the lm object

## [1] 0.6360622

# We can calculate R^2 using our own codes.
y = cars2$dist # actual values of y
y_hat = cars_lm$fitted.values #fitted values of y
SST   = sum((y - mean(y)) ^ 2)
SSR = sum((y_hat - mean(y)) ^ 2)
SSE   = sum((y - y_hat) ^ 2)

SSR/SST #R_squared

## [1] 0.6360622

1-SSE/SST #Same

## [1] 0.6360622

SSR/SST * 100 #of the variation in the response variable is explained by the
fitted model

## [1] 63.60622

#(g) Consider the statement:
#"If someone is driving at 100mph, according to the fitted model, the
distance taken to stop will be exactly 389.5114ft."
predict(cars_lm, newdata=data.frame(speed=100))

##        1
## 389.5114

#The fitted regression model is only valid for the range of the predictors.
Since speed 100 mph is far beyond
#the range of the observed speed values, the current model should not be used
for prediction of the car. In addition,
#even if the speed of the car is within the range, the actual distance will
not be exactly the same as the predicted value.

#(h) Construct a 90% confidence interval for beta_1
confint(cars_lm, level = 0.9)[2, ]

##      5 %     95 %
## 3.235322 4.923142
```

```r
#(i) Construct a 95% confidence interval for E(Y|x = 15)
# Confidence interval for the mean response at speed = 15
predict(cars_lm, newdata = data.frame(speed = 15), interval =
"confidence",level = 0.95)
```

```
##        fit     lwr      upr
## 1 42.77671 37.6773 47.87612
```

```r
#(j) Is the linear relationship between Y and x significant?
summary(cars_lm)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.094 -10.638  -4.014  11.263  41.303
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.4118     8.3470  -2.206   0.0335 *
## speed         4.0792     0.5006   8.149 7.26e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.66 on 38 degrees of freedom
## Multiple R-squared:  0.6361, Adjusted R-squared:  0.6265
## F-statistic: 66.41 on 1 and 38 DF,  p-value: 7.26e-10
```

```r
#Since the p-value is much less than 0.05, we reject the null hypothesis that
beta_1 = 0.
#Hence there is a significant relationship between the variables in the
linear regression model of the data set.


#(k) Test H0: beta_1 = 5 vs H1: beta_1 < 5 at alpha = 0.05 (one sided)
se_beta_hat_1 = sqrt(unbiased_estimate)/sqrt(sum((cars2$speed -
mean(cars2$speed))^2))
t.stats = (beta_1_hat - 5)/se_beta_hat_1
t.stats
```

```
## [1] -1.839501
```

```r
#summary(cars_lm)
#summary(cars_lm)$coefficients[2,2]

#Calculate the p_value
p_val = pt(t.stats,df=38)
p_val
```

```
## [1] 0.03683199
```

```r
p_val < 0.05
```

```
## [1] TRUE
```

```r
#Since p_val < alpha = 0.05 we reject H0.

#Compare t_stats with the critical value
cv = qt(0.975,8) # Gives t_value at which the cdf (left side) becomes 0.975
cv
```

```
## [1] 2.306004
```

```r
t.stats > cv  #FALSE means t.stats was smaller than the critical value -> No
evidence against H0
```

```
## [1] FALSE
```