SS3859A Assignment 1

Bradley Assaly-Nesrallah

17/09/2020

Question 1

a.We compute the LS estimates of B0 and B1 as follows:

SSxy = sum_xy - (sum_x $*$ sum_y)/n = 25825 - (517$*$346)/14 = 13047.71

SSxx = sum_x^2 - $(\text{sum\_x})^2$/n = 39095 - $(517)^2$/14 = 20002.92

SSyy = sum_^{2} - $(\text{sum\_y})^2$/n = 17354 - 346^2/14 = 8802.86

B1_hat_LS= SSxy/SSxx= 13047.71/20002.92 =0.652

B0_hat_LS= sum_y/n - b1_hat$*$sum_x/n = 346/14 - 0.652$*$517/14 =0.637

Thus the LS estimates of B0 and B1 are 0.637 and 0.652 respectively.

b.Using the LS estimates we obtain the fitted value of x =120 as follows:

y_hat = B0_hat_LS + b1_hat_LS * x = 0.637 + 0.652$*$120 = 78.877

Thus the fitted value of y at x=120 is 78.877

c.We compute an unbiased estimate of sigma^2 as follows

We know that the LS estimate is unbiased with sigma^2_hat=sum_ei^2/(n-2)

So sigma^2_hat=sum_ei^2/(n-2)= 391.8257/12=32.65

Thus an unbiased estimate of sigma^2 is 32.65

d.We compute the proportion of observed variation in y explained by the linear relationship between the two values by computing the R^2 Value

We know that R^2 (Note: SS values obtained from part a) = SSxy^2 2/SSxx $*$SSyy=13047.71^2 / (20002.92 $*$ 8802.86) = 0.9668

Thus there is a 0.9668 proportion of observed variation in y explained by the linear relationship between variables.

e.We compute a 95% CI for E(Y|x=120) as follows

the 95% CI is given by E(Y|x=120)+- t0.025,12$*$sqrt(sigma^2_hat)$*$sqrt(1/n+$(120\text{-sum_x/n})^2$/SSxx) = 78.877

+- 2.179$*$sqrt(32.65)$*$sqrt(1/14+$(120\text{-}517/14)^2$/20002.92)= 78.877 =- 8.0346 = (70.842,86.912)

Thus a 95% CI for E(Y|x=120) is given by (70.842,86.912)

Question 2

a.We show that the regression line passes through the point (x_bar,y_bar) as follows:

y_bar = sum_yi/n = 1/n∗sum(yi_hat+eps_i_hat) = 1/n∗(sum(y_hat)+sum(eps_i_hat))=

1/n∗(sum(B0_hat+b1_hat∗)+0)=1/n∗(nB0_hat+nB1_hat_x∗xi)

=B0_hat+B1_hat∗xi)=B0_hat + B1_hat∗x_bar

So y_bar=B0_hat+B1_hat∗x_bar, hence the regression line goes through

the point (x_bar,y_bar) as required, so we are done.

b.We show that SST=SSE+SSR as follows:

SST=$\sum_{i=1}^{n}$(yi-y_bar)^2=$\sum_{i=1}^{n}$(yi-y_hat+y_hat-y_bar)^2

=$\sum_{i=1}^{n}$(yi-yhat)^2+2$\sum_{i=1}^{n}$(yi-y_hat)(y_hat-y_bar)+$\sum_{i=1}^{n}$(yi-y_bar)^2

Note that $\sum_{i=1}^{n}$(yi-yhat)^2 = SSE and $\sum_{i=1}^{n}$(yi-y_bar)^2=SSR so we must just prove that the term 2$\sum_{i=1}^{n}$(yi-y_hat)(y_hat-y_bar)=0,

$\sum_{i=1}^{n}$(yi-y_hat)(y_hat-y_bar)=$\sum_{i=1}^{n}$(yi-y_hat)-y_bar$\sum_{i=1}^{n}$(yi-y_hat), and since $\sum_{i=1}^{n}$(yi-y_hat)=$\sum_{i=1}^{n}$(eps_i)=0 the whole term is equal to zero,

Thus SST=$\sum_{i=1}^{n}$(yi-y_bar)^2=$\sum_{i=1}^{n}$(yi-yhat)^2+$\sum_{i=1}^{n}$(yi-y_bar)^2 = SSE + SSR as required so we are done.

Question 3

```r
hw1_data=read.csv("https://raw.githubusercontent.com/hgweon2/ss3859/master/hw1_data1.csv")

#a.We count the observations with x1<4 as follows:
newdata=hw1_data[which(hw1_data$x1<4),]
nrow(newdata)
```

```
## [1] 40
```

```r
#thus there are 40 obs with x1<4

#b.We compute
newdata2=hw1_data[which(hw1_data$x1<4 & hw1_data$x2=='L'),]
nrow(newdata2)
```

```
## [1] 32
```

```r
#So we have 32 obs with x1<4 and x2==L

#c.We create a subset A with x2==L and find the mean, median and std of
#the x1 vals as follows
newdata3=hw1_data[which(hw1_data$x2=='L'),]
mean(newdata3$x1)
```

```
## [1] 2.581517
```

```r
median(newdata3$x1)
```

```
## [1] 2.567377
```

```r
sd(newdata3$x1)
```

```
## [1] 2.152698
```

```r
#Thus the mean,median,sd are 2.581517,2.567377,2.152698 respectively
mean(hw1_data$x1)
```

```
## [1] 4.434816
```

```r
#d.We test H0: u=4 v Ha:u!=4 at alpha=0.05 with a t test as follows
n=nrow(hw1_data)
sample_mean = mean(hw1_data$x1) # x_bar
sample_sd = sd(hw1_data$x1) # s
t_stat = (sample_mean - 4)/(sample_sd/sqrt(n))
t_stat # The t stat is 1.719151
```

```
## [1] 1.719151
```

```r
A = 1 - pt(abs(t_stat),df=n-1)
p_val = 2*A
p_val # the p-value is computed to be 0.08871225
```

```
## [1] 0.08871225
```

```r
#We do not reject H0 as p_value=0.08871225>0.05=alpha thus there is not statistical
#evidence to claim u is not equal to 4
newdata3=hw1_data[which(hw1_data$x2=='L'),]
n=nrow(newdata3)
sample_mean = mean(newdata3$x1) # x_bar
sample_sd = sd(newdata3$x1) # s
t_stat = (sample_mean - 4)/(sample_sd/sqrt(n))
t_stat # The t stat is -4.32
```

```
## [1] -4.320911
```

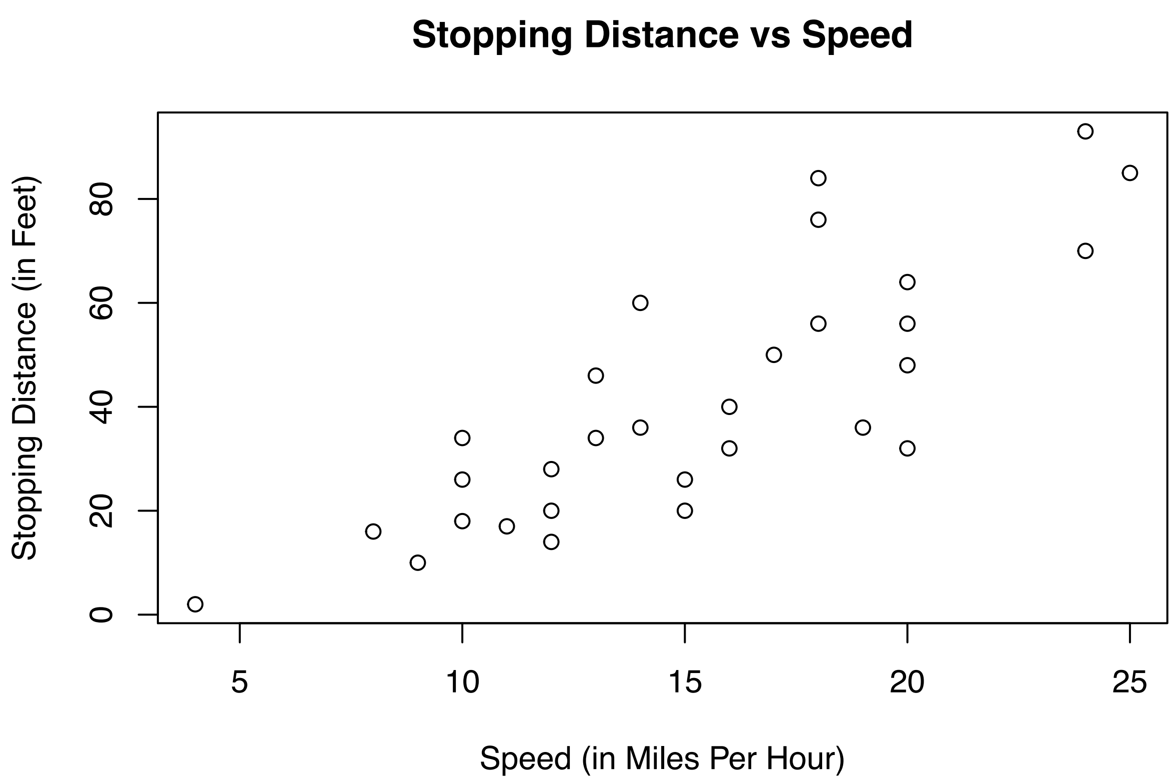
```r
A = 1 - pt(abs(t_stat),df=n-1)
p_val = 2*A
p_val #p -value is 9.319577e-05
```

```
## [1] 9.319577e-05
```

```r
#e.e do reject H0 as p_value=9.319577e-05<0.05=alpha thus there is
#statistical evidence to claim u is not equal to 4 when x2==L
```

```r
set.seed (10)
idx=sample(nrow(cars),30,replace=FALSE)
cars2=cars[idx,]

#a.We plot the relationship between x and Y as follows
plot(dist~speed,data=cars2,
     xlab = "Speed (in Miles Per Hour)",
     ylab = "Stopping Distance (in Feet)",
     main = "Stopping Distance vs Speed")
```




```r
#There appears to be a linear relationship between x and Y from the scatterplot

#b.We obtain the LS estimates for B0 and B1 as follows
cars2_lm=lm(dist~speed, data=cars2)
#Thus the LS estimates are B0_hat=8.162 and b1_hat=0.1726

#c.We obtain epsilon
resid(cars2_lm)[5] #5th residual
```

```
##        8
## 5.451101
```

```r
resid(cars2_lm)[10] #10th residual
```

```
##        42
```

```
## -3.563742
```

```r
resid(cars2_lm)[20] #20th residual
```

```
##       33
## 4.239227
```

```r
#d.We find and plot the residuals with value greater than 20 on the s
#Scatterplot with a different color and shape
idx=which(abs(resid(cars2_lm))>20)
col_idx = rep(2,nrow(cars2))
pch_idx = rep(2,nrow(cars2))
col_idx[idx] = 3
pch_idx[idx] = 3
plot(dist~speed,data=cars2,
     xlab = "Speed (in Miles Per Hour)",
     ylab = "Stopping Distance (in Feet)",
     main = "Stopping Distance vs Speed",
     col= col_idx,
     pch=pch_idx)

#e.We compute the sum of the residuals as follows:
sum(resid(cars2_lm))
```

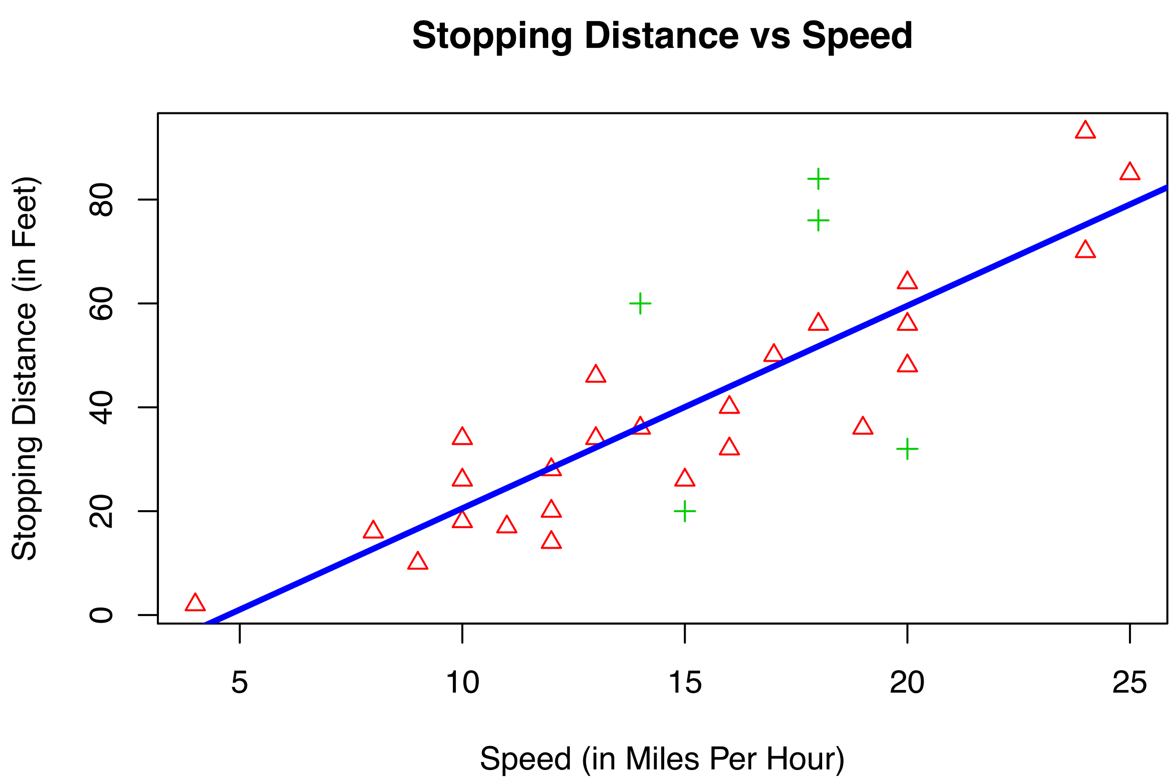
```
## [1] -5.662137e-15
```

```r
#the sum of the residuals is equal to -5.662e-15

#f.We report the fitted model and add the fitted regression line to the scatterplot,
#and predict when speed=21 using the fitted
summary(cars2_lm)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.564  -8.126  -0.253   5.303  32.239
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.4659     8.2289  -2.244   0.0329 *
## speed         3.9015     0.5135   7.598 2.82e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14 on 28 degrees of freedom
## Multiple R-squared:  0.6734, Adjusted R-squared:  0.6617
## F-statistic: 57.73 on 1 and 28 DF,  p-value: 2.816e-08
```

```r
abline(cars2_lm, lwd = 3, col = "blue")
```




```r
predict(cars2_lm, newdata=data.frame(speed=21))
```

```
##        1
## 63.46523
```

```r
#The predicted distance when speed is 21 is 63.46523

#g.We state the goodness of fit of the model
summary(cars2_lm)$r.squared
```

```
## [1] 0.6734058
```

```r
#thus the R^2 value is 0.6734058 so 67.34058% of the variation is explained
#by the fitted model

#h.The model is based on data with values up to a speed of 25, so our model
#cannot predict for values outside of the range,
#meanwhile the claim predicts an exact value with certainty. For a linear
#model outside of
#the range of data we cannot be certain if the linear relationship persists
#beyond the range
```

```r
#of values used to produce the model, hence the claim cannot hold, so we are done.

#i.We obtain a 95% confidence interval for B1 as follows
confint(cars2_lm)
```

```
##                  2.5 %    97.5 %
## (Intercept) -35.322104 -1.609783
## speed         2.849685  4.953284
```

```r
#thus a 95%CI for B1 is (2.849685,4.953284)

#j.We obtain a 90%CI for E(Y|x=21) as follows
new.dat <- data.frame(speed=21)
predict(cars2_lm, newdata = new.dat, interval = 'confidence',level = 0.90)
```

```
##        fit      lwr      upr
## 1 63.46523 56.81107 70.11938
```

```r
#Therefore a 90%CI for E(Y|x=21) is (56.81107,70.11938)
```