Hw1

Philip Sweet

9/21/2020

1

Derive the least square estimators for the coefficients of a simple linear regression

$$Q = \sum_{i=1}^{n} (y_i - (b + b_1 x))^2$$

$$0 = \Sigma y_i - (n b_1 \Sigma x_1 + n b_0)$$

$$\Sigma y_i = n b_0 + b_1 \Sigma x_1$$

$$\Sigma x_i y_i = b_0 \Sigma x_i + b_1 \Sigma x_i^2$$

$$b_1 \Sigma x_i^2 = b_0 \Sigma x_i - \Sigma x_i y_i$$

$$b_1 = \frac{b_0 \Sigma x_i - \Sigma x_i y_i}{\Sigma x_i^2}$$

$$b_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}$$

$$b_0 = \Sigma x_i y_i - b_1 \Sigma x_i^2$$

$$b_0 = \frac{1}{n} (\Sigma y - b \Sigma x)$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$\Sigma y_i = n b_0 + b_1 \Sigma x_i$$

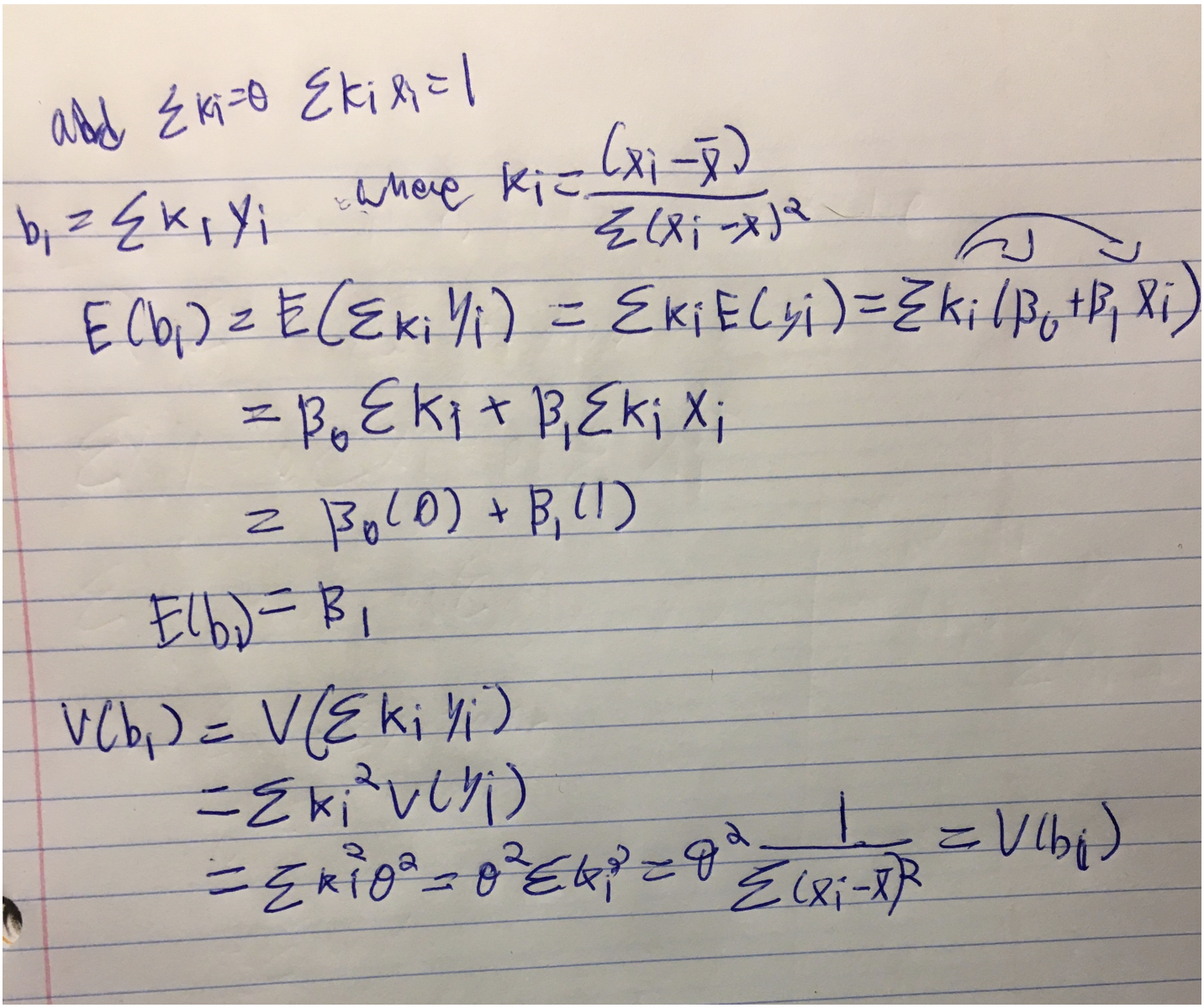
$$\frac{n b_0}{n} = \frac{\Sigma y_i - b_1 \Sigma x_i}{n}$$

$$b_0 = \frac{1}{n} (\Sigma y_i - \Sigma x_i) = \bar{y} - b_1 \bar{x} = b_0$$

Drawing.

2

derive the Expectation and Variance of b1



Drawing.

3

Consider the normal error regression model…

3. $\beta_0 = 200$

$\beta_1 = 5$

$\theta = 4$

$y = \beta_1(x) + \beta_0$

$y = 5(10) + 200$

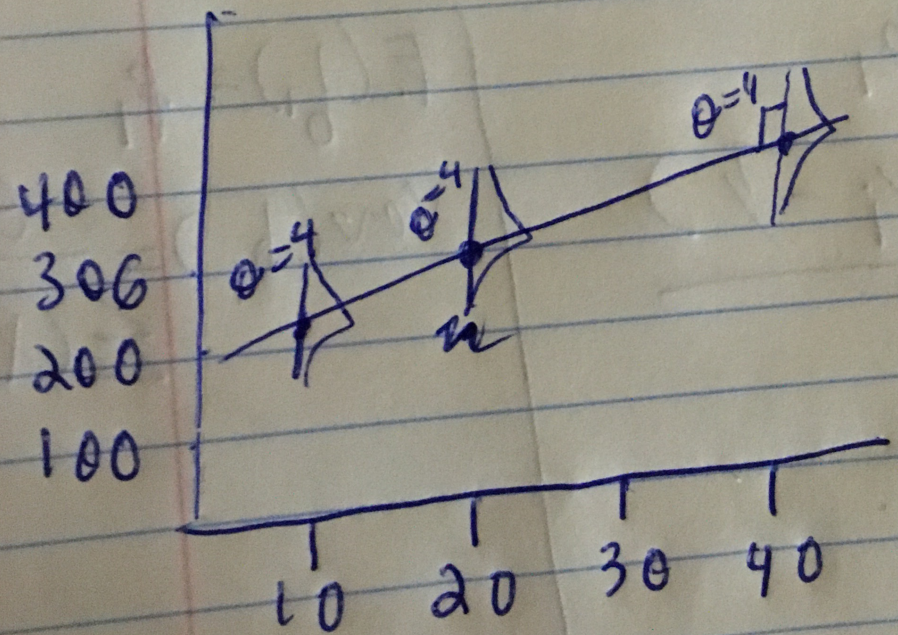
$y = 250$ @ $x = 10$

(x, y)

$(10, 250)$

$(20, 300)$

$(40, 400)$



4

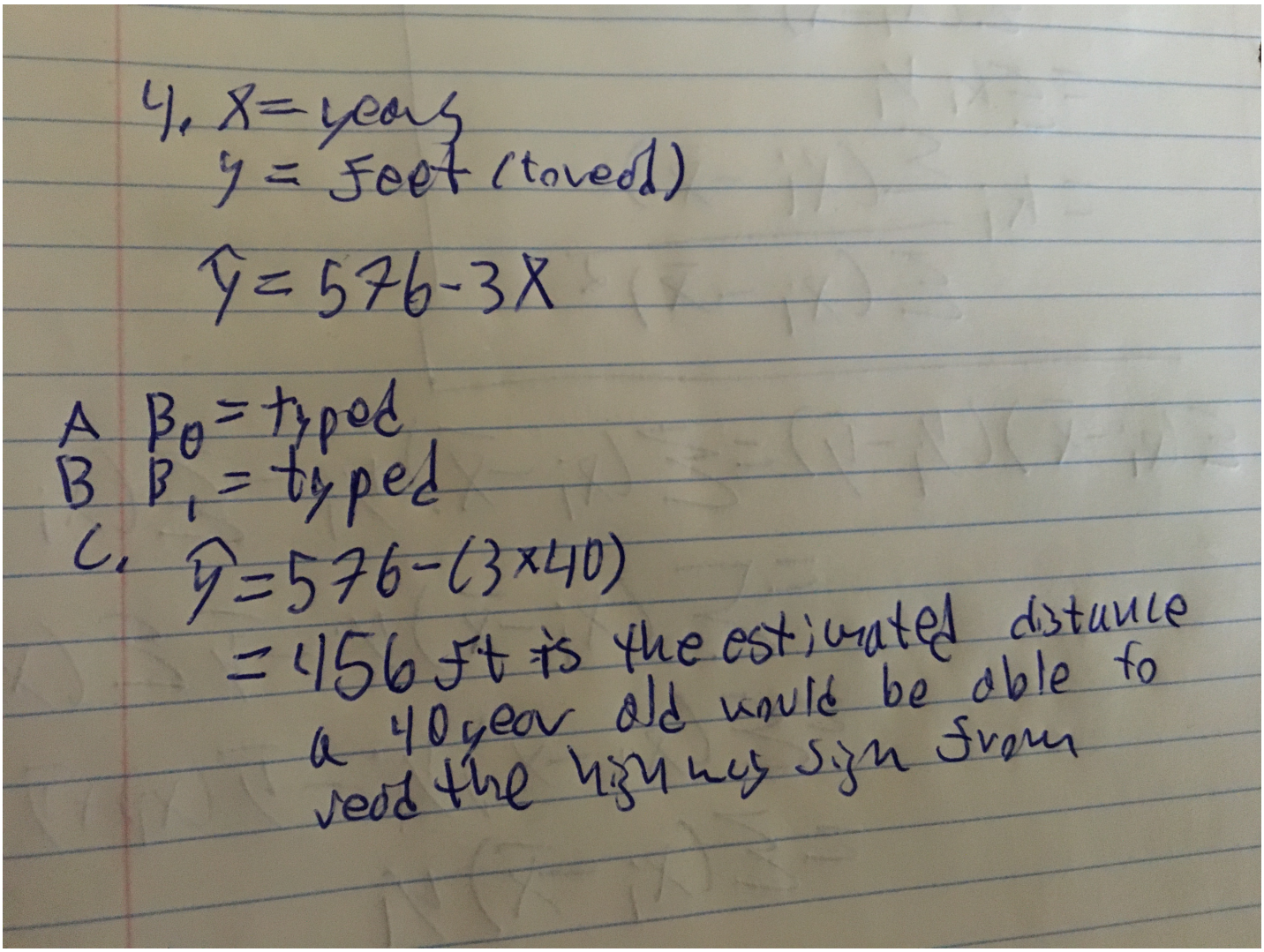
A

In this situation, the B0 would be relatively meaningless because B0 represents how far away a person who was zero year old would be able to read the a highway sign. No one who is zero years old is driving, however, if they were, this model predicts that they would be able to read it from 576 feet away.

B

In this situation, the B1 represents the negative change in distance (in feet) per year of age from which a driver can read a highway sign. B1's value of 3 means that for each unit increase in age, the distance a driver can read a highway sign from decreases

C



Drawing.

D

Residual = 44

E

D was an under-estimate

5

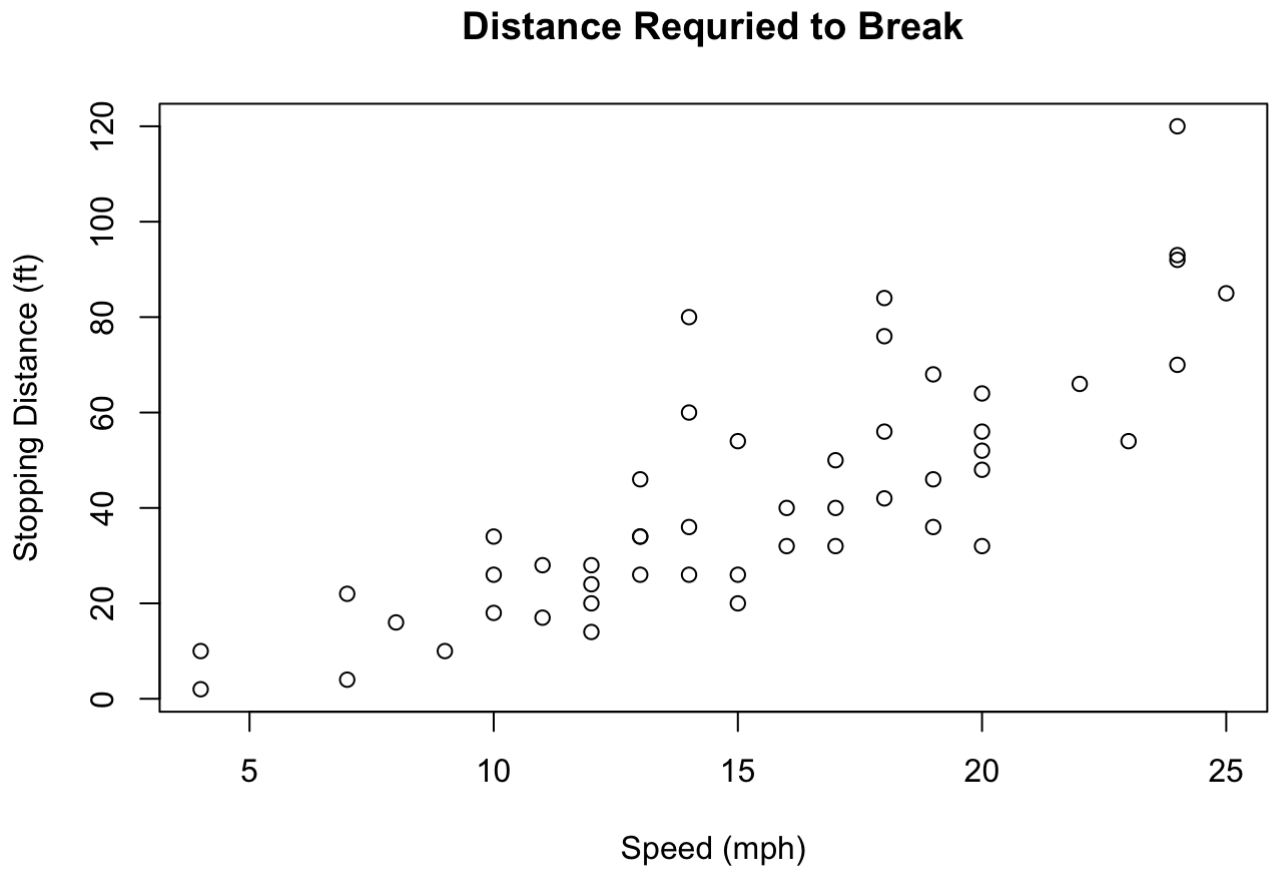
Stop has data on the speed (X, in mph) and stopping distance (Y, in ft) of 50 cars.

```
read.csv("Stop.csv", header = TRUE) -> data
```

A.

In this scatter plot, you can see that there is a positive linear relationship between current speed and breaking distance.

```
plot(data$speed,
     data$dist,
     main="Distance Requried to Break",
    xlab="Speed (mph)",
    ylab="Stopping Distance (ft)")
```



B.

Here we will calculate the sum of squares

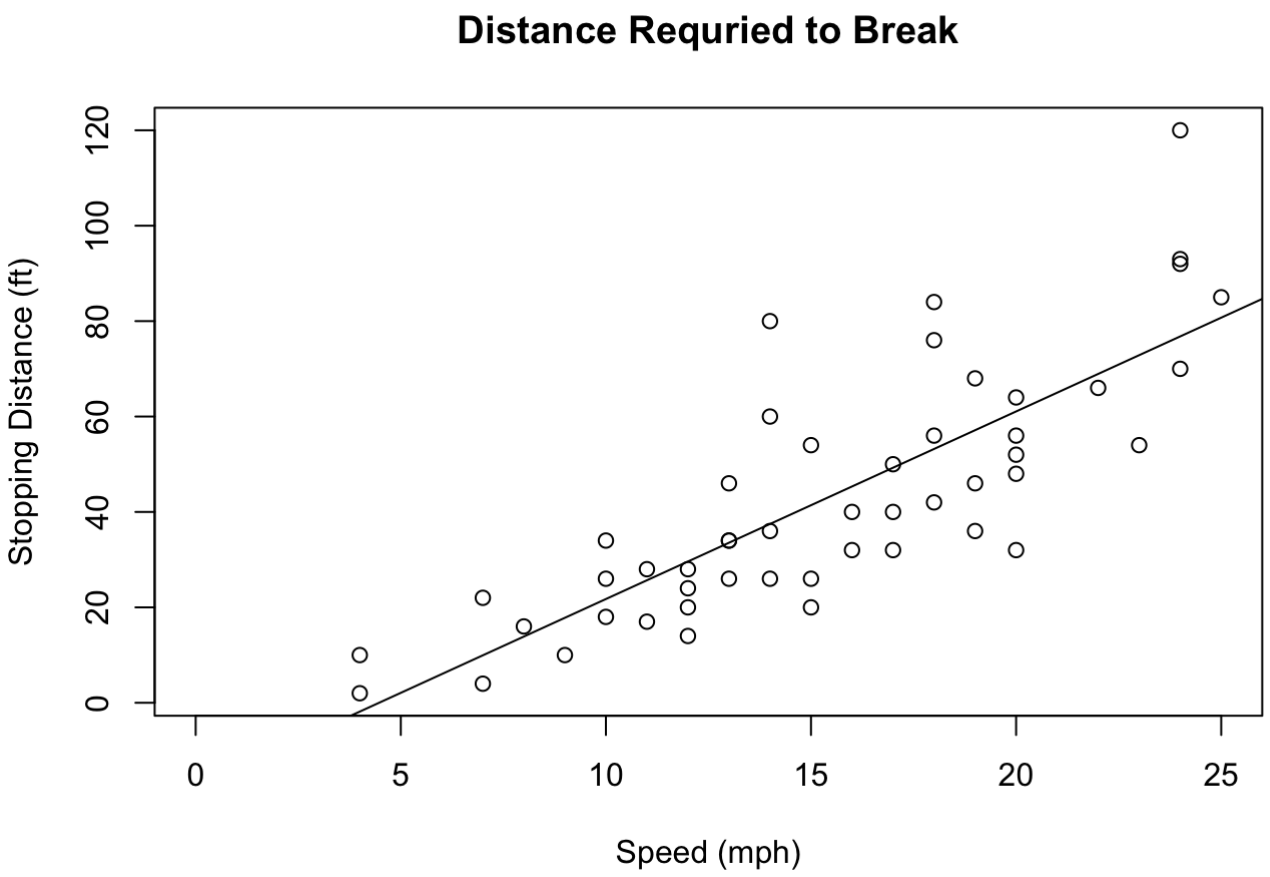
```
n <-length(data)
X <-data$speed
Y <-data$dist

## find the means of both vars
mean_x <-mean(X)
mean_y <-mean(Y)

## find the variance of each var
var_x <-var(X)
var_y <-var(Y)

cov_xy <-cov(X,Y)

# finbd the sum of squares
SS_xx <-(n-1)*var_x
SS_xy <-(n-1)*cov_xy
SS_yy <-(n-1)*var_y

## solve for estimaters
b1 <-SS_xy/SS_xx
b0 <-mean_y -b1*mean_x
yhat <-b0 + b1*X
e <-Y-yhat
SSE <-sum(e^2)
MSE <-SSE/(n-2)
s <-sqrt(MSE)
```

The slope (b1) = 3.9324088 and the intercept (b0) = -17.5790949

Thus the est. regression equation is y = 3.9324088x -17.5790949

C.

```
plot(X,Y,
    xlim=c(0,25),
    main="Distance Requried to Break",
    xlab="Speed (mph)",
    ylab="Stopping Distance (ft)")
abline(a=b0,b=b1)
```



When we lay the regression line overe the data, we can see that line seems to estiamte the stopping distance well at ll speeds provided in the data.

D

When using the linear model function in R (lm) we can see that …

```
lm_a <- lm(Y ~ X)

summary(lm_a)
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## 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 *
## X             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
```

… which (in the Coefficients column) generates the same estimates for B1 and B0 is what we had manually calculated above.

E

In this context the slope (b1) represents an increased spotting distance of 3.9 feet for every extra mile per hout in speed. The intercept (b0) in this case is of no meaning, as a car that was not moving (speed = 0) would require no distance to stop, but it does suggest that the model may be less informative at lower speeds.

F

```
conf <- confint(lm_a, 'X', level=0.95)
```

The 95% confidence interval for the slope is (3.0969643, 4.7678532), suggesting that there is a postive linear relationship since 0 is not within the interval.

G

To a conduct a hypothesis test for a significant linear relationship between starting speed and stopping distance, we can use…

Ho: b1 = 0

Ha: b1 ≠ 0

This test produces a p-value of 1.49e-12 that well below the 0.05 cut off. Thus we can reject the null hypthesis (Ha) that b1 = 0 and state that there is a linear relationship between speed and stopping distance.