M440B HW 13

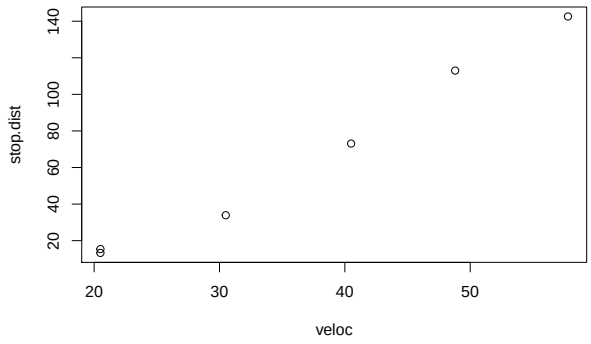
Erick Castillo

4/15/2021

Extra Problem 1: I'll begin by inputting the data:

```
veloc = c(20.5,20.5,30.5,40.5,48.8,57.8)
stop.dist = c(15.4,13.3,33.9,73.1,113,142.6)
```

a. Create a scatterplot of the data.



b. Find the LS line with the formula that has means for terms. Recall the formula is given by

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

where,

$$\hat{\beta}_1 = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2}$$

Below I calculate the terms required to plug into the formula:

```
mean(veloc*stop.dist) #xy-bar
```

```
## [1] 3056.588
```

```
mean(veloc)*mean(stop.dist) #xbar*ybar
```

```
## [1] 2376.061
```

```
mean(veloc^2) #x^2-bar
```

```
## [1] 1522.213
```

```
mean(veloc)^2 #xbar squared
```

```
## [1] 1327.388
```

```
mean(veloc) #xbar
```

```
## [1] 36.43333
```

```
mean(stop.dist) #ybar
```

```
## [1] 65.21667
```

Compiling the above values and plugging them into the formula yields $\hat{y} = 65.217 + 15.567(x - 36.43)$, which simplifies to

$$\hat{y} = 3.495x_i - 62.049$$

c. Find the slope and intercept of the LS line with the other formula discussed in class. Recall the formula is

$$\hat{y} = \bar{y} + r\frac{s_y}{s_x}(x_i - \bar{x})$$

Now, using the S.D's and r that were given and the $\bar{x}$ and $\bar{y}$ values from before, we can construct a LS line of the form (worked out on a separate sheet of paper):

$$\hat{y} = 3.493x_i - 62.047$$

d. Interpret the slope coefficient.

This means that for every one mph increase in velocity, the stopping distance increases by a factor of 3.493 ft.

e. Predict the stopping distances for 25 mph and 75 mph. Which of these predictions is most reliable?

- $\hat{y} = 3.493(25) - 62.047 = 25.28$
- $\hat{y} = 3.493(75) - 62.047 = 199.93$

The most reliable value would be 25, because it's close to the mean value of the velocities recorded. In this case, predicting the stopping distance for a vehicle going 75 mph would be unreliable as it is far from $\bar{x}$, and as discussed in class, extrapolating leads to dubious results.

f. The $\hat{\beta}_0 = -62.047$ does not make sense. It can be interpreted that the stopping distance of an object traveling 0 mph is -62 feet.

g. Compute the residuals and the fitted values.

```
mod1 = lm(stop.dist~veloc)
mod1$residuals
```

```
##          1          2          3          4          5          6
##   5.838639   3.738639 -10.591469  -6.321577   4.586433   2.749335
```

```
mod1$fitted.values
```

```
##          1          2          3          4          5          6
##   9.561361   9.561361  44.491469  79.421577 108.413567 139.850665
```

h. The residual value at 30.5 mph is -10.59. This means that the difference between the fitted value and the actual value is 10.59. The negative sign indicates that the actual value is below the fitted line.

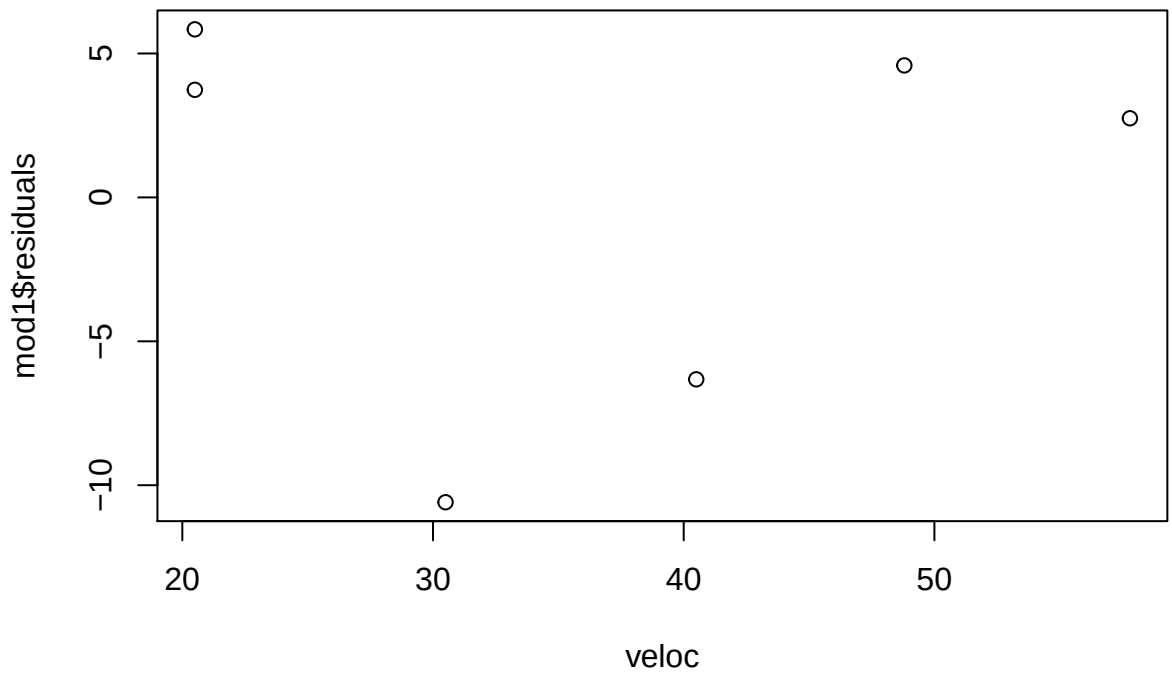
i. Calculate the residual variance.

```
var(stop.dist)*(1-(0.992)^2)
```

```
## [1] 46.18683
```

j. Make the residual plot.

```
plot(mod1$residuals~veloc)
```



Notice that there appears to be a violation of normality as the residuals follow a "parabola" shape.

k. Notice that the response variable is transformed.

```
stop.dist2 = sqrt(stop.dist)
mean(stop.dist2) #ybar of transformed
```

```
## [1] 7.419182
```

Plugging in the known values grants $\hat{y} = 7.419 + 0.997(\frac{3.49}{15.29})(x_i - 36.433)$. Upon simplifying, this grants $\hat{y} = -0.8776 + 0.227x_i$.

l. Repeat parts g, i, and j.

```
#compute the residuals and fitted values.
mod2 = lm(stop.dist2~veloc)
mod2$residuals
```

```
##          1          2          3          4          5          6
##  0.1335061 -0.1438608 -0.2456483  0.2045931  0.3947745 -0.3433646
```

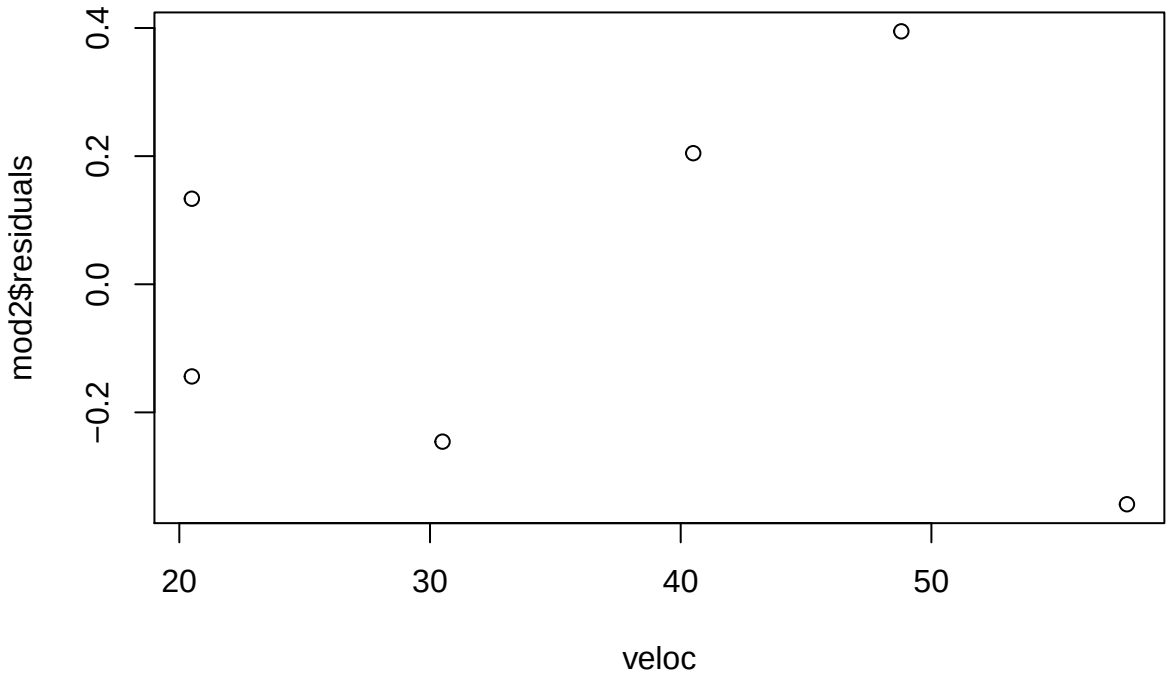
```
mod2$fitted.values
```

```
##         1         2         3         4         5         6
##  3.790777  3.790777  6.068019  8.345261 10.235371 12.284889
```

```
#calculate the residual variance.
var(stop.dist2)*(1-(0.997)^2)
```

```
## [1] 0.07313141
```

```
#construct the residual plot.
plot(mod2$residuals~veloc)
```



Notice that the transformation resulted in a residual that does not appear to have a pattern. The transformation was indeed helpful.

m. Now to test the hypothesis that the true intercept is 0 for the transformed model.

```
summary(mod2)
```

```
##
## Call:
## lm(formula = stop.dist2 ~ veloc)
##
## Residuals:
##       1       2       3       4       5       6
##  0.1335 -0.1439 -0.2456  0.2046  0.3948 -0.3434
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.877568   0.367328  -2.389   0.0752 .
## veloc        0.227724   0.009415  24.188 1.73e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3219 on 4 degrees of freedom
## Multiple R-squared:  0.9932, Adjusted R-squared:  0.9915
## F-statistic:   585 on 1 and 4 DF,  p-value: 1.733e-05
```

It is clear from the above p-value that at a 5% level of significance, we would fail to reject $H_0 : \hat{\beta}_0 = 0$. In other words, there is insufficient evidence to suggest that the intercept value is not zero.

n. Compute the 95% confidence interval for the slope of the LS line of the transformed data.

```
confint(mod2)
```

```
##                   2.5 %    97.5 %
## (Intercept) -1.8974350 0.1422987
## veloc        0.2015842 0.2538642
```

The slope is calculated as accurately as it could be from the given data. The confidence interval for $\hat{\beta}_1$ does not include 0.

Extra Problem 2: This question presents a scenario where a study noticed that the blood pressure of individuals who initially had high values, fell after a period of time where they twiddled their thumbs. Does this demonstrate that the therapy works for patients who have a high BP?

Response: This does not demonstrate that therapy works for patients who have high blood pressure because there may not be a perfect association between the response and the predictor variable. This could be a form of the regression effect in play, where blood pressure tends to go towards the mean population value. If this is the case, we could expect to see individuals with lower blood pressure have an increase after the therapy.

Extra Problem 3: It's given $r = 0.76$, which is the correlation between ft^2 of lots for a house and the price in a neighborhood. If I want to estimate the price of a particular house in this neighborhood, how important is the lot size? What does r^2 mean in this question?

Response: The lot size is an important factor in estimating the price of a house because the correlation value is very high between both these factors. $r^2 = 0.5776$, this is typically referred to as the coefficient of determination. This represents how much of the variation is explained by the model.

Problem 23: A midterm and final exam both had the following characteristics:

- $r = 0.5$ between each of them.
- $\bar{x} = \bar{y} = 75$
- $s_x = s_y = 10$

Where X_i is the set of scores for the midterm and Y_i for the final, $i = \{1, 2, ..., n\}$.

a. If a student's score on the midterm is 95, what would I predict their final score to be?

Using the formula mentioned in part c of the extra problem, we can plug in the given information to construct the following LS line: $\hat{y} = 0.5x_i + 37.5$. Plugging in 95 yields $\hat{y} = 85$. That is, the student will get a score of 85 on their final.

b. If a student's score is 85 on the final, what would the estimated midterm score be?

Notice that both the sets have the same mean and standard deviation. Thus the formula for the LS line with midterm score as the response variable would be similar to the one in part a. It can be expressed as $\hat{x} = 0.5y_i + 37.5$, thus plugging in 85 for y yields $\hat{x} = 80$.