DATA 605 Assignment Week 11

CUNY Spring 2021

Philip Tanofsky

18 April 2021

Prompt

Using the "cars" dataset in R, build a linear model for stopping distance as a function of speed and replicate the analysis of your textbook chapter 3 (visualization, quality evaluation of the model, and residual analysis.)

Initial Data Summary

The dataset contains 50 instances of 2 variables, speed and dist (stopping distance). The data appears to be numeric, containing only positive integer values. The summary function indicates there are no missing values. Speed is the independent variable, and stopping distance represents the dependent variable.

```
dim(cars)
```

```
## [1] 50  2
```

```
head(cars)
```

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

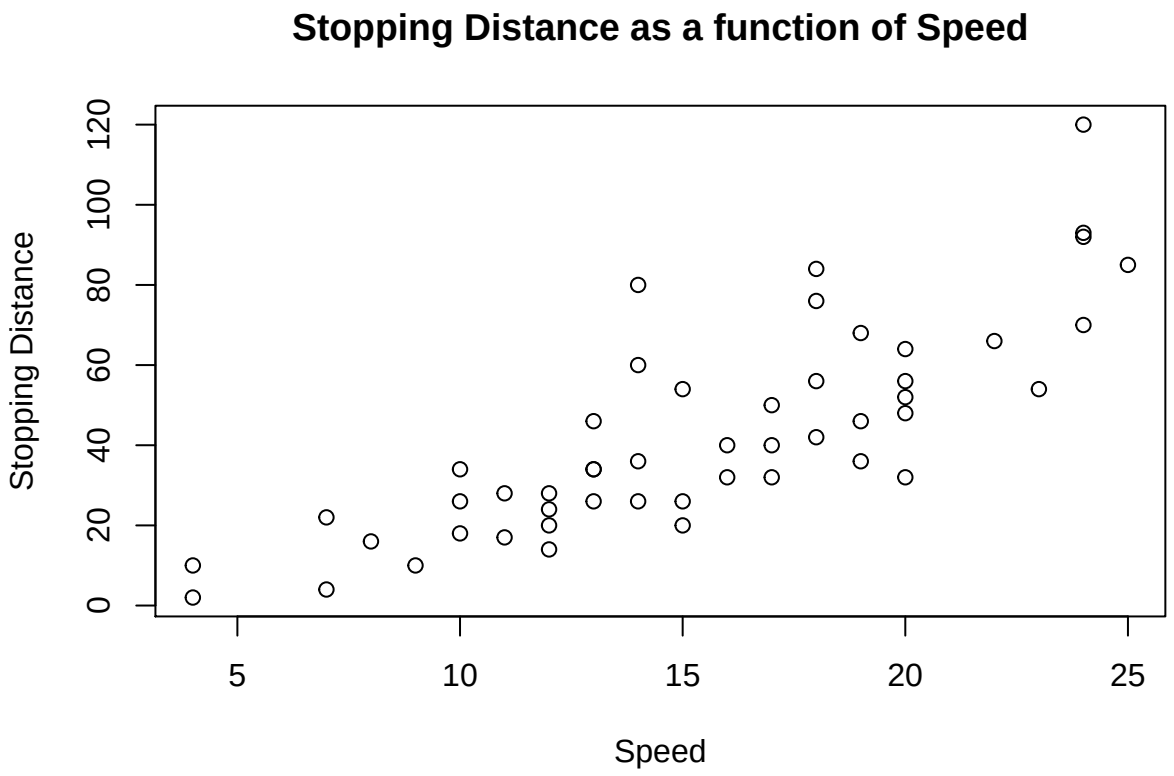
```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Visualize the Data

Initial scatterplot of the stopping distance as a function of speed indicates the stopping distance tends to increase as the speed increases, as is expected. The plot does show the relationship is likely linear.

```
plot(cars$speed,cars$dist, main="Stopping Distance as a function of Speed",
    xlab="Speed", ylab="Stopping Distance")
```



Create Model

To determine the degree of linearity in the relationship between the independent variable speed and the dependent variable stopping distance, a linear model is calculated using the lm() function in R.

```
cars_lm <- lm(cars$dist ~ cars$speed)
cars_lm
```

```
##
## Call:
## lm(formula = cars$dist ~ cars$speed)
##
## Coefficients:
## (Intercept)   cars$speed
##     -17.579        3.932
```

The output of the model indicates a linear function as:

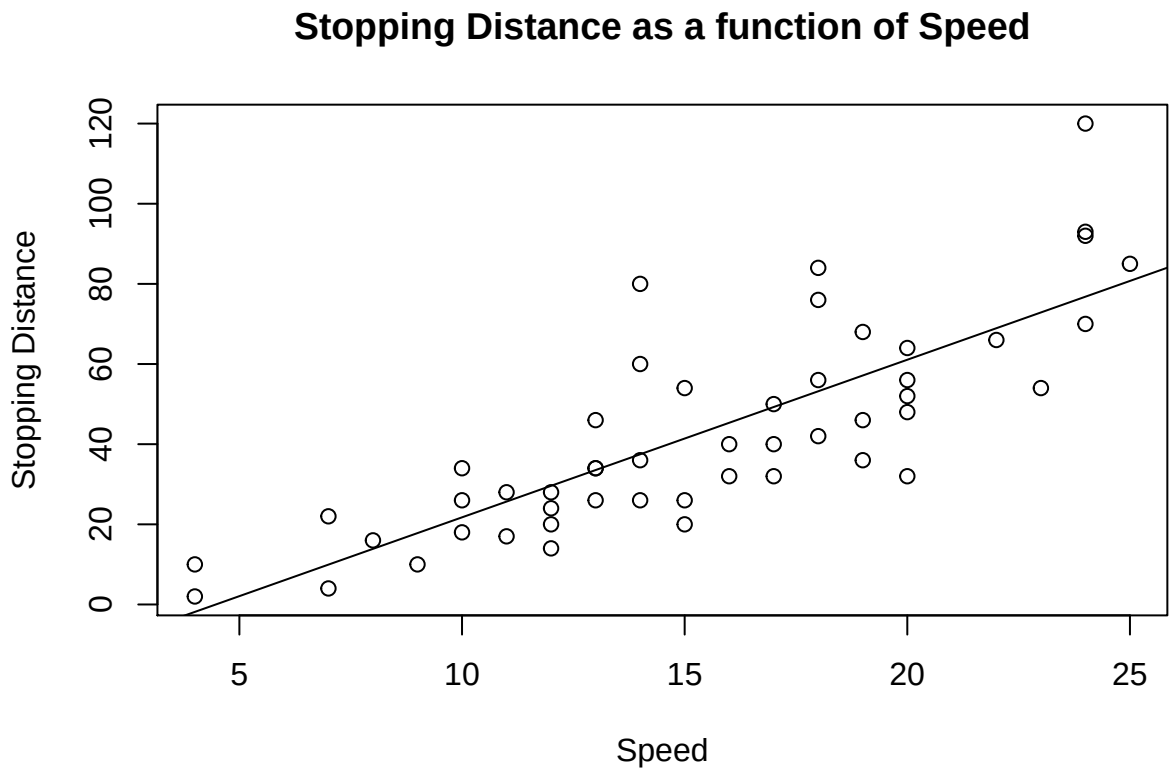
$$stopping\ distance = -17.579 + (3.932 * speed)$$

A y-intercept of -17.579 does seem peculiar. Based on the linear model, this would indicate a car at very low speeds would actual stop in less than 0 feet, which is impossible. The slope of 3.932 based on the speed appears reasonable.

Visualization with Linear Regression Model

The scatterplot with the linear regression superimposed does align with the trend of the data, as the speed increases, so does the stopping distance.

```
plot(cars$speed, cars$dist, main="Stopping Distance as a function of Speed",
    xlab="Speed", ylab="Stopping Distance")
abline(cars_lm)
```



Evaluation of the Linear Regression Model

```
summary(cars_lm)
```

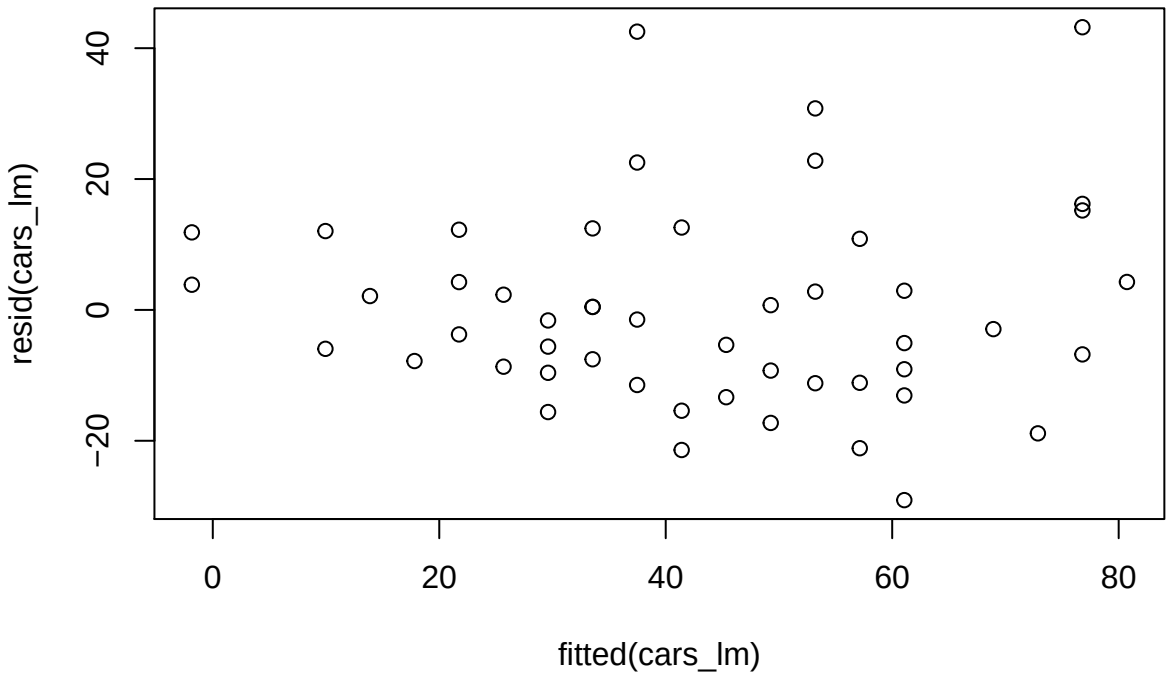
```
##
## Call:
## lm(formula = cars$dist ~ cars$speed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## cars$speed    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
```

- Residuals: A median value of -2.272 indicates that overall the actual values are a bit lower than the calculated values, which is noticeable in the above plot, but not too far from zero. The first quartile (-9.525) and the third quartile (9.215) are nearly balanced around zero, which is indicative of a normal distribution. The minimum value (-29.069) and maximum value (43.201) aren't too different in magnitude either, again, a good sign of a normal distribution of the residuals.
- Coefficients: The t-value for speed (9.464) indicates the standard error of speed is roughly 9-10 times smaller than the coefficient for speed (3.9324). According to the textbook, a good model will have coefficient standard errors of at least five to ten times smaller than the coefficient itself. This large ratio would indicate there is little variability in the slope estimate. Per the intercept, the standard error (6.7584) against the estimate (-17.5791) results in a t-value of -2.601. The resulting t-value for the intercept indicates this value is expected to vary. This understanding makes sense as the negative intercept implies negative stopping distances for very low speeds. In reality, the intercept would be 0, as a car at rest would not need any distance to stop. The intercept p-value is slightly above .01, which would indicate the intercept is not relevant 1 out of 100 times. (Note: the dataset only contains 50 instances.) The p-value of speed is 1.49×10^{-12}, a vary small value, denoting the speed variable is rarely not relevant.
- Residual standard error: Measures the total variation in the residual values, the result of 15.38. According to the book, for a normal distribution the first and third quartiles of the previous residuals should be about 1.5 times this standard error. For this model, that is not true. Both the first and third quartile values are lower than the residual standard error. Not a good sign for a potential normal distribution.
- Degrees of freedom: A calculated value of 48 is expected given the 50 instances and the two coefficients, slope and intercept.
- Multiple R-squared: A calculated value of 0.6511 indicates the model explains 65.11% of the data's variation.
- Adjusted R-squared: A calculated value of 0.6438 is expected to be smaller than the R-squared value, so this value is reasonable.
- F-statistic: Not useful because this linear regression model has only one independent variable.

Residual Analysis

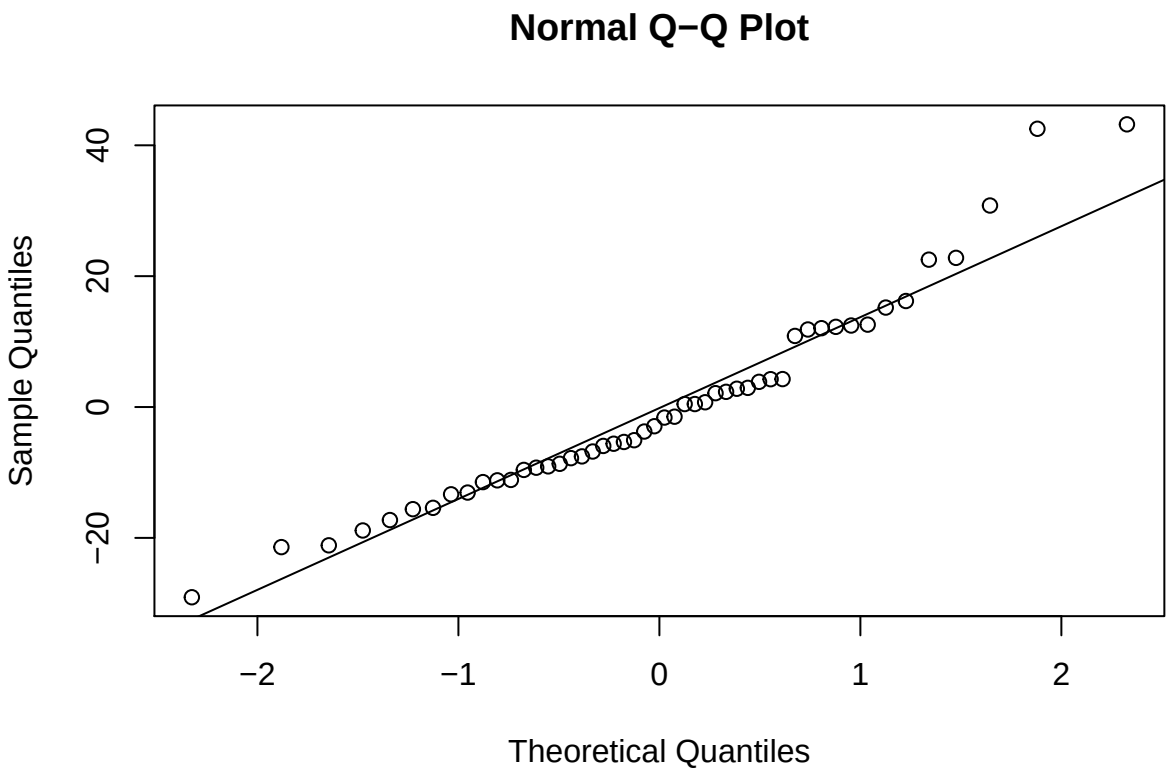
First, plot the residual values against the input values.

```
plot(fitted(cars_lm),resid(cars_lm))
```



The residual plot above shows a relatively even separation of residuals above zero and below zero. The residuals do appear uniform in the plot. A trend may exist in which the greater the input value, a higher potential for greater residual value may occur. Considering the scenario of car speed and stopping distance, this understanding makes sense. From a higher speed, one would expect greater variability in stopping distance as compare to lower speeds. Overall, this residual plot would indicate speed is a good predictor of stopping distance.

```
qqnorm(resid(cars_lm))
qqline(resid(cars_lm))
```



The quantile-versus-quantile plot (Q-Q plot) above indicates that the high end contains a few examples diverging from the line whereas the lower end shows little divergence from the line. A normal distribution would expect all the points to follow the straight line. Because of the minimal divergence, except at the higher end, the Q-Q plot would indicate the speed variable a good predictor of stopping distance as the residuals likely follow a normal distribution.

Conclusion

Overall, the car speed would appear to be a good predictor of stopping distance. The linear regression model does contain some flaws, particularly in the intercept value and the predictions at higher speeds. As noted above, a linear model of stopping distance against speed would expect an intercept at 0. As for the greater variability at higher speeds, the divergence in residuals is expected. To improve the model, road conditions, tire conditions, car make and model should be considered. It should be noted, the data were recorded in the 1920s.