HW11

Zachary Palmore

4/15/2021

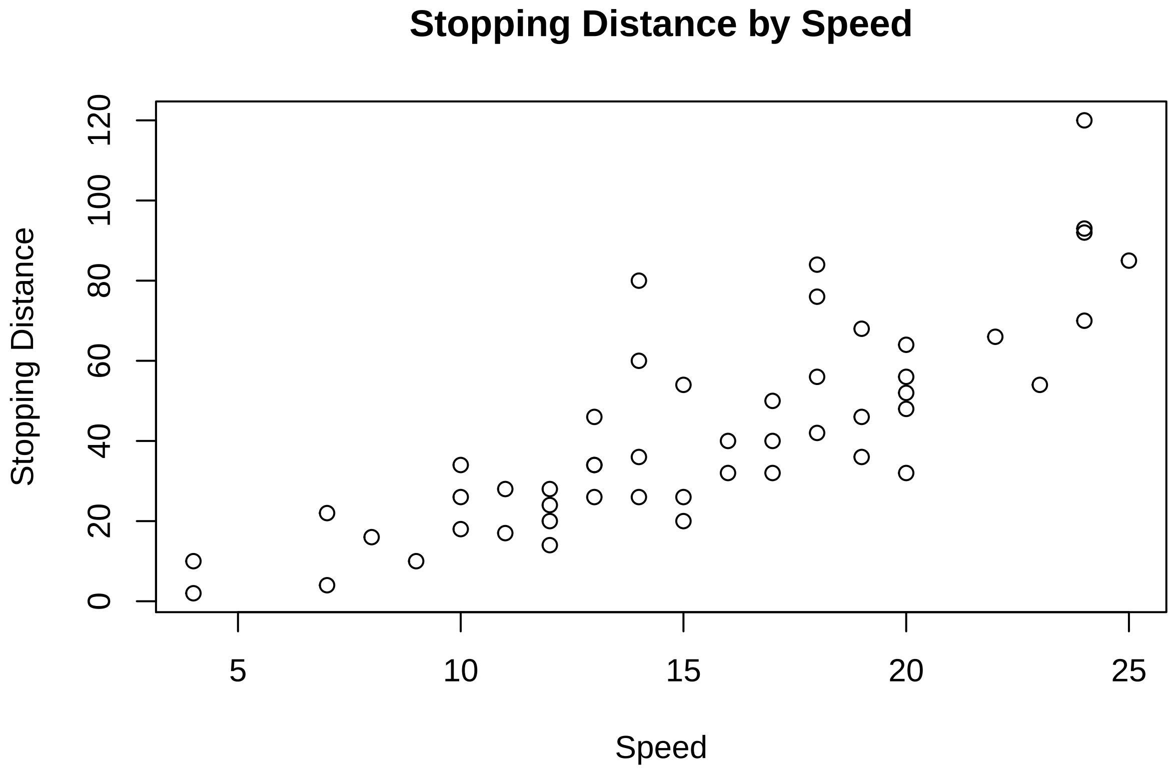
Prompt

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.)

Model

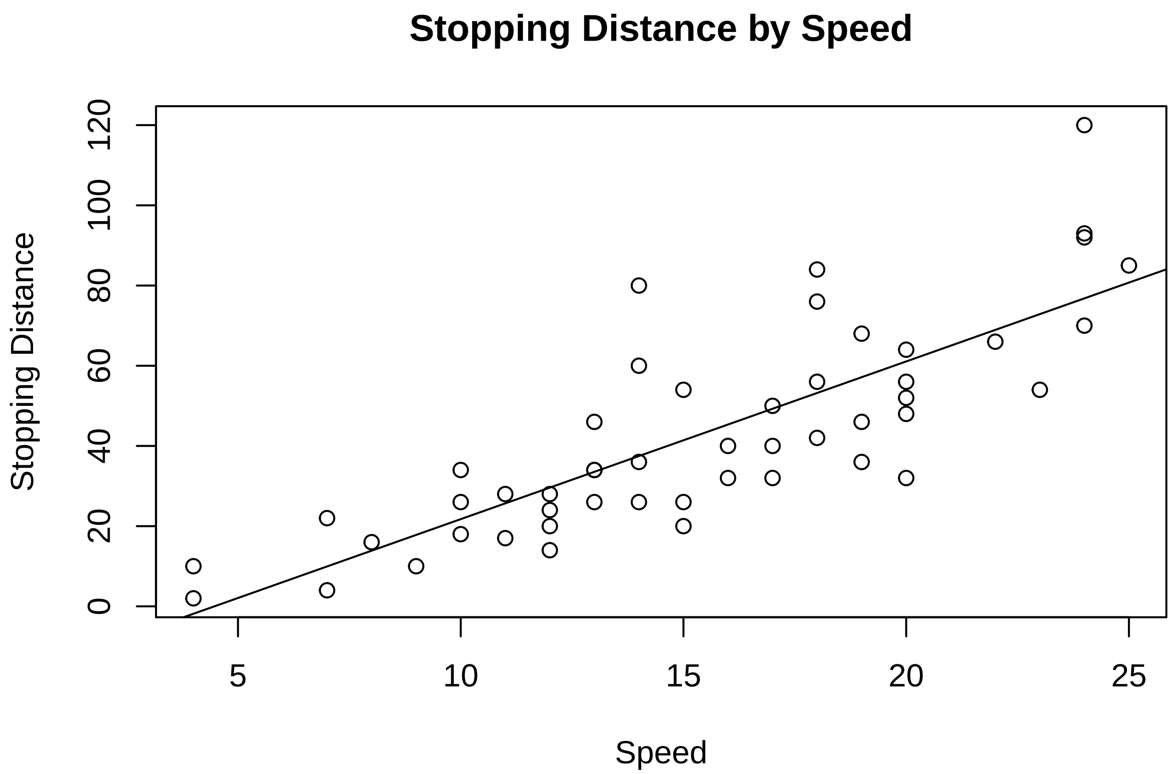
In the spirit of replication, before we begin with the model building we should take a look at the scatterplot of stopping distance as a function of speed. When we do, it should be examined for the presence or absence of linearity. This is shown below.

```
plot(cars$speed, cars$dist, main = "Stopping Distance by Speed", xlab = "Speed", ylab = "Stopping Distan
```



There is a presence of linearity although it might not be ideal. Speaking generally, as speed increases the stopping distance also increases. However, there are many points that deviated from this trend. Because linearity is present, we can continue with modeling. Next, we build a linear model to test the fit.

```
lm.cars <- lm(cars$dist ~ cars$speed)
plot(cars$speed, cars$dist, main = "Stopping Distance by Speed", xlab = "Speed", ylab = "Stopping Distan
abline(lm.cars)
```



It looks decent enough. There are still many points that stray from the diagonal line of best fit but their differences are small enough to say the model is linear. The question now is how well those points fit the line. We can evaluate this with a summary and some diagnostic plots.

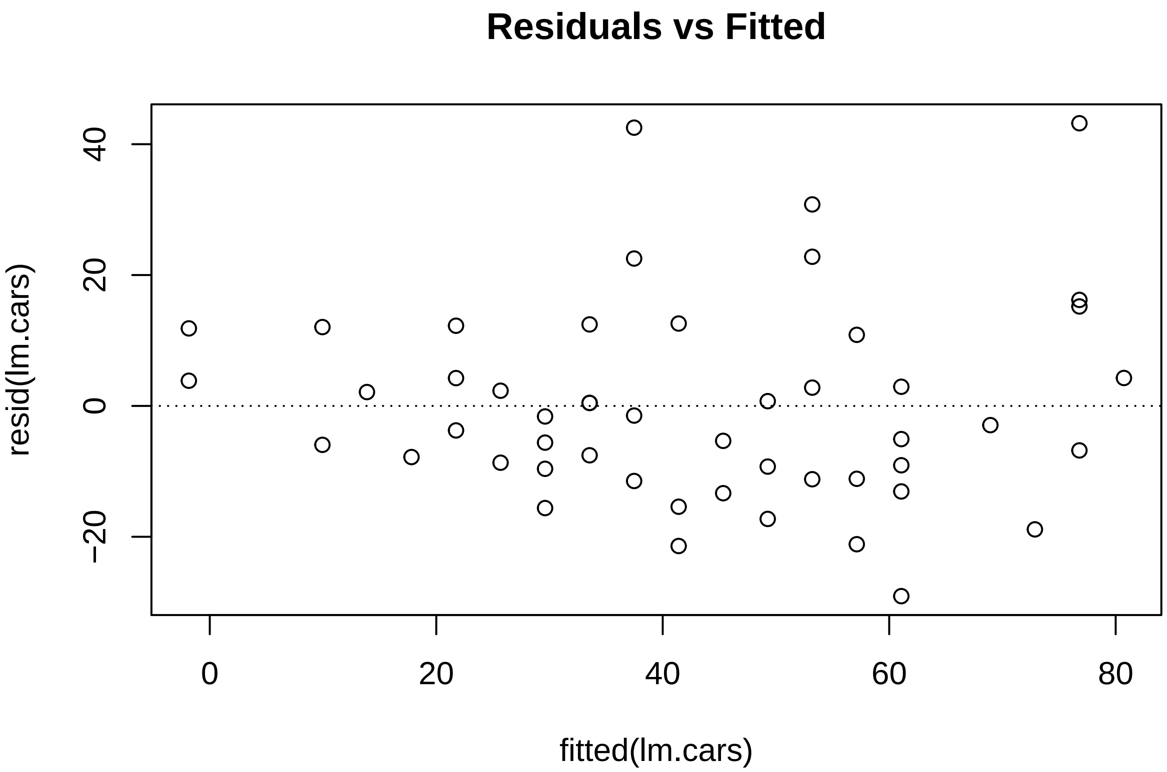
```
summary(lm.cars)
```

```
##
## 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
```

From this summary we can see that we selected the correct model, stopping distance modeled by speed, and the residuals have a median distance of -2.272. The coefficients also provide the slope and intercept of our predictor speed along with the significance of the predictor for this model. There is a moderately strong positive correlation and about 65% of the variation in the data is explained by the model.

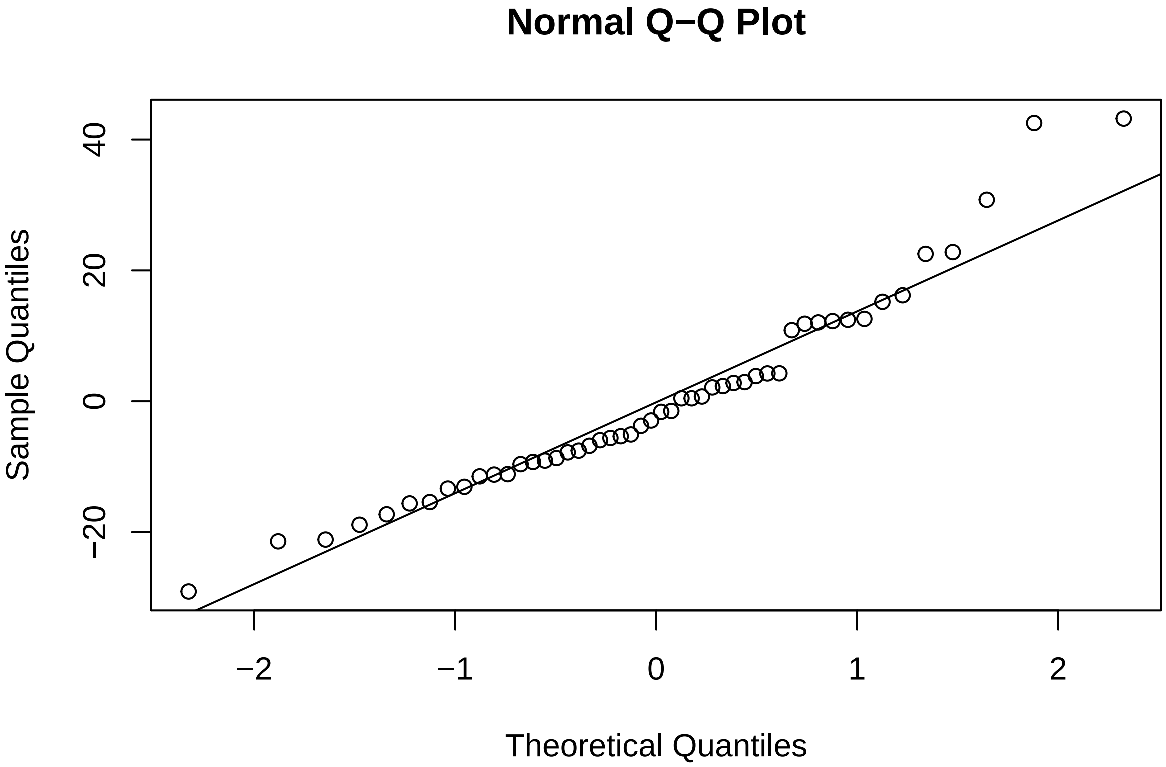
Our standard error in our residuals is 15.38 using 48 degrees of freedom which describes the total variation in our data. It is not the worst it could be and also indicates approximate normality given the first and third quantiles are about 1.5 times this standard error. However, we can confirm this with a Normal QQ plot and review the residuals vs fitted values.

```
plot(fitted(lm.cars), resid(lm.cars), main="Residuals vs Fitted")
abline(h=0, lty=3)
```



In the residuals vs fitted plot we can see that there are no obvious patterns in the residuals. The points are randomly scattered around the dotted line in the middle. This is a good indication of an acceptable model based soley on the residuals vs fitted plot. Another option is evaluating normality with a QQ-plot.

```
qqnorm(resid(lm.cars))
qqline(resid(lm.cars))
```



With the Normal Q-Q Plot we begin to distinguish that some points might not be normally distributed. This is shown in the deviation from the straight diagonal line. We can say that the plot is weighted towards the ends and is especially heavy towards the higher end of the distribution. Normality may not be present in this model which undermines its ability to predict with confidence.

Ultimately, there may be other factors to include in this model that could explain the relationship between the variables. This may include predictors like road condition, tire tred, brake pad wear, and others. If we wanted to improve this model, those factors might prove useful in predicting stopping distance. It also might be fruitful to try other model types to ensure the best possible model in prediction stopping distance.