DATA 605 Week 11 Homework

Ilya Kats

November 12, 2017

Task

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).

Analysis

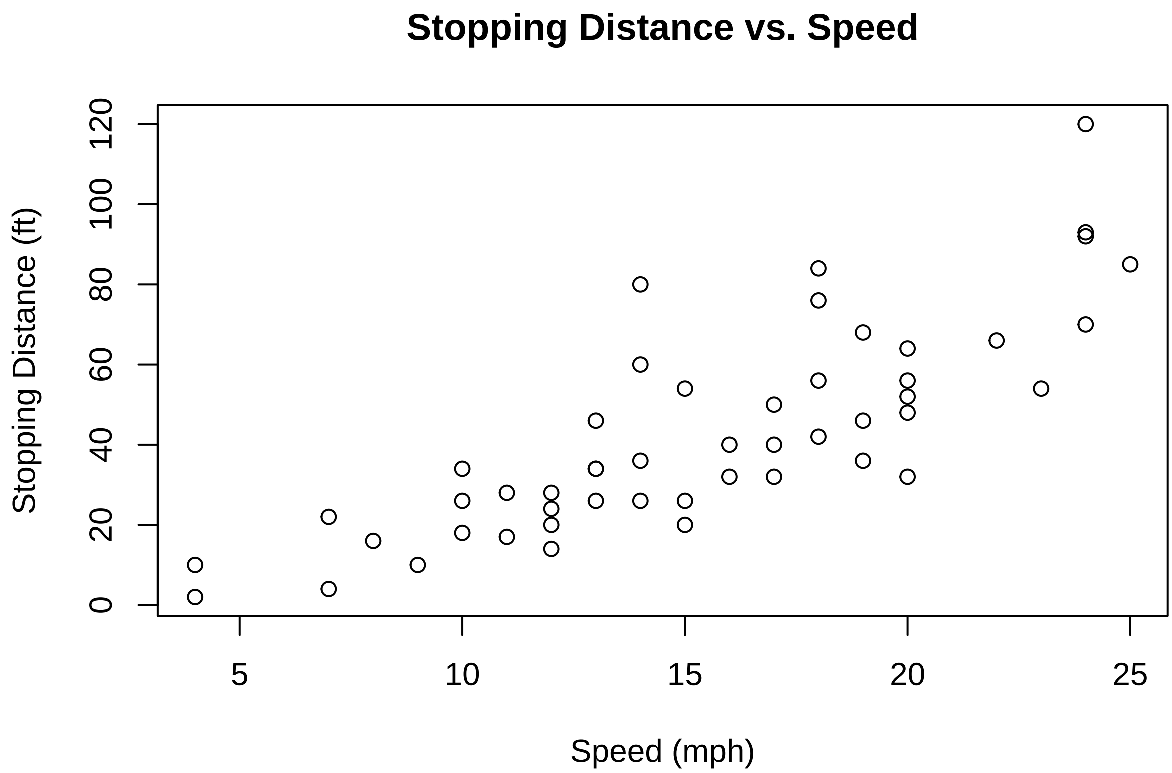
Dataset `cars` includes 50 observations with 2 variables - `dist` containing stoppping distance in feet and `speed` containing speed of a car before applying the brakes in miles per hour. Data were recorded in the 1920s.

```
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
```

Let us look at the plot of two variables. Speed is the **explanatory** variable and stopping distance is the **response** one.

```
plot(cars$speed, cars$dist, xlab='Speed (mph)', ylab='Stopping Distance (ft)',
     main='Stopping Distance vs. Speed')
```

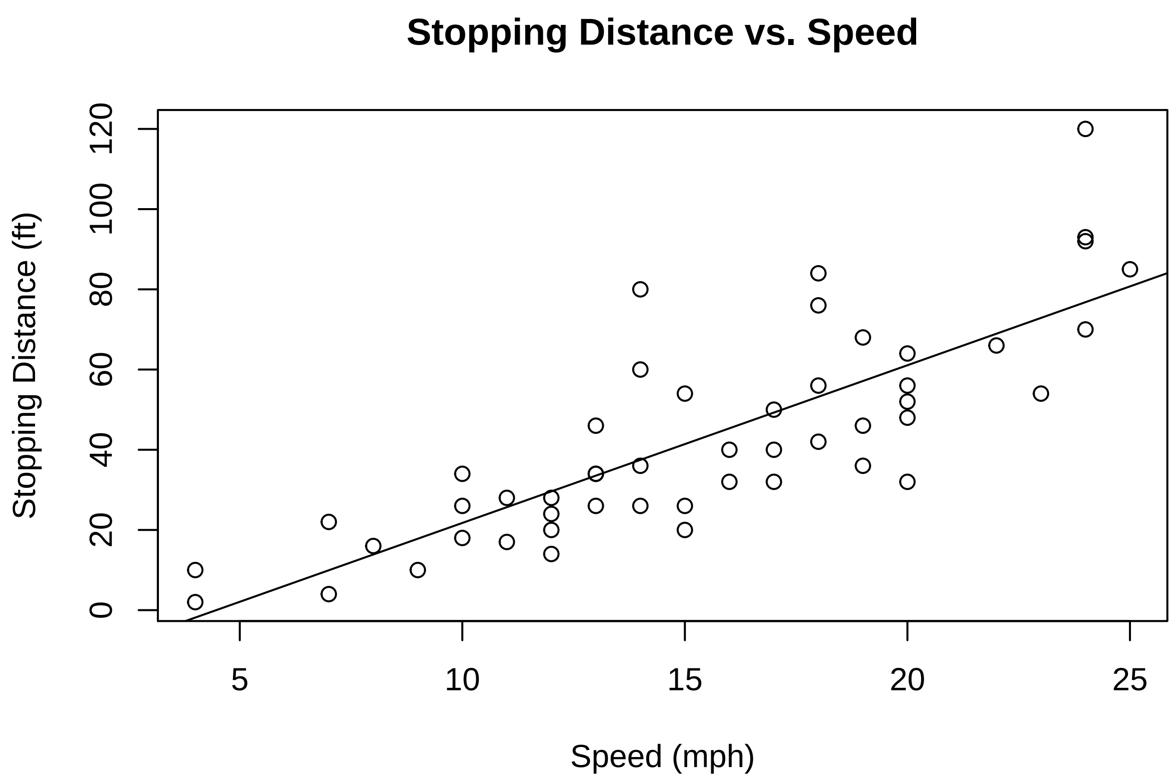


Let us build a linear model and find the best fitting line.

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

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

```
plot(cars$speed, cars$dist, xlab='Speed (mph)', ylab='Stopping Distance (ft)',
     main='Stopping Distance vs. Speed')
abline(cars_lm)
```



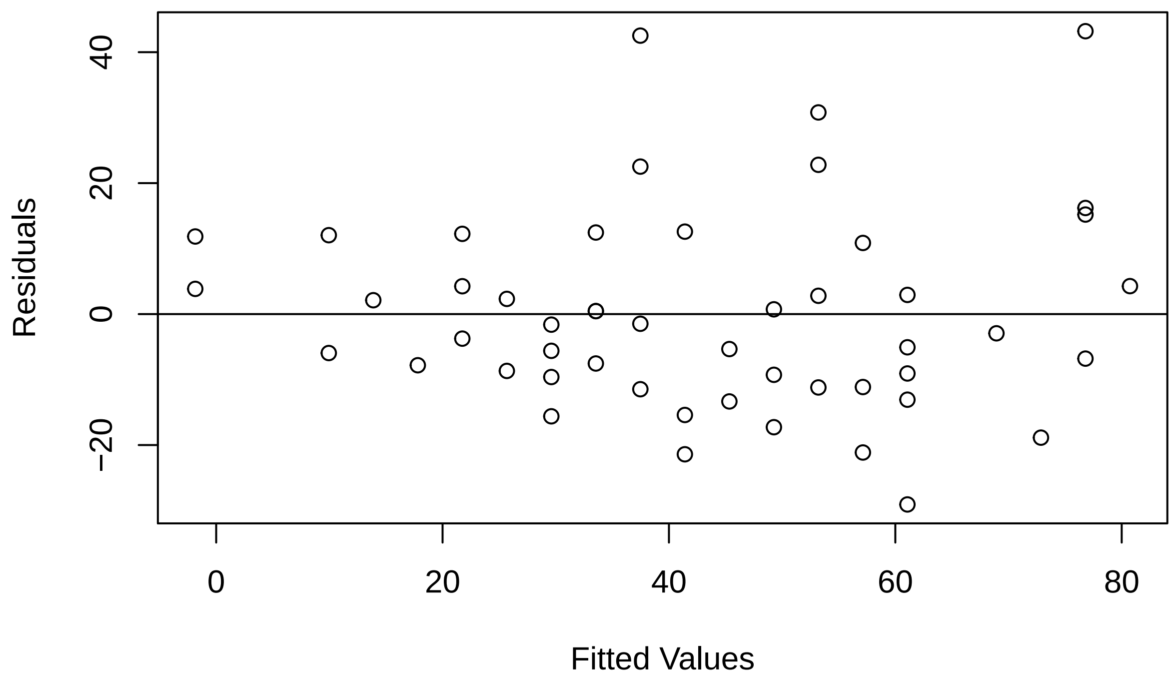
There appears to be some correlation between two variables, but let us evaluate the linear model we have.

```
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
```

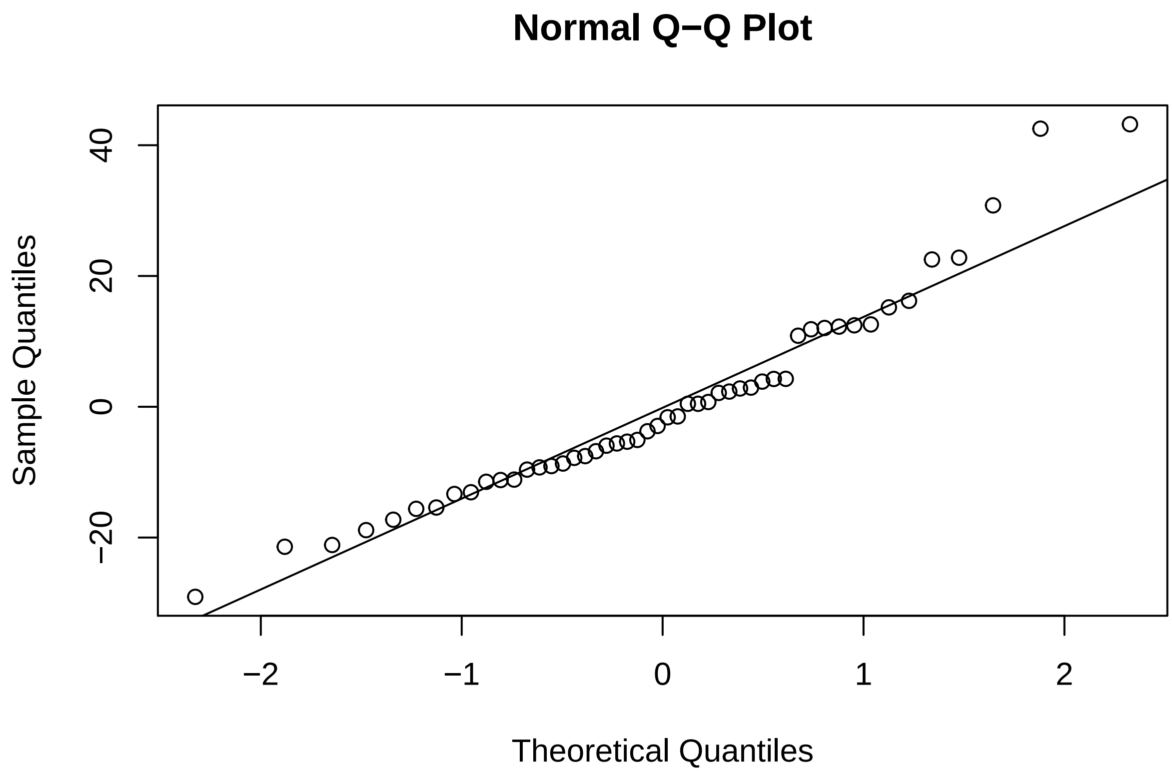
The median value of the residuals is somewhat close to zero and quartiles and min/max values are roughly the same magnitude. The standard error of the `speed` variable is more than 9 times smaller than the corresponding coefficient. There should not be a lot of variability in this coefficient. On the other hand, the difference between the intercept estimate and standard error is less significant, so there may be more variability. The `speed` coefficient is highly significant. The intercept coefficient is less significant, but it is still relevant depending on the confidence interval desired. Finally, R^2 explains about 65.11% of the data's variation.

```
plot(cars_lm$fitted.values, cars_lm$residuals, xlab='Fitted Values', ylab='Residuals')
abline(0,0)
```



It is possible to say that the outlier values do not show the same variance of the residuals; however, it is not very clear. I think it is reasonable to continue with the analysis and assume similar variance of residuals.

```
qqnorm(cars_lm$residuals)
qqline(cars_lm$residuals)
```



Althought again there are some problems at the outlier levels, the normal Q-Q plot of the residuals appears to follow the theoretical line. Residuals are reasonably normally distributed.

Conclusion

I believe the linear model does a good job at explaining the data. There appears to be some slight curvature in the main plot and in the residuals plot, so I decided to try a simple quadratic model (see below). It has it's own problems - again varability of residuals is not constant enough, q-q plot has some deviations, coefficients are not very significant and R^2 is not increased by much. I don't think it's an improvement over a simplier linear model.

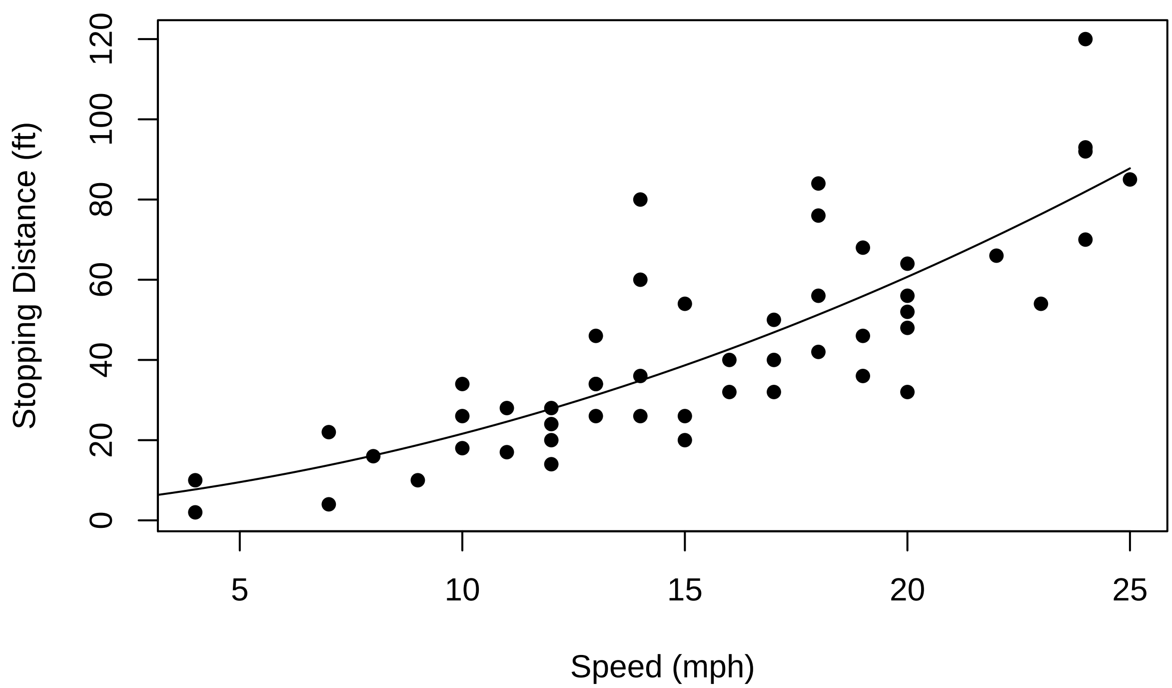
Quadtratic Model

```
speed <- cars$speed
speed2 <- speed^2
dist <- cars$dist

cars_qm <- lm(dist ~ speed + speed2)
summary(cars_qm)
```

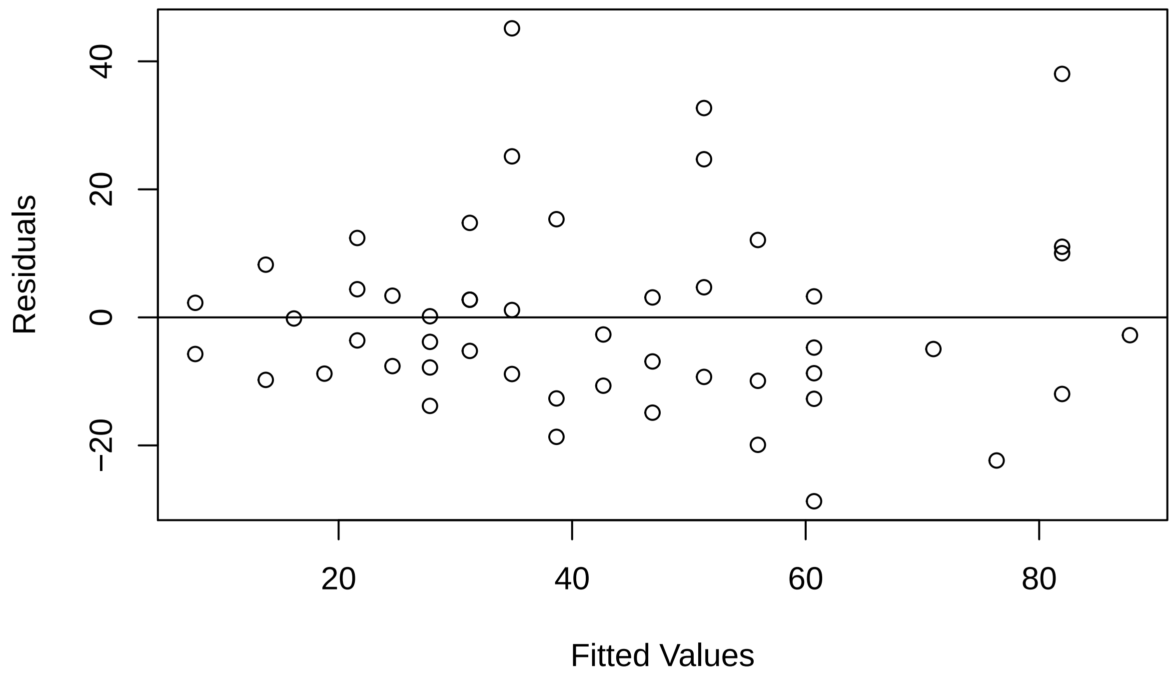
```
##
## Call:
## lm(formula = dist ~ speed + speed2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -28.720  -9.184  -3.188   4.628  45.152
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.47014   14.81716   0.167    0.868
## speed        0.91329    2.03422   0.449    0.656
## speed2       0.09996    0.06597   1.515    0.136
##
## Residual standard error: 15.18 on 47 degrees of freedom
## Multiple R-squared:  0.6673, Adjusted R-squared:  0.6532
## F-statistic: 47.14 on 2 and 47 DF,  p-value: 5.852e-12
```

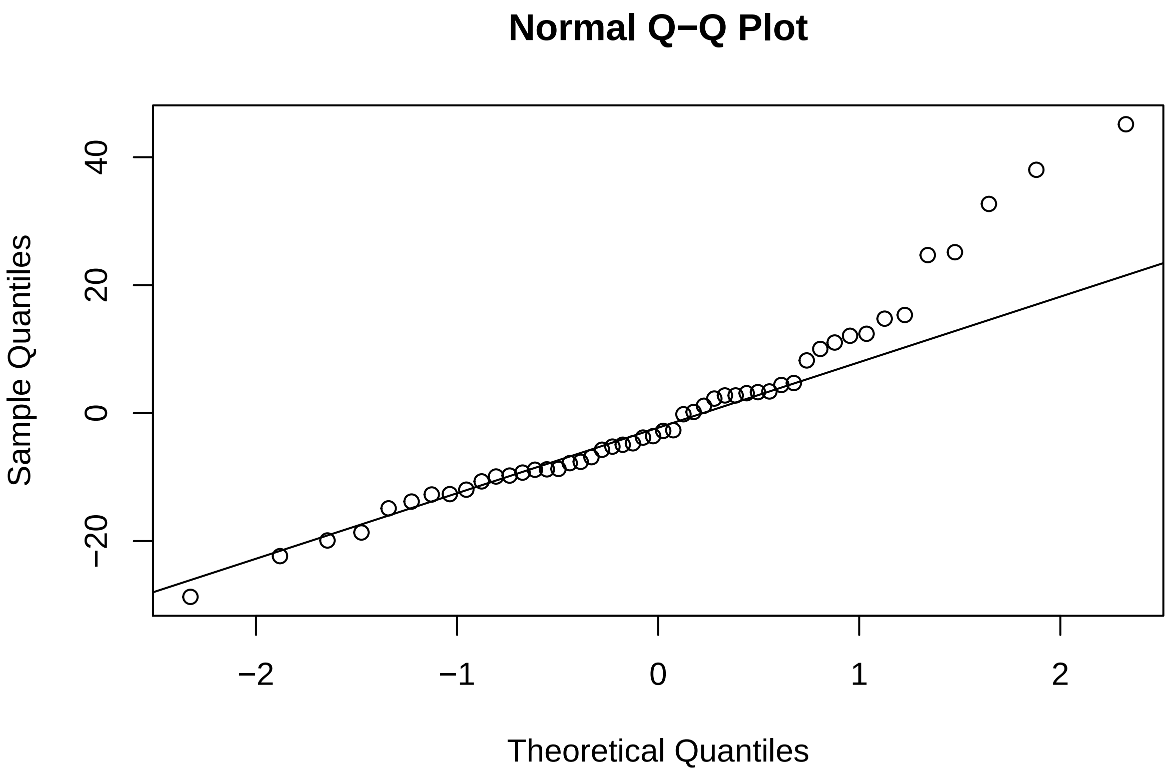
```
speedvalues <- seq(0, 25, 0.1)
predictedcounts <- predict(cars_qm,list(speed=speedvalues, speed2=speedvalues^2))

plot(speed, dist, pch=16, xlab='Speed (mph)', ylab='Stopping Distance (ft)')
lines(speedvalues, predictedcounts)
```




```
plot(cars_qm$fitted.values, cars_qm$residuals, xlab='Fitted Values', ylab='Residuals')
abline(0,0)
```




```
qqnorm(cars_qm$residuals)
qqline(cars_qm$residuals)
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−20
0
20
40
−2
−1
0
1
2