Data 605 HW-11

Vishal Arora

November 10, 2019

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.)

Loading cars dataset and getting a glimpse of the dataset by using the str() function.

```
data(cars)
str(cars)
```

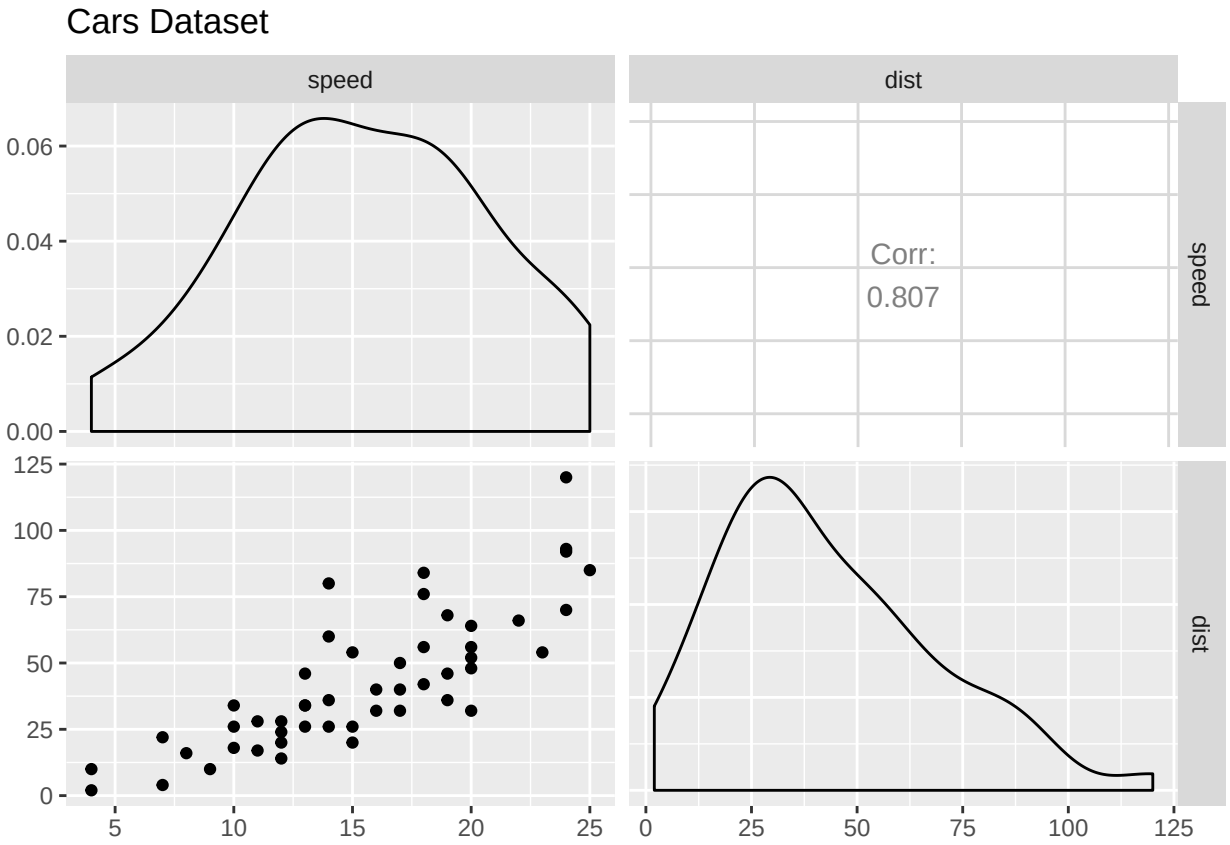
```
## 'data.frame':    50 obs. of  2 variables:
##  $ speed: num  4 4 7 7 8 9 10 10 10 11 ...
##  $ dist : num  2 10 4 22 16 10 18 26 34 17 ...
```

Cars dataset has 50 observations with 2 variables (dist & speed). Now we will see how these 2 variables relate to each other .

To decide whether we can make a predictive model, the first step is to see if there appears to be a relationship between our predictor and response variables (in this case speed and dist). Let's do some exploratory data visualization using the ggpairs() function from the GGally package to create a plot matrix to see how the variables relate to one another.

The ggpairs() function gives us scatter plots for each variable combination, as well as density plots for each variable and the strength of correlations between variables.

```
ggpairs(data=cars, columns=1:2, title="Cars Dataset")
```



The correlation coefficients provide information about how close the variables are to having a relationship; the closer the correlation coefficient is to 1, the stronger the relationship is. In our case the correlation between both variables (i.e. speed & dist) is quiet strong. The scatter plots let us visualize the relationships between pairs of variables.In the scatter plot between speed(Predictor) & dist(response) variable we can clearly sense a linear relationship, with few outliers but those can be ignored as of now.

Forming a hypothesis and rejecting one of the hypothesis using summary function and then plot it accordingly.

$$H_0 \quad : \quad Speed \quad \& \quad Dist \quad variables \quad are \quad not \quad related \quad to \quad each \quad other.$$

$$H_A \quad : \quad Speed \quad \& \quad Dist \quad variables \quad have \quad some \quad relation \quad with \quad each \quad other.$$

```
model1 <- lm(dist ~ speed, data = cars)
summary(model1)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## 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 *
## 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 intercept(-17.5791) in our example is the expected car distance if the value of speed was zero. Of course we cannot have a car having distance traveled in negative.

The slope in our example is the effect of car speed on distance travelled by car. We see that for each additional increase of speed(mph), the distance travelled by car increases by 3.9324 mtrs.

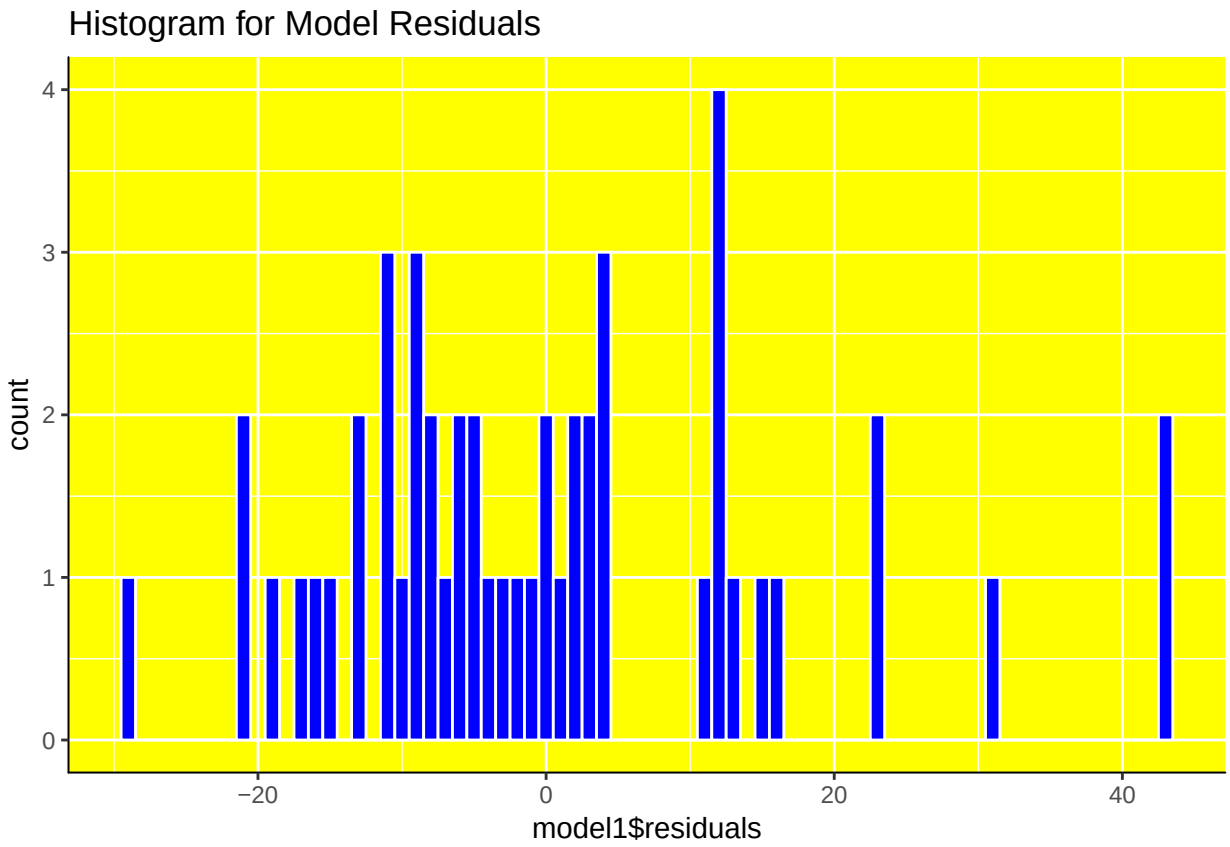
The coefficient standard errors tell us the average variation of the estimated coefficients from the actual average of our response variable.

Pr(>|t|): This depicts the p-value, using 95% confidence level, if the -value is less than 0,05 we reject the Null hypothesis. In our case also the p-vaue is less than 0.05 hence we can say that our variables (Speed & Dist) has some kind of relations between themselves.

Now to check how well our Data fits the model.

Residuals (recall that these are the distances between our observation and the model), which tells us something about how well our model fit our data. The residuals should have a pretty symmetrical distribution around zero.
Using ggplot2 to visualise through plotting histogram

```
ggplot(data=cars, aes(model1$residuals)) +
geom_histogram(binwidth = 1, color = "white", fill = "blue") +
theme(panel.background = element_rect(fill = "yellow"),
axis.line.x=element_line(),
axis.line.y=element_line()) +
ggtitle("Histogram for Model Residuals")
```



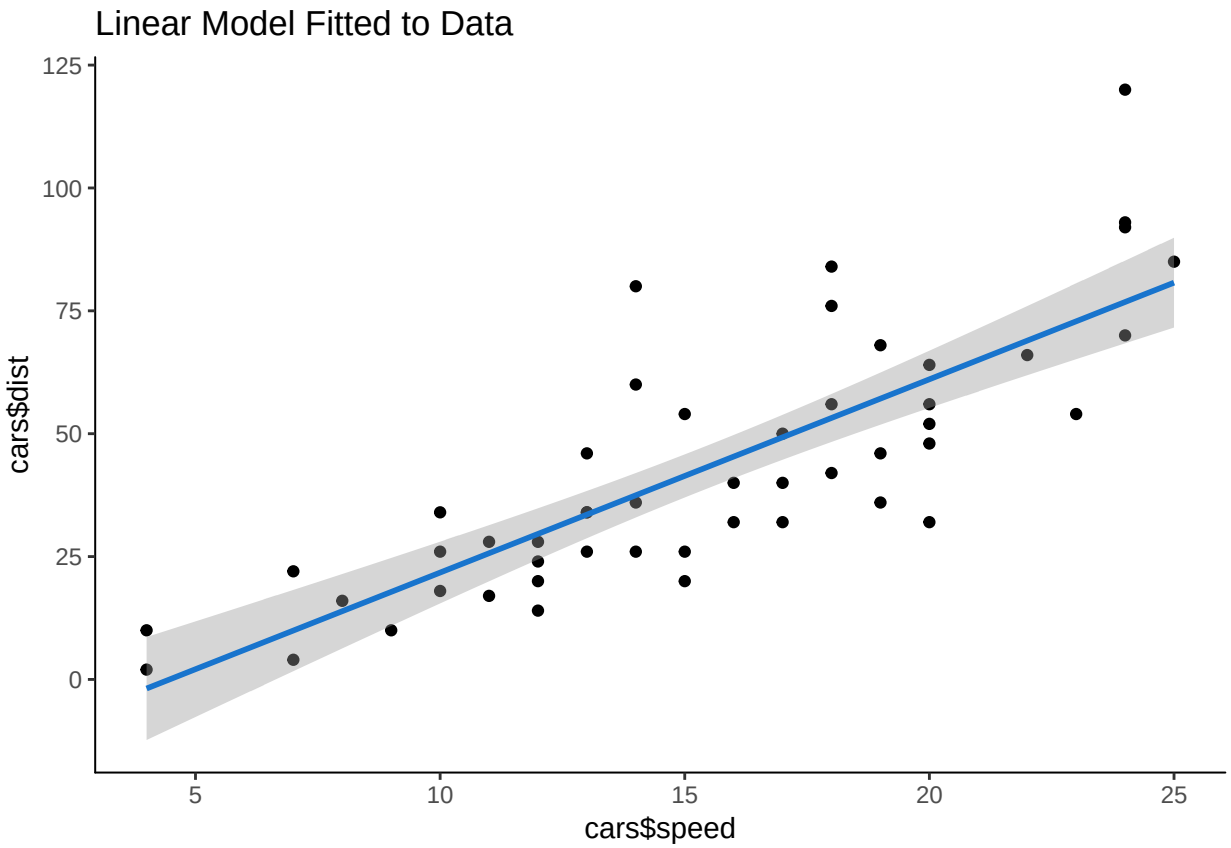
Our residuals look pretty symmetrical around 0, suggesting that our model fits the data well. Residual standard error. This term represents the average amount that our response variable measurements deviate from the fitted linear model (the model error term).There is a little bit of left skewness.

Multiple R-squared, The R2 value is a measure of how close our data are to the linear regression model. R2 values are always between 0 and 1.Numbers closer to 1 represent well-fitting models. In our case the R2 value is 0.6511 which is close to 1 hence which means our data fits the linear regression model.

A higher F-statistic value indicates that there is relationship between the independent & depedent(response) variable that we are testing. In our case the value seems to be quiet high hence we can safely assume that our variables have a linear relationship.

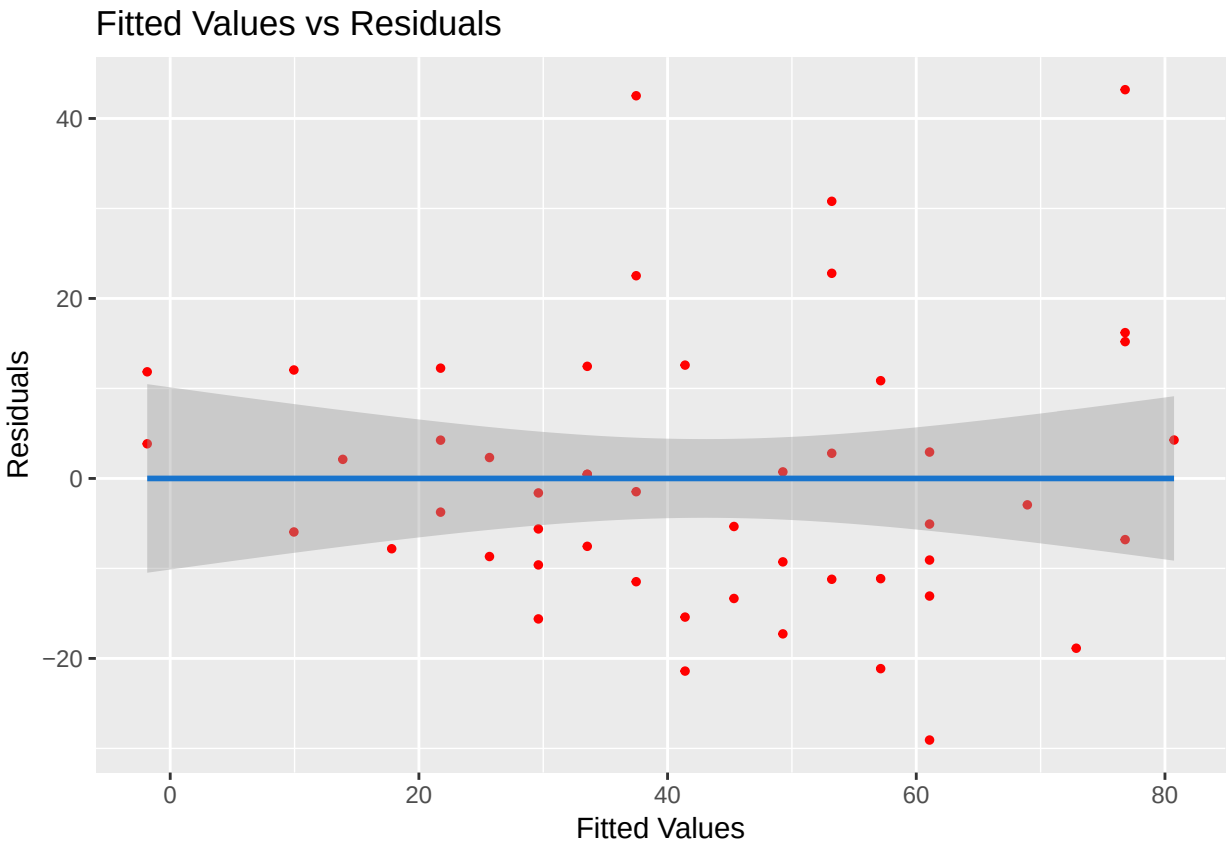
Let's have a look at our model fitted to our data for Speed and Dist by using ggplot function.

```
ggplot(data = cars, aes(x = cars$speed, y = cars$dist)) +
geom_point() +
stat_smooth(method = "lm", col = "dodgerblue3") +
theme(panel.background = element_rect(fill = "white"),
axis.line.x=element_line(),
axis.line.y=element_line()) +
ggtitle("Linear Model Fitted to Data")
```



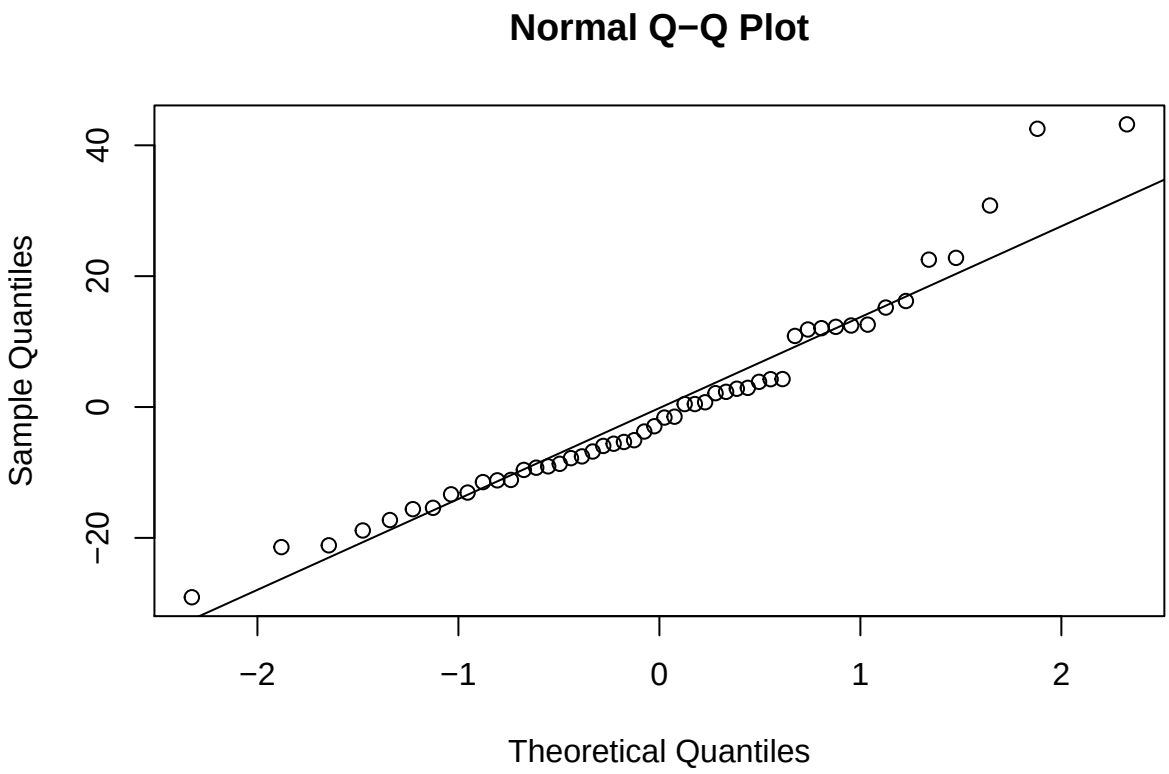
Residual Analysis

```
ggplot(model1, aes(.fitted, .resid)) +
  geom_point(color = "red", size=1) +
  stat_smooth(method = "lm", col = "dodgerblue3") +
  labs(title = "Fitted Values vs Residuals") +
  labs(x = "Fitted Values") +
  labs(y = "Residuals")
```



The plot below shows that the residuals look uniformly distributed around zero. The residuals appear to be uniformly scattered above and below zero, with few outliers.

```
qqnorm(resid(model1))
qqline(resid(model1))
```



Conclusion :- We can say that Speed & Dist are linearly related, but there can be other factors are which might influence the relationship which need to be explored.