Ginorio\_Assignment\_11

MGinorio

11/7/2021

### Regression Analysis

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.)

library(tidyverse)  
library(moderndive)  
library(skimr)

Let’s consider a simple example of how the speed of a car affects its stopping distance, that is, how far it travels before it comes to a stop. To examine this relationship, we will use the ‘cars’ dataset.

cars\_regression <- cars %>%   
 select(speed, dist)  
  
cars\_regression %>% skim()

Data summary

|  |  |
| --- | --- |
| Name | Piped data |
| Number of rows | 50 |
| Number of columns | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| numeric | 2 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: numeric**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| speed | 0 | 1 | 15.40 | 5.29 | 4 | 12 | 15 | 19 | 25 | ▂▅▇▇▃ |
| dist | 0 | 1 | 42.98 | 25.77 | 2 | 26 | 36 | 56 | 120 | ▅▇▅▂▁ |

### Question

How the speed of a car affects its stopping distance, that is, how far it travels before it comes to a stop?

### EDA

#### Variables

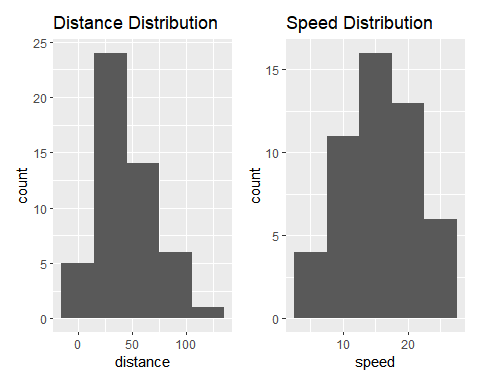
Exploratory analysis of variables

Dependent Variable -

Independent Variable -

Investigate correlation between this variables

* distance
* speed

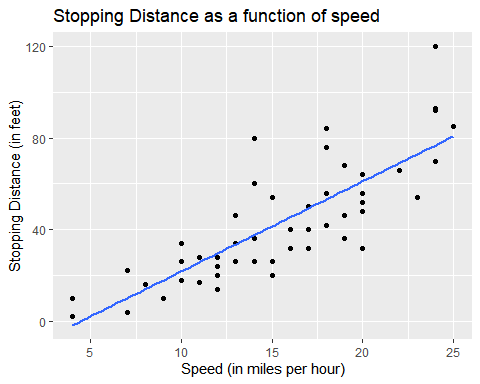


#### Visualize Data

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.

## `geom\_smooth()` using formula 'y ~ x'



#### Create Model

The output of the model indicates a linear function as:

A y-intercept of -17.579 does seem peculiar.

Based on the linear model, this would indicate that a car at speeds 0 or closer to 0 (negative) it would stop in less than 0 feet, which is accurate since a car that is not in movement thus it does not need to stop.

The slope of 3.932 based on the speed.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| term | estimate | std\_error | statistic | p\_value | lower\_ci | upper\_ci |
| intercept | -17.579 | 6.758 | -2.601 | 0.012 | -35.707 | 0.548 |
| speed | 3.932 | 0.416 | 9.464 | 0.000 | 2.818 | 5.047 |

**Coefficient**

also seen under estimate - This portion of the output shows the estimated coefficient values

**Std. Error**

For a good model, we typically would like to see a standard error that is at least five to ten times smaller than the corresponding coefficient

For Example:

the SD error for is 9.45 times smaller then the coefficient value.(3.932/0.416)

**P-value**

shows the probability that the corresponding coefficient is not relevant in the model. This value is also known as the significance or p-value of the coefficient

The probability that the intercept is not relevant is 0.012.

## # A tibble: 50 x 5  
## ID dist speed dist\_hat residual  
## <int> <dbl> <dbl> <dbl> <dbl>  
## 1 1 2 4 -1.85 3.85  
## 2 2 10 4 -1.85 11.8   
## 3 3 4 7 9.95 -5.95  
## 4 4 22 7 9.95 12.1   
## 5 5 16 8 13.9 2.12  
## 6 6 10 9 17.8 -7.81  
## 7 7 18 10 21.7 -3.74  
## 8 8 26 10 21.7 4.26  
## 9 9 34 10 21.7 12.3   
## 10 10 17 11 25.7 -8.68  
## # ... with 40 more rows

**Residuals**

The residuals are the differences between the actual measured values and the corresponding values on the fitted regression line.

Residual values are normally distributed around a mean of zero in this case we see that even though the values are not exactly zero we can still expect a normal distribution.

That is, a good model’s residuals should be roughly balanced around and not too far away from the mean of zero.

**RSquared & Residual Standard Error RSE**

These final few lines in the output provide some statistical information about the quality of the regression model’s fit to the data

summary(speed\_model)$r.squared

## [1] 0.6510794

summary(speed\_model)$sigma

## [1] 15.37959

**Minimum Maximum**

minimum and maximum values of roughly the same magnitude, and first and third quartile values of roughly the same magnitude.

summary(speed\_model\_points$residual)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -29.06900 -9.52550 -2.27200 -0.00004 9.21450 43.20100

**Residual Visual Relationship**

Distribution of Residuals Investigate potential relationships between the residuals and all explanatory/predictor variables

**Residual Vs Fitter**

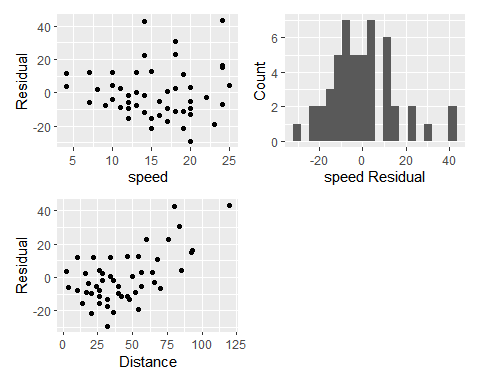
we may be able to construct a model that produces tighter residual values and better predictions.

Residual values greater than zero mean that the regression model predicted a value that was too small compared to the actual measured value, and negative values indicate that the regression model predicted a value that was too large

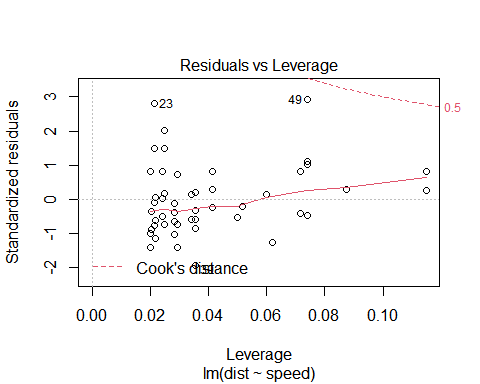
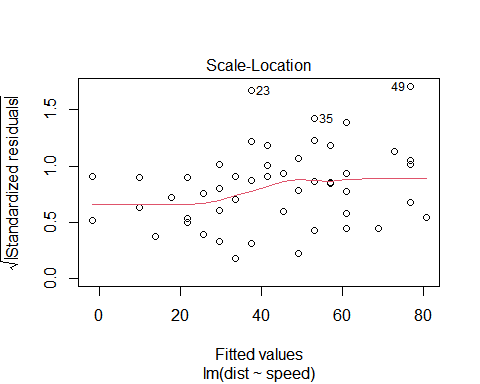
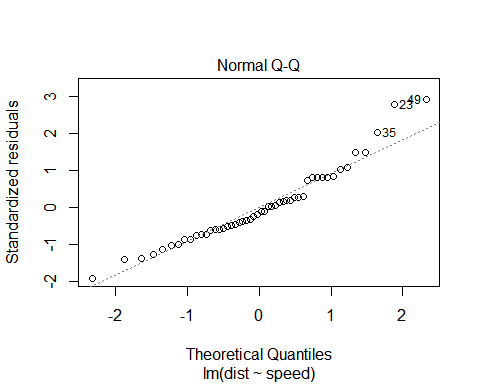
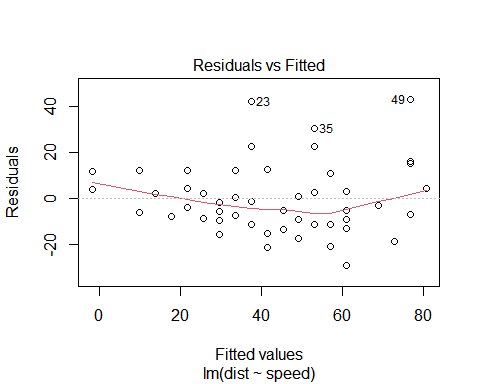
**QQ Plot**

If the residuals were normally distributed, we would expect the points plotted in this figure to follow a straight line. Which in this case we do see a straight line forming.

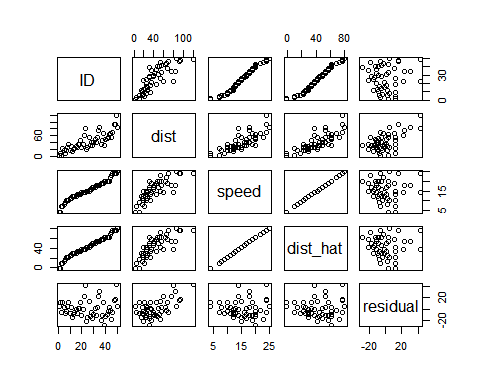
This test could confirm that the speed as a predictor in the model may be sufficient to explain the data.



plot(speed\_model)



plot(speed\_model\_points)



### Predictions

We do this so that we can specify that 8 is a value of speed, so that predict knows how to use it with the model stored in speed\_model

predict(speed\_model, newdata = data.frame(speed = c(8, 21, 50)))

## 1 2 3   
## 13.88018 65.00149 179.04134

\*\* note 50 is out of range

### 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