data 605 regression

Warner Alexis

2024-04-07

Cars Regression Analysis

We are going to a regression analysis on the car data set.

```r
# import car dataset
require(carData)
```

```
## Loading required package: carData
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
head(cars)
```

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

```r
dim(cars)
```

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

```r
require(ResourceSelection)
```

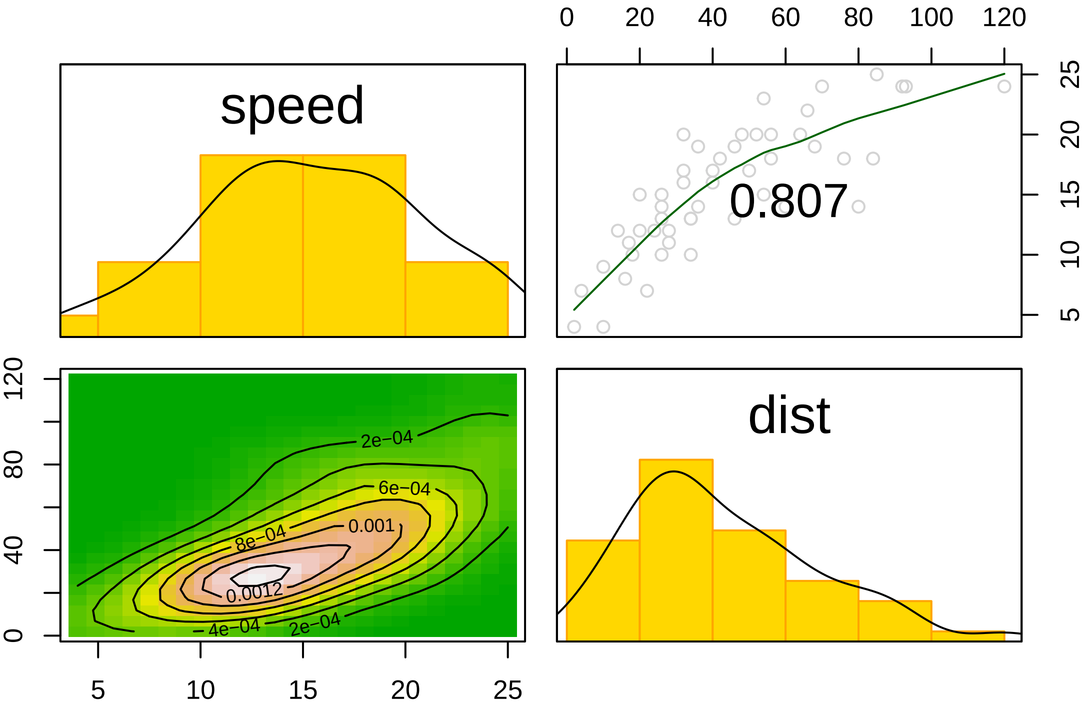
```
## Loading required package: ResourceSelection

## ResourceSelection 0.3-6   2023-06-27
```

```r
kdepairs(cars)
```

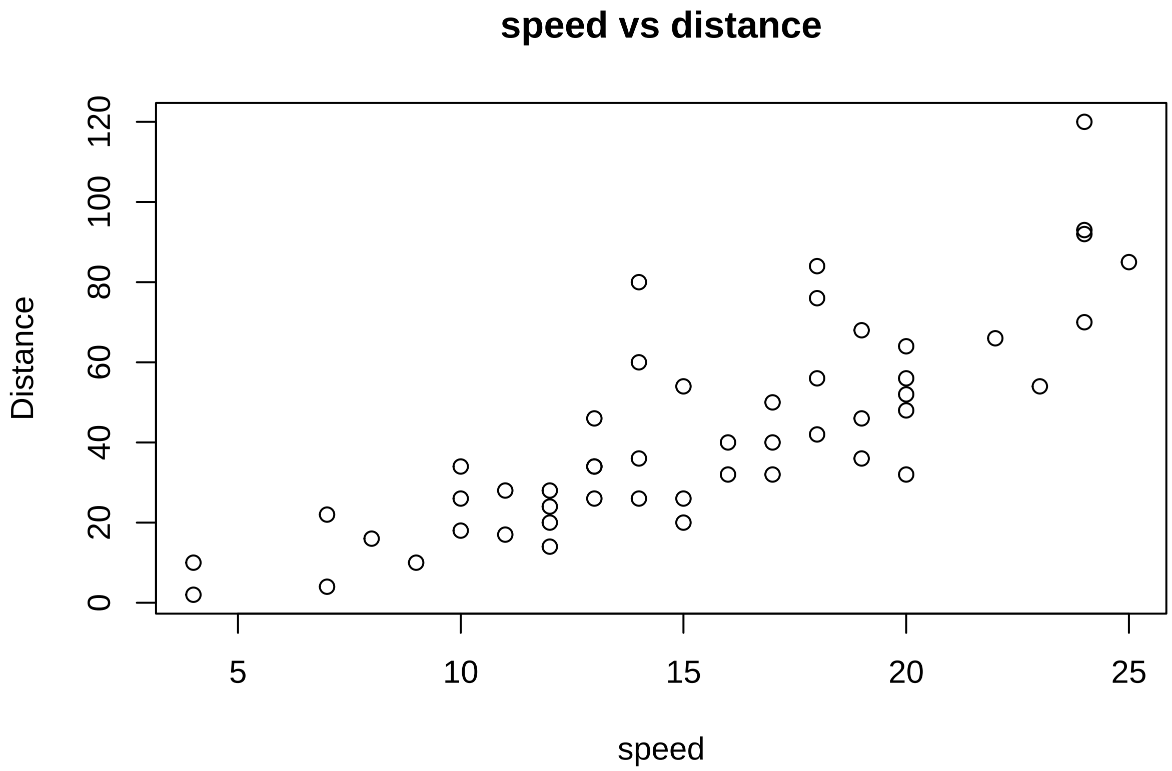
```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```

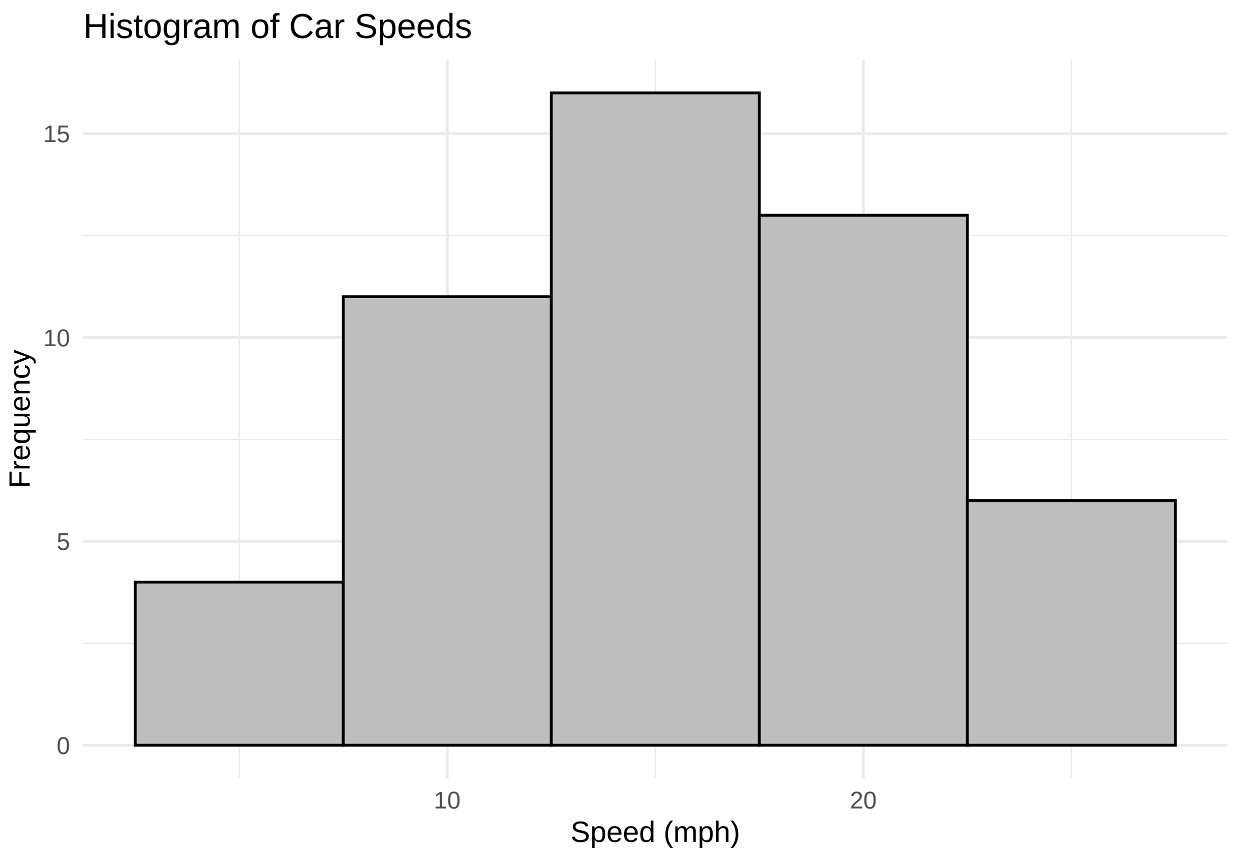


There 50 observations and 2 features from the data set. speed seems to increase with driving long distance. the speed variable is distributed in the center while the dist variable is distributed to the left. variables are well correlated with each other. they have a correlation of 0.807

```r
plot(cars[,"speed"],cars[,"dist"], main= "speed vs distance",
     xlab="speed", ylab="Distance")
```



```r
# see how data is skewed
ggplot(cars, aes(x=speed)) +
  geom_histogram(binwidth = 5, fill="grey", color="black") +
  labs(title="Histogram of Car Speeds", x="Speed (mph)", y="Frequency") +
  theme_minimal()
```

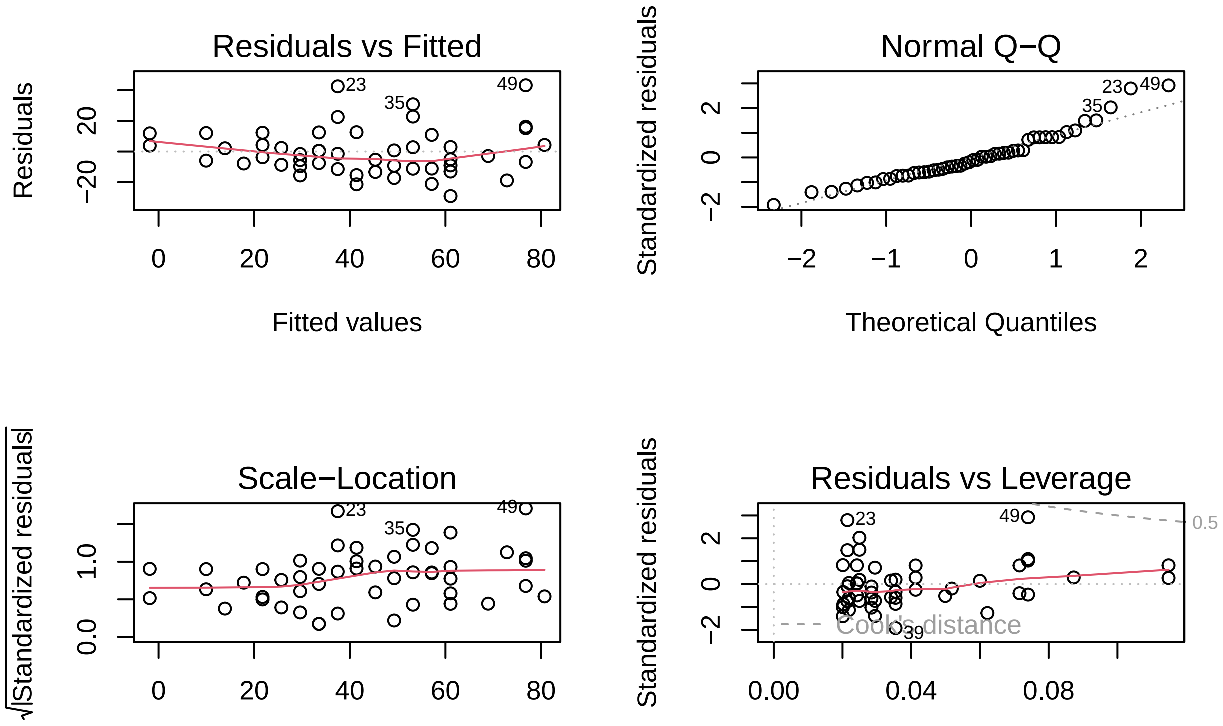


we are going to use regression to predict the cars speed using the distance variables. The correlation already shows that there is a good relationship among the variables. Based on the summary, it takes a car going 0 mph to stop - 17 feet to stop. every time need need to stop, it need to an additional to 3.9324 feet to stop.

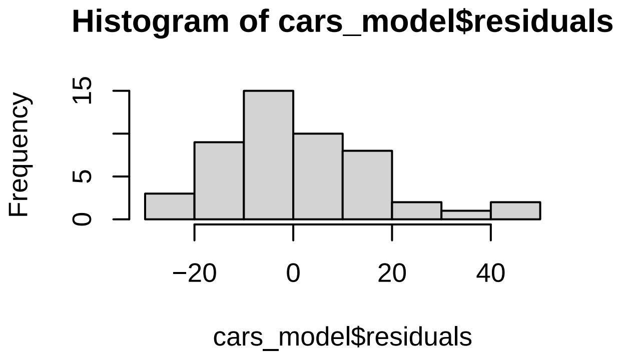
```
cars_model <- lm(dist ~ speed, data = cars)
summary(cars_model)
```

```
##
## 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
```

```r
par(ask=F)
par(mfrow=c(2,2))
plot(cars_model)
```

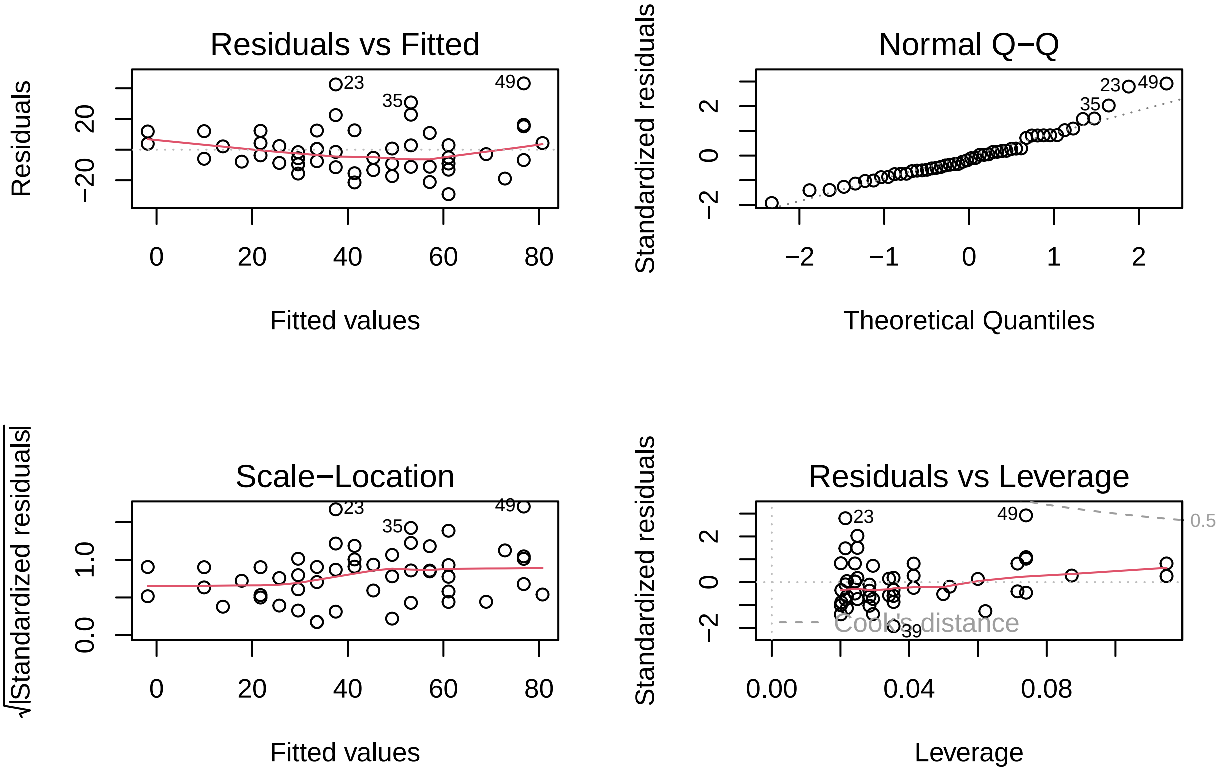


```r
hist(cars_model$residuals)
```

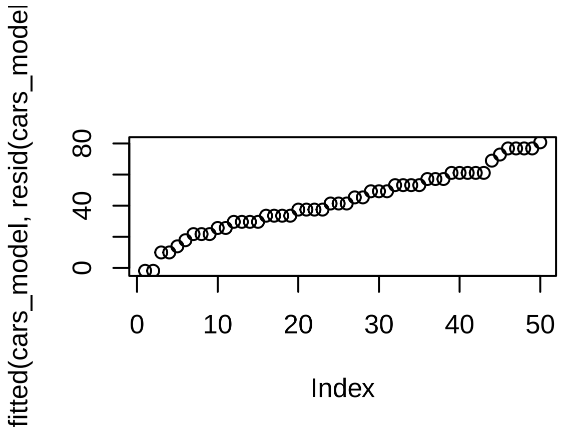


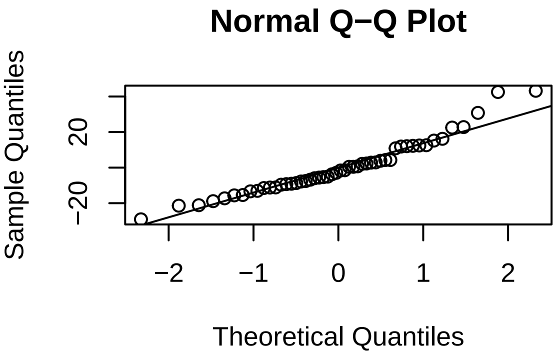
This fitted values graph shows a good relations between fitted values and and residual values. there are only 3 outliers which are 23, 35, 49. The data appear to be well modeled by a linear relationship between speed and dist, and the points appear to be randomly spread out about the line, with no discernible non-linear trends or indications of non-constant variance.

```r
par(ask=F)
par(mfrow=c(2,2))
plot(cars_model)
```



```r
plot(fitted(cars_model, resid(cars_model)))
qqnorm(resid(cars_model))
qqline(resid(cars_model))
```





Both features have good relationship , but we can use other evaluation techniques to improve the R^2 from 0.6511 to better score.