PShaji_Assignment11

Priya Shaji

11/9/2019

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.)

Step 1) Load the built-in dataset of R: `cars`

Step 2) Display first few rows of `cars` dataset

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
```

```
## ✔ ggplot2 3.1.1     ✔ purrr   0.3.2
## ✔ tibble  2.1.3     ✔ dplyr   0.8.3
## ✔ tidyr   0.8.3     ✔ stringr 1.4.0
## ✔ readr   1.3.1     ✔ forcats 0.4.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
head(cars)
```

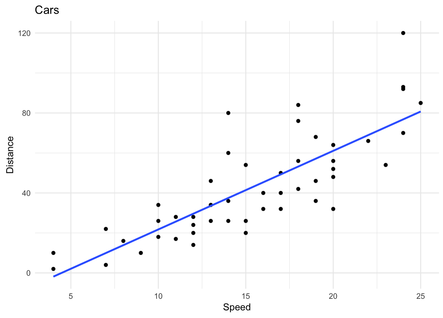
	speed <dbl>	dist <dbl>
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10

6 rows

Visualization

Step 3) Let us now visualize our data

```
cars %>%
  ggplot(aes(speed,dist)) +
  geom_point() +
  geom_smooth(method = lm,se = F)+
  labs(title = "Cars",
       x = "Speed", y = "Distance") +
  theme_minimal()
```



In the scatter plot above: there is positive linear trend between speed and distance.

Step 4) Build a linear model for stopping distance as a function of speed

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

```
##
## Call:
## lm(formula = speed ~ dist, data = cars)
##
## Coefficients:
## (Intercept)         dist
##      8.2839       0.1656
```

Quality evaluation of the model

Step 5) Summary of the linear model : `cars_lm`

```
summary(cars_lm)
```

```
##
## Call:
## lm(formula = speed ~ dist, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5293 -2.1550  0.3615  2.4377  6.4179
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.28391    0.87438   9.474 1.44e-12 ***
## dist         0.16557    0.01749   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.156 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 summary of the linear regression model shows that min-max and 1Q-3Q has approximately similar magnitudes and the median is close to zero.

Therefore, this model is good but lets do some more evaluation. The standard error is 49 times smaller than the corresponding coefficient.

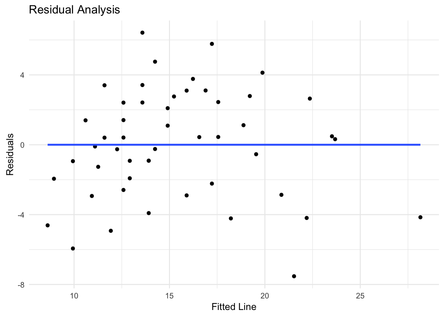
The p-value shows that the probability of this variable is very low to be considered irrelevant.

R-squared is 0.65, which means this model explains 65% of the data’s variation. Therefore,this is a good model.

Residual analysis

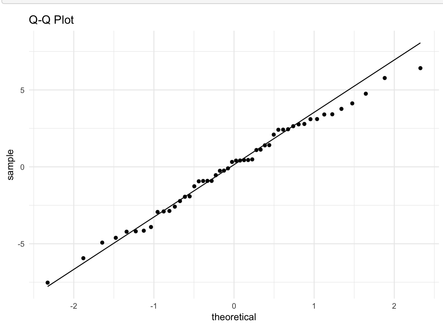
Step 6) Let’s build a residual plot for our linear model

```
cars_lm %>%
  ggplot(aes(fitted(cars_lm), resid(cars_lm))) +
  geom_point() +
  geom_smooth(method = lm, se =F) +
  labs(title = "Residual Analysis",
       x = "Fitted Line", y = "Residuals") +
  theme_minimal()
```



Step 7) Plot the Q-Q plot

```
cars_lm %>%
  ggplot(aes(sample = resid(cars_lm))) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Q-Q Plot") +
  theme_minimal()
```



As we see in the residual plot above, variance of residuals are not uniform whcih tells that our explanatory variable does not fully explain the data. But if we look at the quartile-quartile plot, we see that the residulas are normally distributed. Therefore, overall this is a good model.