Week 11

Shaya Engelman

2024-04-04

Instructions

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.)

Solution

```
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Warning: package 'stringr' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

The cars dataset comes built into R, so we can load it directly. We will first visualize the data we are working with and then build a linear model to predict stopping distance based on speed. We will then evaluate the model and perform residual analysis to check the quality of the model.

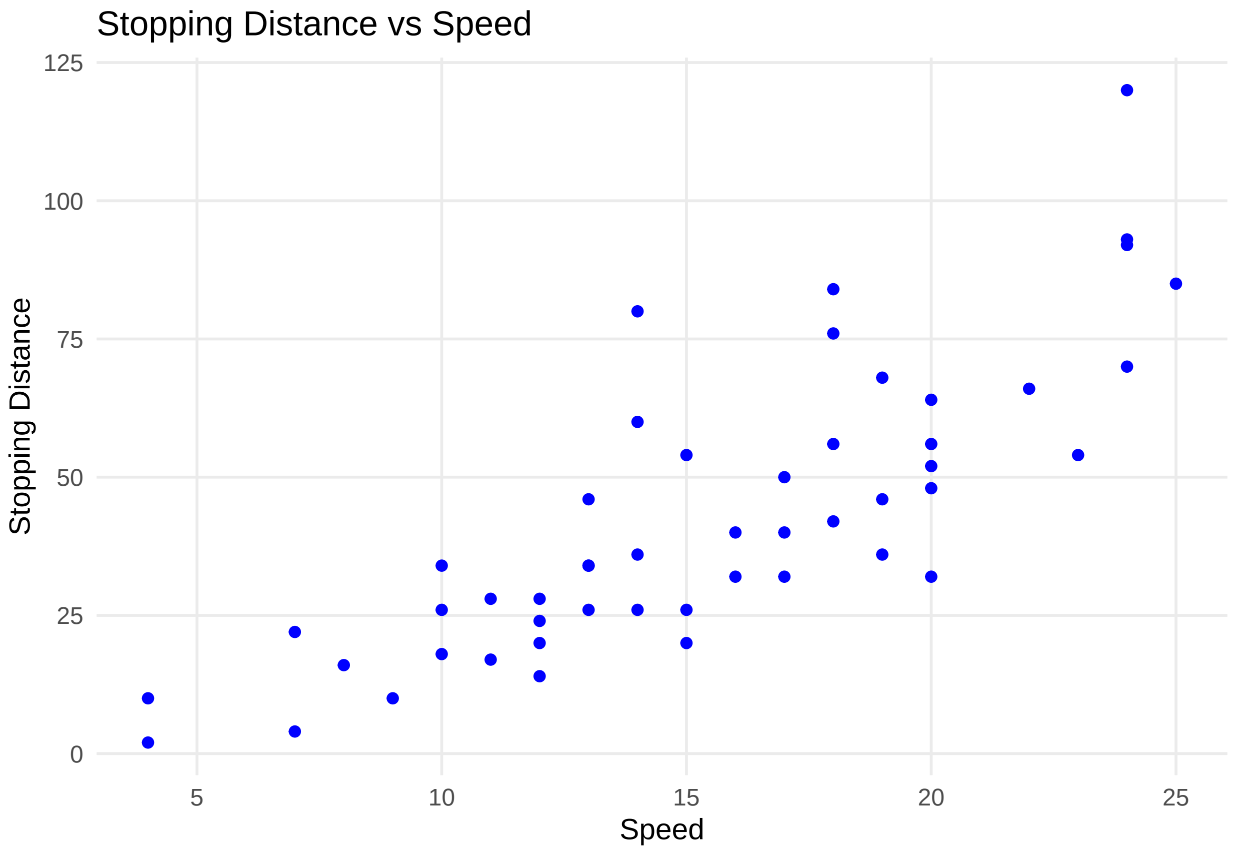
```
# Load the cars dataset
data(cars)
glimpse(cars)
```

```
## Rows: 50
## Columns: 2
## $ speed <dbl> 4, 4, 7, 7, 8, 9, 10, 10, 10, 11, 11, 12, 12, 12, 12, 13, 13, 13~
## $ dist  <dbl> 2, 10, 4, 22, 16, 10, 18, 26, 34, 17, 28, 14, 20, 24, 28, 26, 34~
```

There are only the two variables we are interested in in this dataset, there are 50 observations, each with a speed and stopping distance.

We can visualize the data using a scatter plot. The plot shows a clearly positive linear relationship between speed and stopping distance.

```
# Visualize the data
ggplot(cars, aes(x = speed, y = dist)) +
  geom_point(color = "blue") +
  labs(title = "Stopping Distance vs Speed",
       x = "Speed",
       y = "Stopping Distance") +
  theme_minimal() +
  theme(panel.grid.minor = element_blank())
```



Now we can build a linear model to predict stopping distance based on speed. We also recreate the above plot with the regression line drawn on it.

The coefficient of the speed variable in the model is 3.932, which means that for every unit increase in speed, the stopping distance increases by 3.9324 units. The intercept of the model is -17.5791, which is the stopping distance when the speed is 0.

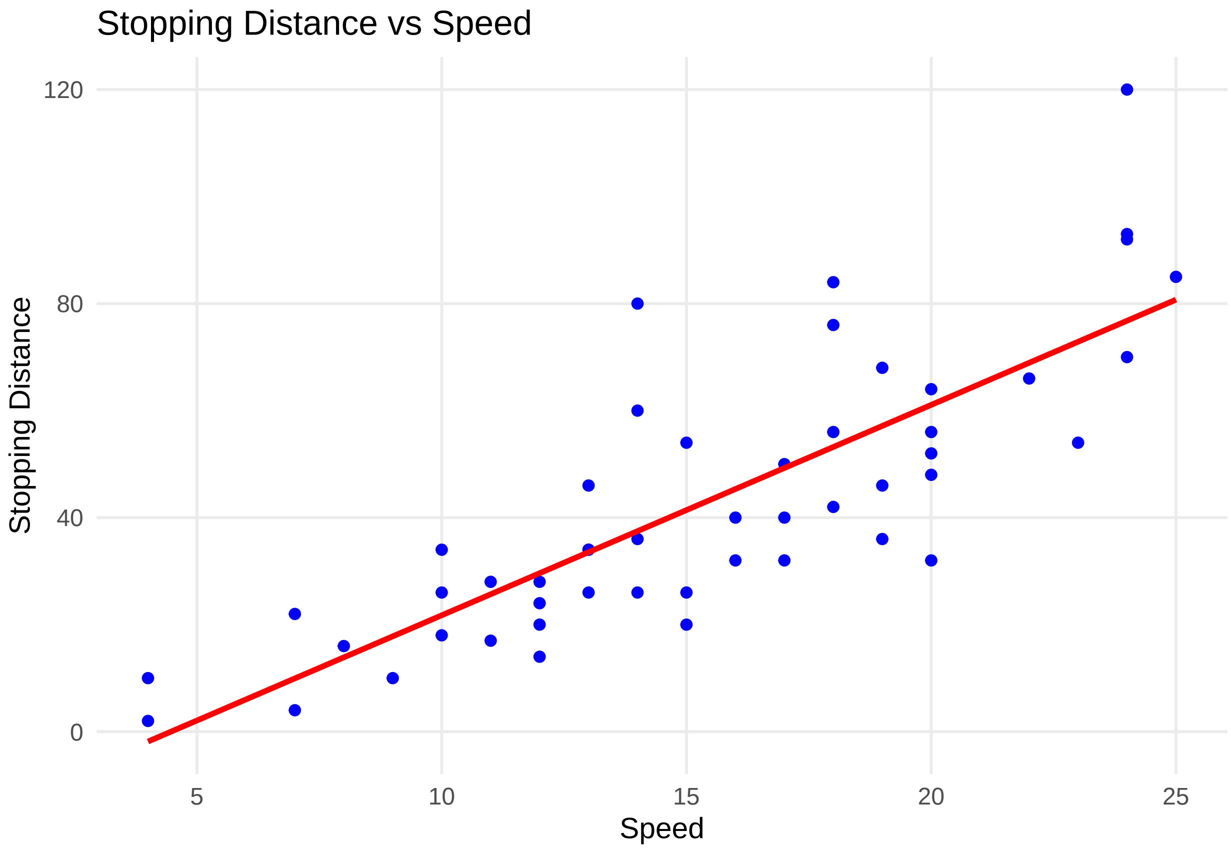
```
# Build the linear model
model <- lm(dist ~ speed, data = cars)
model
```

```
##
```

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

```
# Visualize the data with the regression line
ggplot(cars, aes(x = speed, y = dist)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Stopping Distance vs Speed",
       x = "Speed",
       y = "Stopping Distance") +
  theme_minimal() +
  theme(panel.grid.minor = element_blank())
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



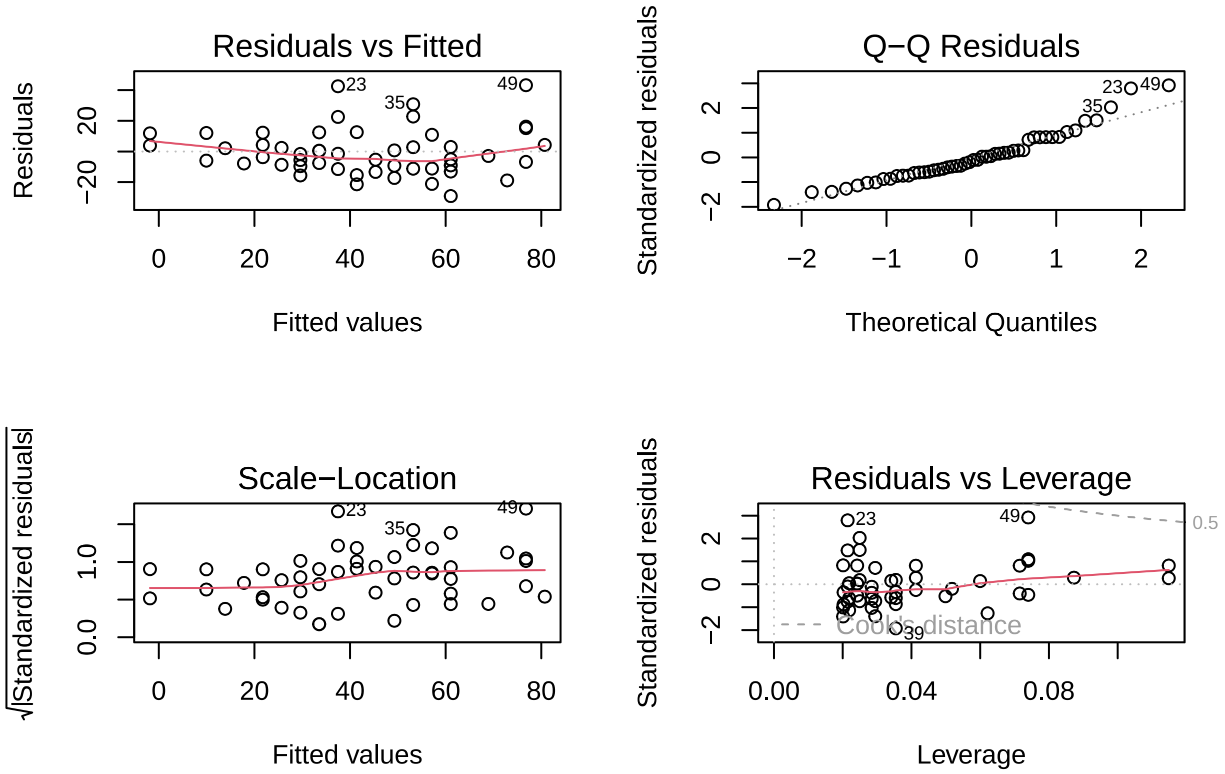
The model is built, and we can now evaluate its quality. We can do this by looking at the summary of the model. The summary below shows that the model has an R-squared value of 0.6511, which means that 65.11% of the variance in stopping distance is explained by the speed variable. The p-value of the F-statistic is less than 0.05, which means that the model is statistically significant. Finally, the median of the residuals is close to 0, the 1st and 3rd quartiles of a similar magnitude as are the minimum and maximum values. These all imply a good model fit.

```r
# Evaluate the model
summary(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
```

At first glance, the model seems to be a decent for the data. However, we need to perform residual analysis to check the quality of the model and whether the assumptions of linear regression are met.

```r
# Residual analysis
par(mfrow=c(2, 2))
plot(model)
```



The residuals vs fitted plot shows that the residuals are randomly distributed around 0, which is a good sign. The Q-Q plot shows that the residuals are normally distributed. These both imply that the model is a good fit.

In conclusion, we have built a linear model to predict stopping distance based on speed. The model is statistically significant and explains 65.11% of the variance in stopping distance. The residual analysis shows that the model is a good fit for the data.