605-HW11-Regression

Michael Y.

November 10, 2019

Contents

HW11 - Regression

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.)

```
cars <- datasets::cars
attach(cars)
# Summary of cars dataset
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

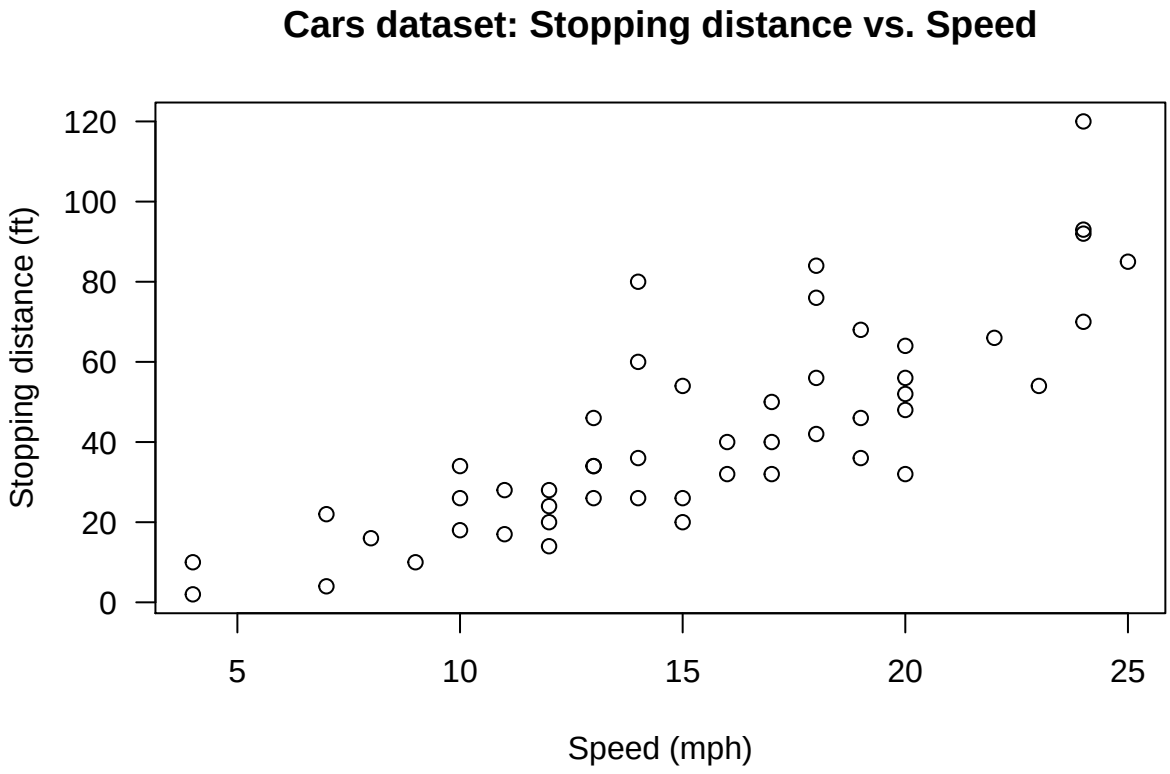
```
numrows = dim(cars)[1]
```

There are 50 observations of speed and stopping distance.

3.1 Visualize the Data

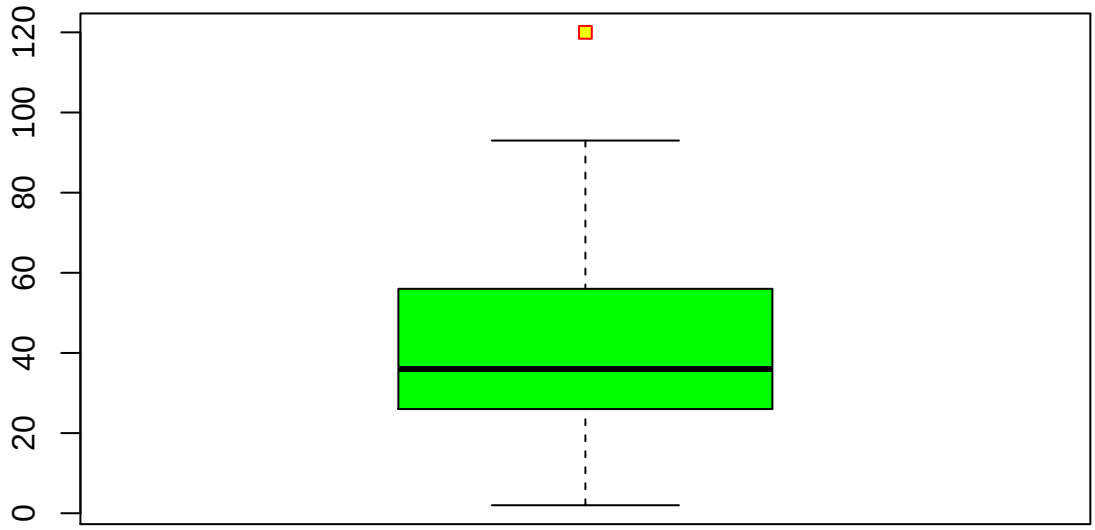
Scatterplot

```
# scatterplot
require(stats); require(graphics)
plot(cars, xlab = "Speed (mph)", ylab = "Stopping distance (ft)", las = 1)
title(main = "Cars dataset: Stopping distance vs. Speed")
```



Boxplot

```
bplot = boxplot(cars$dist,col = "green",outpch=22, outcol="red", outbg="yellow", plot = T)
```




```
outliers = bplot$out
outliers
```

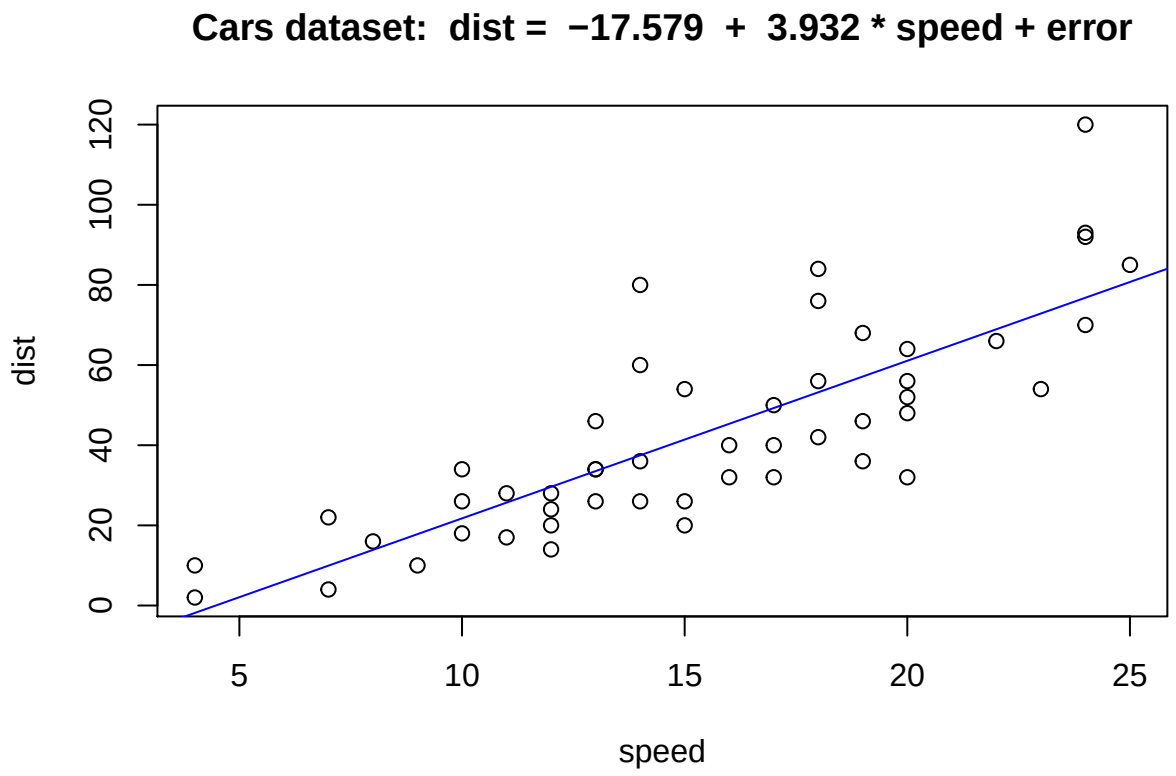
```
## [1] 120
```

There is an outlier at Stopping Distance = 120.

3.2 The Linear Model Function

Linear model: dist ~ speed

```
LinearModel <- lm(dist ~ speed, data = cars)
intercept = round(LinearModel$coefficients[1], 3)
slope =     round(LinearModel$coefficients[2], 3)
formula = paste ("dist = ", intercept, " + ", slope, "*", "speed + error")
plot(x = speed, y = dist,
     main=paste("Cars dataset: ", formula))
abline(reg = LinearModel,col="blue")
```



The intercept is negative, which means that at very low speeds, the predicted stopping distance would be negative.
(Of course, in reality, this is not possible. . .)

3.3 Evaluating the Quality of the Model

```
summary(LinearModel)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -29.06908  -9.52532  -2.27185   9.21472  43.20128
##
## Coefficients:
##               Estimate Std. Error  t value           Pr(>|t|)
## (Intercept) -17.579095   6.758440 -2.60106           0.012319 *
## speed         3.932409   0.415513  9.46399 0.0000000000014898 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.3796 on 48 degrees of freedom
## Multiple R-squared:  0.651079,   Adjusted R-squared:  0.64381
## F-statistic: 89.5671 on 1 and 48 DF,  p-value: 0.00000000000148984
```

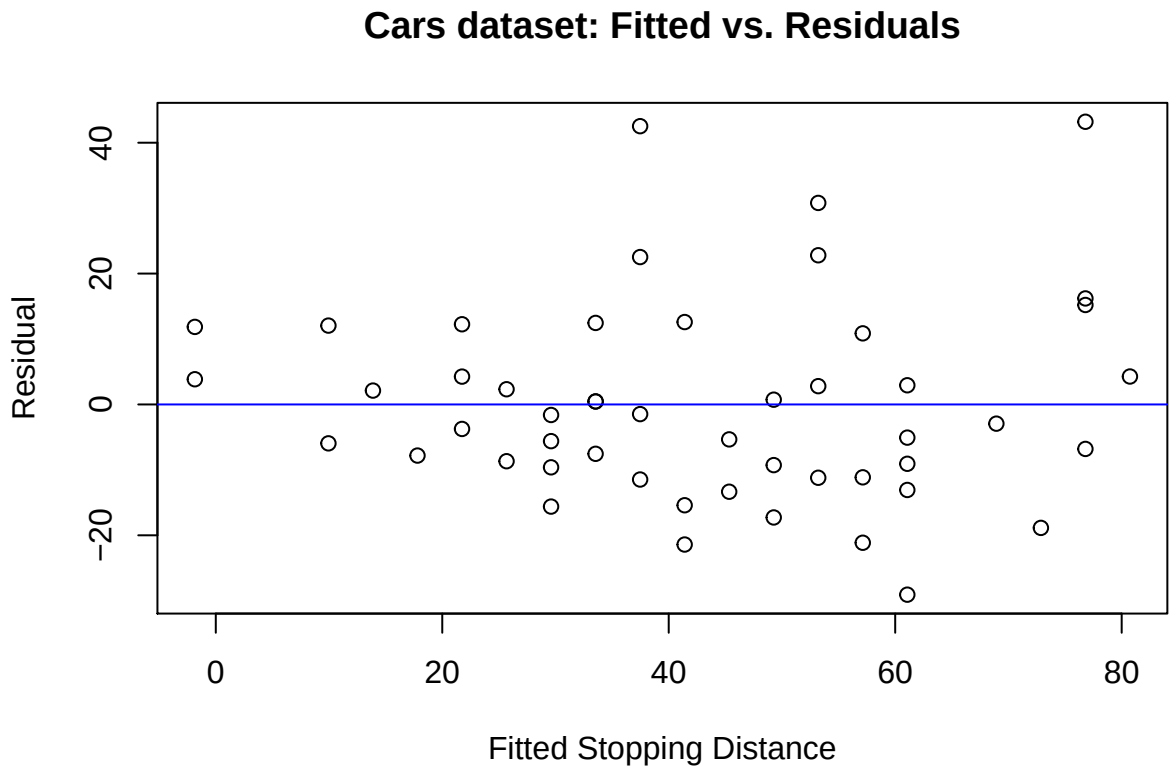
The model shows strong significance at 99% confidence on the slope of the speed parameter, but the intercept is significant only at 95% confidence.

The R^2 indicates that the model explains about 65 percent of the variance, which is reasonably good. This figure corresponds to correlation, R, of about 80 percent between the variables.

3.4 Residual Analysis

Plot residuals

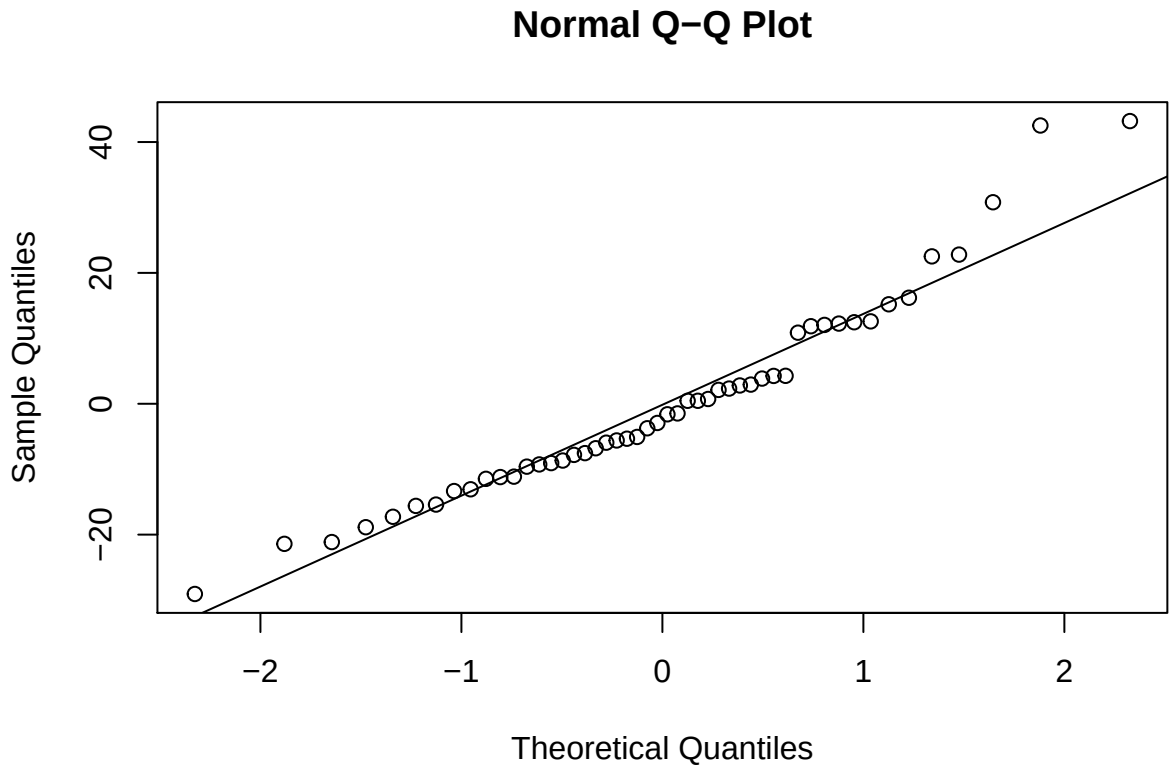
```
Residual = resid(LinearModel)
Fitted = fitted(LinearModel)
plot(Fitted,Residual, main="Cars dataset: Fitted vs. Residuals", xlab="Fitted Stopping Distance")
abline(h=0, col="blue")
```



The residuals do not show any recognizable pattern, though the variance at higher speeds does appear to be larger than that at lower speeds, which may indicate heteroscedasticity.

QQ plot

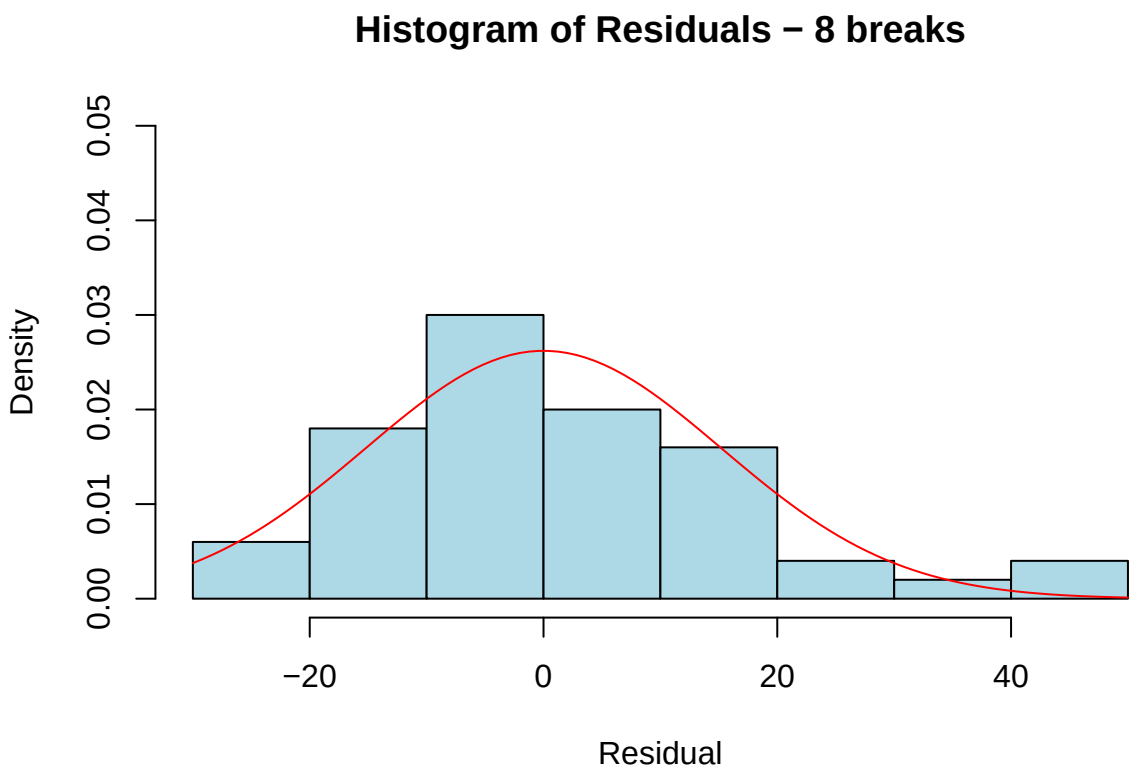
```
qqnorm(Residual)
qqline(Residual)
```



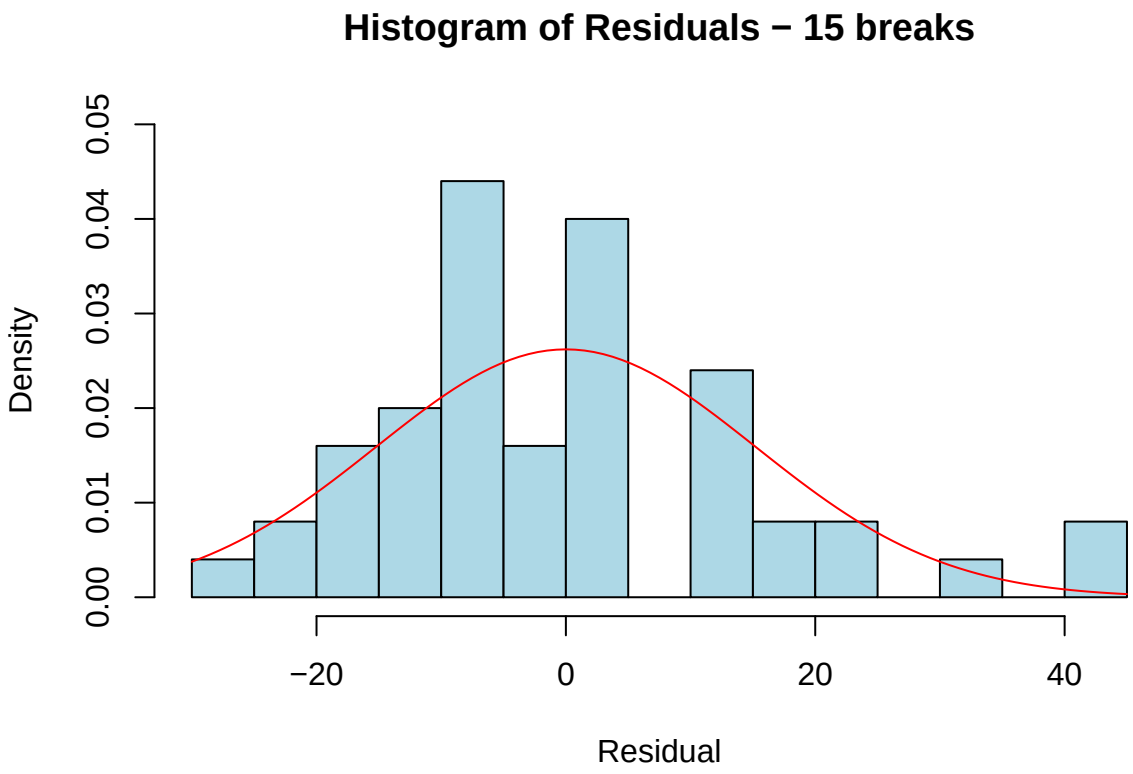
The QQ plot indicates some outliers at the upper end (actual stopping distance well above model prediction), which may call normality into question.

Plot histograms of Residuals

```
Residual = resid(LinearModel)
hist(Residual, main = "Histogram of Residuals - 8 breaks", ylab = "Density",
     ylim = c(0, 0.05),prob = TRUE,breaks=8, col="lightblue")
curve(dnorm(x, mean = mean(Residual), sd = sd(Residual)), col="red", add=TRUE)
```




```
hist(Residual, main = "Histogram of Residuals - 15 breaks", ylab = "Density",
     ylim = c(0, 0.05),prob = TRUE,breaks=15, col="lightblue")
curve(dnorm(x, mean = mean(Residual), sd = sd(Residual)), col="red", add=TRUE)
```



While the mean of the residuals is, by definition, zero, the **median** is **-2.27** .
The number of observations for which the residual is negative is **27**,
while the number of cases for which the residual is positive is **23** .
These figures are consistent with the graphs shown above.
Because the sample size is so small, it is difficult to determine from the above whether or not Normality is achieved.

Shapiro-Wilks test

H_0 : The residuals *are* normal H_A : The residuals are *not* normal

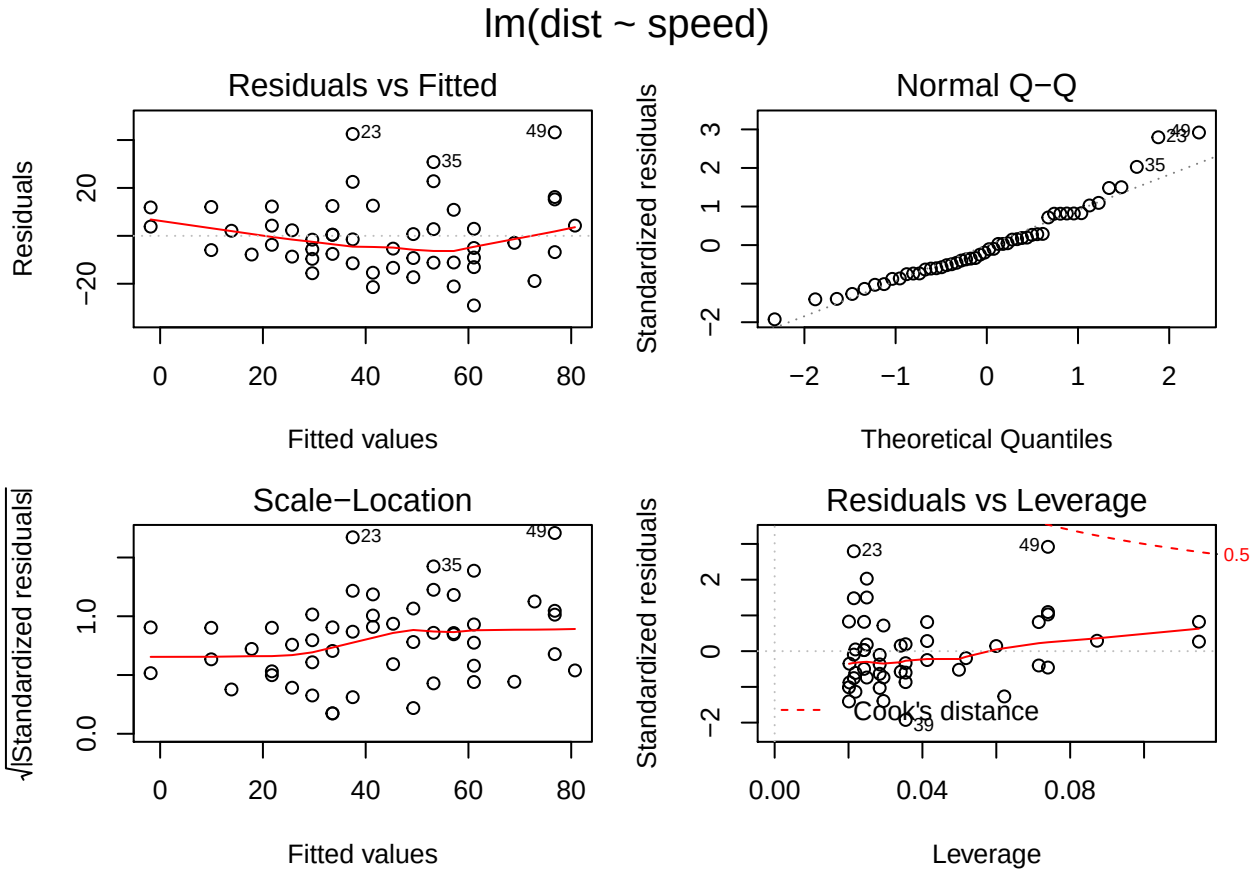
```
shapiro.test(Residual)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Residual
## W = 0.9450906, p-value = 0.0215246
```

Because the p-value (0.0215) is low, the null hypothesis is **rejected** at 95% confidence. This indicates that the residuals are **not** sufficiently close to the normal distribution to meet the conditions of linear regression.

More Plots

```
opar1 <- par(mfrow = c(2, 2),
             oma = c(0, 0, 1.1, 0),
             mar = c(4.1, 4.1, 2.1, 1.1))
plot(LinearModel)
```




```
par(opar1)
```

Conclusion

Although the conditions for linear regression are questionable due to a few outliers resulting in failure of the normality test on the residuals, the model is significant and the results seem adequate in explaining that about $\frac{2}{3}$ of the variance in stopping distance is attributable to speed, while the other $\frac{1}{3}$ of the variance remains unexplained by this model and thus must be attributable to other factors which have not been modeled.