Linear Regression

- Regression analysis is a very widely used statistical tool to establish a relationship model between two variables.
- Linear Regression is the most commonly used predictive modelling techniques.
- One of these variable is called predictor variable whose value is gathered through experiments.
- The other variable is called response variable whose value is derived from the predictor variable.
- Two variables are related through an equation where the power of both these variables is 1.
- Mathematically, a linear relationship represents a straight line when plotted as a graph.
- The general mathematical equation for linear regression is **Y=ax+b**

y is the response variable

x is the predictor variable

a and b are constants which are coefficients

Problem

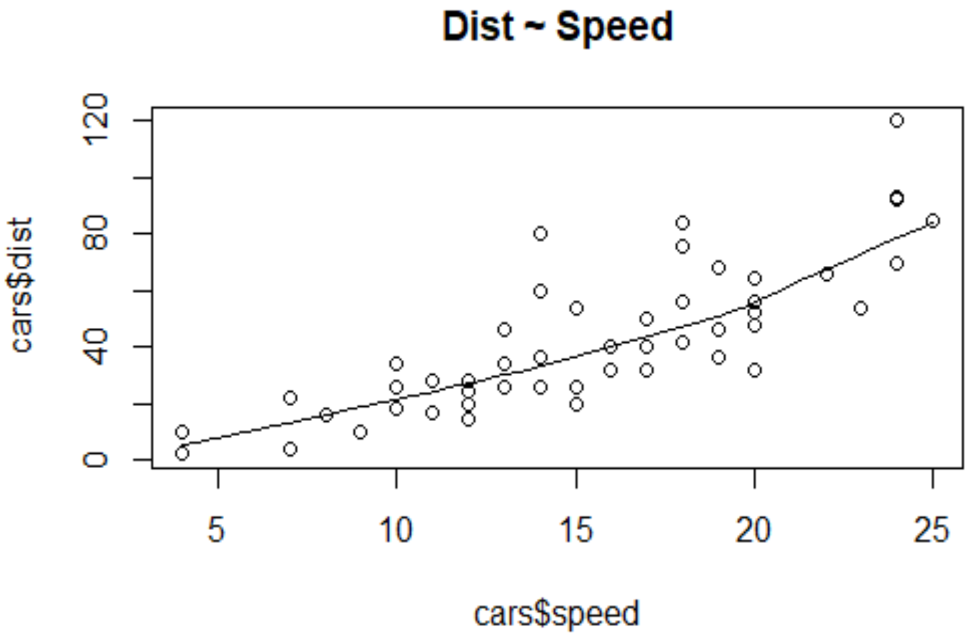
- The goal here is to establish a mathematical equation for dist as a function of speed, so you can use it to predict dist when only the speed of the car is known.
- So it is desirable to build a linear regression model with the response variable as dist and the predictor as speed.
- It is a good practice to analyse and understand the variables.
- The graphical analysis and correlation study below will help with this.

Graphical Analysis

- **Scatter plot:** Visualise the linear relationship between the predictor and response
- **Box plot:** To spot any outlier observations in the variable.
- **Density plot:** To see the distribution of the predictor variable.

Using Scatter Plot To Visualise The Relationship

- Scatter plots can help visualise linear relationships between the response and predictor variables.
- Ideally, if you have many predictor variables, a scatter plot is drawn for each one of them against the response, along with the line of best fit as seen below.

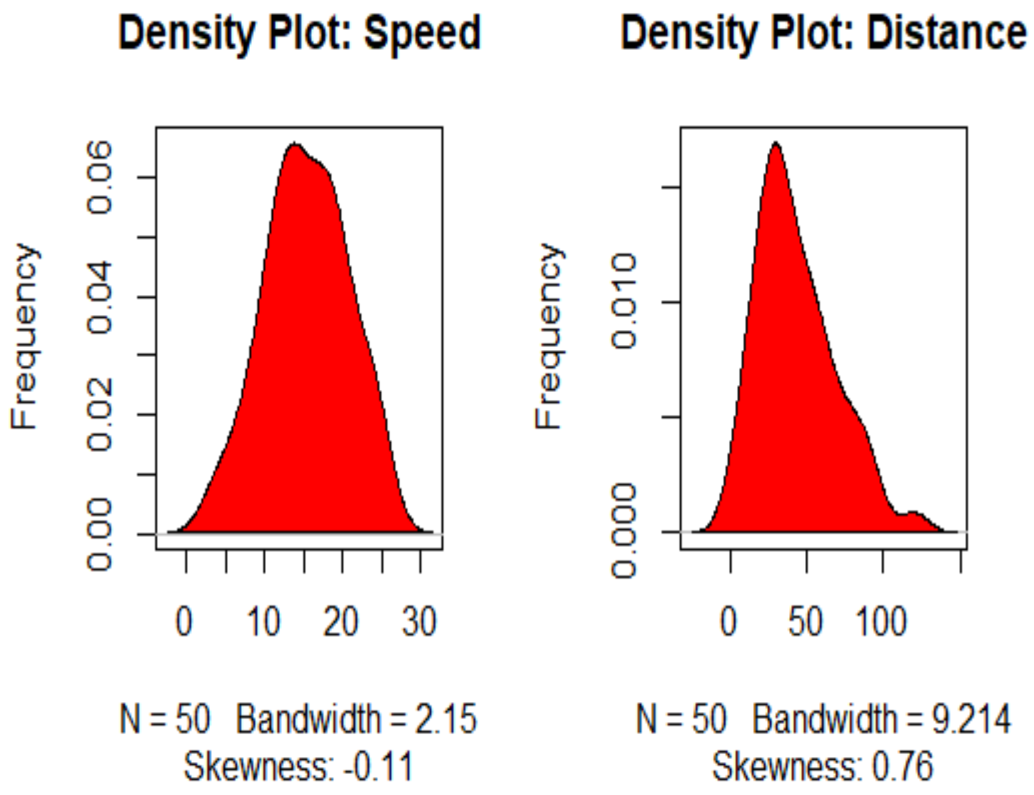


Using BoxPlot To Check For Outliers

- An outlier is any datapoint that lies outside the 1.5 * inter quartile range (IQR).
- IQR is calculated as the distance between the 25th percentile and 75th percentile values for that variable.

Using Density Plot To Check If Response Variable Is Close To Normal

- A **density plot** is a representation of the distribution of a numeric variable.
- It is used for density estimate to show the probability density function of the variable.



Correlation Analysis

- Correlation analysis studies the strength of relationship between two continuous variables.
- It involves computing the correlation coefficient between the two variables
- Correlation is a statistical measure that shows the degree of linear dependence between two variables.
- If one variables consistently increases with increasing value of the other, then they have a strong positive correlation (value close to +1).
- Similarly, if one consistently decreases when the other increase, they have a strong negative correlation (value close to -1).
- A value closer to 0 suggests a weak relationship between the variables.
- cor(cars$speed,cars$dist)
- [1]0.8068949

Building the Linear Regression Model

The function used for building linear models is `lm()`.
The lm() function takes in two main arguments:
1. Formula
2. Data

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

Mathematical Formula

By building the linear regression model, we have established the relationship between the predictor and response in the form of a mathematical formula.

Distance (dist) as a function for speed. the ‘Coefficients’ part having two components:
Intercept: -17.579, speed: 3.932.

- dist = Intercept + (β ∗ speed)
- dist = −17.579 + 3.932∗speed

Is this model statistically significant

- Printing the summary statistics for linearMod
- Summary(linearMod)

Summary

```
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
```

Using P-value to check for statistical significance

- The p-Values are very important.
- Because, we can consider a linear model to be statistically significant only when both these p-Values are less than the pre-determined statistical significance level of 0.05.
- This can visually interpreted by the significance stars at the end of the row against each X variable.
- The more the stars beside the variable’s p-Value, the more significant the variable.

Null and Alternate Hypotheses

- Whenever there is a p-value, there is always a Null and Alternate Hypothesis associated.

t-value

- When p Value is less than significance level (< 0.05), we can safely *reject the null hypothesis* that the co-efficient β of the predictor is zero.
- In our case, `linearMod`, both these p-Values are well below the 0.05 threshold.
- So, we can reject the null hypothesis and conclude the model is indeed statistically significant.
- It is very important for the model to be statistically significant before we go ahead and use it to predict the dependent variable.
- Otherwise, the confidence in predicted values from that model reduces and may be constructed as an event of chance.

To calculate t-statistic and P-values

- When the model co-efficients and standard error are known, the formula for calculating t Statistic and p-Value is
- t-value=coefficients/stderror

Residuals

- Residuals are essentially the difference between the actual observed response values (distance to stop dist in our case) and the response values that the model predicted.
- The Residuals section of the model output breaks it down into 5 summary points.
- When assessing how well the model fit the data, you should look for a symmetrical distribution across these points on the mean value zero (0).
- We can see that the distribution of the residuals do not appear to be strongly symmetrical.
- That means that the model predicts certain points that fall far away from the actual observed points

Coefficients

In simple linear regression, the coefficients are two unknown constants that represent the *intercept* and *slope* terms in the linear model.

Coefficient-Estimate

Coefficient-Std.Error

Estimate

- Estimated value is -17.5791. It’s the average value for y when x=0.
- It does not mean anything for inference.
- Practical significance is $\beta 1$: Given one unit increase in X this is expected to change in Y on an average.
- For every 1mph increase in the speed of a car the required distance to stop goes by 3.9324 feet.

Standard Error

- The coefficient Standard Error measures the average amount that the coefficient estimates vary from the actual average value of our response variable.
- The Standard Error can be used to compute an estimate of the expected difference in case we ran the model again and again.
- The required distance for a car to stop can vary by **0.4155128** feet.
- The Standard Errors can also be used to compute confidence intervals and to statistically test the hypothesis of the existence of a relationship between speed and distance required to stop.

Coefficient - t value

- The coefficient t-value is a measure of how many standard deviations our coefficient estimate is far away from 0.
- The value we want it to be far away from zero as this would indicate we could reject the null hypothesis - that is, we could declare a relationship between speed and distance exist.
- In our example, the t-statistic values are relatively far away from zero and are large relative to the standard error, which could indicate a relationship exists.
- In general, t-values are also used to compute p-values.

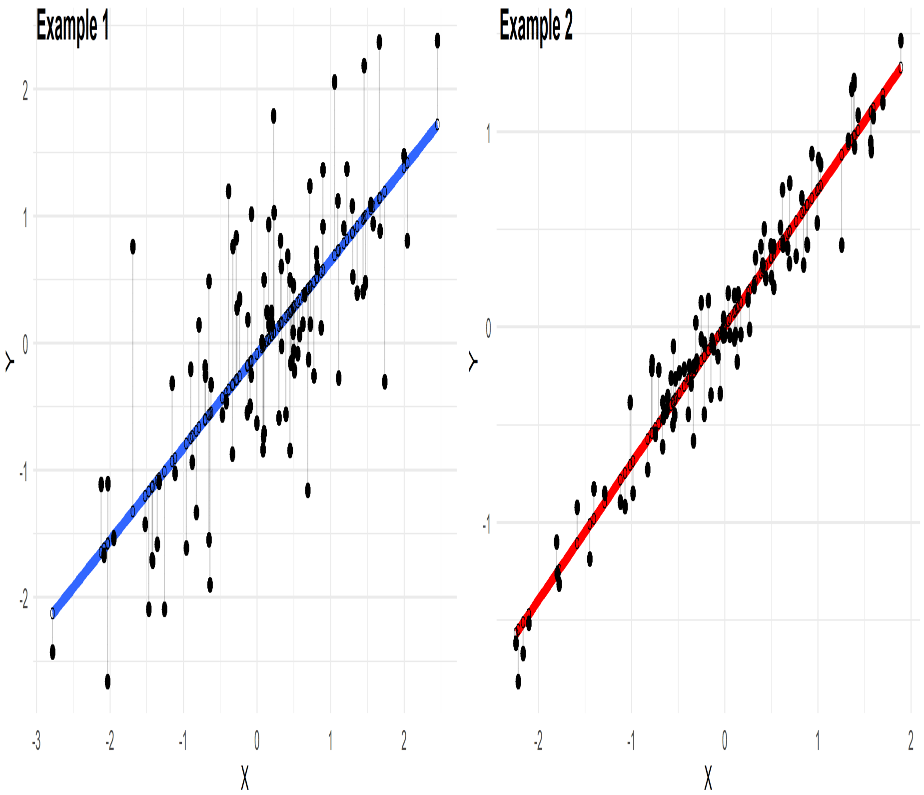
Residual Standard error(RSS)

- The **residual standard deviation** (or **residual standard error**) is a measure used to assess how well a linear regression model fits the data.
- Example 2 fits the data better than example 1 because the points are closer to the regression line.
- The gray vertical lines represent the error terms- the difference between the model and the true values of y.
- to quantify how far the data points are from the regression line, is to calculate the average distance from this line.

$$Average\ distance = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)}{\text{n}}$$

calculate the sum of these squared distances for all data points, and then take the square root of this sum to obtain the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\text{n}}}$$



Residual Standard Error

- Residual Standard Error is measure of the *quality* of a linear regression fit.
- every linear model is assumed to contain an error term *E*.
- Due to the presence of this error term, we are not capable of perfectly predicting our response variable (dist) from the predictor (speed) one.
- The Residual Standard Error is the average amount that the response (dist) will deviate from the true regression line.
- The actual distance required to stop can deviate from the true regression line by approximately **15.3795867** feet, on average.
- The mean distance for all cars to stop is **42.98** and that the Residual Standard Error is **15.3795867**, we can say that the percentage error is **35.78%**.
- The Residual Standard Error was calculated with 48 degrees of freedom.

Multiple R-squared, Adjusted R-squared

- The R-squared statistic provides a measure of how well the model is fitting the actual data.

- It takes the form of a proportion of variance. R is a measure of the linear relationship between our predictor variable (speed) and our response / target variable (dist).

- It always lies between 0 and 1

F-Statistic

- F-statistic is a good indicator of whether there is a relationship between our predictor and the response variables.
- The further the F-statistic is from 1 the better it is.
- When the number of data points is large, an F-statistic that is only a little bit larger than 1 is already sufficient to reject the null hypothesis (H0 : There is no relationship between speed and distance).

Diagnostic plots for Linear Regression Analysis

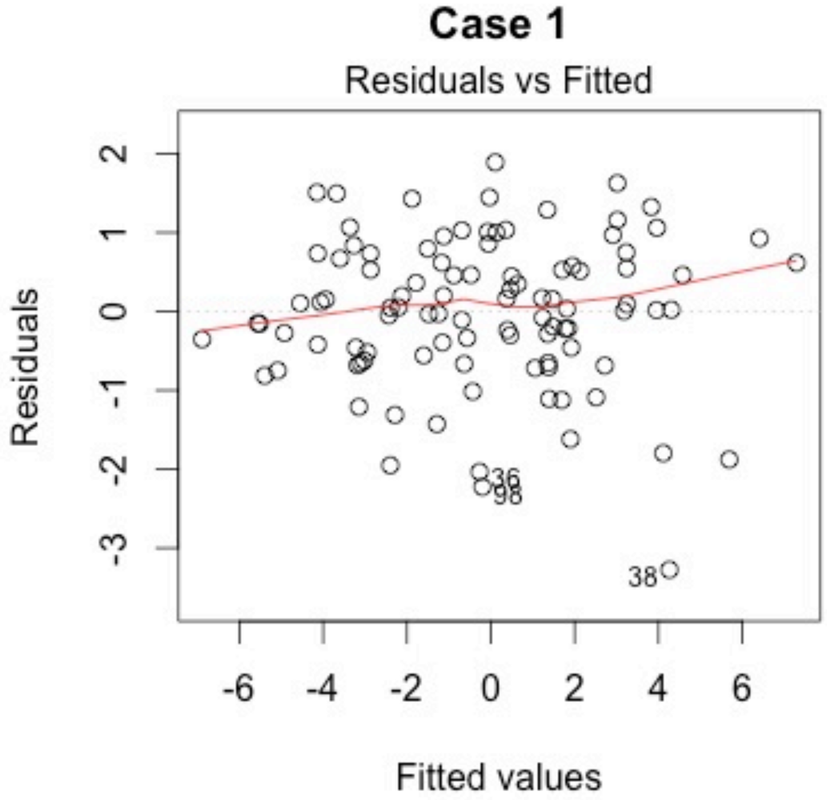
- To check if a model works well for data I can be done in many different ways.
- We pay great attention to regression results, such as slope coefficients, p-values, or R^2 that tell us how well a model represents given data.
- Residuals could show how poorly a model represents data.
- Residuals are leftover of the outcome variable after fitting a model (predictors) to data and they could reveal unexplained patterns in the data by the fitted model.
- Using this information, not only could you check if linear regression assumptions are met, but you could improve your model in an exploratory way.

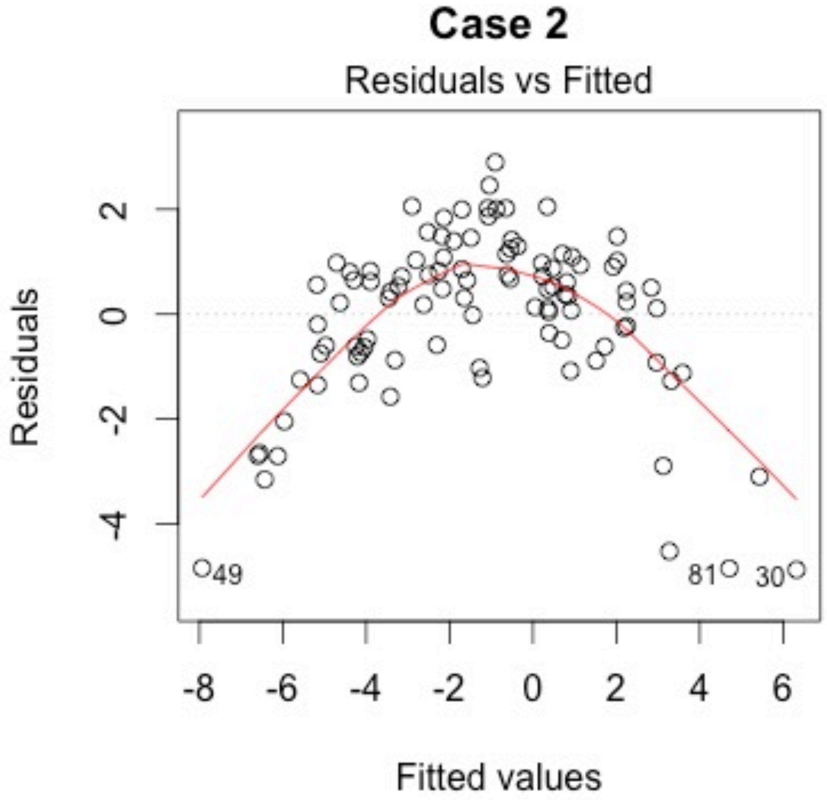
Residual plot

- A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis.
- If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data else , a non-linear model is more appropriate.

- This plot shows if residuals have non-linear patterns.
- There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model doesn’t capture the non-linear relationship.
- If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don’t have non-linear relationships.

Residuals vs Fitted

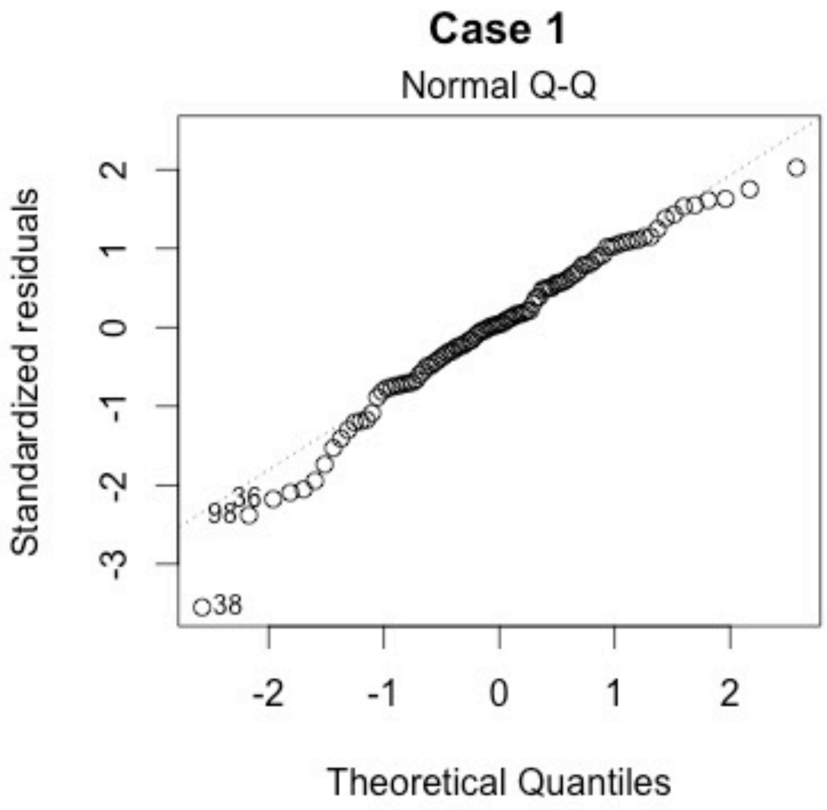




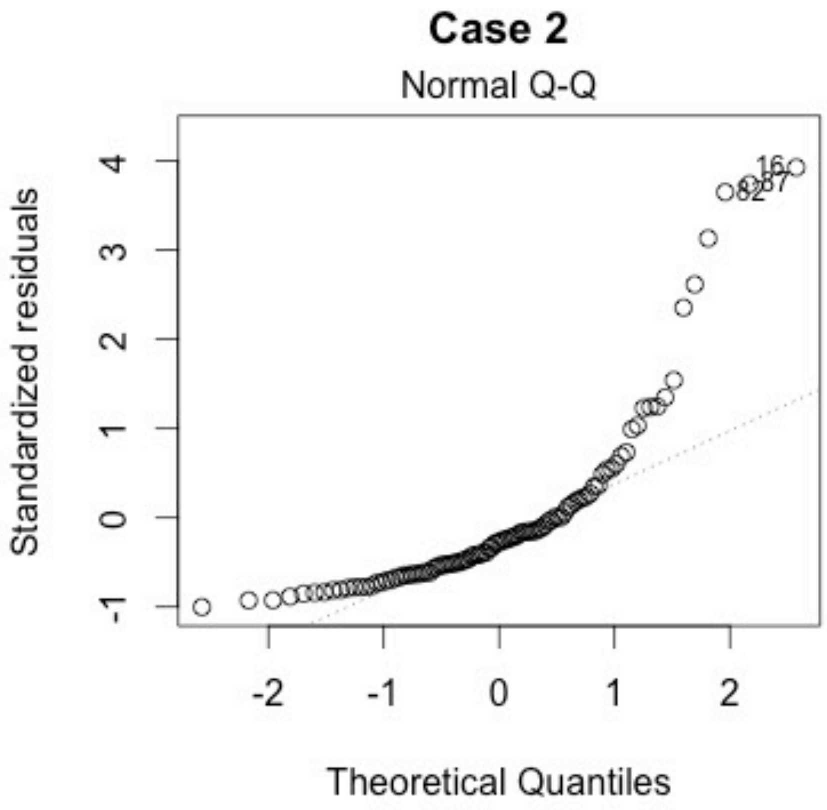
Normal Q-Q Plot

- This plot shows if residuals are normally distributed.
- It’s good if residuals are lined well on the straight dashed line.

Normal Q-Q



Case 1
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-3
-2
-1
0
1
2
98
36
38

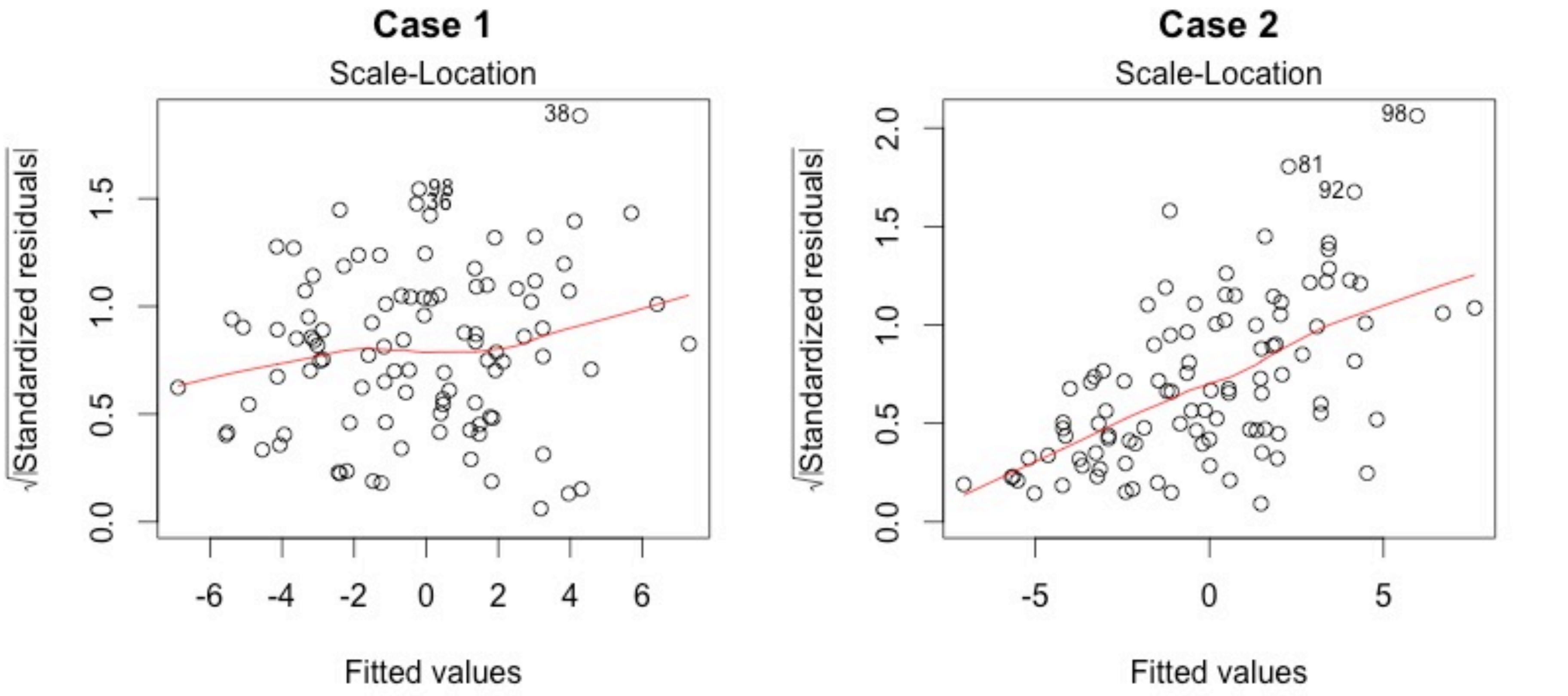


Case 2
Normal Q-Q
Standardized residuals
Theoretical Quantiles
-1
0
1
2
3
4
16
82
87

Scale-Location

- It’s also called Spread-Location plot.
- This plot shows if residuals are spread equally along the ranges of predictors.
- This is how you can check the assumption of equal variance (homoscedasticity).
- It’s good if you see a horizontal line with equally (randomly) spread points.

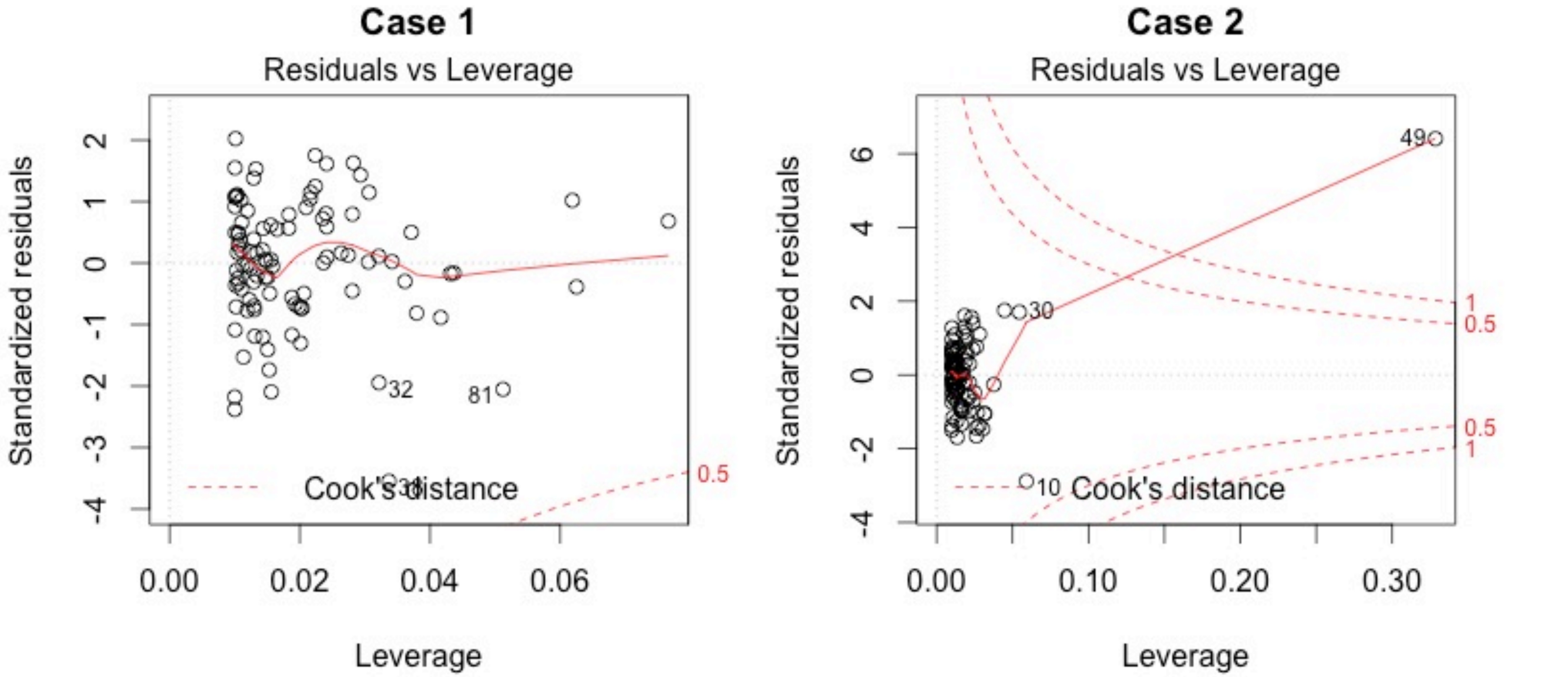
Scale Location



Residuals Vs Leverage

- This plot helps us to find influential cases if any.
- Not all outliers are influential in linear regression analysis .
- Even though data have extreme values, they might not be influential to determine a regression line.
- That means, the results wouldn’t be much different if we either include or exclude them from analysis.
- They follow the trend in the majority of cases and they don’t really matter; they are not influential.
- On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis.

Residuals Vs Leverage



Cook’s distance

- In statistics, **Cook's distance** or **Cook's *D*** is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis.

What does having patterns in residuals mean to your research?

- It tells you about your model and data.
- Your current model might not be the best way to understand your data if there’s so much good stuff left in the data.
- In that case, you may want to go back to your theory and hypotheses.
- Is it really a linear relationship between the predictors and the outcome
- Then a log transformation may better represent the phenomena that you’d like to model.
- Or, is there any important variable that you left out from your model?
- Other variables you didn’t include may play an important role in your model and data.
- Or, maybe, your data were systematically biased when collecting data. You may want to redesign data collection methods.
- ***Checking residuals is a way to discover new insights in your model and data!***

Regression model is best fit for the data

Statistic	Criteria
R-Squared	Higher the better
Adjusted R-Sqaured	Higher the better
F-statistic	Higher the better
Std Error	Close to zero the better
MSE	Lower the better

Predicting Linear Models

- Split your dataset into a 80:20 sample (training:test), then, build the model on the 80% sample and then use the model built to predict the dependent variable on test data.
- set.seed(100) # setting seed to reproduce results of random sampling
- trainingRowIndex <- sample(1:nrow(cars), 0.8*nrow(cars)) # row indices for training data
- trainingData <- cars[trainingRowIndex,] # model training data
- testData <- cars[-trainingRowIndex,]

set.seed()

- set.seed(n) function is used to produce results where n is the seed number which is an integer variable.

Fit the model on training data and predict on test data

- lmMod <- lm(dist ~ speed, data=trainingData) # build the model
- distPred <- predict(lmMod, testData) # predict distance
- distPred
- summary(lmMod)

Review diagnostic measures.

- From the model summary, the model p value and predictor’s p value are less than the significance level.
- So you have a statistically significant model.
- The R-Sq and Adj R-Sq are comparative to the original model built on full data.