Linear Regression

April 15, 2017

What is Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

General expression of linear regression:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + a_3 X_3 + \epsilon$$

Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables.

Consider the most simplest case: A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

A note about sample size. In Linear regression the sample size rule of thumb is that regression analysis requires at least 20 cases per independent variable in the analysis.

Response Variable

The response variable Y must be a continuous variable.

Predictor Variables

The predictors can be continuous, discrete or categorical variables.

Assumptions of Linear Regression (Not too important, but please keep in mind)

Linear regression is an analysis that assesses whether one or more predictor variables explains the dependent (criterion) variable. The regression has five key assumptions:

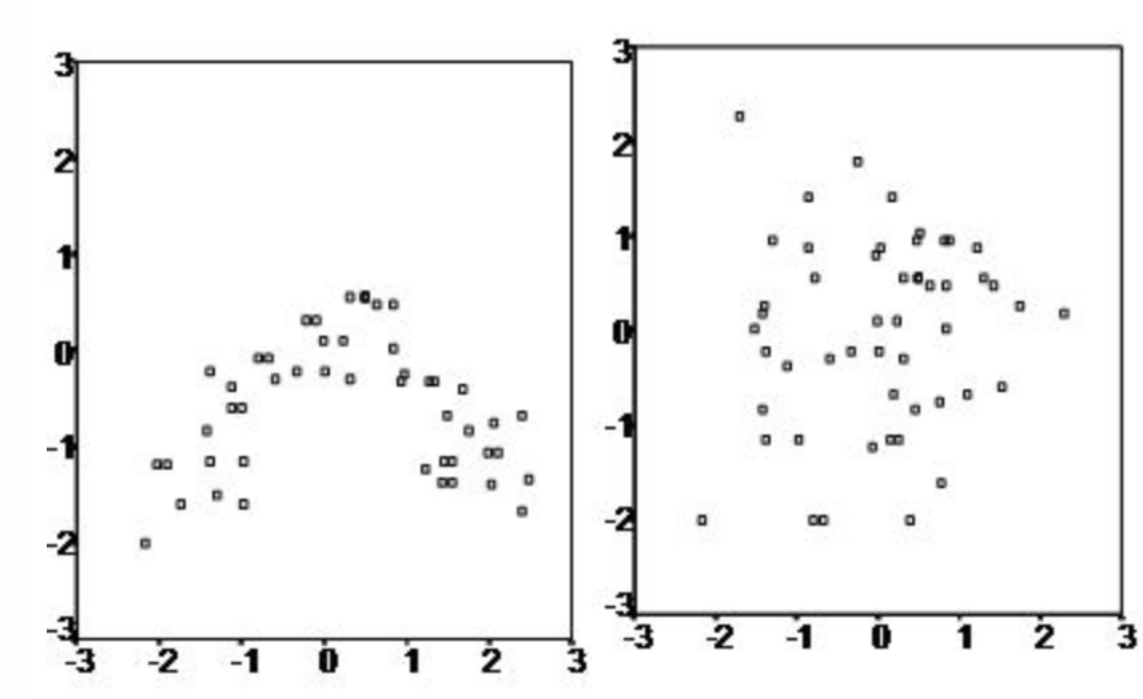


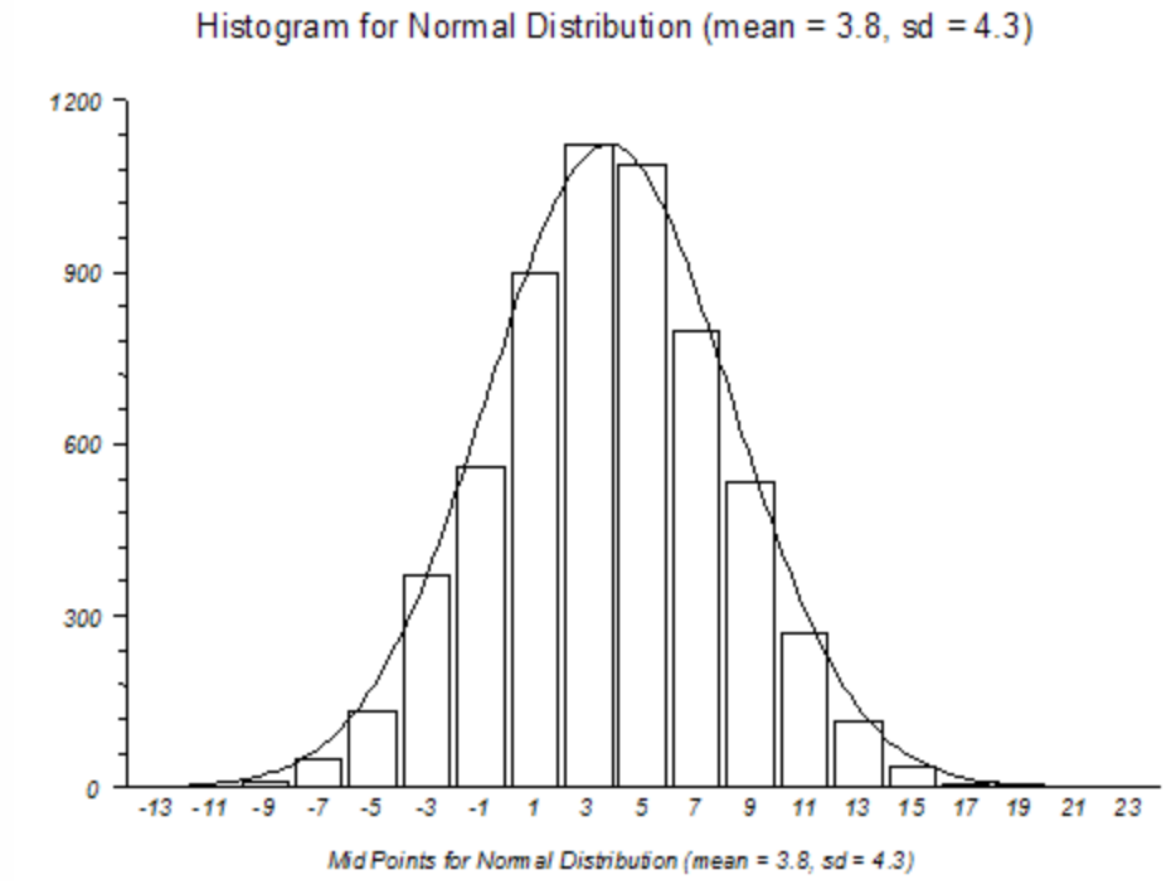
Figure 1:

Linear Relationship

Firstly, linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects. The linearity assumption can best be tested with scatter plots, the following two examples depict two cases, where no and little linearity is present.

Multivariate normality

Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram and a fitted normal curve or a Q-Q-Plot. Normality can be checked with a goodness of fit test, e.g., the Kolmogorov-Smirnof test. When the data is not normally distributed a non-linear transformation, e.g., log-transformation might fix this issue, however it can introduce effects of multicollinearity.

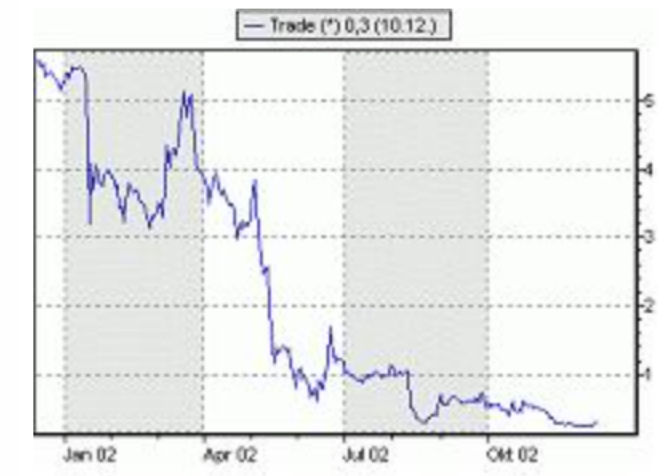


No or little multicollinearity

Thirdly, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are not independent from each other. A second important independence assumption is that the error of the mean has to be independent from the independent variables.
One of the way to test multicolonearity is Variance Inflation Factor (VIF).

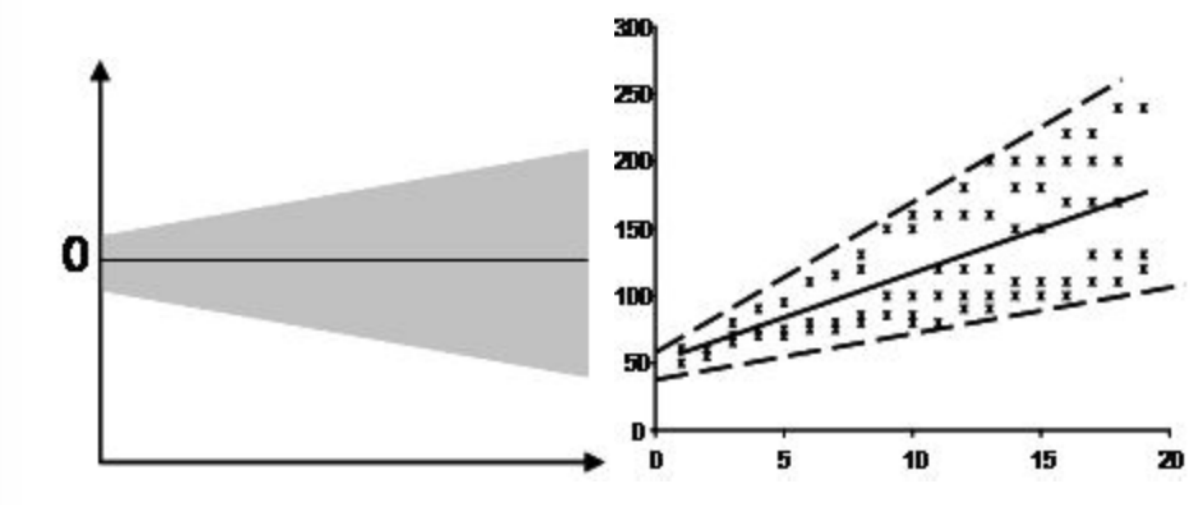
No auto-correlation

Fourthly, linear regression analysis requires that there is little or no autocorrelation in the data. Autocorrelation occurs when the residuals are not independent from each other. In other words when the value of y(x+1) is not independent from the value of y(x). This for instance typically occurs in stock prices, where the price is not independent from the previous price.



Constant Variances

The last assumption the linear regression analysis makes is homoscedasticity. The scatter plot is good way to check whether homoscedasticity (that is the error terms along the regression are equal) is given. If the data is heteroscedastic the scatter plots looks like the following examples:



Residuals

Once a regression model has been fit to a group of data, examination of the residuals (the deviations from the fitted line to the observed values) allows the modeler to investigate the validity of his or her assumption that a linear relationship exists. Plotting the residuals on the y-axis against the explanatory variable on the x-axis reveals any possible non-linear relationship among the variables, or might alert the modeler to investigate other variables.

Interpretation of Linear Regression

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant.

The key to understanding the coefficients is to think of them as slopes, and they're often called slope coefficients.

Let's See an Example

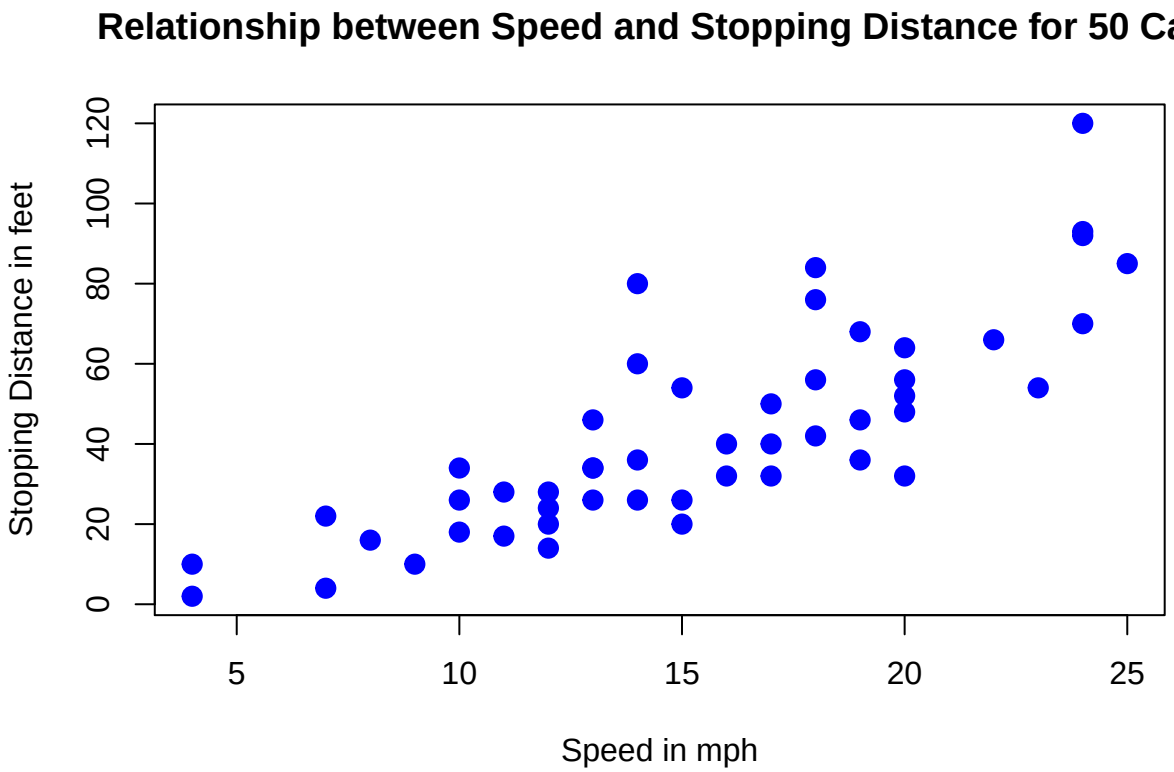
In the example below, we'll use the cars dataset found in the datasets package in R (for more details on the package you can call: library(help = "datasets")):

```
data(cars)
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
```

The cars dataset gives Speed and Stopping Distances of Cars. This dataset is a data frame with 50 rows and 2 variables. The rows refer to cars and the variables refer to speed (the numeric Speed in mph) and dist (the numeric stopping distance in ft.). As the summary output above shows, the cars dataset's speed variable varies from cars with speed of 4 mph to 25 mph (the data source mentions these are based on cars from the '20s! - to find out more about the dataset, you can type ?cars). When it comes to distance to stop, there are cars that can stop in 2 feet and cars that need 120 feet to come to a stop.
Below is a scatterplot of the variables:

```
plot(cars, col='blue',
     pch=20,
     cex=2,
     main="Relationship between Speed and Stopping Distance for 50 Cars",
     xlab="Speed in mph",
     ylab="Stopping Distance in feet")
```



From the plot above, we can visualise that there is a somewhat strong relationship between a cars' speed and the distance required for it to stop (i.e.: the faster the car goes the longer the distance it takes to come to a stop).

```
mod1 = lm(formula = dist ~ speed, data = cars)
summary(mod1)
```

```
##
## 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
```

Note that for this example we are not too concerned about actually fitting the best model but we are more interested in interpreting the model output - which would then allow us to potentially define next steps in the model building process The model above is achieved by using the lm() function in R and the output is called using the summary() function on the model.

Residuals

The next item in the model output talks about the 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). In our example, 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. We could take this further consider plotting the residuals to see whether this normally distributed, etc. but will skip this for this example.

Coefficient - Estimate

The coefficient Estimate contains two rows; the first one is the intercept. The intercept, in our example, is essentially the expected value of the distance required for a car to stop when we consider the average speed of all cars in the dataset. In other words, it takes an average car in our dataset 42.98 feet to come to a stop. The second row in the Coefficients is the slope, or in our example, the effect speed has in distance required for a car to stop. The slope term in our model is saying that for every 1 mph increase in the speed of a car, the required distance to stop goes up by 3.9324088 feet.

Coefficient - Pr(>|t|)

The Pr(>|t|) acronym found in the model output relates to the probability of observing any value equal or larger than |t|. A small p-value indicates that it is unlikely we will observe a relationship between the predictor (speed) and response (dist) variables due to chance. Typically, a p-value of 5% or less is a good cut-off point. In our model example, the p-values are very close to zero. Note the 'signif. Codes' associated to each estimate. Three stars (or asterisks) represent a highly significant p-value. Consequently, a small p-value for the intercept and the slope indicates that we can reject the null hypothesis which allows us to conclude that there is a relationship between speed and distance.

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. The R-squared 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 (i.e.: a number near 0 represents a regression that does not explain the variance in the response variable well and a number close to 1 does explain the observed variance in the response variable). In our example, the R-squared we get is 0.6510794. Or roughly 65% of the variance found in the response variable (dist) can be explained by the predictor variable (speed).

Nevertheless, it's hard to define what level of R-squared is appropriate to claim the model fits well. Essentially, it will vary with the application and the domain studied.

Categorical Variables

People often wonder how they can include categorical variables in their regression models. With R this is extremely easy. Just include the categorical variable in your regression formula and R will take care of the rest. R calls categorical variables factors. A factor has a set of levels, or possible values. These levels will show up as variables in the model summary.

Miltivariate Linear Regression

Regression with more than 1 predictoes called multivariate linear regression. Recall the general formula:

$$Y = a_0 + a_1X_1 + a_2X_2 + \epsilon^{**}$$

One example would be a model of the height of a shrub (Y) based on the amount of bacteria in the soil X_1 and whether the plant is located in partial or full sun X_2. Height is measured in cm, bacteria is measured in thousand per ml of soil, and type of sun = 0 if the plant is in partial sun and type of sun = 1 if the plant is in full sun.
Let's say it turned out that the regression equation was estimated as follows:

$$Y = 42 + 2.3X_1 + 11X_2$$

Interpreting the Intercept

a_0, the Y-intercept, can be interpreted as the value you would predict for Y if both $X_1 = 0$ and $X_2 = 0$. We would expect an average height of 42 cm for shrubs in partial sun with no bacteria in the soil. However, this is only a meaningful interpretation if it is reasonable that both X_1 and X_2 can be 0, and if the data set actually included values for X_1 and X_2 that were near 0. If neither of these conditions are true, then a_0 really has no meaningful interpretation. It just anchors the regression line in the right place. In our case, it is easy to see that X_2 sometimes is 0, but if X_1, our bacteria level, never comes close to 0, then our intercept has no real interpretation.

Interpreting Coefficients of Continuous Predictor Variables

Since X_1 is a continuous variable, a_1 represents the difference in the predicted value of Y for each one-unit difference in X_1, if X_2 remains constant. This means that if X_1 differed by one unit (and X_2 did not differ) Y will differ by a_1 units, on average.
In our example, shrubs with a 5000 bacteria count would, on average, be 2.3 cm taller than those with a 4000/ml bacteria count, which likewise would be about 2.3 cm taller than those with 3000/ml bacteria, as long as they were in the same type of sun.
(Don't forget that since the bacteria count was measured in 1000 per ml of soil, 1000 bacteria represent one unit of X_1).

Interpreting Coefficients of Categorical Predictor Variables

Similarly, a_2 is interpreted as the difference in the predicted value in Y for each one-unit difference in X_2, if X_1 remains constant. However, since X_2 is a categorical variable coded as 0 or 1, a one unit difference represents switching from one category to the other. a_2 is then the average difference in Y between the category for which $X_2 = 0$ (the reference group) and the category for which $X_2 = 1$ (the comparison group). So compared to shrubs that were in partial sun, we would expect shrubs in full sun to be 11 cm taller, on average, at the same level of soil bacteria.

Model Diagnostics

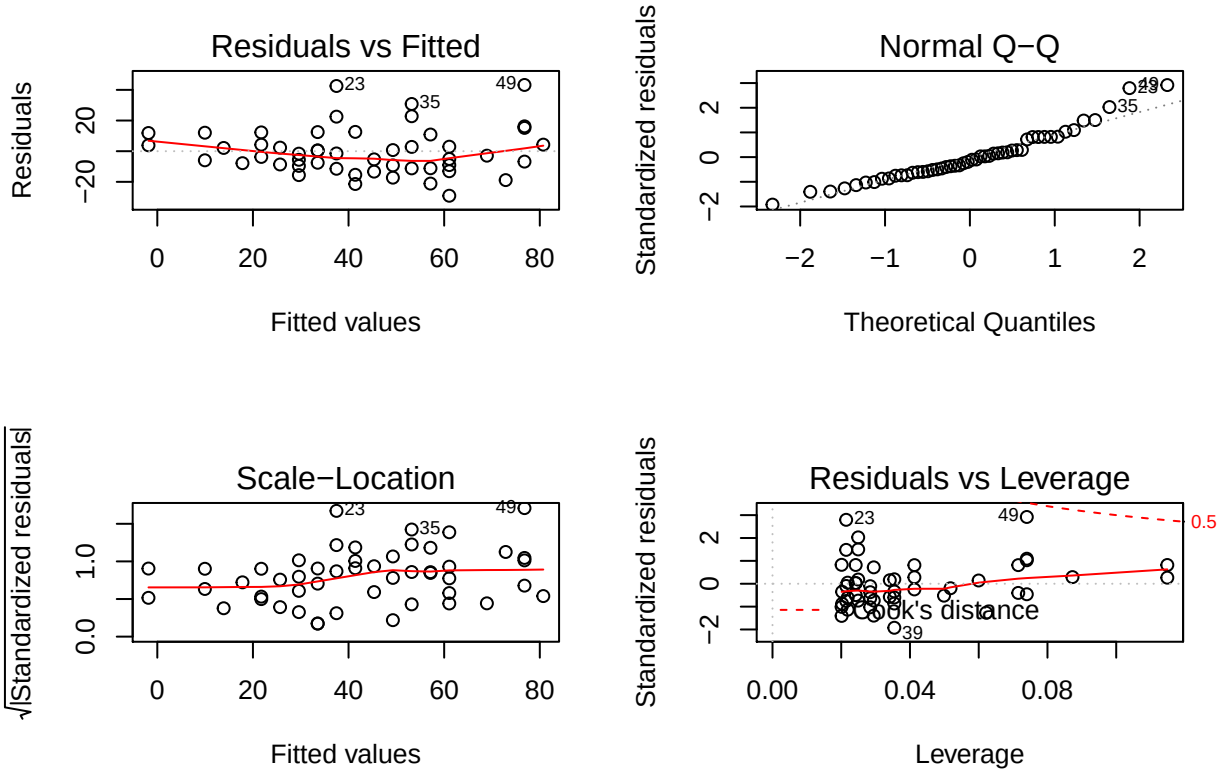
Linearity

Non-normal distribution

Non-constant variance

Outliers

```
par(mfrow = c(2, 2))
plot(mod1)
```



Application of Linear Regression

Prediction (Sales Forecast)

Inference (MMM, Price Elasticity)