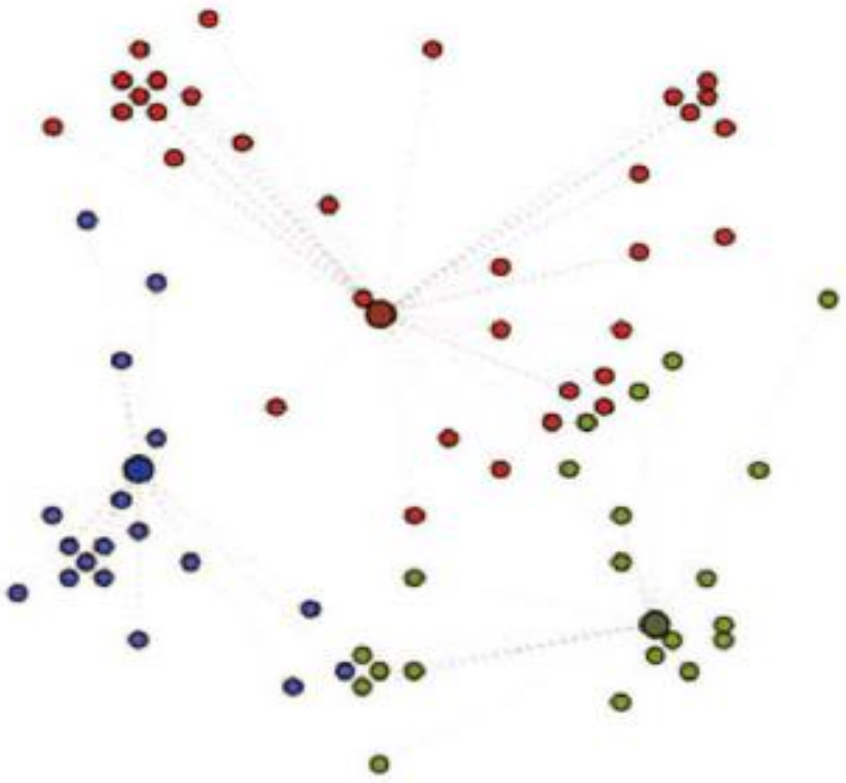
1.1-Overview of Clustering

In general, clustering is the use of *unsupervised* techniques for grouping similar objects. In machine learning, unsupervised refers to the problem of finding hidden structure within unlabeled data. Clustering techniques are unsupervised in the sense that the data scientist does not determine, in advance, the labels to apply to the clusters. The structure of the data describes the objects of interest and determines how best to group the objects.

Clustering is a method often used for exploratory analysis of the data. In clustering, there are no predictions made. Rather, clustering methods find the similarities between objects according to the object attributes and group the similar objects into clusters. Clustering techniques are utilized in marketing, economics, and various branches of science. A popular clustering method is k-means.

1.2-K-means

Given a collection of objects each with n measurable attributes, *k-means* is an analytical technique that, for a chosen value of k, identifies k clusters of objects based on the objects' proximity to the center of the k groups. The center is determined as the arithmetic average (mean) of each cluster's n-dimensional vector of attributes. Below figure illustrates three dusters of objects with two attributes. Each object in the dataset is represented by a small dot color-coded to the closest large dot, the mean of the cluster.



Possible k-means clusters for k=3

1.2.1-Use Cases

Clustering is often used as a lead-in to classification. Once the clusters are identified, labels can be applied to each cluster to classify each group based on its characteristics. Some specific applications of k-means are image processing, medical and customer segmentation.

Image Processing

Video is one example of the growing volumes of unstructured data being collected. Within each frame of a video, k-means analysis can be used to identify objects in the video. For each frame, the task is to determine which pixels are most similar to each other. The attributes of each pixel can include brightness, color, and location, the x and y coordinates in the frame. With security video images, for example, successive frames are examined to identify any changes to the clusters. These newly identified dusters may indicate unauthorized access to a facility.

Medical

Patient attributes such as age, height, weight, systolic and diastolic blood pressures, cholesterol level, and other attributes can identify naturally occurring clusters. These dusters could be used to target individuals for specific preventive measures or clinical trial participation. Clustering, in general, is useful in biology for the classification of plants and animals as well as in the field of human genetics.

Customer Segmentation

Marketing and sales groups use k-means to better identify customers who have similar behaviors and spending patterns. For example, a wireless provider may look at the following customer attributes: monthly bill, number of text messages, data volume consumed, minutes used during various daily periods, and years as a customer. The wireless company could then look at the naturally occurring clusters and consider tactics to increase sales or reduce the customer *churn rate,* the proportion of customers who end their relationship with a particular company.

1.2.2-Overview of the Method

To illustrate the method to find k clusters from a collection of M objects with n attributes, the two-dimensional case (n = 2) is examined. It is much easier to visualize the k-means method in two dimensions.

Because each object in this example has two attributes, it is useful to consider each object corresponding to the point (x_i, y_i), where x and y denote the two attributes and i = 1, 2 ... M. For a given cluster of m points (m ≤ M), the point that corresponds to the cluster's mean is called a *centroid.*

The k-means algorithm to find k dusters can be described in the following four steps.

1. Choose the value of k and the k initial guesses for the centroids.

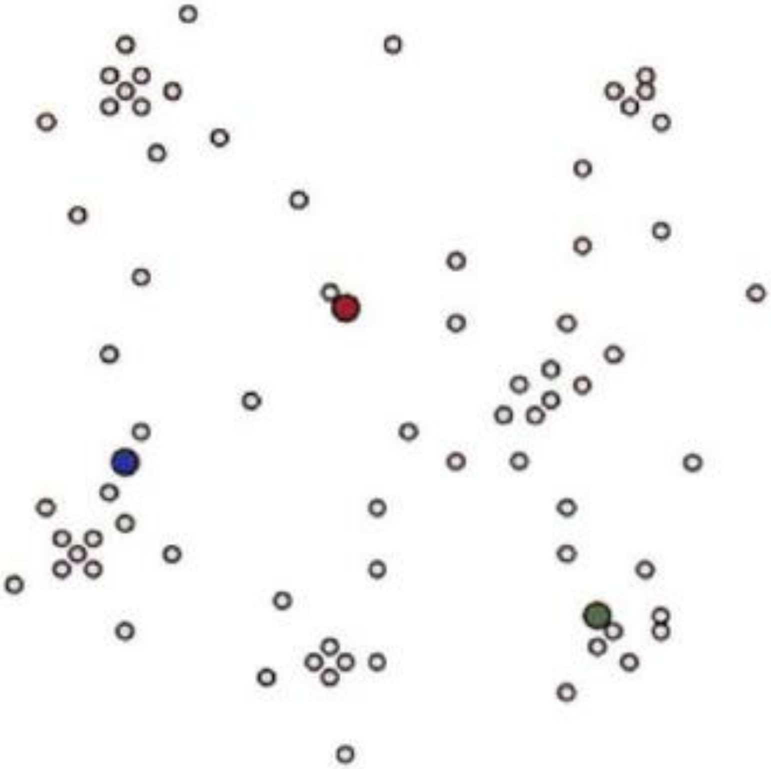
   In this example, k = 3, and the initial centroids are indicated by the points shaded in red, green, and blue in figure 2.2.

2. Compute the distance from each data point (x_i ,y_i) to each centroid. Assign each point to the closest centroid. This association defines the first k dusters.

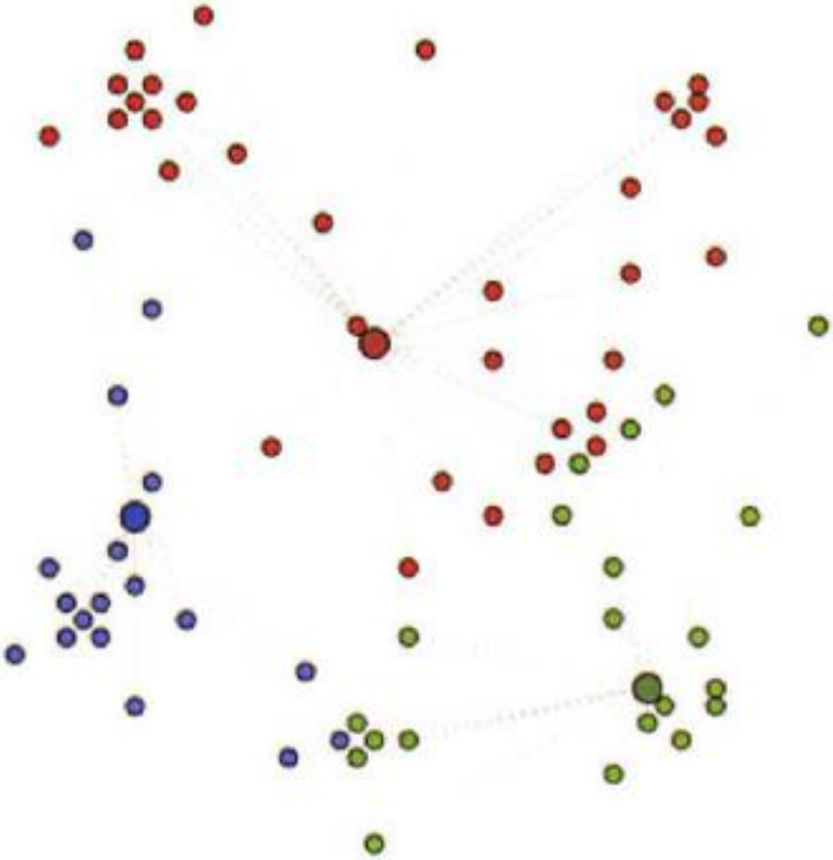
   In two dimensions, the distance, *d,* between any two points, (x_1,y_1) and (x_2,y_2), in the Cartesian plane is typically expressed by using the Euclidean distance measure provided in Equation.

   $$d= \sqrt{(x1-y1)^2+(x2-y2)^2}$$

   In figure 2.3, the points closest to a centroid are shaded the corresponding color.



2.2-Initial starting points for the centroids



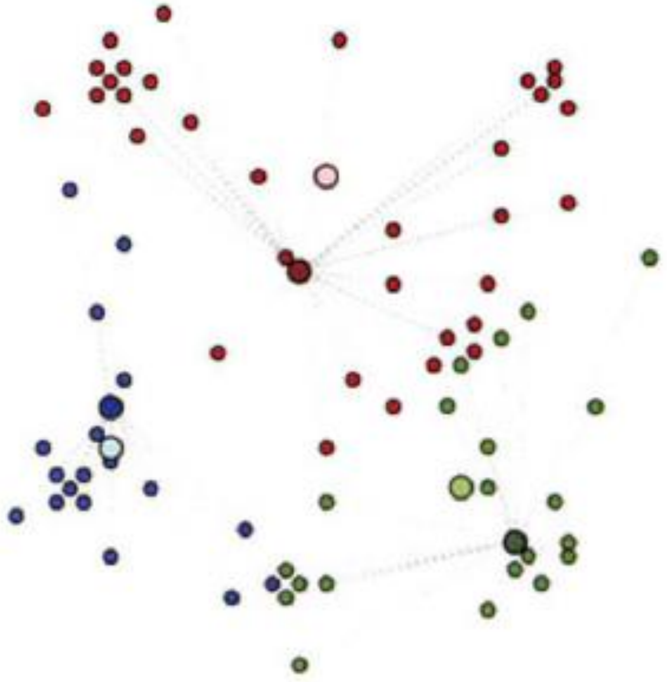
2.3-Points are assigned to the closest centroid

3. Compute the centroid, the center of mass, of each newly defined cluster from Step *2.*

   In Figure 2-4, the computed centroids in Step 3 are the lightly shaded points of the corresponding color. In two dimensions, the centroid (x_c,y_c) of the m points in a k-means duster is calculated as follows in Equation.

$$(x_c, y_c) = \left(\frac{\sum_{i=1}^{m} x_i}{m}, \frac{\sum_{i=1}^{m} y_i}{m} \right)$$

   Thus, (x_c,y_c) is the ordered pair of the arithmetic means of the coordinates of the m points in the cluster. In this step, a centroid is computed for each of the k clusters.



2.4-Compute the mean of each cluster

4. Repeat Steps 2 and 3 until the algorithm converges to an answer

   a. Assign each point to the closest centroid computed in Step 3.
   b. Compute the centroid of newly defined clusters.
   c. Repeat until the algorithm reaches the final answer.

1.2.3-Determining the Number of Clusters

In k-means, k clusters can be identified in a given dataset, but what value of k should be selected? The value of k can be chosen based on a reasonable guess or some predefined requirement. However, even then, it would be good to know how much better or worse having k clusters versus k-1 or k+1 cluster would be in explaining the structure of the data. Next, a heuristic using the Within Sum of Squares (WSS) metric is examined to determine a reasonably optimal value of k. Using the distance function, WSS is defined as shown below.

$$WSS = \sum_{i=1}^{M} d(p_i, q^{(i)})^2 = \sum_{i=1}^{M}\sum_{j=1}^{n}\left(p_{ij} - q_j^{(i)}\right)^2$$

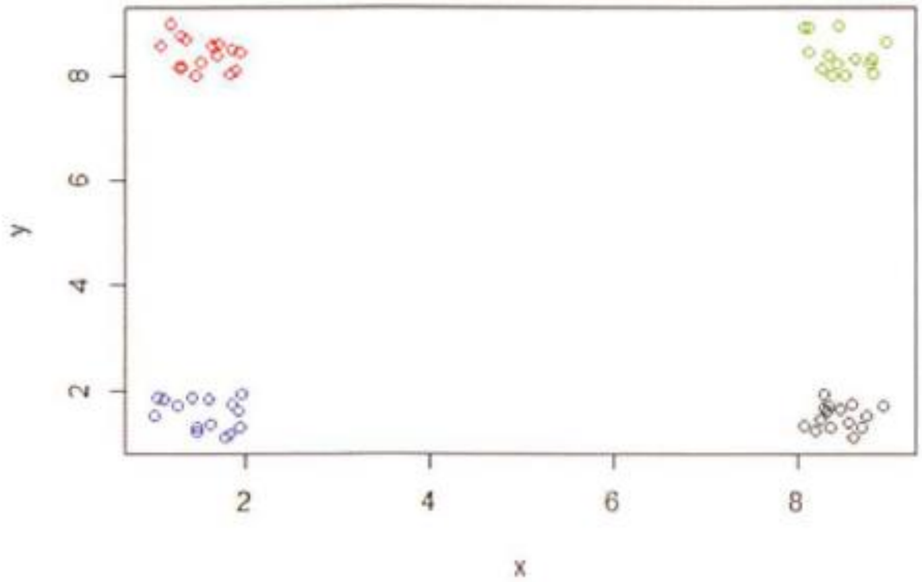
In other words, WSS is the sum of the squares of the distances between each data point and the closest centroid. The term $q^{(i)}$ indicates the closest centroid that is associated with the i^{th} point. If the points are relatively close to their respective centroids, the WSS is relatively small. Thus, if k +1 clusters do not greatly reduce the value of WSS from the case with only k clusters, there may be little benefit to adding another cluster.

1.2.4-Diagnostics

The heuristic using WSS can provide at least several possible k values to consider. When the number of attributes is relatively small, a common approach to further refine the choice of k is to plot the data to determine how distinct the identified clusters are from each other. In general, the following questions should be considered.

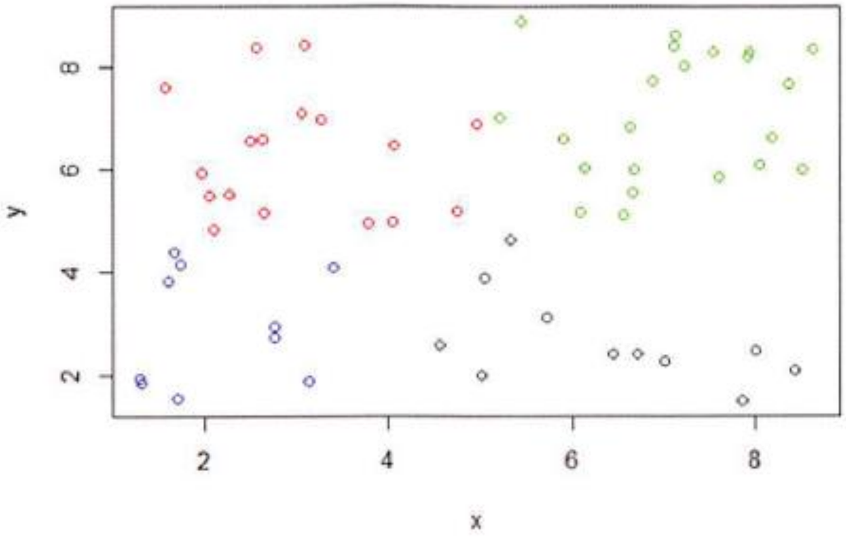
- Are the dusters well separated from each other?
- Do any of the dusters have only a few points?
- Do any of the centroids appear to be too close to each other?

In the first case, ideally the plot would look like the one shown in below figure, when n = 2.



Example of distinct clusters

The clusters are well defined, with considerable space between the four identified clusters. However, in other cases, such as in below figure, the clusters may be close to each other, and the distinction may not be so obvious.



Example of less obvious clusters

1.2.5-Reasons to Choose and Cautions

K-means is a simple and straightforward method for defining clusters. Once clusters and their associated centroids are identified, it is easy to assign new objects (for example, new customers) to a cluster based on the object's distance from the closest centroid. Because the method is unsupervised, using k-means helps to eliminate subjectivity from the analysis.

Although k-means is considered an unsupervised method, there are still several decisions that the practitioner must make:

a. What object attributes should be included in the analysis?
b. What unit of measure (for example, miles or kilometers) should be used for each attribute?
c. Do the attributes need to be rescaled so that one attribute does not have a disproportionate effect on the results?
d. What other considerations might apply?

a-Object Attributes

Regarding which object attributes (for example, age and income) to use in the analysis, it is important to understand what attributes will be known at the time a new object will be assigned to a cluster. For example, information on existing customers' satisfaction or purchase frequency may be available, but such information may not be available for potential customers.

The Data Scientist may have a choice of a dozen or more attributes to use in the clustering analysis. Whenever possible and based on the data, it is best to reduce the number of attributes to the extent possible. Too many attributes can minimize the impact of the most important variables. Also, the use of several similar attributes can place too much importance on one type of attribute. For example, if five attributes related to personal wealth are included in a clustering analysis, the wealth attributes dominate the analysis and possibly mask the importance of other attributes, such as age.

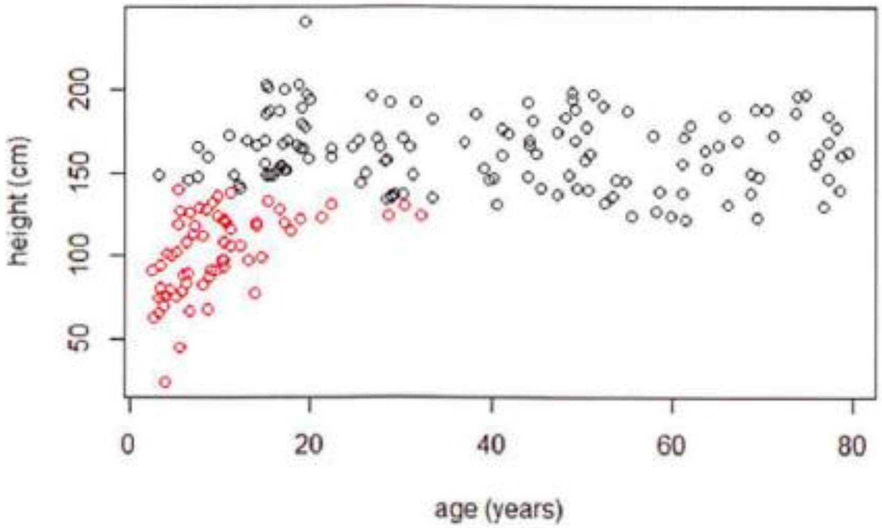
When dealing with the problem of too many attributes, one useful approach is to identify any highly correlated attributes and use only one or two of the correlated attributes in the clustering analysis.

Another option to reduce the number of attributes is to combine several attributes into one measure. For example, instead of using two attribute variables, one for Debt and one for Assets, a Debt to Asset ratio could be used, This option also addresses the problem when the magnitude of an attribute is not of real interest, but the relative magnitude is a more important measure.

b-Units of Measure

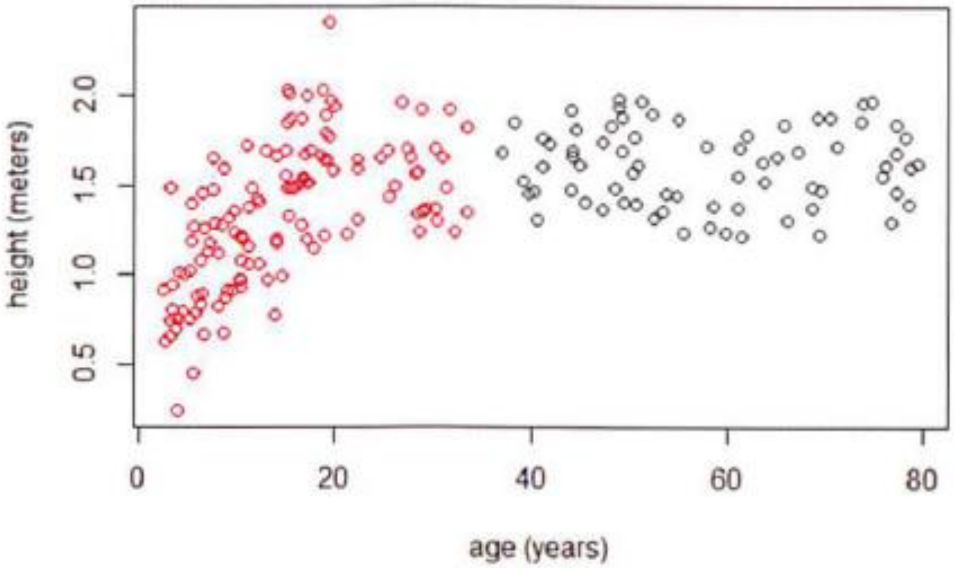
From a computational perspective, the k-means algorithm is somewhat indifferent to the units of measure for a given attribute (for example, meters or centimeters for a patient's height). However, the algorithm will identify different clusters depending on the choice of the units of measure.

For example, suppose that k-means is used to cluster patients based on age in years and height in centimeters. For k=2, below figure illustrates the two clusters that would be determined for a given dataset.



Clusters with height expressed in centimeters

But if the height was rescaled from centimeters to meters by dividing by 100, the resulting dusters would be slightly different, as illustrated in below Figure.



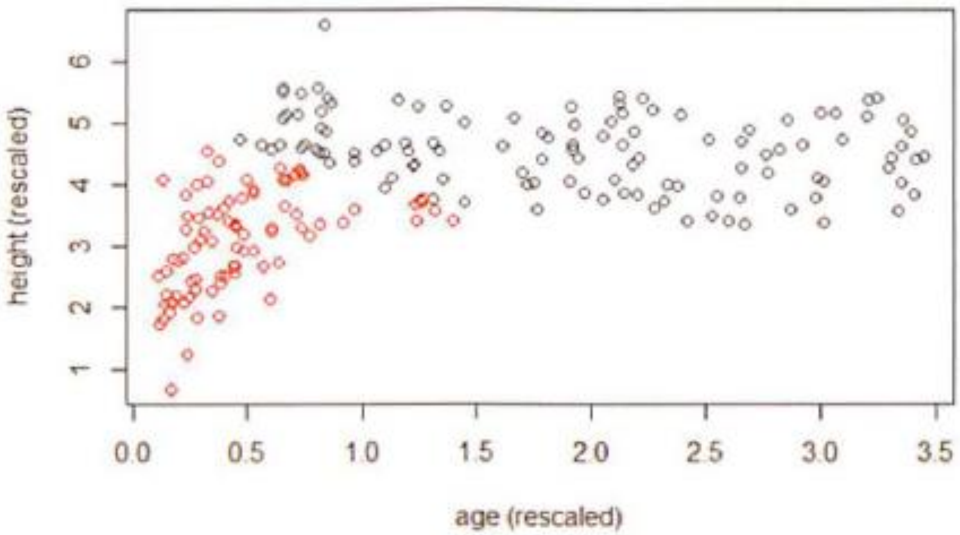
Clusters with height expressed in meters

c-Rescaling

Attributes that are expressed in dollars are common in clustering analyses and can differ in magnitude from the other attributes. For example, if personal income is expressed in dollars and age is expressed in years, the income attribute, often exceeding 510,000, can easily dominate the distance calculation with ages typically less than 100 years.

Although some adjustments could be made by expressing the income in thousands of dollars (for example, 10 for 510,000), a more straightforward method is to divide each attribute by the attribute's standard deviation. The resulting attributes will each have a standard deviation equal to 1 and will be without units.

Returning to the age and height example, the standard deviations are 23.1 years and 36.4 cm, respectively. Dividing each attribute value by the appropriate standard deviation and performing the k-means analysis yields the result shown in Figure 4-13.



Clusters with rescaled attributes

In many statistical analyses, it is common to transform typically skewed data, such as income, with long tails by taking the logarithm of the data. Such transformation can also be applied in k-means, but the Data Scientist needs to be aware of what effect this transformation will have.

d-Additional Considerations

The k-means algorithm is sensitive to the starting positions of the initial centroid. Thus, it is important to rerun the k-means analysis several times for a particular value of k to ensure the cluster results provide the overall minimum WSS. As we know, this task is accomplished in R by using the nstart option in the kmeans () function call.

K-means clustering is applicable to objects that can be described by attributes that are numerical with a meaningful distance measure. Interval and ratio attribute types can certainly be used. However, k-means does not handle categorical variables well. For example, suppose a clustering analysis is to be conducted on new car sales. Among other attributes, such as the sale price, the color of the car is considered important. Although one could assign numerical values to the color, such as red = 1, yellow = 2, and green = 3, it is not useful to consider that yellow is as close to red as yellow is to green from a clustering perspective. In such cases, it may be necessary to use an alternative clustering methodology.

2-Association Rules

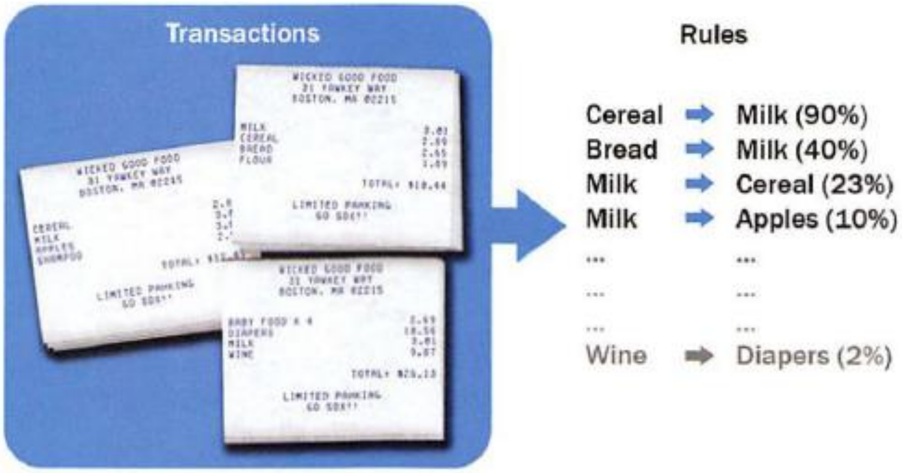
An unsupervised learning method called association rules. This is a descriptive, not predictive, method often used to discover interesting relationships hidden in a large dataset. The disclosed relationships can be represented as rules or frequent item sets. Association rules are commonly used for mining transactions in databases.

Here are some possible questions that association rules can answer:

- Which products tend to be purchased together?
- Of those customers who are similar to this person, what products do they tend to buy?
- Of those customers who have purchased this product, what other similar products do they tend to view or purchase?

2.1- Overview

Below figure shows the general logic behind association rules. Given a large collection of transactions (depicted as three stacks of receipts in the figure), in which each transaction consists of one or more items, association rules go through the items being purchased to see what items are frequently bought together and to discover a list of rules that describe the purchasing behavior. The goal with association rules is to discover interesting relationships among the items. The relationships that are interesting depend both on the business context and the nature of the algorithm being used for the discovery.



The general logic behind association rules

Each of the uncovered rules is in the form X —> Y, meaning that when item X is observed, item Y is also observed. In this case, the left-hand side (LHS) of the rule is X, and the right-hand side (RH5) of the rule is Y.

Using association rules, patterns can be discovered from the data that allow the association rule algorithms to disclose rules of related product purchases. The uncovered rules are listed on the right side of Figure. The first three rules suggest that when cereal is purchased, 90% of the time milk is purchased also. When bread is purchased, 40% of the time milk is purchased also. When milk is purchased, 23% of the time cereal is also purchased.

In the example of a retail store, association rules are used over transactions that consist of one or more items. In fact, because of their popularity in mining customer transactions, association rules are sometimes referred to as ***market basket analysis.*** Each transaction can be viewed as the shopping basket of a customer that contains one or more items. This is also known as an itemset. The term ***itemset*** refers to a collection of items or individual entities that contain some kind of relationship. This could be a set of retail items purchased together in one transaction, a set of hyperlinks clicked on by one user in a single session, or a set of tasks done in one day. An itemset containing ***k*** items is called a ***k-itemset*** denoted by {item1,item 2**,** . . . item k}**.**

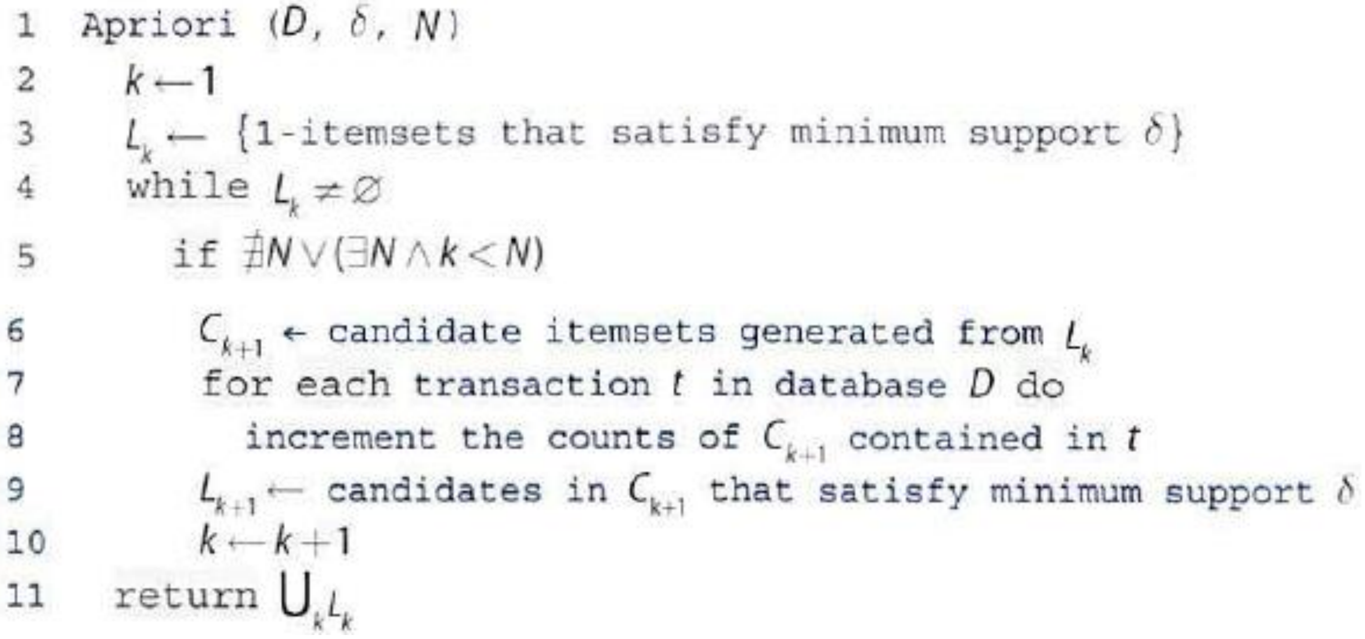
2.2-Apriori Algorithm

The Apriori algorithm takes a bottom-up iterative approach to uncovering the frequent itemsets by first determining all the possible items (or 1-itemsets, for example {bread}, ***{eggs}, {milk},*** …) and then identifying which among them are frequent.

Assuming the minimum support threshold (or the minimum support criterion) is set at 0.5, the algorithm identifies and retains those itemsets that appear in at least 50% of all transactions and discards (or "prunes away") the itemsets that have a support less than 0.5 or appear in fewer than 50% of the transactions.

In the next iteration of the Apriori algorithm, the identified frequent 1-itemsets are paired into 2-itemsets (for example, {bread, ***eggs***}, {bread, ***milk}, {eggs, milk},***…) and again evaluated to identify the frequent 2-itemsets among them.

At each iteration, the algorithm checks whether the support criterion can be met; if it can, the algorithm grows the itemset, repeating the process until it runs out of support or until the itemsets reach a predefined length. Let variable C_k be the set of candidate k-itemsets and variable L_k be the set of k-itemsets that satisfy the minimum support. Given a transaction database D, a minimum support threshold δ, and an optional parameter N indicating the maximum length an itemset could reach, Apriori iteratively computes frequent itemsets L_{k+1} , based on $L_{k.}$



```
Apriori ($D$, $\delta$, $N$)
  $k \leftarrow 1$
  $L_k \leftarrow$ {1-itemsets that satisfy minimum support $\delta$}
  while $L_k \neq \emptyset$
    if $\nexists N \vee (\exists N \wedge k < N)$
      $C_{k+1} \leftarrow$ candidate itemsets generated from $L_k$
      for each transaction $t$ in database $D$ do
        increment the counts of $C_{k+1}$ contained in $t$
      $L_{k+1} \leftarrow$ candidates in $C_{k+1}$ that satisfy minimum support $\delta$
      $k \leftarrow k+1$
  return $\bigcup_k L_k$
```

2.3-Evaluation of Candidate Rules

Frequent itemsets from the previous section can form candidate rules such as X implies Y (X —> Y). *Confidence* is defined as the measure of certainty or trustworthiness associated with each discovered rule. Mathematically, confidence is the percent of transactions that contain both X and Y out of all the transactions that contain X

$$Confidence(X \rightarrow Y) = \frac{Support(X \wedge Y)}{Support(X)}$$

For example, if ***{bread, eggs, milk}*** has a support of 0.15 and ***{bread, eggs}*** also has a support of 0.15, the confidence of rule ***{bread, eggs }->{ milk}*** is 1, which means 100% of the time a customer buys bread and eggs, milk is bought as well. The rule is therefore correct for 100% of the transactions containing bread and eggs.

A relationship may be thought of as interesting when the algorithm identifies the relationship with a measure of confidence greater than or equal to a predefined threshold. This predefined threshold is called the *minimum confidence.* A higher confidence indicates that the rule (X —> Y) is more interesting or more trustworthy, based on the sample dataset.

Even though confidence can identify the interesting rules from all the candidate rules, it comes with a problem. Given rules in the form of X -> Y, confidence considers only the antecedent (X) and the cooccurrence of X and Y; it does not take the consequent of the rule (Y) into concern. Therefore, confidence cannot tell if a rule contains true implication of the relationship or if the rule is purely coincidental. X and Y can be statistically independent yet still receive a high confidence score. Other measures such as lift and leverage are designed to address this issue.

Lift measures how many times more often X and Y occur together than expected if they are statistically independent of each other. Lift is a measure of how X and Y are really related rather than coincidentally happening together

$$Lift(X \rightarrow Y) = \frac{Support(X \wedge Y)}{Support(X) * Support(Y)}$$

Lift is 1 if X and Y are statistically independent of each other. In contrast, a lift of X —> Y greater than 1 indicates that there is some usefulness to the rule. A larger value of lift suggests a greater strength of the association between X and Y.

Assuming 1,000 transactions, with (milk, eggs} appearing in 300 of them, {milk} appearing in 500, and {eggs} appearing in 400, then *Lift(milk—>eggs)* = 0.3/(0.5*0.4) = 1.5. If {bread} appears in 400 transactions and {milk, bread} appears in 400, then *Lift(milk —>bread)* = 0.4/(0.5*0.4) = 2. Therefore it can be concluded that milk and bread have a stronger association than milk and eggs.

Leverage is a similar notion, but instead of using a ratio, leverage uses the difference. Leverage measures the difference in the probability of X and Y appearing together in the dataset compared to what would be expected if X and Y were statistically independent of each other.

Leverage(X —> Y) = Support(X ^Y) - Support(X)* Support(Y)

In theory, leverage is 0 when X and Y are statistically independent of each other. If X and Y have some kind of relationship, the leverage would be greater than zero. A larger leverage value indicates a stronger relationship between X and Y. For the previous example, *Leverage{milk —> eggs)* = 0.3-(0.5*0.4) = 0.1 and *Leverage(milk -> bread*)=0.4 - (0.5 * 0.4) = 0.2. It again confirms that milk and bread have a stronger association than milk and eggs.

Confidence is able to identify trustworthy rules, but it cannot tell whether a rule is coincidental.

2.4-Applications of Association Rules

The term *market basket analysis* refers to a specific implementation of association rules mining that many companies use for a variety of purposes, including these:

- Broad-scale approaches to better merchandising—what products should be included in or excluded from the inventory each month
- Cross-merchandising between products and high-margin or high-ticket items
- Physical or logical placement of product within related categories of products
- Promotional programs—multiple product purchase incentives managed through a loyalty card program

Besides market basket analysis, association rules are commonly used for recommender systems and clickstream analysis.

Many online service providers such as Amazon and Netflix use recommender systems. Recommender systems can use association rules to discover related products or identify customers who have similar interests. For example, association rules may suggest that those customers who have bought product A have also bought product B, or those customers who have bought products A, B, and C are more similar to this customer. These findings provide opportunities for retailers to cross-sell their products.

Clickstream analysis refers to the analytics on data related to web browsing and user clicks, which is stored on the client or the server side. Web usage log files generated on web servers contain huge amounts of information, and association rules can potentially give useful knowledge to web usage data analysts. For example, association rules may suggest that website visitors who land on page X click on links A, B, and C much more often than links D, E, and F. This observation provides valuable insight on how to better personalize and recommend the content to site visitors.

2.5-Validation and Testing

After gathering the output rules, it may become necessary to use one or more methods to validate the results in the business context for the sample dataset. The first approach can be established through statistical measures such as confidence, lift, and leverage. Rules that involve mutually independent items or cover few transactions are considered uninteresting because they may capture spurious relationships.

Confidence measures the chance that X and Y appear together in relation to the chance X appears. Confidence can be used to identify the interestingness of the rules.

Lift and leverage both compare the support of X and Y against their individual support. While mining data with association rules, some rules generated could be purely coincidental. For example, if 95% of customers buy X and 90% of customers buy Y, then X and Y would occur together at least 85% of the time, even if there is no relationship between the two.

Another set of criteria can be established through subjective arguments. Even with a high confidence, a rule may be considered subjectively uninteresting unless it reveals any unexpected profitable actions. For example, rules like {***paper}->{pencil***} may not be subjectively interesting or meaningful despite high support and confidence values. In contrast, a rule like ***{diaper}->{beer}*** that satisfies both minimum support and minimum confidence can be considered subjectively interesting because this rule is unexpected and may suggest a cross-sell opportunity for the retailer.

2.6-Diagnostics

Although the Apriori algorithm is easy to understand and implement, some of the rules generated are uninteresting or practically useless. Additionally, some of the rules may be generated due to coincidental relationships between the variables. Measures like confidence, lift, and leverage should be used along with human insights to address this problem.

Another problem with association rules is that, in Phase 3 and 4 of the Data Analytics Lifecycle, the team must specify the minimum support prior to the model execution, which may lead to too many or too few rules. In related research, a variant of the algorithm can use a predefined target range for the number of rules so that the algorithm can adjust the minimum support accordingly.

Apriori algorithm is one of the earliest and the most fundamental algorithms for generating association rules. The Apriori algorithm reduces the computational workload by only examining itemsets that meet the specified minimum threshold. However, depending on the size of the dataset, the Apriori algorithm can be computationally expensive. For each level of support, the algorithm requires a scan of the entire database to obtain the result. Accordingly, as the database grows, it takes more time to compute in each run. Here are some approaches to improve Apriori's efficiency:

- **Partitioning:** Any itemset that is potentially frequent in a transaction database must be frequent in at least one of the partitions of the transaction database.
- **Sampling:** This extracts a subset of the data with a lower support threshold and uses the subset to perform association rule mining.
- **Transaction reduction:** A transaction that does not contain frequent fc-itemsets is useless in subsequent scans and therefore can be ignored.
- **Hash-based itemset counting:** If the corresponding hashing bucket count of a fc-itemset is below a certain threshold, the/c-itemset cannot be frequent.
- **Dynamic itemset counting:** Only add new candidate itemsets when all of their subsets are estimated to be frequent.

3-Regression Analysis

In general, regression analysis attempts to explain the influence that a set of variables has on the outcome of another variable of interest. Often, the outcome variable is called a *dependent variable* because the outcome depends on the other variables. These additional variables are sometimes called the *input variables* or the *independent variables.* Regression analysis is useful for answering the following kinds of questions:

- What is a person's expected income?
- What is the probability that an applicant will default on a loan?

Linear regression is a useful tool for answering the first question, and logistic regression is a popular method for addressing the second.

Regression analysis is a useful explanatory tool that can identify the input variables that have the greatest statistical influence on the outcome. With such knowledge and insight, environmental changes can be attempted to produce more favorable values of the input variables. For example, if it is found that the reading level of 10-year-old students is an excellent predictor of the students' success in high school and a factor in their attending college, then additional emphasis on reading can be considered, implemented, and evaluated to improve students' reading levels at a younger age.

3.1-Linear Regression

Linear regression is an analytical technique used to model the relationship between several input variables and a continuous outcome variable. A key assumption is that the relationship between an input variable and the outcome variable is linear. Although this assumption may appear restrictive, it is often possible to properly transform the input or outcome variables to achieve a linear relationship between the modified input and outcome variables.

A linear regression model is a probabilistic one that accounts for the randomness that can affect any particular outcome. Based on known input values, a linear regression model provides the expected value of the outcome variable based on the values of the input variables, but some uncertainty may remain in predicting any particular outcome.

3.1.1-Use Cases

Linear regression is often used in business, government, and other scenarios. Some common practical applications of linear regression in the real world include the following:

- **Real estate:** A simple linear regression analysis can be used to model residential home prices as a function of the home's living area. Such a model helps set or evaluate the list price of a home on the market. The model could be further improved by including other input variables such as number of bathrooms, number of bedrooms, lot size, school district rankings, crime statistics, and property taxes
- **Demand forecasting:** Businesses and governments can use linear regression models to predict demand for goods and services. For example, restaurant chains can appropriately prepare for the predicted type and quantity of food that customers will consume based upon the weather, the day of the week, whether an item is offered as a special, the time of day, and the reservation volume. Similar models can be built to predict retail sales, emergency room visits, and ambulance dispatches.
- **Medical:** A linear regression model can be used to analyze the effect of a proposed radiation treatment on reducing tumor sizes. Input variables might include duration of a single radiation treatment, frequency of radiation treatment, and patient attributes such as age or weight.

3.1.2-Model Description

As the name of this technique suggests, the linear regression model assumes that there is a linear relationship between the input variables and the outcome variable. This relationship can be expressed as shown in Equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots\ldots\ldots + \beta p_{-1} x\, p_{-1} + \varepsilon$$

where:

y is the outcome variable

x_j are the input variables, for j=1,2,...,p-1

β_0 is the value of y when each x_j equals zero

β_j is the change in y based on a unit change in x_j for j=1,2,...,p-1

ε is a random error term that represents the difference in the linear model and a particular observed value for y.

Suppose it is desired to build a linear regression model that estimates a person's annual income as a function of two variables—age and education—both expressed in years. In this case, income is the outcome variable, and the input variables are age and education.

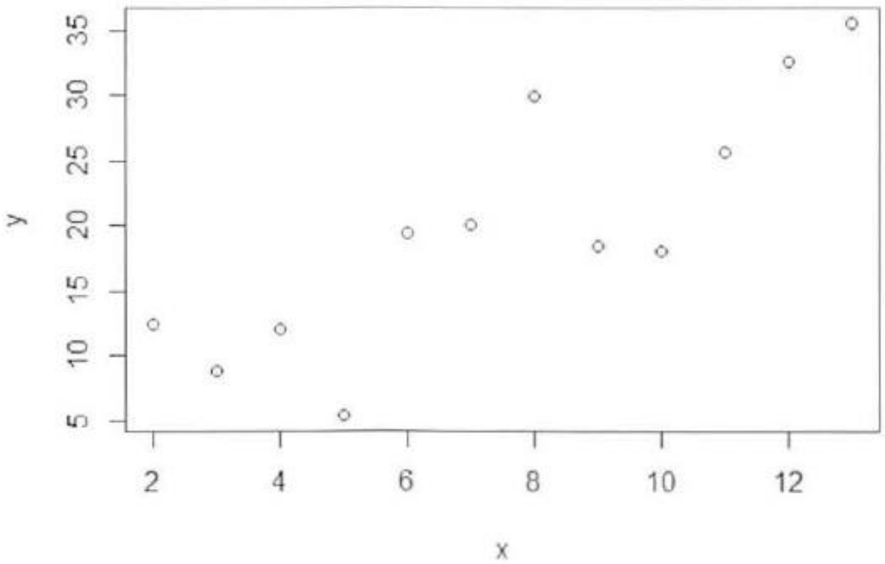
However, it is also obvious that there is considerable variation in income levels for a group of people with identical ages and years of education. This variation is represented by Ɛin the model. So, in this example, the model would be expressed as shown in Equation.

$$\text{Income} = \beta_0 + \beta_1\text{Age} + \beta_2\text{Education} + \varepsilon$$

Linear Regression Model (Ordinary Least Squares)

In the linear model, the β_2s represent the unknown p parameters. The estimates for these unknown parameters are chosen so that, on average, the model provides a reasonable estimate of a person's income based on age and education. In other words, the fitted model should minimize the overall error between the linear model and the actual observations. Ordinary Least Squares (OLS) is a common technique to estimate the parameters.

To illustrate how OLS works, suppose there is only one input variable, x, for an outcome variable y. Furthermore, *n* observations of (x, *y)* are obtained and plotted in below Figure.

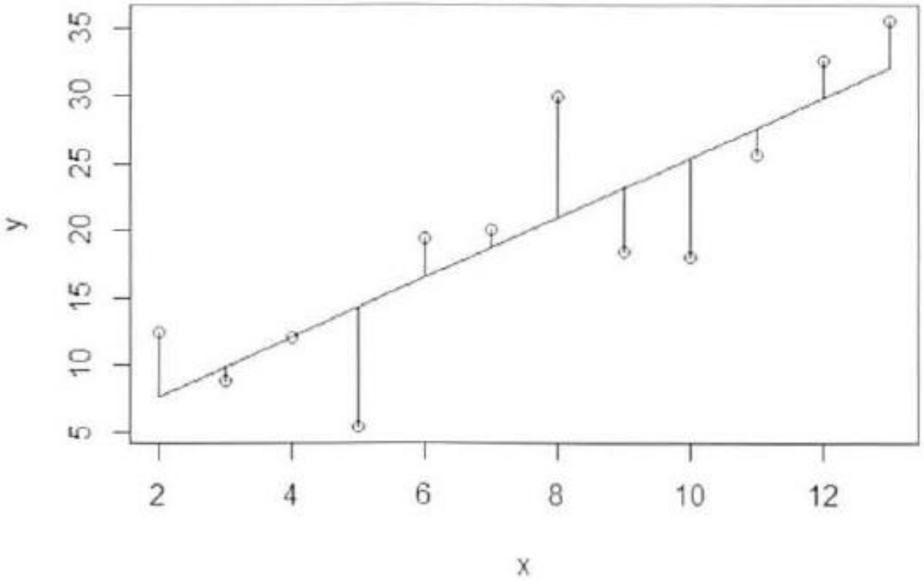


Scatterplot of y versus x

The goal is to find the line that best approximates the relationship between the outcome variable and the input variables. With OLS, the objective is to find the line through these points that minimizes the sum of the squares of the difference between each point and the line in the vertical direction. In other words, find the values of β_0 and β_1, such that the summation shown in Equation is minimized.

$$\sum_{i=1}^{n}[y_i - (\beta_0 + \beta_1 x_i)]^2$$

The n individual distances to be squared and then summed are illustrated in below figure. The vertical lines represent the distance between each observed y value and the line y **= β_0 + $\beta_1 x_1$**



Scatterplot of y versus x with vertical distances from the observed points to a fitted line

Linear Regression Model (with Normally Distributed Errors)

In the normal model description, there were no assumptions made about the error term; no additional assumptions were necessary for OLS to provide estimates of the model parameters. However, in most linear regression analyses, it is common to assume that the error term is a normally distributed random variable with mean equal to zero and constant variance. Thus, the linear regression model is expressed as shown in Equation.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 \; X_2 \ldots\ldots\ldots + \beta p_{-1} \; X \; p_{-1} + \varepsilon$$

where:

y is the outcome variable

x_j are the input variables, for j=1,2,...,p-1

β_0 is the value of y when each x_j equals zero

β_j is the change in y based on a unit change in x_j for j=1,2,...,p-1

$\varepsilon \sim N(0,\sigma^2)$ and the Ɛs are independent of each other

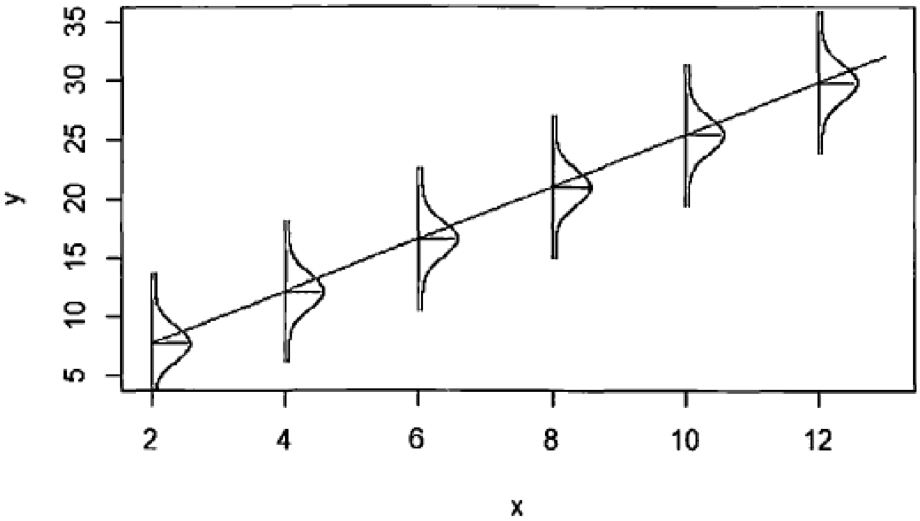
This additional assumption yields the following result about the expected value of y, E(y) for given $(x_1, x_2, \ldots x_{p-1})$:

$$\begin{aligned} E(y) &= E(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1} + \varepsilon) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1} + E(\varepsilon) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1} \end{aligned}$$

Because β_j and x_j are constants, the E(y) is the value of the linear regression model for the given $(x_1, x_2, \ldots x_{p-1})$. Furthermore, the variance of y, V(y), for given $(x_1, x_2, \ldots x_{p-1})$ is this:

$$\begin{aligned} V(y) &= V(\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1} + \varepsilon) \\ &= 0 + V(\varepsilon) = \sigma^2 \end{aligned}$$

Thus, for a given (x_1, x_2,... x_{p-1}), y is normally distributed with mean $\beta_0 + \beta_1 x_1 + \beta_2 \, x_2 \ldots\ldots\ldots + \beta p_{-1} \, x \, p_{-1}$ and variance σ^2. For a regression model with just one input variable, below figure illustrates the normality assumption on the error terms and the effect on the outcome variable, y, for a given value of x.



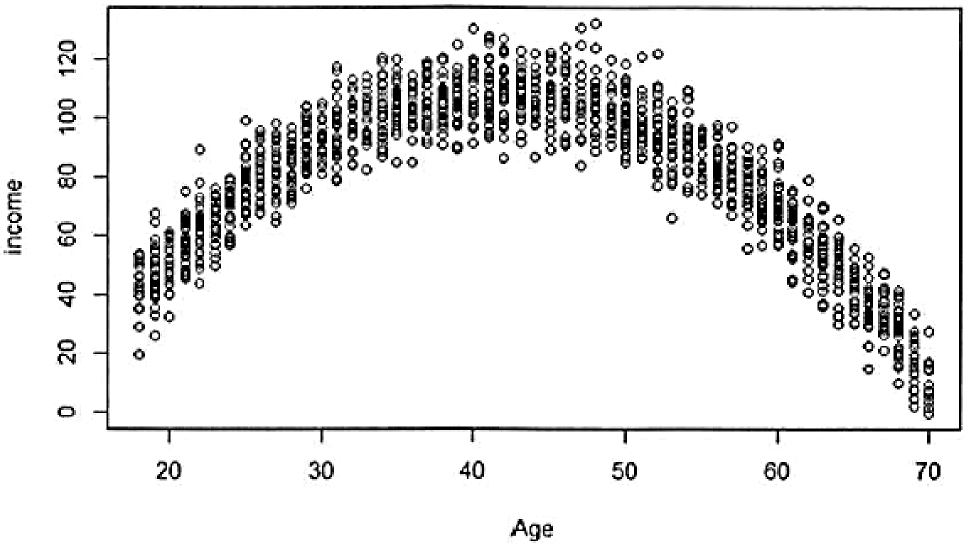
Normal distribution about y for a given value of x

3.1.3-Diagnostics

The use of hypothesis tests, confidence intervals, and prediction intervals is dependent on the model assumptions being true. Following are Some tools and techniques that can be used to validate a fitted linear regression model.

a-Evaluating the Linearity Assumption

A major assumption in linear regression modeling is that the relationship between the input variables and the outcome variable is linear. The most fundamental way to evaluate such a relationship is to plot the outcome variable against each input variable. If the relationship between *Age* and *Income* is represented as illustrated in Figure 6-5, a linear model would not apply.



Income as a quadratic function of Age

In such a case, it is often useful to do any of the following:

- Transform the outcome variable.

- Transform the input variables.
- Add extra input variables or terms to the regression model.

Common transformations include taking square roots or the logarithm of the variables. Another option is to create a new input variable such as the age squared and add it to the linear regression model to fit a quadratic relationship between an input variable and the outcome.

b-Evaluating the Residuals

As stated previously, it is assumed that the error terms in the linear regression model are normally distributed with a mean of zero and a constant variance. If this assumption does not hold, the various inferences that were made with the hypothesis tests, confidence intervals, and prediction intervals are suspect.

c-Evaluating the Normality Assumption

The residual plots are useful for confirming that the residuals were centered on zero and have a constant variance. However, the normality assumption still has to be validated.

d-N-Fold Cross-Validation

To prevent overfitting a given dataset, a common practice is to randomly split the entire dataset into a training set and a testing set. Once the model is developed on the training set, the model is evaluated against the testing set. When there is not enough data to create training and testing sets, an N-fold cross-validation technique may be helpful to compare one fitted model against another. In N-fold cross-validation, the following occurs:

- The entire dataset is randomly split into N datasets of approximately equal size.
- A model is trained against N - 1 of these datasets and tested against the remaining dataset. A measure of the model error is obtained.
- This process is repeated a total of N times across the various combinations of N datasets taken N - 1 at a time. Recall:

$$\binom{N}{N-1} = N$$

- The observed N model errors are averaged over the N folds.

The averaged error from one mode! is compared against the averaged error from another model. This technique can also help determine whether adding more variables to an existing model is beneficial or possibly overfitting the data.

3.2-Logistic Regression

In linear regression modeling, the outcome variable is a continuous variable. When the outcome variable is categorical in nature, logistic regression can be used to predict the likelihood of an outcome based on the input variables. Although logistic regression can be applied to an outcome variable that represents multiple values, but we will examine the case in which the outcome variable represents two values such as true/false, pass/fail, or yes/no.

For example, a logistic regression model can be built to determine if a person will or will not purchase a new automobile in the next 12 months. The training set could include input variables for a person's age, income, and gender as well as the age of an existing automobile. The training set would also include the outcome variable on whether the person purchased a new automobile over a 12-month period. The logistic regression model provides the likelihood or probability of a person making a purchase in the next 12 months.

3.2.1-Use Cases

The logistic regression model is applied to a variety of situations in both the public and the private sector. Some common ways that the logistic regression model is used include the following:

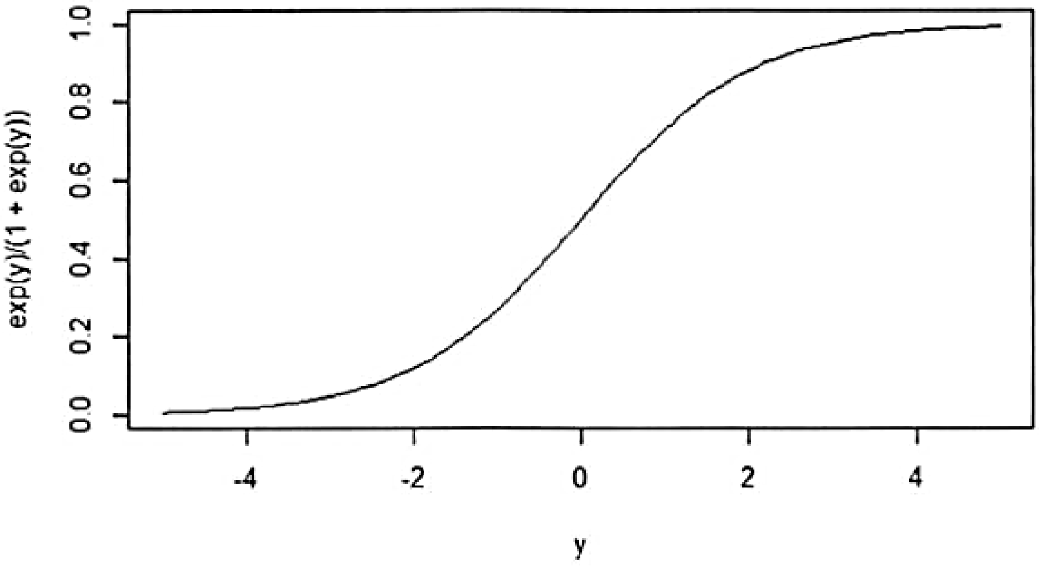
- **Medical:** Develop a model to determine the likelihood of a patient's successful response to a specific medical treatment or procedure. Input variables could include age, weight, blood pressure, and cholesterol levels.
- ***Finance:*** Using a loan applicant's credit history and the details on the loan, determine the probability that an applicant will default on the loan. Based on the prediction, the loan can be approved or denied, or the terms can be modified.
- ***Marketing:*** Determine a wireless customer's probability of switching carriers (known as churning) based on age, number of family members on the plan, months remaining on the existing contract, and social network contacts. With such insight, target the high-probability customers with appropriate offers to prevent churn.
- ***Engineering:*** Based on operating conditions and various diagnostic measurements, determine the probability of a mechanical part experiencing a malfunction or failure. With this, probability estimate, schedule the appropriate preventive maintenance activity.

3.2.2-Model Description

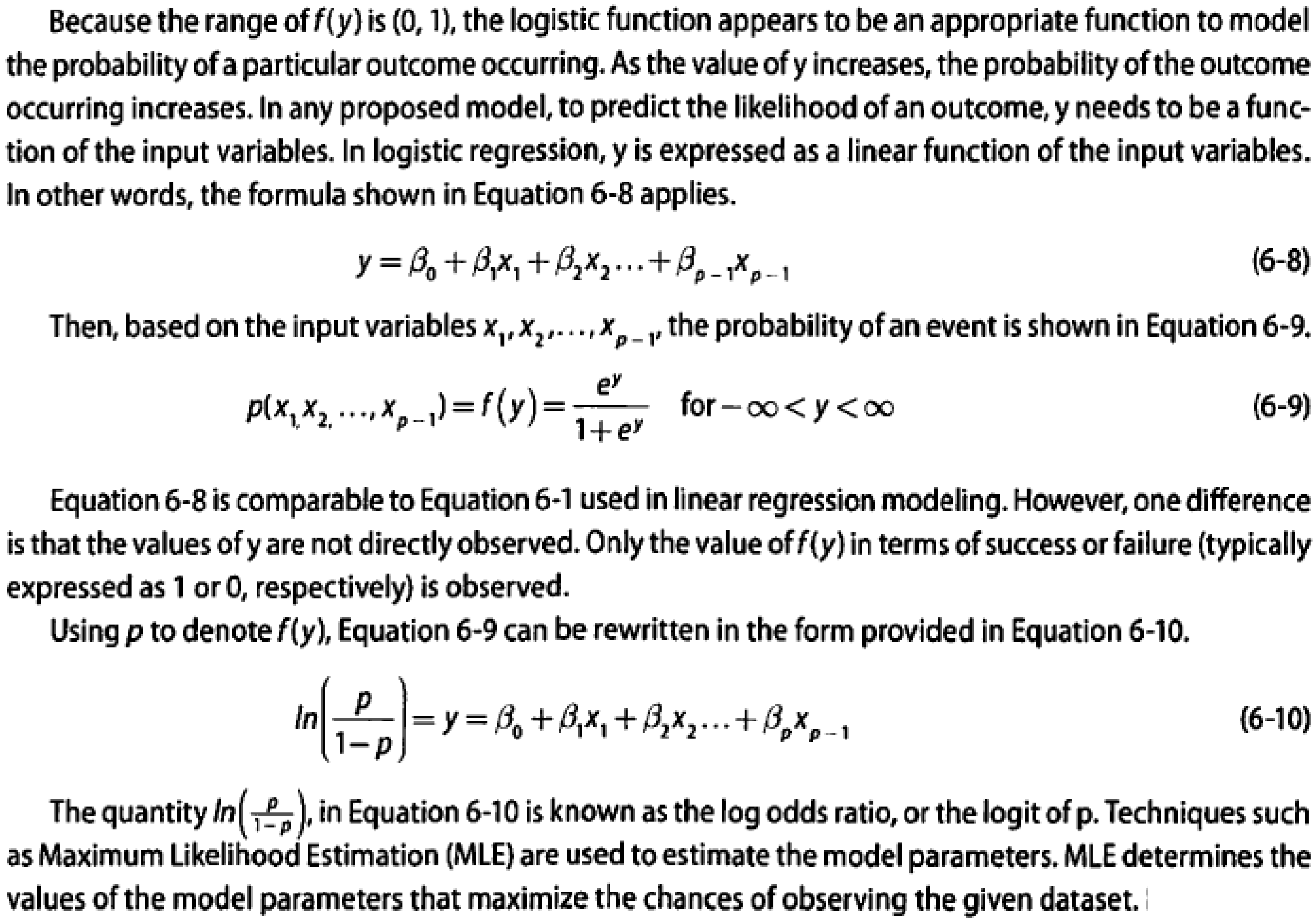
Logistic regression is based on the logistic function $f(y)$, as given in Equation 6-7.

$$f(y) = \frac{e^y}{1+e^y} \quad \text{for} -\infty < y < \infty \tag{6-7}$$

Note that as $y \to \infty$, $f(y) \to 1$, and as $y \to -\infty$, $f(y) \to 0$. So, as Figure 6-14 illustrates, the value of the logistic function $f(y)$ varies from 0 to 1 as y increases.



RE 6-14 *The logistic function*



Because the range of $f(y)$ is (0, 1), the logistic function appears to be an appropriate function to model the probability of a particular outcome occurring. As the value of y increases, the probability of the outcome occurring increases. In any proposed model, to predict the likelihood of an outcome, y needs to be a function of the input variables. In logistic regression, y is expressed as a linear function of the input variables. In other words, the formula shown in Equation 6-8 applies.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_{p-1} x_{p-1} \qquad \text{(6-8)}$$

Then, based on the input variables $x_1, x_2, \ldots, x_{p-1}$, the probability of an event is shown in Equation 6-9.

$$p(x_1, x_2, \ldots, x_{p-1}) = f(y) = \frac{e^y}{1+e^y} \quad \text{for} -\infty < y < \infty \qquad \text{(6-9)}$$

Equation 6-8 is comparable to Equation 6-1 used in linear regression modeling. However, one difference is that the values of y are not directly observed. Only the value of $f(y)$ in terms of success or failure (typically expressed as 1 or 0, respectively) is observed.

Using p to denote $f(y)$, Equation 6-9 can be rewritten in the form provided in Equation 6-10.

$$ln\left(\frac{p}{1-p}\right) = y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_p x_{p-1} \qquad \text{(6-10)}$$

The quantity $ln\left(\frac{p}{1-p}\right)$, in Equation 6-10 is known as the log odds ratio, or the logit of p. Techniques such as Maximum Likelihood Estimation (MLE) are used to estimate the model parameters. MLE determines the values of the model parameters that maximize the chances of observing the given dataset.

Customer Churn Example

A wireless telecommunications company wants to estimate the probability that a customer will churn (switch to a different company) in the next six months. With a reasonably accurate prediction of a person's likelihood of churning, the sales and marketing groups can attempt to retain the customer by offering various incentives. Data on 8,000 current and prior customers was obtained. The variables collected for each customer follow:

o Age (years)
o *Married*(true/false)
o Duration as a customer (years)
o *Churned_contacts*(count)—Number of the customer's contacts that have churned (count)
o churned (true/false)—Whether the customer churned

After analyzing the data and fitting a logistic regression model, *Age* and *Churned_contacts* were selected as the best predictor variables. Equation 6-11 provides the estimated model parameters.

y = 3.50 - 0.16 * ***Age*** + 0.38 * ***Churned _ contacts*** *(*6.11)

Using the fitted model from Equation 6-11, below table provides the probability of a customer churning based on the customer's age and the number of churned contacts.

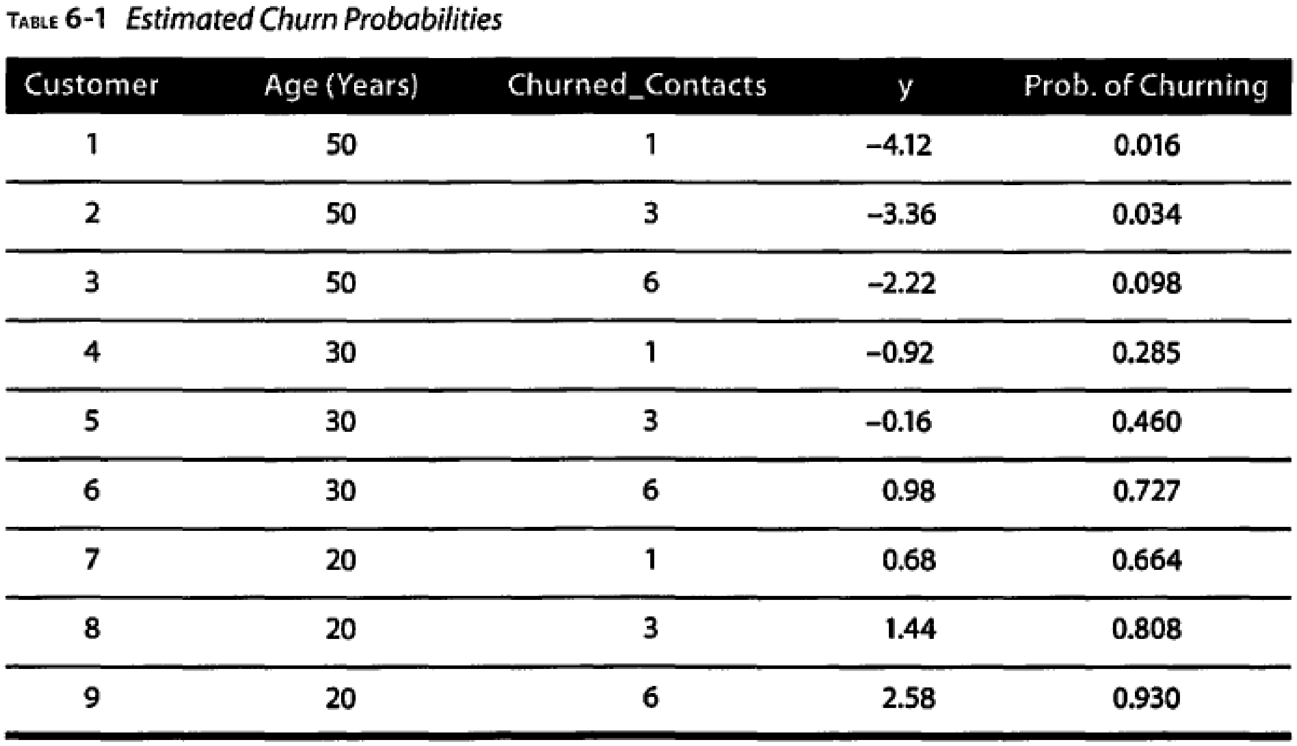
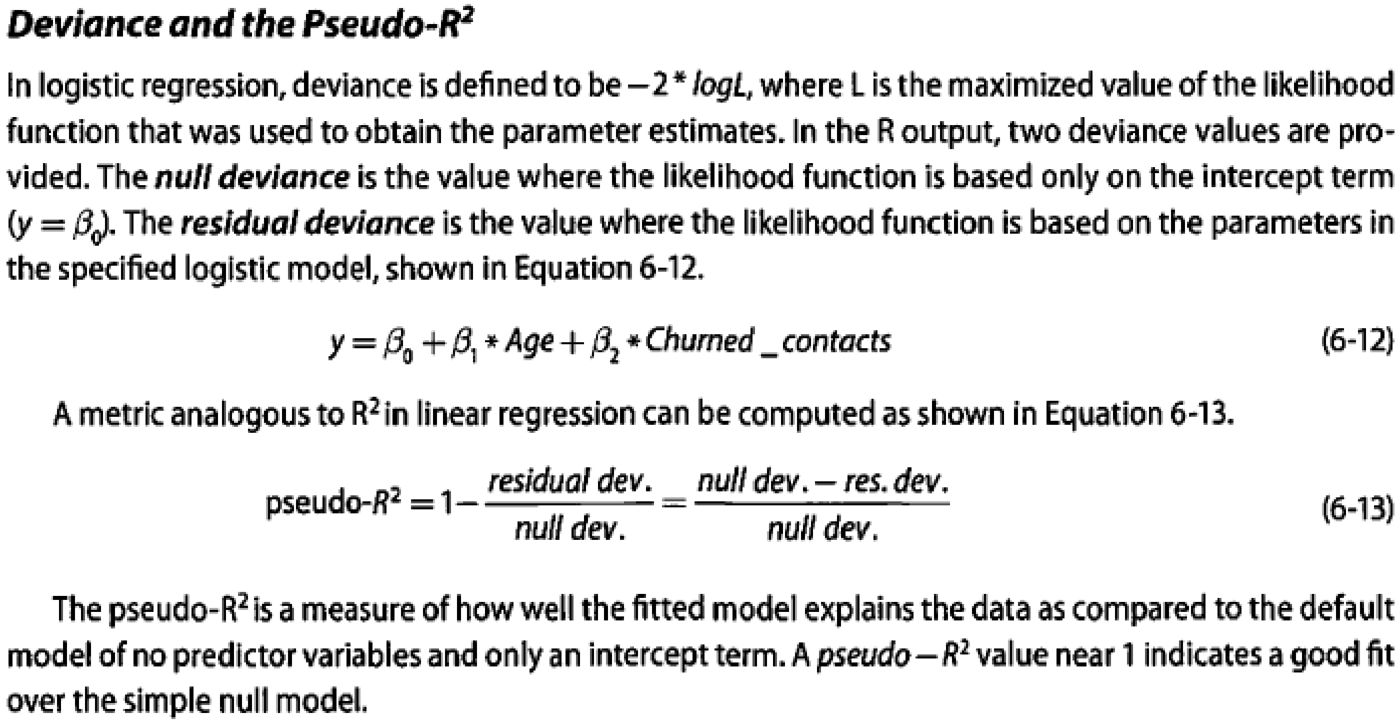


Table 6-1 *Estimated Churn Probabilities*

Customer	Age (Years)	Churned_Contacts	y	Prob. of Churning
1	50	1	−4.12	0.016
2	50	3	−3.36	0.034
3	50	6	−2.22	0.098
4	30	1	−0.92	0.285
5	30	3	−0.16	0.460
6	30	6	0.98	0.727
7	20	1	0.68	0.664
8	20	3	1.44	0.808
9	20	6	2.58	0.930

Based on the fitted model, there is a 93% chance that a 20-year-old customer who has had six contacts churn will also churn.

3.2.3-Diagnostics



Deviance and the Pseudo-R^2

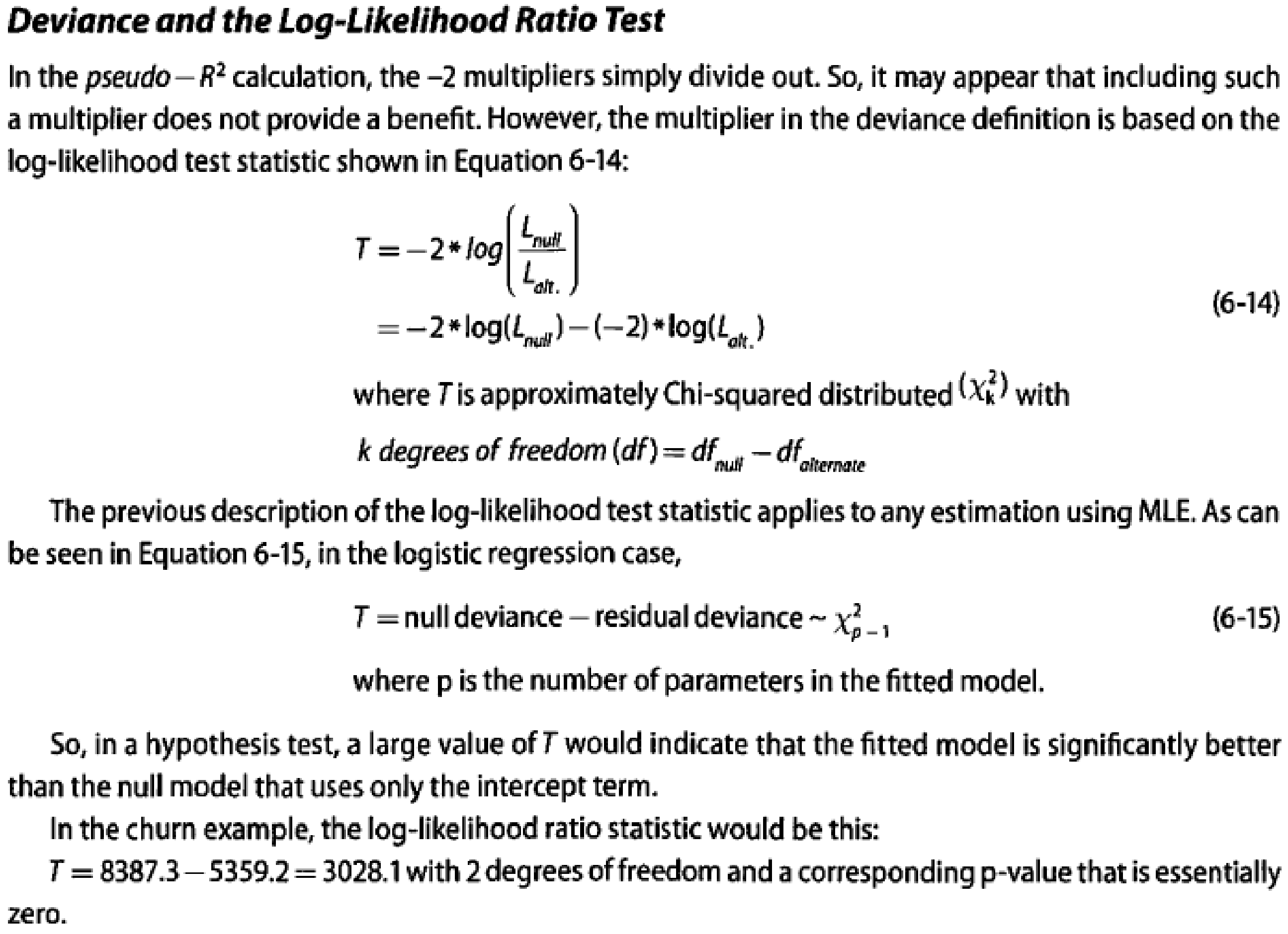
In logistic regression, deviance is defined to be $-2 * logL$, where L is the maximized value of the likelihood function that was used to obtain the parameter estimates. In the R output, two deviance values are provided. The ***null deviance*** is the value where the likelihood function is based only on the intercept term ($y = \beta_0$). The ***residual deviance*** is the value where the likelihood function is based on the parameters in the specified logistic model, shown in Equation 6-12.

$$y = \beta_0 + \beta_1 * Age + \beta_2 * Churned\_contacts \tag{6-12}$$

A metric analogous to R^2 in linear regression can be computed as shown in Equation 6-13.

$$\text{pseudo-}R^2 = 1 - \frac{residual\ dev.}{null\ dev.} = \frac{null\ dev. - res.\ dev.}{null\ dev.} \tag{6-13}$$

The pseudo-R^2 is a measure of how well the fitted model explains the data as compared to the default model of no predictor variables and only an intercept term. A $pseudo - R^2$ value near 1 indicates a good fit over the simple null model.



Deviance and the Log-Likelihood Ratio Test

In the $pseudo-R^2$ calculation, the −2 multipliers simply divide out. So, it may appear that including such a multiplier does not provide a benefit. However, the multiplier in the deviance definition is based on the log-likelihood test statistic shown in Equation 6-14:

$$\begin{aligned} T &= -2 * log\left(\frac{L_{null}}{L_{alt.}}\right) \\ &= -2 * \log(L_{null}) - (-2) * \log(L_{alt.}) \end{aligned} \tag{6-14}$$

where T is approximately Chi-squared distributed (χ_k^2) with

k *degrees of freedom* $(df) = df_{null} - df_{alternate}$

The previous description of the log-likelihood test statistic applies to any estimation using MLE. As can be seen in Equation 6-15, in the logistic regression case,

$$T = \text{null deviance} - \text{residual deviance} \sim \chi_{p-1}^2 \tag{6-15}$$

where p is the number of parameters in the fitted model.

So, in a hypothesis test, a large value of T would indicate that the fitted model is significantly better than the null model that uses only the intercept term.

In the churn example, the log-likelihood ratio statistic would be this:

$T = 8387.3 - 5359.2 = 3028.1$ with 2 degrees of freedom and a corresponding p-value that is essentially zero.

So far, the log-likelihood ratio test discussion has focused on comparing a fitted model to the default model of using only the intercept. However, the log-likelihood ratio test can also compare one fitted model to another.

Receiver Operating Characteristic (ROC) Curve

Logistic regression is often used as a classifier to assign class labels to a person, item, or transaction based on the predicted probability provided by the model. In the Churn example, a customer can be classified with the label called **Churn** if the logistic model predicts a high probability that the customer will churn. Otherwise, a **Remain** label is assigned to the customer. Commonly, 0.5 is used as the default probability threshold to distinguish between any two class labels. However, any threshold value can be used depending on the preference to avoid false positives (for example, to predict **Churn** when actually the customer will Remain) or false negatives (for example, to predict **Remain** when the customer will actually **Churn).**

Histogram of the Probabilities

It can be useful to visualize the observed responses against the estimated probabilities provided by the logistic regression. Figure 6-17 provides overlaying histograms for the customers who churned and for the customers who remained as customers. With a proper fitting logistic model, the customers.who remained tend to have a low probability of churning. Conversely, the customers who churned have a high probability of churning again. This histogram plot helps visualize the number of items to be properly classified or mis- dassified. In the Churn example, an ideal histogram plot would have the remaining customers grouped at the left side of the plot, the customers who churned at the right side of the plot, and no overlap of these two groups.

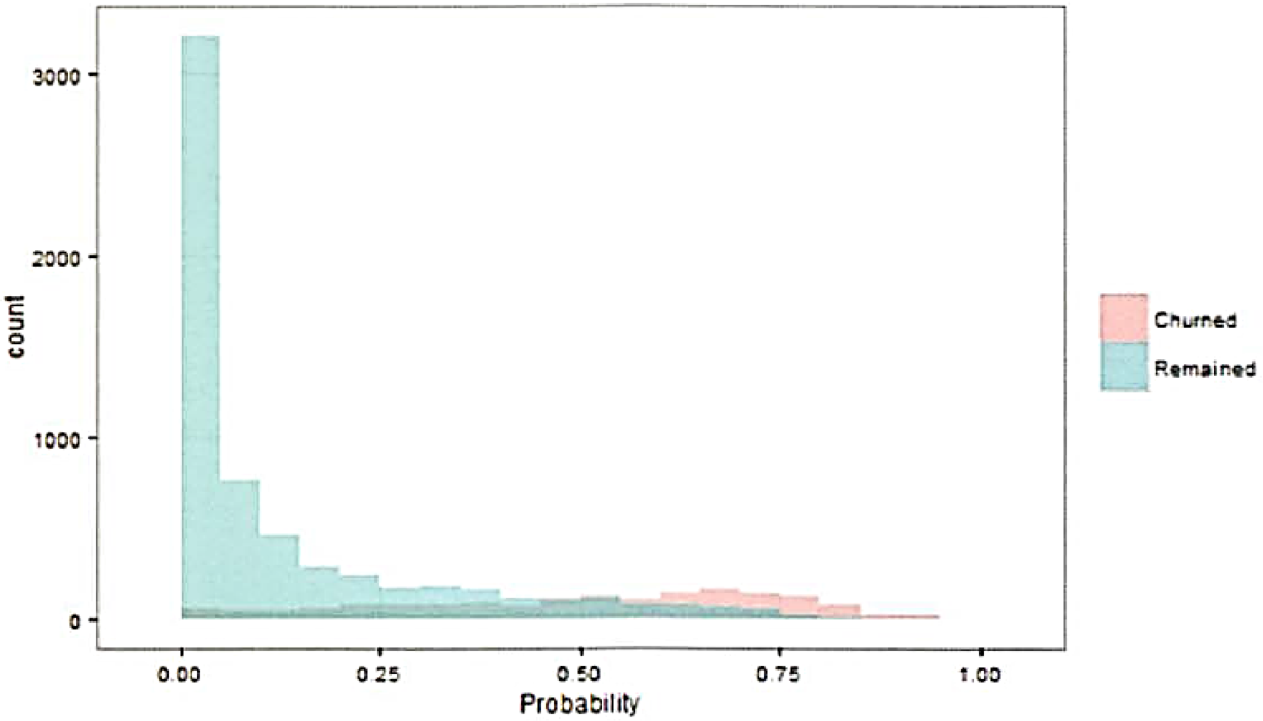


FIGURE 6-17 *Customer counts versus estimated churn probability*

3.3-Reasons to Choose and Cautions

Linear regression is suitable when the input variables are continuous or discrete, including categorical data types, but the outcome variable is continuous. If the outcome variable is categorical, logistic regression is a better choice.

Both models assume a linear additive function of the input variables. If such an assumption does not hold true, both regression techniques perform poorly, Furthermore, in linear regression, the assumption of normally distributed error terms with a constant variance is important for many of the statistical inferences that can be considered. If the various assumptions do not appear to hold, the appropriate transformations need to be applied to the data.

Although a collection of input variables may be a good predictor for the outcome variable, the analyst should not infer that the input variables directly cause an outcome. For example, it may be identified that those individuals who have regular dentist visits may have a reduced risk of heart attacks. However, simply sending someone to the dentist almost certainly has no effect on that person's chance of having a heart attack. It is possible that regular dentist visits may indicate a person's overall health and dietary choices, which may have a more direct impact on a person's health.

Use caution when applying an already fitted model to data that falls outside the dataset used to train the model. The linear relationship in a regression model may no longer hold at values outside the training dataset. For example, if income was an input variable and the values of income ranged from $35,000 to $90,000, applying the model to incomes well outside those incomes could result in inaccurate estimates and predictions.

If several of the input variables are highly correlated to each other, the condition is known as *multicollinearity*. Multicollinearity can often lead to coefficient estimates that are relatively large in absolute magnitude and may be of inappropriate direction (negative or positive sign). When possible, the majority of these correlated variables should be removed from the model or replaced by a new variable that is a function of the correlated variables.