Hierarchical Clustering, Regression

ECE/CS 498 DS U/G

Lecture 9

Ravi K. Iyer

Electrical and Computer Engineering

University of Illinois

Announcements

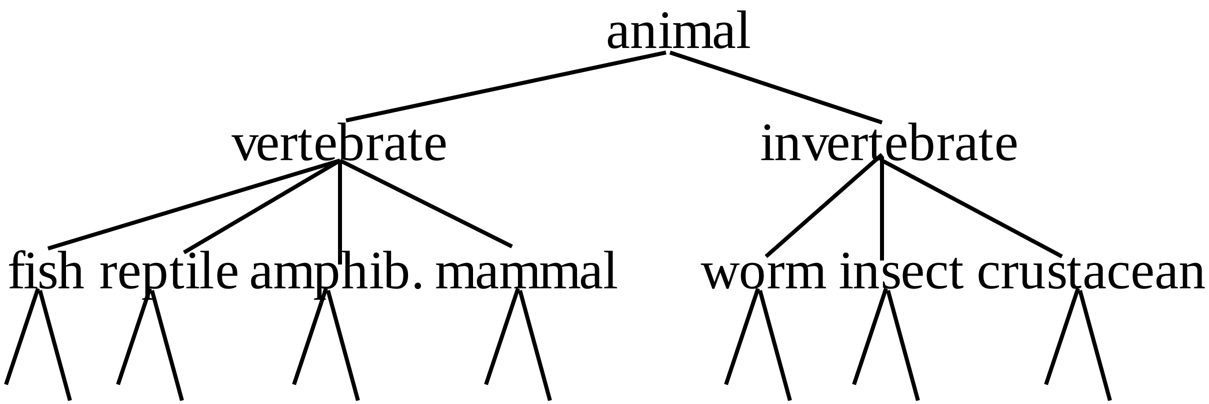
- MP1 Checkpoint 3 due on today
- Poll to sign up for MP1 slot released; please sign up if you haven’t
- MP2 will be released This Wednesday (Feb. 20)
- HW 2 due Friday Feb 22
- Today; Hierarchical Clustering, Regression\=]

Ways to do clustering

- Agglomerative vs Divisive
  - *Agglomerative*: each instance is its own cluster and the algorithm merges clusters
  - *Divisive*: begins with all instances in one cluster and divides it up
- Hard vs Fuzzy
  - Hard clustering assigns each instance to one cluster whereas in fuzzy clustering assigns degree of membership

Hierarchical Clustering

- Build a tree-based hierarchical taxonomy (*dendrogram*) from a set of documents.



- *One approach*: recursive application of a partitional clustering algorithm.

Hierarchical Clustering

- Agglomerative approach

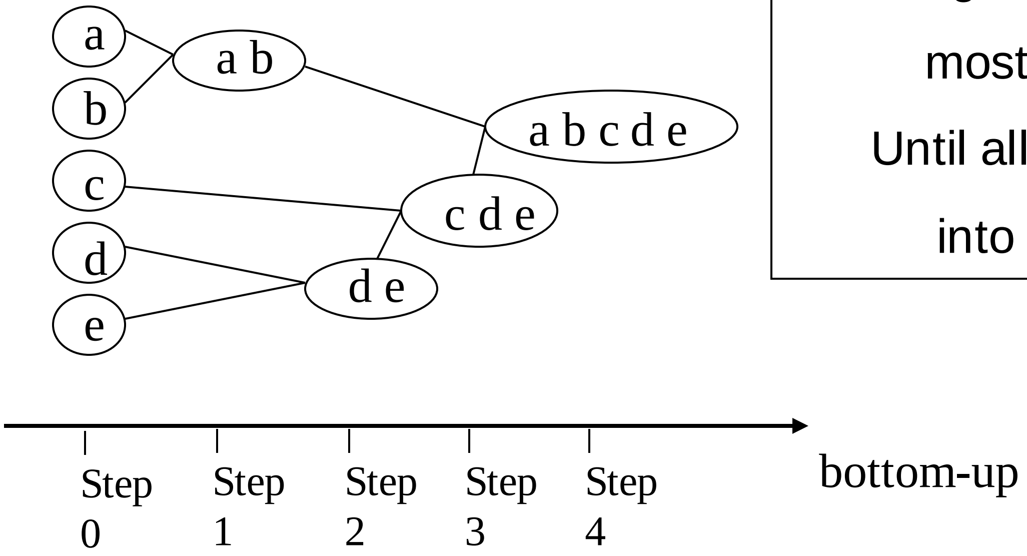
Initialization:

Each object is a cluster

Iteration:

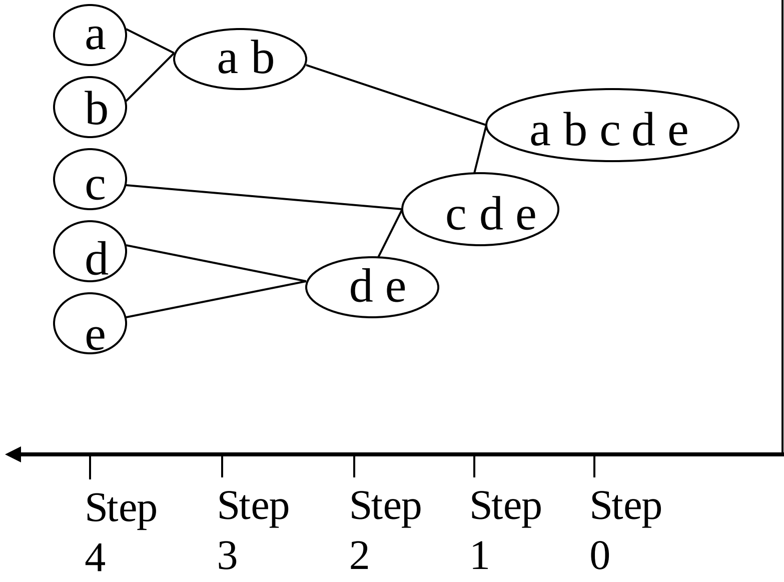
Merge two clusters which are most similar to each other;

Until all objects are merged into a single cluster



Hierarchical Clustering

- Divisive Approaches



Initialization:

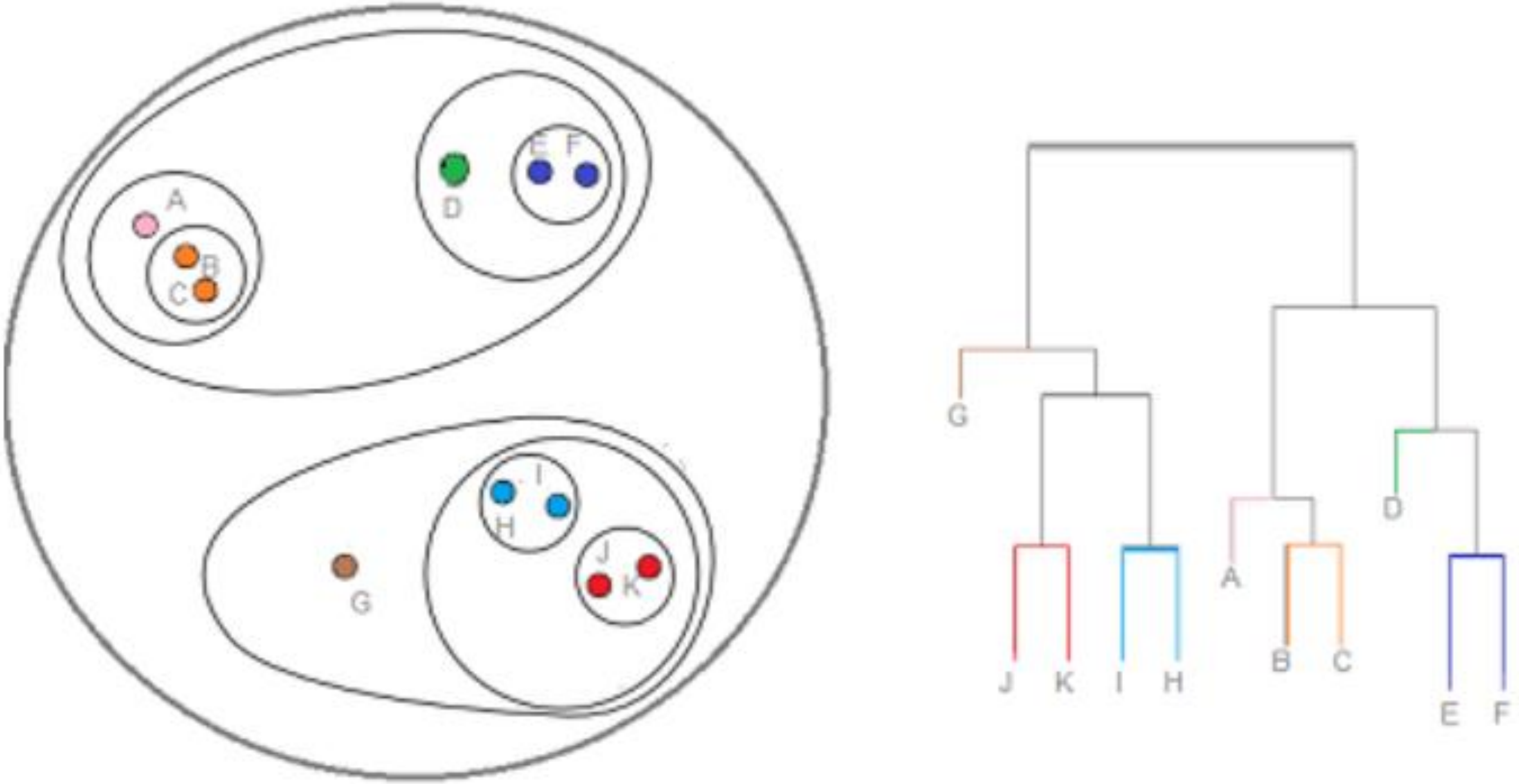
All objects stay in one cluster

Iteration:

Select a cluster and split it into two sub clusters

Until each leaf cluster contains only one object

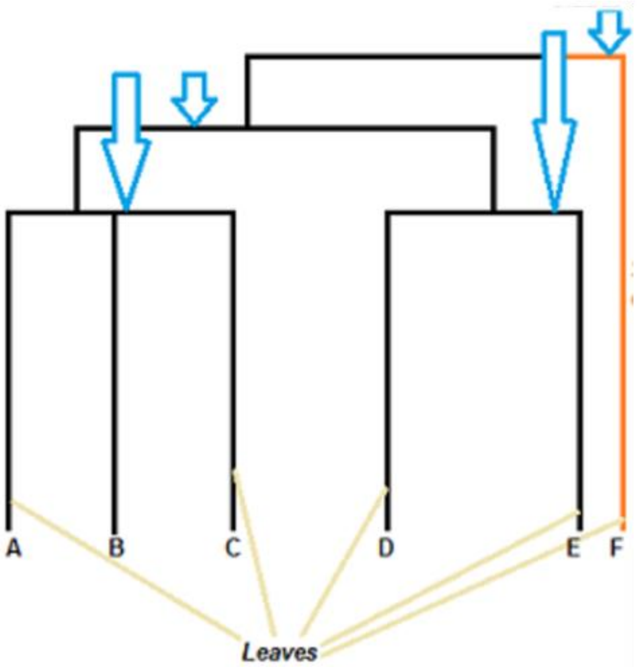
Dendrogram



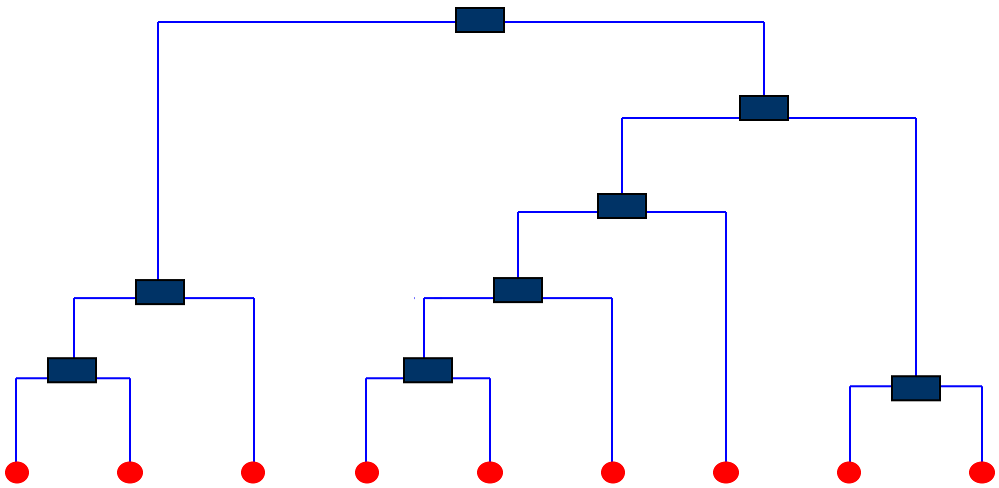
A dendrogram represents nested clusters

Dendrogram

- A binary tree that shows how clusters are merged/split hierarchically
- Each node on the tree is a cluster; each leaf node is a singleton cluster



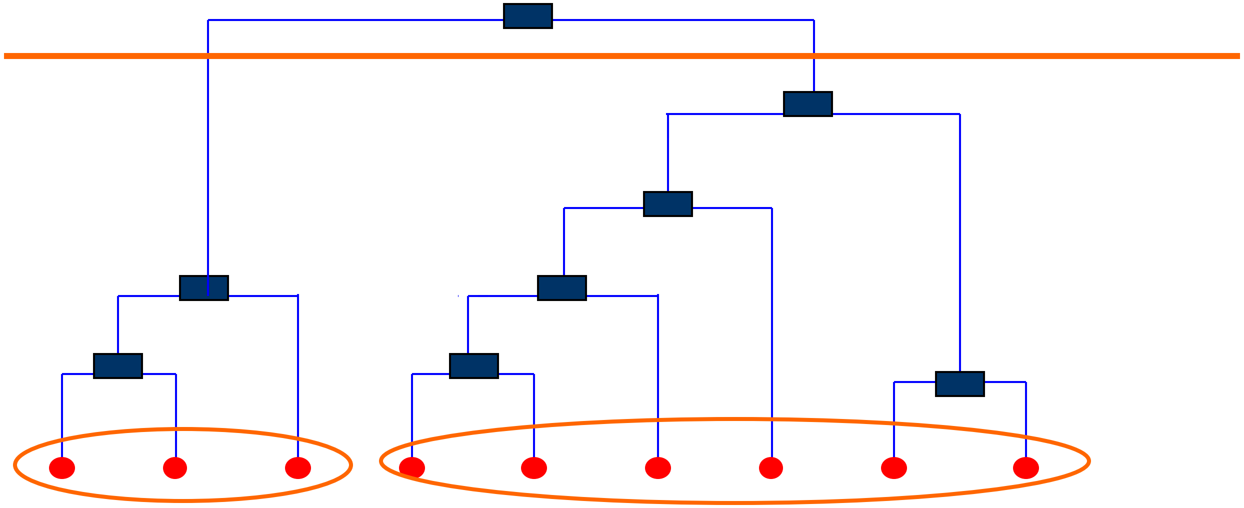
Example 1: How points are clustered



Example 2: How points are clustered

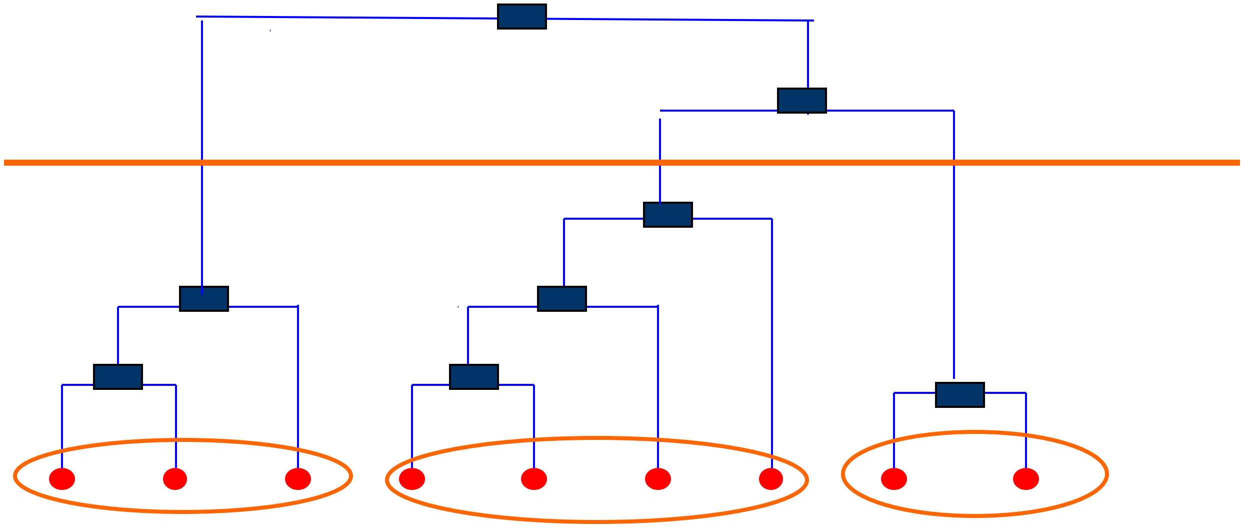
Dendrogram

- A clustering of the data objects is obtained by cutting the *dendrogram* at the desired level, then each connected component forms a cluster



Dendrogram

- A clustering of the data objects is obtained by cutting the *dendrogram* at the desired level, then each connected component forms a cluster



How to Merge Clusters?

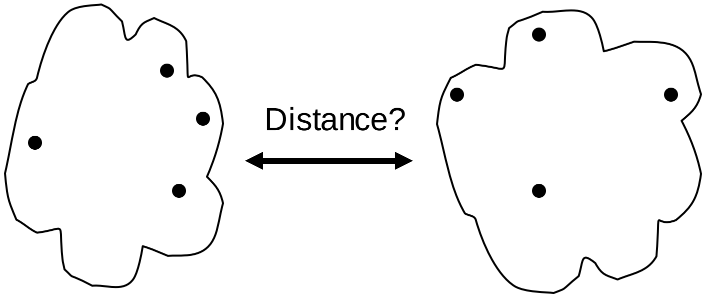
- How to measure the distance between clusters?

Single-link

Complete-link

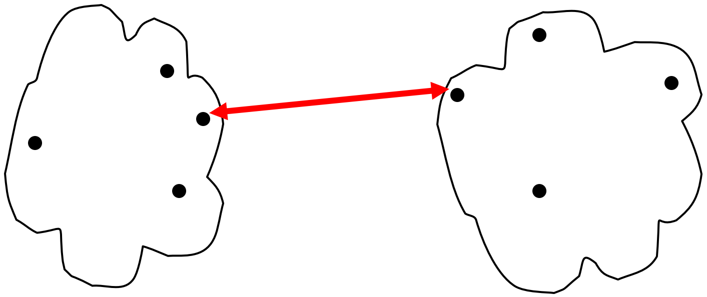
Average-link

Centroid distance



Hint: *Distance between clusters* is usually defined on the basis of *distance between objects.*

How to Define Inter-Cluster Distance



Single-link

Complete-link

Average-link

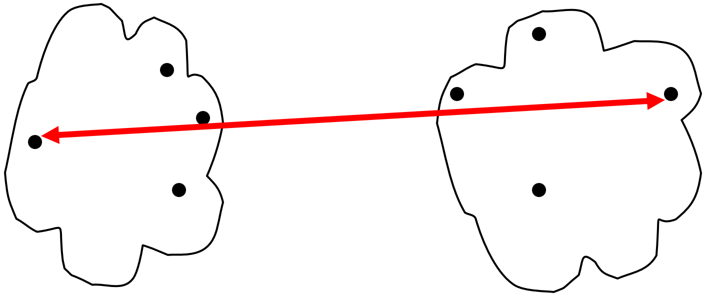
Centroid distance

$$d_{min}(C_i, C_j) = \min_{p \in C_i, q \in C_j} d(p,q)$$

- The distance between two clusters is represented by the distance of the *closest pair of data objects* belonging to different clusters.

- Can result in “straggly” (long and thin) clusters due to chaining effect

How to Define Inter-Cluster Distance



Single-link

Complete-link

Average-link

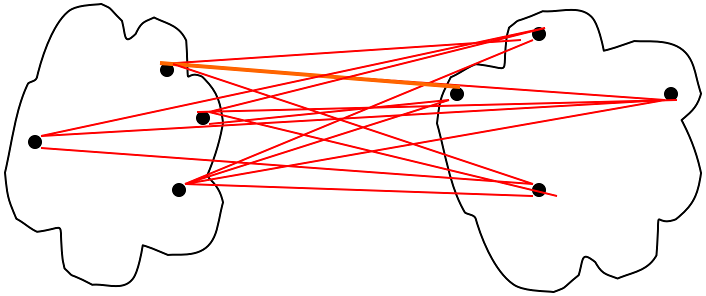
Centroid distance

$$d_{max}(C_i, C_j) = \max_{p \in C_i, q \in C_j} d(p, q)$$

- The distance between two clusters is represented by the distance of the *farthest pair of data objects* belonging to different clusters.

- Makes tighter spherical clusters that are typically preferred

How to Define Inter-Cluster Distance



Single-link

Complete-link

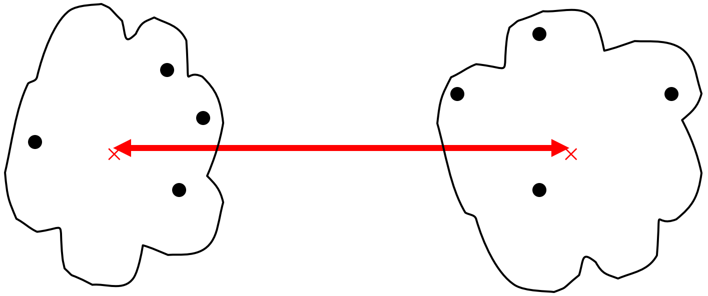
Average-link

Centroid distance

$$d_{avg}(C_i, C_j) = \underset{p \in C_i, q \in C_j}{\text{avg}} d(p, q)$$

The distance between two clusters is represented by the *average* distance of *all pairs of data objects* belonging to different clusters.

How to Define Inter-Cluster Distance



m_i,m_j are the means of C_i, C_j,

Single-link

Complete-link

Average-link

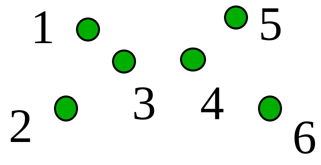
Centroid distance

$$d_{mean}(C_i, C_j) = d(m_i, m_j)$$

The distance between two clusters is represented by the distance between *the means of the cluters.*

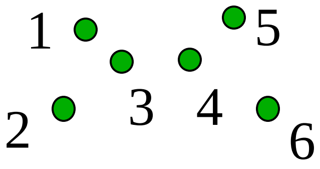
An Example of the Agglomerative Hierarchical Clustering Algorithm

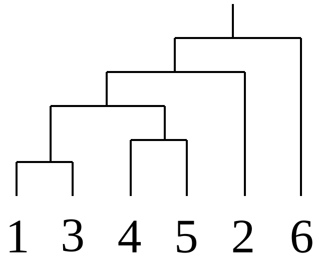
- For the following data set, we will get different clustering results with the single-link and complete-link algorithms.



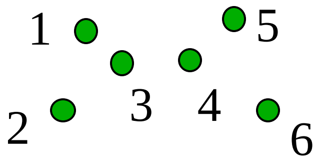
Results

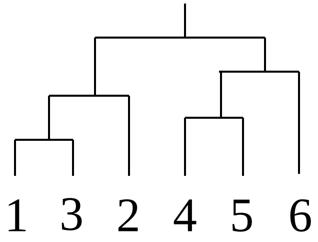
Single Link algorithm





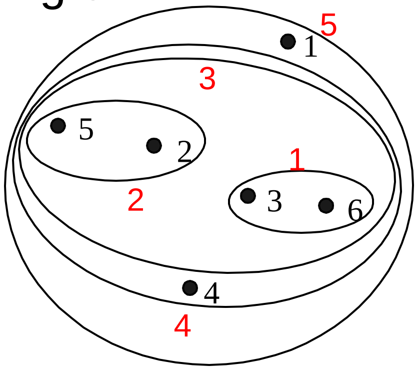
Complete Link algorithm



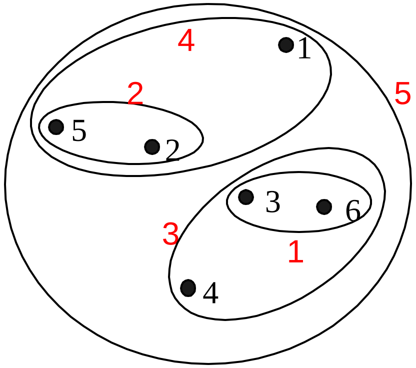


Hierarchical Clustering: Comparison

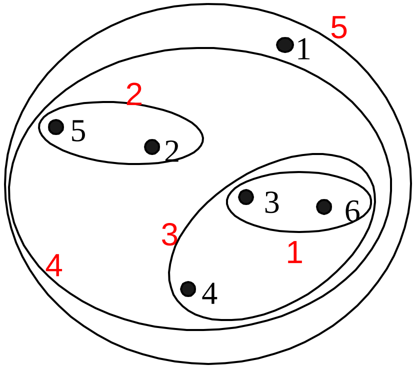
Single-link



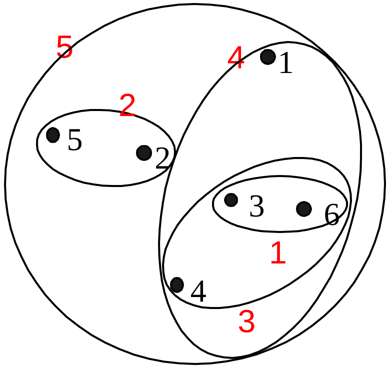
Complete-link



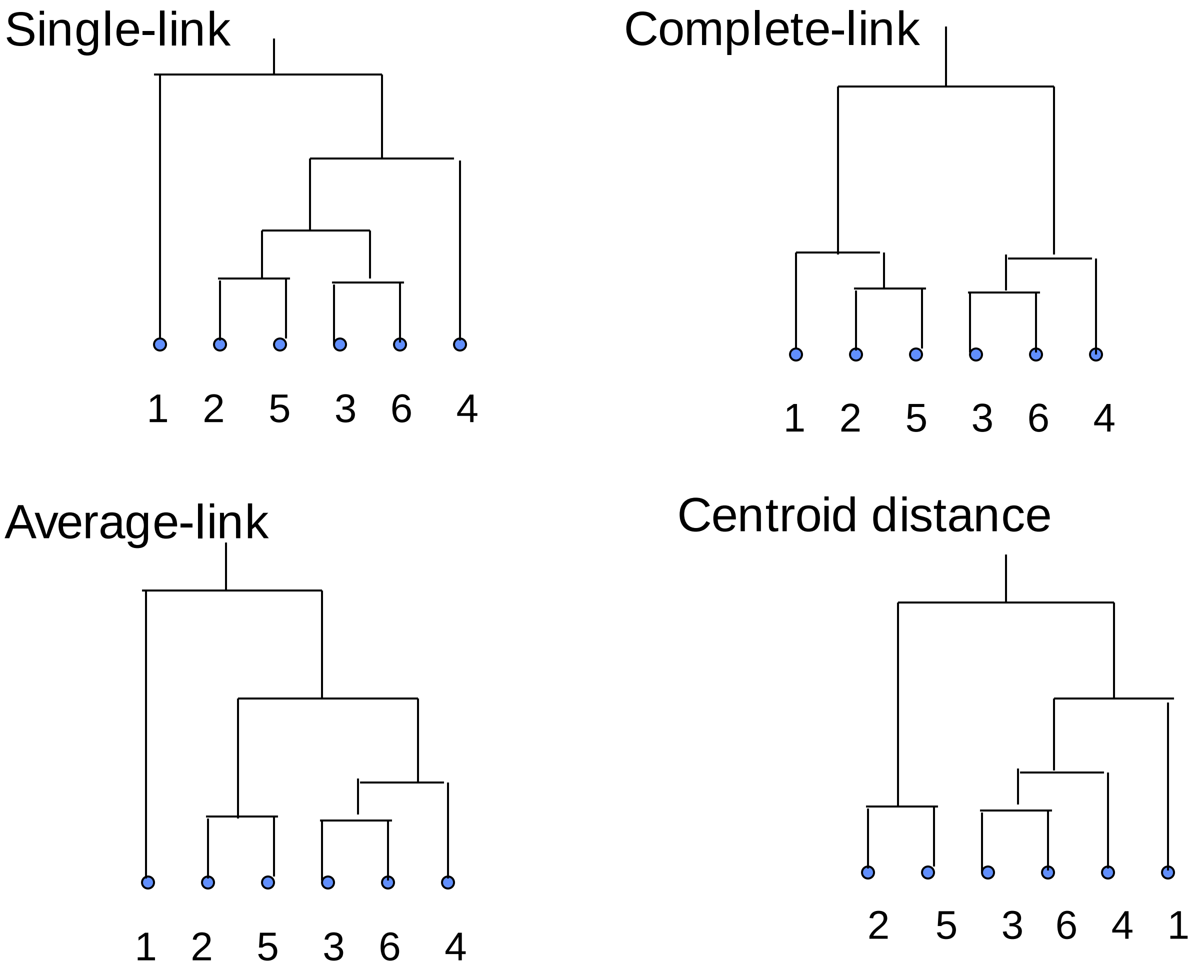
Average-link



Centroid distance



Compare Dendrograms


Single-link

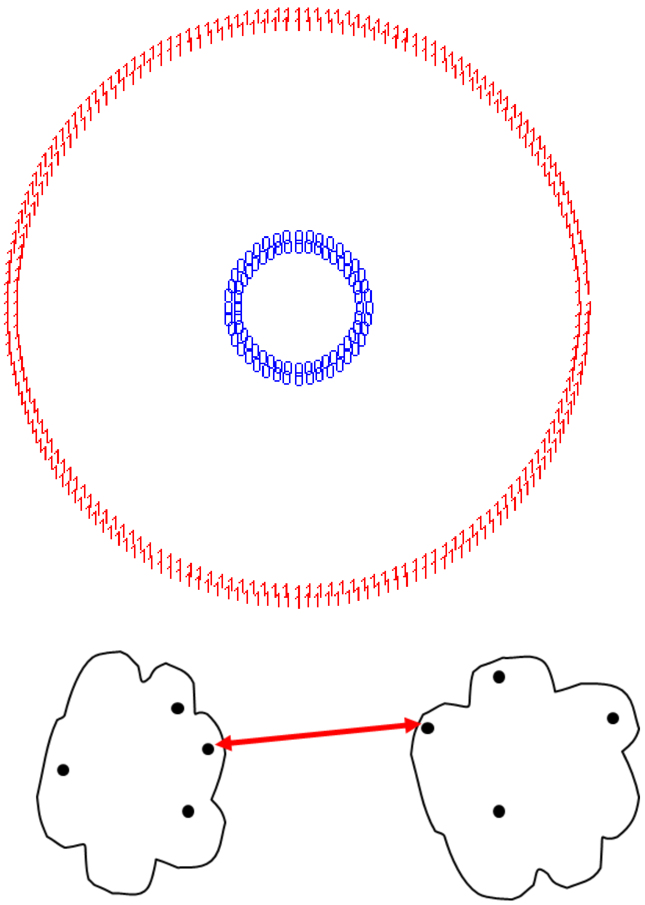
Complete-link

Average-link

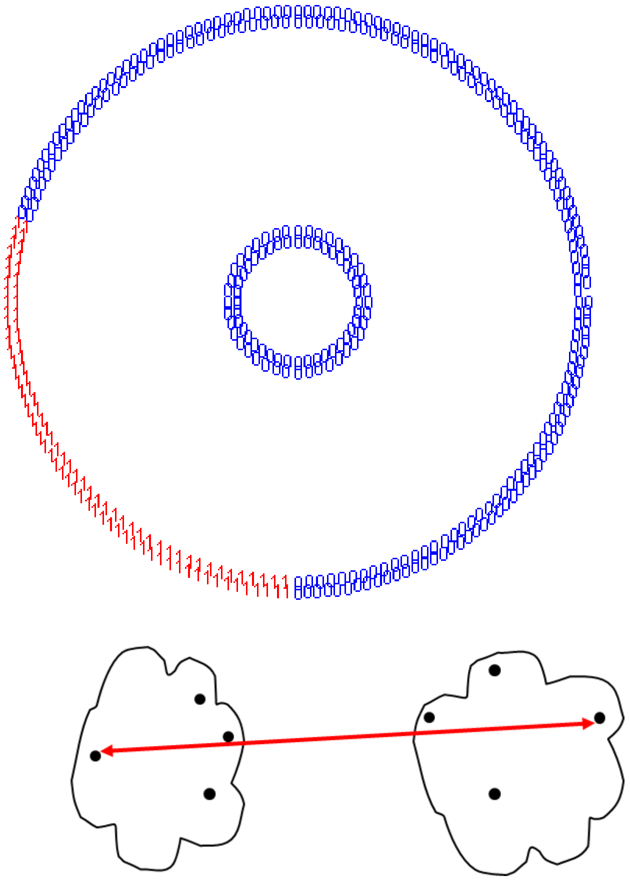
Centroid distance

Effect of Bias towards Spherical Clusters

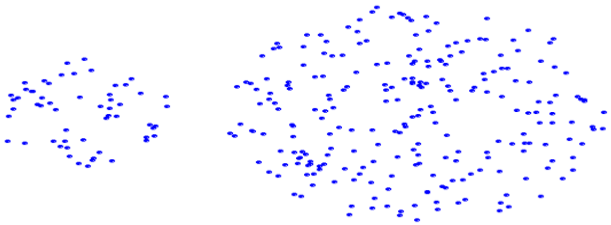
Single-link (2 clusters)



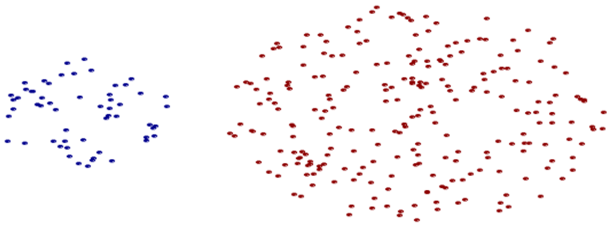
Complete-link (2 clusters)



Strength of Single-link



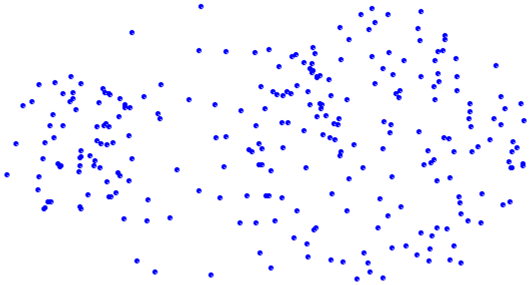
Original Points



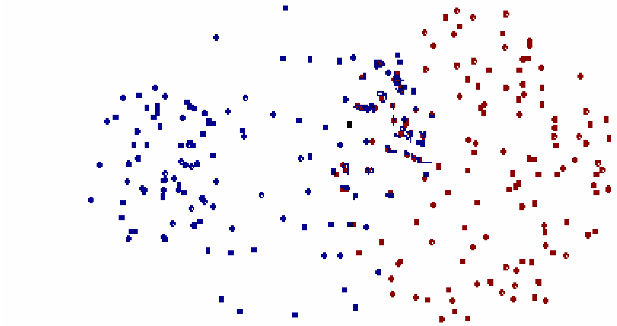
Two Clusters

- Can handle non-global shapes

Limitations of Single-Link



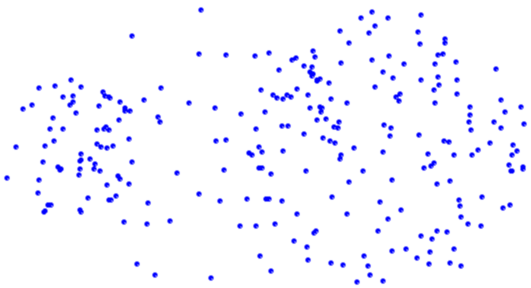
Original Points



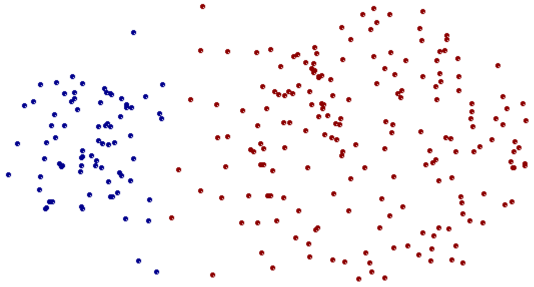
Two Clusters

- Sensitive to noise and outliers

Strength of Complete-link



Original Points



Two Clusters

- Less susceptible to noise and outliers

Which Method is Better?

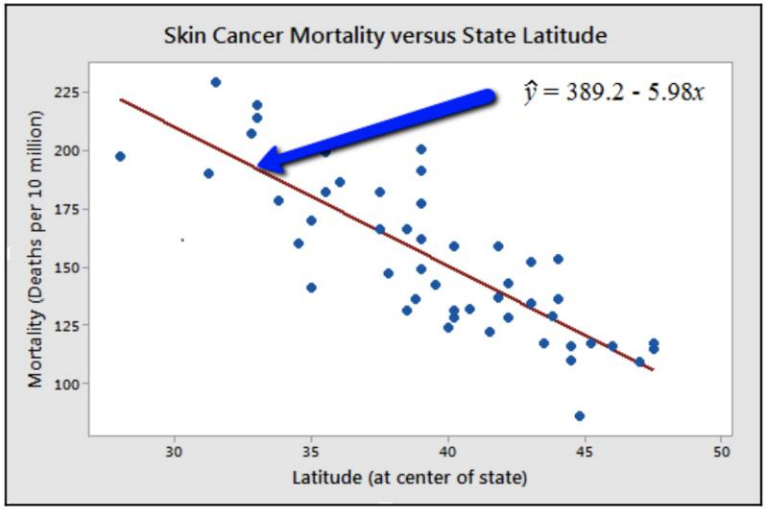
- Each method has its own advantages and disadvantages; application-dependent, single-link and complete-link are the most common methods
- Single-link
  - Can find irregular-shaped clusters
  - Sensitive to outliers, suffers the so-called chaining effects
- Complete-link, Average-link, and Centroid distance
  - Robust to outliers
  - Tend to break large clusters
  - Prefer spherical clusters

Other similarity measure

- In the examples described above, we used Euclidean distance to find the distance between points/clusters
- Depending on the type of the data, other similarity measures (measures of distance) might be preferred such as **correlation-based distance**
- Correlation-based distance considers two observations to be similar if their features are highly correlated, even though the observed values may be far apart in terms of Euclidean distance
- If Euclidean distance is chosen, then observations with high values of features will be clustered together. The same holds true for observations with low values of features.

Linear Regression: Motivating Example

- **Chances of getting skin cancer based on latitude**
- Questions
  - Relationship between skin-cancer mortality rate and latitude
  - Predicting unknown value
  - Should I move to city X in state Y (interpolate or extrapolate)



- **Other Examples,**
  - Alcohol consumed and blood alcohol content — as alcohol consumption increases, you'd expect one's blood alcohol content to increase, but not perfectly.
  - Vital lung capacity and pack-years of smoking — as amount of smoking increases (as quantified by the number of pack-years of smoking), you'd expect lung function (as quantified by vital lung capacity) to decrease, but not perfectly.
  - Driving speed and gas mileage — as driving speed increases, you'd expect gas mileage to decrease, but not perfectly.

Simple Linear Regression

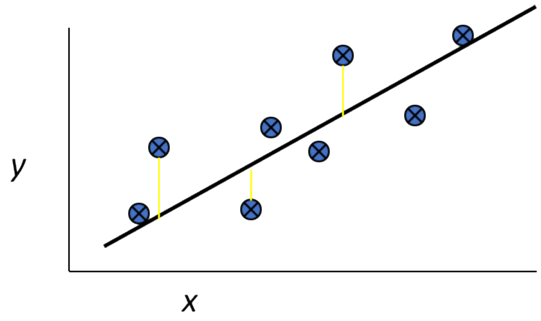
- **Simple linear regression** is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables

- One variable, denoted *x*, is regarded as the **predictor**, **explanatory**, or **independent** variable

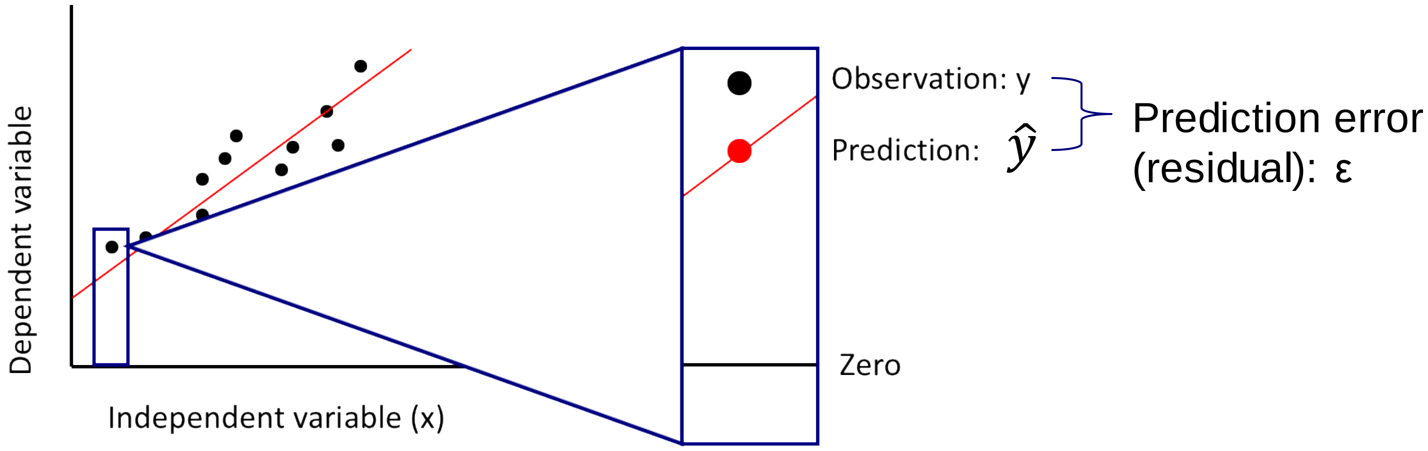
- The other variable, denoted *y*, is regarded as the **response**, **outcome**, or **dependent** variable

- We are interested in summarizing the trend between two quantitative variables, the natural question arises — "what is the best fitting line?"

$$\hat{y} = \alpha + \beta x$$



Simple Linear Regression



The function will make a prediction for each observed data point

The observation is denoted by y and the prediction is denoted by $\hat{y}$

For each observation, the variation can be described as:

$$y = \hat{y} + \varepsilon$$

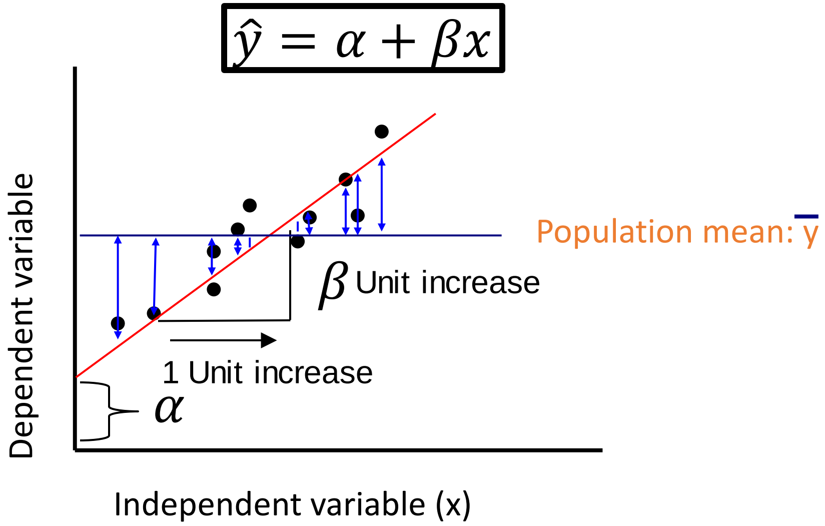
Actual = Explained + Error

Assumptions (or the fine print)

Linear regression assumes that…

- The relationship between X (independent) and Y (dependent) is linear
- Y is distributed normally at each value of X
  - Normality can be checked with a goodness of fit test, e.g., Kolmogorov-Smirnov test
- No or little multicollinearity
  - Multicollinearity occurs when the independent variables are highly correlated with each other.
- No auto-correlation
  - 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)
- Homoscedasticity (same variance)
  - meaning the residuals hve the same varianceacross the regression line

Simple Linear Regression



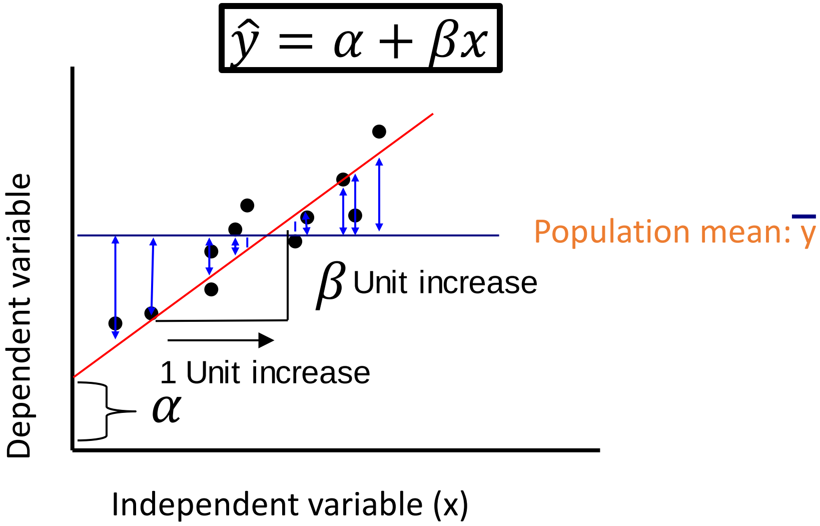
The Total Sum of Squares (SST) is equal to SSR + SSE.

- $SSR = \sum(\hat{y} - \bar{y})^2$ (measure of explained variation)
- $SSE = \sum(y - \hat{y})^2$ (measure of unexplained variation)
- $SST = \sum(y - \bar{y})^2$ (measure of total variation in y)

- A least squares regression selects the line with the lowest total sum of squared prediction errors, which is referred as Sum of Squares of Error, or SSE
- The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean
- The proportion of total variation (SST) that is explained by the regression (SSR) is known as the Coefficient of Determination, and is often referred to as r .

$$r^2 = \frac{SSR}{SST} = \frac{1 - SSE}{SST}$$

Simple Linear Regression



The Total Sum of Squares (SST) is equal to SSR + SSE.

- $SSR = \sum(\hat{y} - \bar{y})^2$ (measure of explained variation)
- $SSE = \sum(y - \hat{y})^2$ (measure of unexplained variation)
- $SST = \sum(y - \bar{y})^2$ (measure of total variation in y)

- The proportion of total variation (SST) that is explained by the regression (SSR) is known as the Coefficient of Determination, and is often referred to as r .

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

The value of r can range between 0 and 1, and the higher its value the more accurate the regression model is. It is often referred to as a percentage.

- If $r^2 = 1$, all of the data points fall perfectly on the regression line. The predictor *x* accounts for *all* of the variation in *y*
- If $r^2 = 0$, the estimated regression line is perfectly horizontal. The predictor *x* accounts for *none* of the variation in *y*

Estimating the intercept and slope: Least squares estimation

- Estimate parameters of the regression model using Least Square Estimation Method
- What’s the constraint? We are trying to minimize the squared distance (hence the “least squares”) between the observations themselves and the predicted values , or (also called the “residuals”, or left-over unexplained variability)

$$residual_i{}^2 = (y_i - (\alpha + \beta x))^2$$

- Find the β that gives the minimum sum of the squared differences. How do you maximize a function? Take the derivative; set it equal to zero; and solve. Typical max/min problem from calculus….

$$\frac{d}{d\beta}\sum_{i=1}^{n}(y_i - (\beta x_i + \alpha))^2 = 2(\sum_{i=1}^{n}(y_i - \beta x_i - \alpha)(-x_i))$$

First (summation) term is SSE

$$2(\sum_{i=1}^{n}(-y_i x_i + \beta x_i{}^2 + \alpha x_i)) = 0...$$

Resulting formulas…

Slope (beta coefficient: $$\boxed{\hat{\beta} = \frac{Cov(x, y)}{Var(x)}}$$

Intercept: $$\boxed{\hat{\alpha} = \overline{\mathrm{y}} - \hat{\beta}\overline{\mathrm{x}}}$$

Relationship with Pearson correlation

The correlation coefficient *r* is directly related to the coefficient of determination *r*^{2}

$$r = \pm\sqrt{r^2}$$

$$\hat{r} = \hat{\beta}\frac{\sigma_x}{\sigma_y}$$

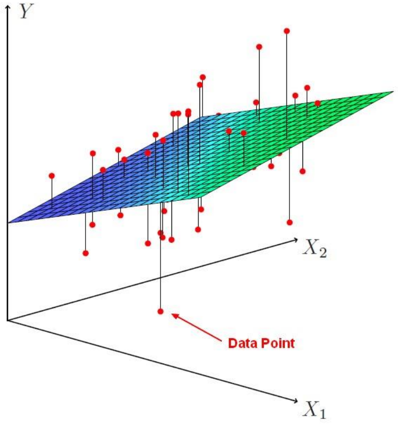
In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable (*X*) and the other the dependent (=outcome) variable *Y*.

Multivariate Regression

- Multiple linear regression model with two or more predictors
- Example: Are a person's brain size and body size predictive of his or her intelligence? [*Willerman, et al, 1991]*
  - Response (*y*): Performance IQ scores (**PIQ**) from the revised Wechsler Adult Intelligence Scale. Measure of the individual's intelligence.
  - Potential predictor (x_1): Brain size based on the count obtained from **MRI** scans (given as count/10,000).
  - Potential predictor (x_2): **Height** in inches
  - Potential predictor (x_3): **Weight** in pounds
- Model: $y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i = \beta^T X_i + \epsilon$
  - $y_i,\ x_{i1},\ x_{i2},\ x_{i3}$ represent PIQ, MRI, Height,

  and Weight of student *i*, respectively

  where, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ and $X_i = (1, x_{i1}, x_{i2}, x_{i3})$ are column vectors



Example

Non-linear Regression

- All of the regression models we have considered (including multiple linear) actually belong to a family of models called **generalized linear models**

- Generalized linear models provides a generalization of ordinary least squares regression that relates the **random term** (the response *Y*) to the **systematic term** (the linear predictor $\beta^T X$) via a **link function** (denoted by g(·)). Specifically, we have the relation

$$E(Y) = \mu = g^{-1}(\beta^T X)$$

Generalized Regression Model

- Different regression models can be derived for different link functions g(.)
  - When $g(\mu) = \mu$, we get linear regression: $\mu = \beta^T X$
  - When $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$, we get logistic regression: $\log(\frac{\mu}{1-\mu}) = \beta^T X$

Logistic Regression

One commonly used algorithm is Logistic Regression

- Assumes that the dependent (output) variable is binary which is often the case in medical and other studies. (Does person have disease or not, survive or not, accepted or not, etc.)
- Logistic Regression does a particular non-linear transform on the data after which it just does linear regression on the transformed data
- Fits the data with a sigmoidal/logistic curve rather than a line and outputs an approximation of the probability of the output given the input

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta^T X$$

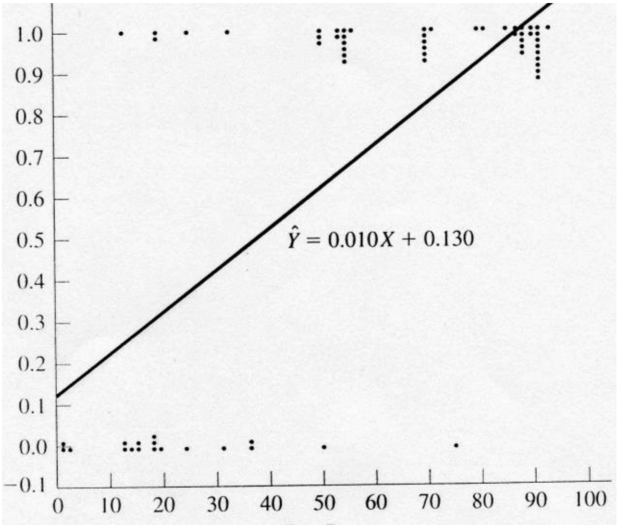
$$1 - \mu = \mu e^{-\beta^T X}$$

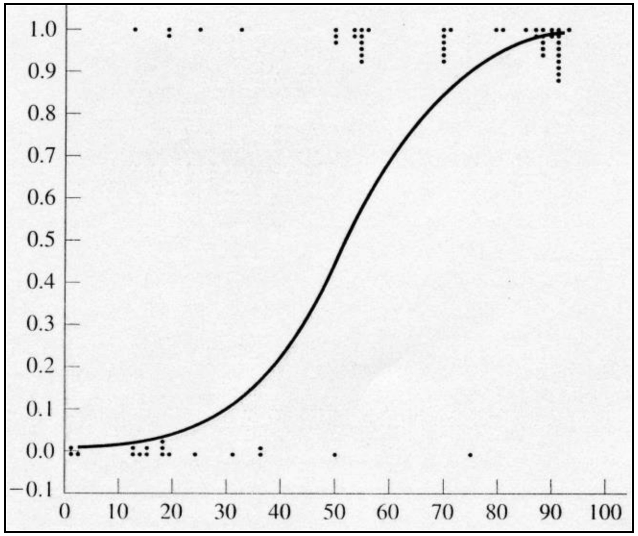
$$\mu = \frac{1}{1 + e^{-\beta^T X}}$$

Sigmoid/logistic function

Logistic Regression Example

- Age (X axis, input variable) – Data is fictional
- Heart Failure (Y axis, 1 or 0, output variable)
- Sigmoidal curve to the right gives empirically good probability approximation and is bounded between 0 and 1





Multinomial Logistic Regression

- Similar to logistic regression, but categorical output variable Y can take more than two values
- Examples
  - Which major will a college student choose, given their grades, stated likes and dislikes, etc.?
  - Which blood type does a person have, given the results of various diagnostic tests?
- Let Y take values from $\{1,2,\dots,K\}$. Multinomial logistic regression model is as follows

$$P(Y=k) \propto e^{\beta_k^T X} \quad \forall\, k \in \{1,\dots,K\}$$

where, $\beta_k = (\beta_{0,k}, \beta_{1,k}, \dots, \beta_{M,k})$ and $X = (1, x_1, \dots, x_M)$

  - Note that the model parameters (β's) are calculated for each value of the categorical variable

Logistic Regression – A Multiclass Example

Binary outcome

- The dependent variable is taking values like
  - Responder /non responder
  - Loss giving / good profile
  - Buyer / non buyer
  - Account holder will make payment / no payment

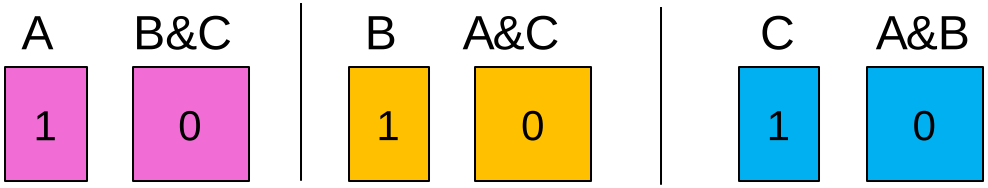
Multiple Classes

- More than two outcome – polytomous / multinomial logit
- At times the dependent variable has
  - More than two possible outcomes
  - They are nominal variables: there is no order in the outcome

Convert multinomial to many binomial

Example – there are three classes of nominal outcome A, B, and C

- Step 1
  - Develop three models separately. Class A vs Rest, Class B vs Rest…



A B&C

1 0

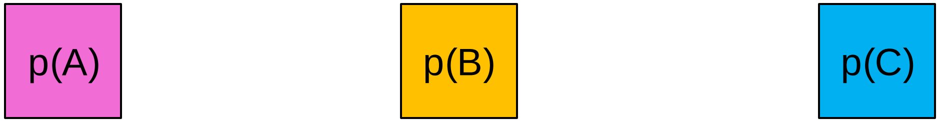
B A&C

1 0

C A&B

1 0

  - Develop equation for three probabilities



p(A) p(B) p(C)

  - Assign any record to the class, based on the input variables, which has the highest probability
  - If p(A) > p(B) and p(A) > p(C) then outcome = class A