MP2 Intro, Regression

Lecture 10:
Mini-Project 2 Introduction,
Linear and Non-Linear Regression

ECE/CS 498 DS
Professor Ravi K. Iyer
Department of Electrical and Computer Engineering
University of Illinois

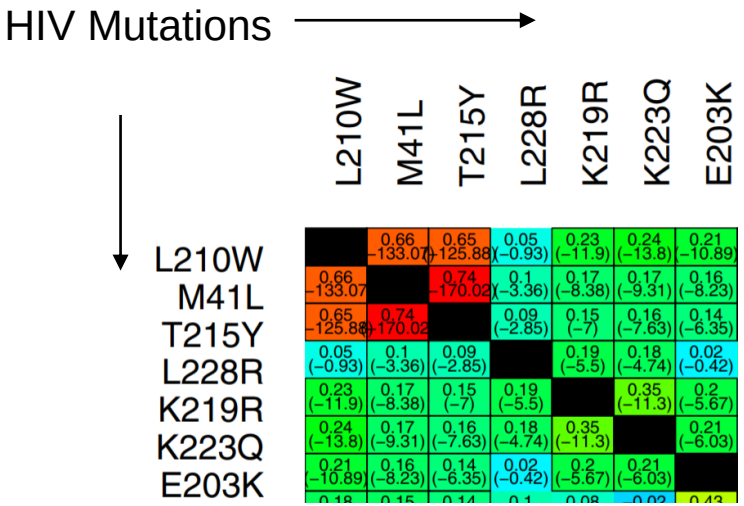
Announcements

- TA Office Hours moved to **MW 2-3 PM in Everitt 1302**
- Guest Lecture this **Wed Feb 26 during class**
  - Dr. Jasmohan Bajaj from Virginia Commonwealth University will provide some background context on the domain of MP 2
  - Attendance in mandatory! Writeup on key takeaways from talk will be submitted via Compass2G by the end of lecture Wed
- MP 2 released today (more information coming in today's lecture)
- Grad Students:
  - Initial grad project ideas due this **Wed Feb 26 @ 11:59 PM**
  - Initial grad project discussions this week: **Wed Feb 26-28**
    - Signup via: https://docs.google.com/spreadsheets/d/1Cd2RxSJWp8Im-K_sMkbuPRwxsTykOM3rdUzexSsYkko/edit?usp=sharing
- HW 2 released, due **Mar 2 @ 11:59 PM on Compass2G**
  - Covers (i) inferencing with Bayesian networks and (ii) clustering with k-means and GMM
- Midterm exam will take place on **Wed March 11th**

Hierarchical Clustering Example

Characterization Novel HIV Drug Resistance Mutations using Clustering.

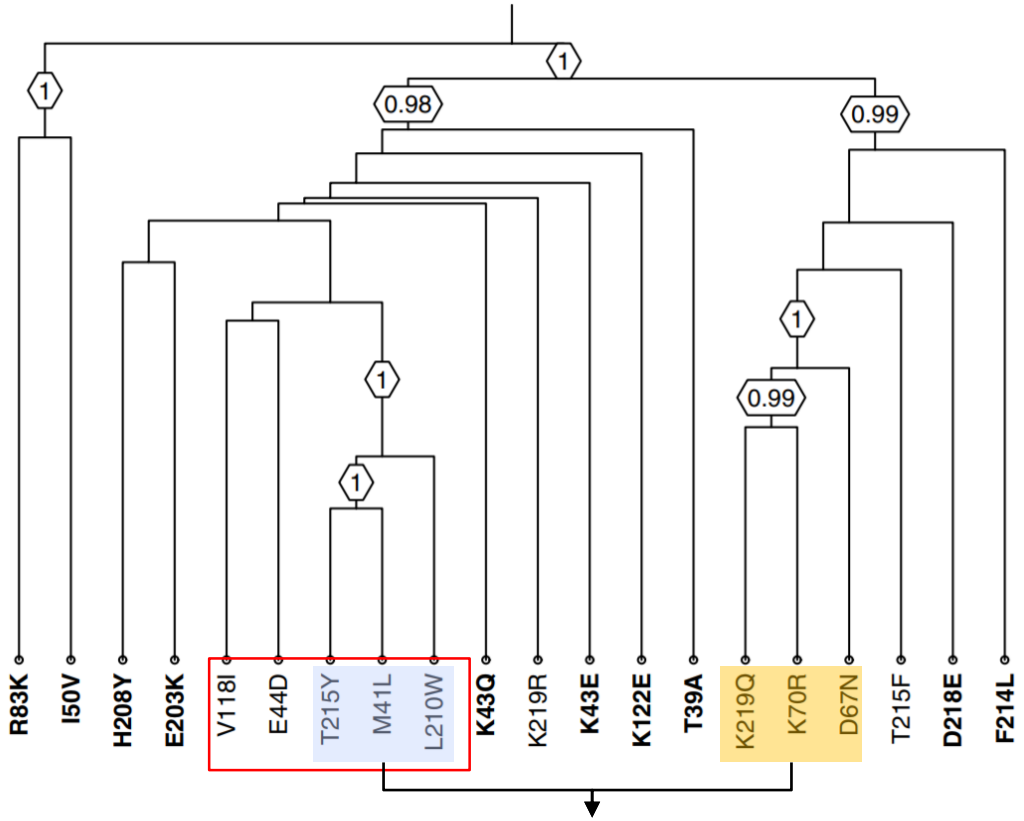
- **Objective:** By clustering new HIV mutations with HIV mutations that have known drug resistance mechanisms, we can infer the possible drug resistance mechanisms of the new mutations
- **Clustering Technique:**
  Agglomerative Hierarchical Clustering using Average-link
- **Distance Metric:**
  Matthews correlation coefficient
  - This coefficient measures how two individual mutations vary together in the population.



Reference: Sing, Tobias, et al. "Characterization of novel HIV drug resistance mutations using clustering, multidimensional scaling and SVM-based feature ranking." *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, Berlin, Heidelberg, 2005

Hierarchical Clustering Example

- **Dendrogram after clustering**



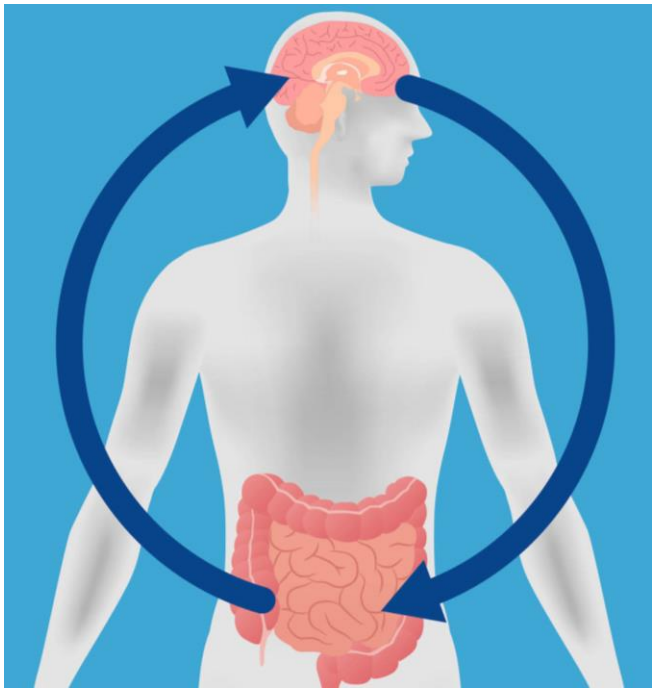
Mutations with known drug resistance mechanisms

- Clustering is performed on HIV mutations with known and unknown drug resistance mechanisms
- Mutation complexes in shaded boxes have known drug resistance mechanisms
- Mutations with unknown drug resistance mechanisms may have similar drug resistance mechanisms with other mutations they are clustered with.
- E.g, Mutations E44D and V118I may have similar resistance mechanisms to the blue mutation complex, for which a drug resistance mechanism is already known.

Mini-Project 2 Introduction

MP 2 – Problem Statement

- Liver cirrhosis is a condition where permanent damage has occurred to the liver
  - When severe enough, the only treatment option is a liver transplant
- Hepatic Encephalopathy (HE) is a condition where disfunction of the liver can result in the accumulation of toxins in the brain
  - HE is usually temporary, but patients suffering from it can see drastic changes to brain function
- Gut Microbiome is the collection of microbes present in the human gut
  - Composition of this microbiome varies from person to person
- We suspect that composition of the gut microbiome can provide clues about occurrence of HE
- Problem statement: **Investigate how the abundance levels of microbes in the gut are related to the occurrence of hepatic encephalopathy in patients with liver cirrhosis**



https://kibm.ucsd.edu/sites/default/files/file_fields/event/KIBM17_Gut_Brain_Axis_WEB_4.6.17%20copy_0.jpg

MP 2- Data

- You are provided with relative abundance data detailing the composition of the gut microbiome for $n \sim 1500$ patients
  - All patients have liver cirrhosis
  - ~ 750 patients don't have HE ("HE0" population), and the remainder have it ("HE1" population)
- For each patient, you are presented with the relative abundance data for ~ 150 different microbes from the gut microbiome
  - The "relative" abundance of microbe X is the proportion of microbe X in the patients gut microbiome
  - Data was collected via stool samples from patients

MP2 - Task Breakdown

- Task 1: Data cleaning and visual inspection
  - Bayesian networks for quality control
  - Data standardization
  - Data visualization with heatmaps
- Task 2: Statistical analysis
  - K-S test and multiple testing for identifying significant differences in abundance levels between HE0 and HE1 populations
- Task 3: Dimensionality reduction and clustering
  - PCA and t-SNE for dimensionality reduction
  - K-means, GMM, and hierarchical clustering
- Task 4: Interpreting your results
  - Engineer your own method to identify how the various clusters are related
  - Relate your findings within a biological context

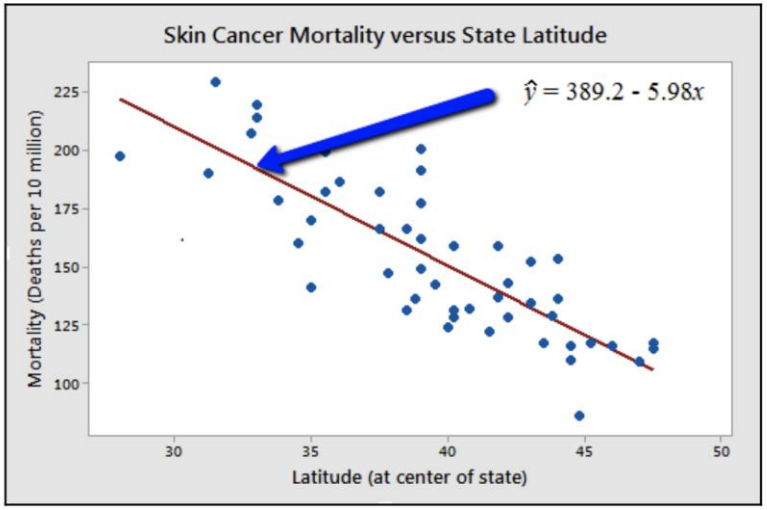
MP 2 Timeline

- Release: **Monday, Feb 24**
- Checkpoint 0.5: **Monday, Mar 2 @ 11:59 PM via Google Form**
  - Quick update on status of work since release
- Checkpoint 1: **Friday, Mar 13 @ 11:59 PM on Compass2G**
  - Single .ipynb notebook with Tasks 1 and 2 completed
  - Single .pdf file with answer to Tasks 1 and 2 (we will provide template)
- Checkpoint 1.5: **Wednesday, Mar 25 @ 11:59 PM via Google Form**
  - Quick update on status of work since checkpoint 1
- Final Submission: **Monday, Mar 30 @ 11:59 PM on Compass2G**
  - Single .ipynb notebook with all tasks completed
  - Single .pdf file with answer to all tasks (we will provide template)
- Reminder: Make sure you are submitting your own work and following the university's academic integrity policy

Linear Regression

Linear Regression: Motivating Example

- **Chances of getting skin cancer based on geographical 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

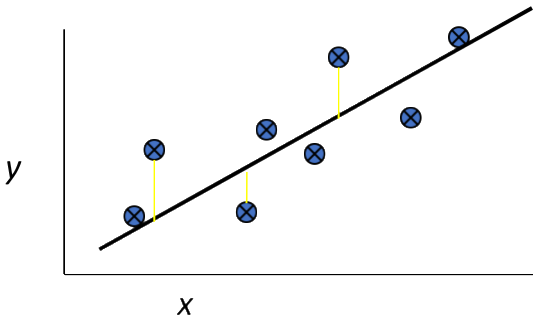
- **Simple linear regression** is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables

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

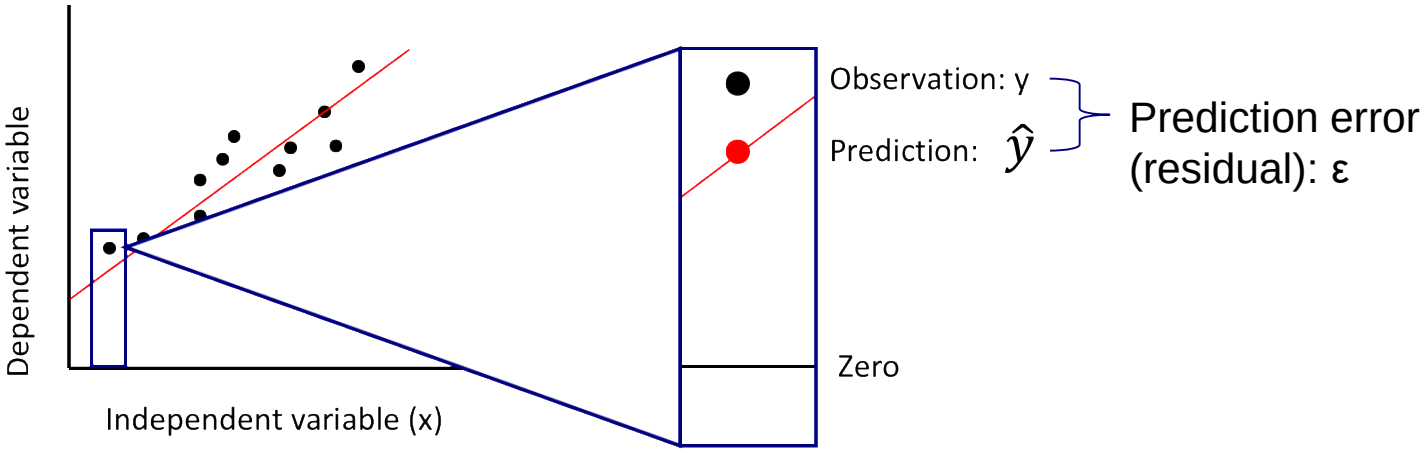
- The other variable, denoted *y*, is regarded as the **response**, **outcome**, or **dependent** variable

- 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:

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

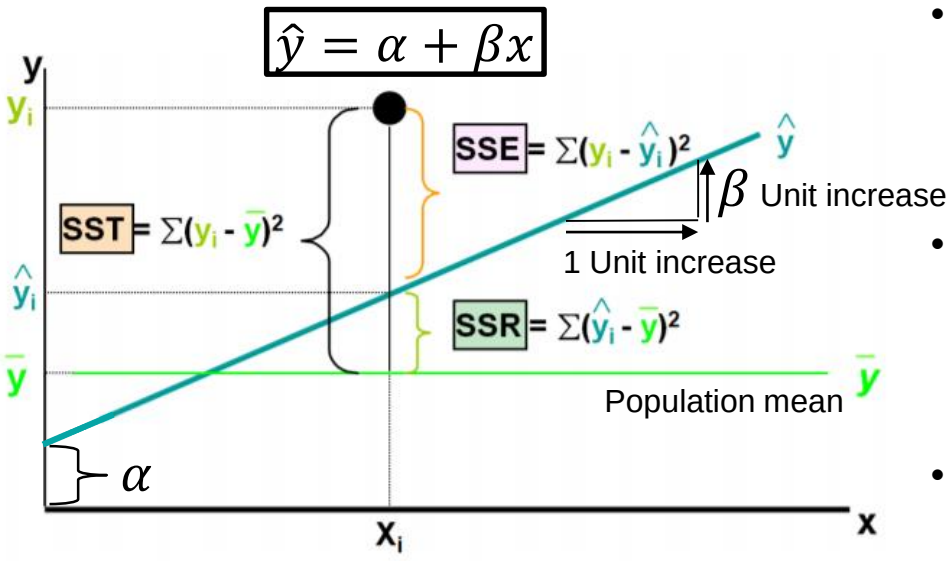
Actual = Explained + Error

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 (errors) 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 have the same variance across the regression line

Simple Linear Regression



The Total Sum of Squares (SST) is equal to SSR + SSE. ($\bar{y}$ is the population mean)

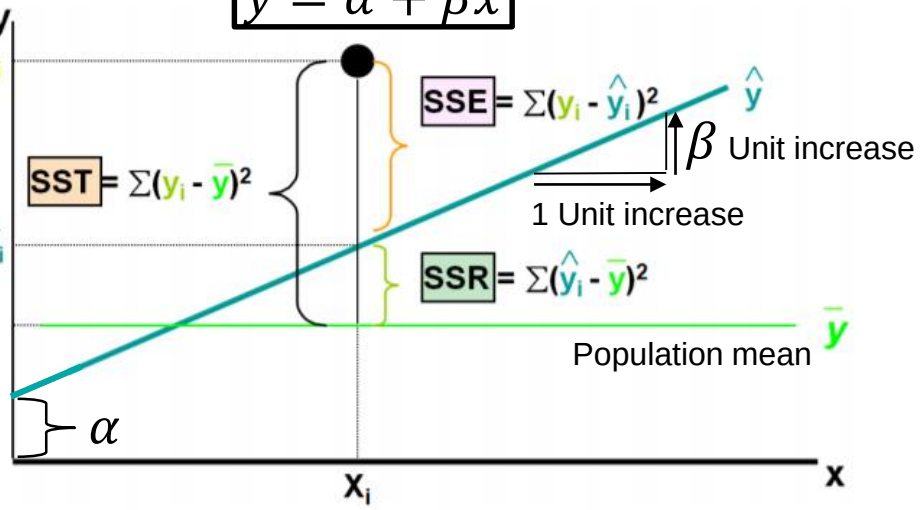
- $SSR = \sum_i (\hat{y}_i - \bar{y})^2$ (measure of explained variation)
- $SSE = \sum_i (y_i - \hat{y}_i)^2$ (measure of unexplained variation)
- $SST = \sum_i (y_i - \bar{y})^2$ (measure of total variation in y)

- 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 (SSE)
- Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation ($\hat{y}_i$) and the population mean ($\bar{y}$)
- Total Sum of Squares (SST) measures the sum of the squared differences of true values (y_i) and the population mean ($\bar{y}$)
- The proportion of total variation that is explained by the regression (SSR/SST) is known as the Coefficient of Determination, and is often referred to as r^2.

$$r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Simple Linear Regression

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



The Total Sum of Squares (SST) is equal to SSR + SSE. ($\bar{y}$ is the population mean)

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

- The proportion of total variation that is explained by the regression (SSR/SST) is known as the Coefficient of Determination, and is often referred to as r^2.

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

- The value of r^2 can range between 0 and 1, and the higher its value the more accurate the regression model is. It is often expressed 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 (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 minimize. a function? Take the derivative; set it equal to zero; and solve. Typical max/min problem from calculus….

First (summation) term is SSE

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

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

Resulting formulas…

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

Intercept: $$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

For samples (x_i, y_i) for $i \in \{1, \ldots, N\}$, the coefficient and slope are given as:

$$\hat{\beta} = \frac{N\left(\sum_{i=1}^{N} x_i y_i\right) - \left(\sum_{i=1}^{N} x_i\right) \cdot \left(\sum_{i=1}^{N} y_i\right)}{N\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2}$$

$$\hat{\alpha} = \frac{\left(\sum_{i=1}^{N} y_i\right)\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} x_i y_i\right)}{N\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2}$$

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*.

Multiple/Multivariate Linear Regression

- **Multiple Linear Regression** ⇒ using multiple features $(X_1, X_2, \dots)$ to predict an outcome Y
  - E.g. $\hat{Y} = \alpha + \beta_1 X_1 + B_2 X_2$

- **Multivariate Linear Regression** ⇒ using a single feature X to independently predict multiple outcomes $(Y_1, Y_2, \dots)$
  - E.g. $\widehat{Y_1} = \alpha_1 + \beta_1 X,$ $\widehat{Y_2} = \alpha_2 + \beta_2 X$

- **Multivariate Multiple Linear Regression** ⇒ using multiple features $(X_1, X_2, \dots)$ to independently predict multiple outcomes $(Y_1, Y_2, \dots)$
  - E.g. $\widehat{Y_1} = \alpha_1 + \beta_{1,1} X_1 + \beta_{1,2} X_2,$ $\widehat{Y_2} = \alpha_2 + \beta_{2,1} X_1 + \beta_{2,2} X_2$

Multivariate Multiple Linear Regression Example

- Suppose we have data on air pollutants from 12 monitoring sites in Beijing from 2013-2017
  - Real data – from UC Irvine ML repository (~394k data points)
  - Features:
    - X_1: Temperature in degrees Celsius
    - X_2: Air pressure in hectopascals (hPa)
  - Outcomes:
    - Y_1: Concentration of carbon monoxide (CO) in micro-gram per cubic meter (μg/m^3)
    - Y_2: Concentration of ozone (O3) in micro-gram per cubic meter (μg/m^3)

Goal: **Given the the temperature (X_1) and the air pressure (X_2), try to predict the concentration of carbon monoxide (Y_1) and of Ozone (Y_2)in the air**

https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data#

Multivariate Multiple Linear Regression Example

Goal: **Given the the temperature (X_1) and the air pressure (X_2), try to predict the concentration of carbon monoxide (Y_1) and of Ozone (Y_2)in the air**

Linear equations to derive:

$$\widehat{Y_1} = \alpha_1 + \beta_{1,1}X_1 + \beta_{1,2}X_2$$

$$\widehat{Y_2} = \alpha_2 + \beta_{2,1}X_1 + \beta_{2,2}X_2$$

After solving with linear regression in Python, we obtain

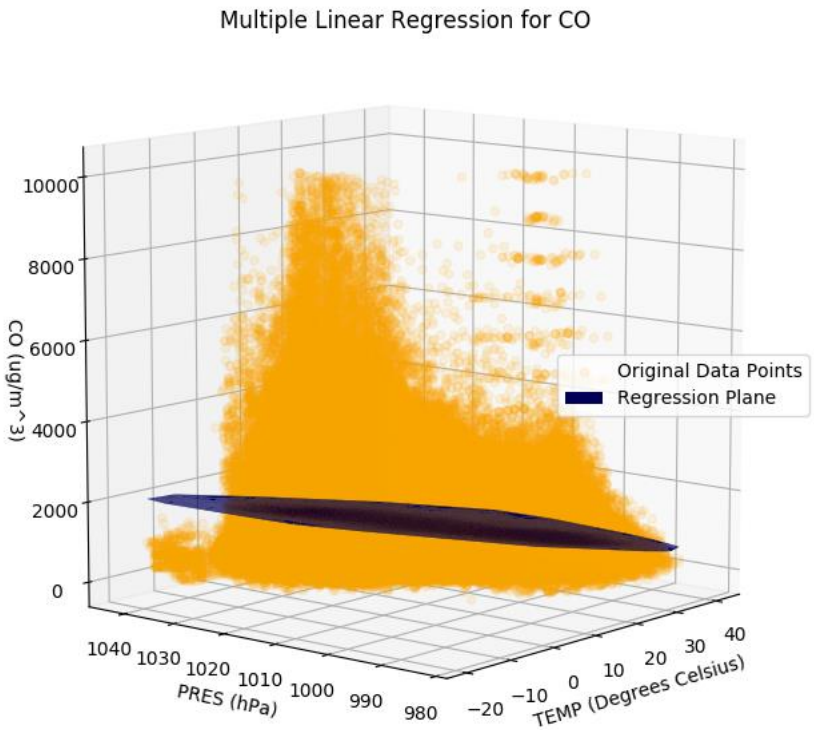
$$\widehat{Y_1} = 27199.32 - 51.35\,X_1 - 25.01\,X_2$$

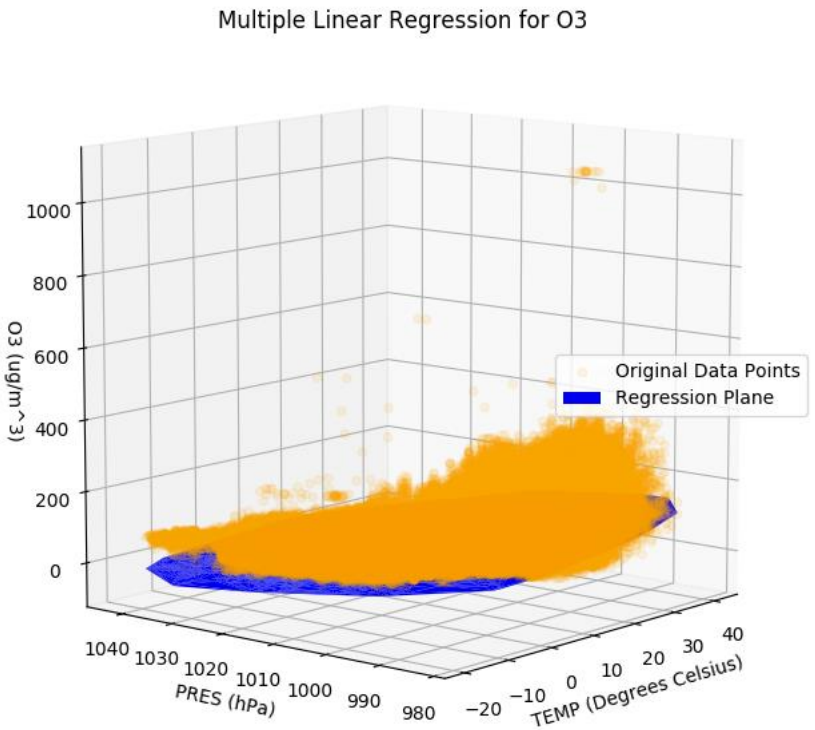
$$\widehat{Y_2} = -601.24 + 3.41\,X_1 + 0.61\,X_2$$

Multivariate Multiple Linear Regression Example

Goal: **Given the the temperature (X_1) and the air pressure (X_2), try to predict the concentration of carbon monoxide (Y_1) and of Ozone (Y_2)in the air**

$$\widehat{Y_1} = 27199.32 - 51.35\,X_1 - 25.01\,X_2$$
$$\widehat{Y_2} = -601.24 + 3.41\,X_1 + 0.61\,X_2$$





Non-Linear Regression

GLM: Generalized Linear Model

- Linear regression and logistic regression belong to a family of models called **generalized linear models (GLMs)**
- All generalized linear models have the following three components:
  1. **Random Component**: A probability distribution describing the outcome/response variable Y
  2. **Systematic Component**: A linear predictor $\eta = \boldsymbol{\beta^T X} = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$
  3. **Link Function**: A function that relates $\mu = E(Y)$ to η

     $$g(\mu) = \eta \quad or \quad \mu = g^{-1}(\eta)$$

     - When $g(\mu) = \mu$, we get linear regression: $\mu = \eta$
     - When $g(\mu) = \log(\frac{\mu}{1-\mu})$, we get logistic regression: $\log(\frac{\mu}{1-\mu}) = \eta$
     - $f = g^{-1}$ is sometimes referred to as the **activation function**

Logistic Regression - Purpose

- Logistic regression is a GLM used to model a **binary categorical variable** using numerical and categorical predictors
  - Examples of output variable: Does the person have disease or not, survive or not, etc.

- We want to model the probability of success p for the outcome variable given a set of predictors/input

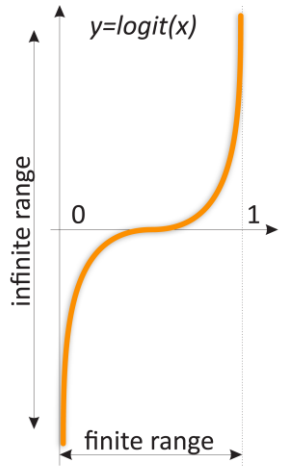
- Logistic regression fits the data with a sigmoid/logistic curve, from which we get this probability p
  - In this case, $p = E[Y]$ since Y is a Bernoulli random variable

- We need to establish a reasonable link function that connects η to p. The most common function for this purpose is the **logit function**

Logistic Regression - Logit Function

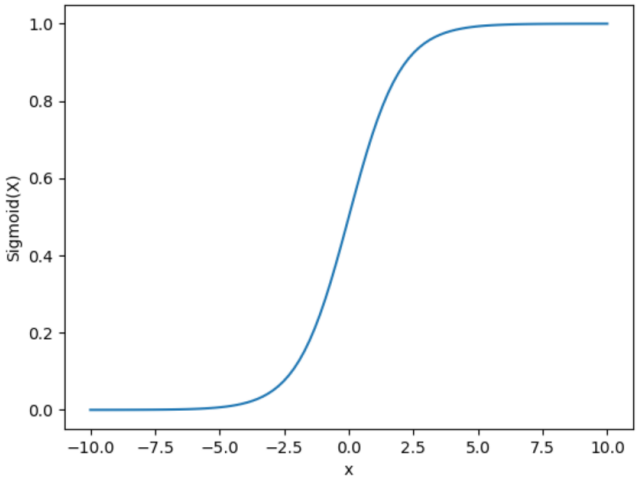
- The **logit function** takes a value p between 0 and 1 and maps it to a continuous scale between $-\infty$ and ∞

$$\eta = logit(p) = ln\left(\frac{p}{1-p}\right) for\ 0 < p < 1$$



- The **activation/sigmoid** function (inverse logit) takes a value between $-\infty$ and ∞ and maps it to a value between 0 and 1

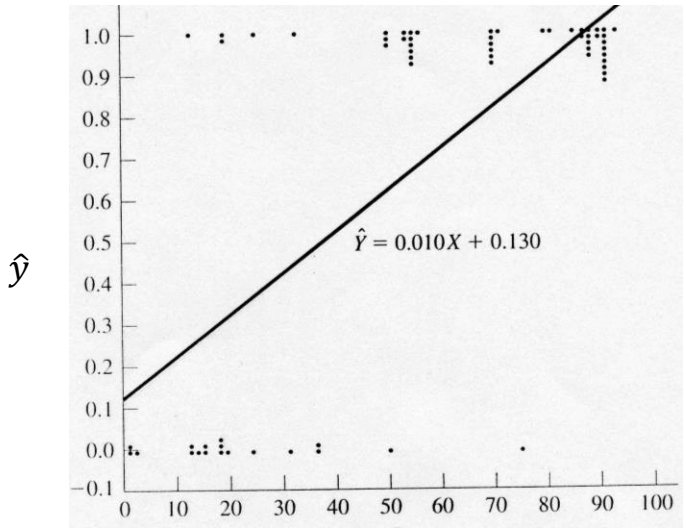
$$\mu = f(\eta) = g^{-1}(\eta) = \frac{\exp(\eta)}{1+\exp(\eta)} = \frac{1}{1+\exp(-\eta)}$$



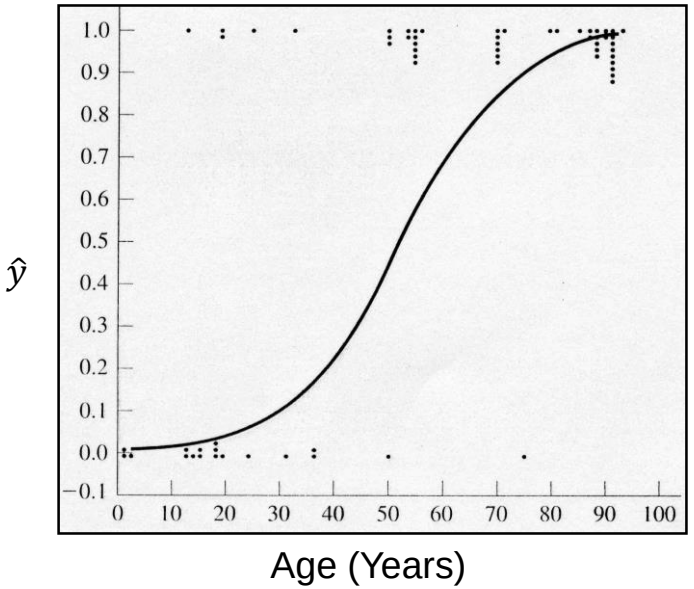
Logistic Regression Example

- 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

Linear Regression to Predict Heart Failure using Age



Logistic Regression to Predict Heart Failure using Age



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

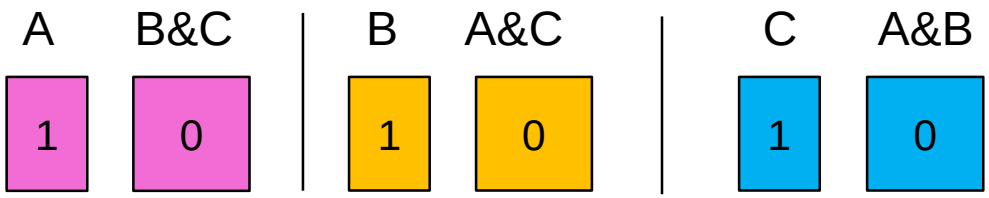
Multiple Classes

- More than two outcome – polytomous / multinomial logit
- The dependent variable
  - Has more than two possible outcomes
  - Is a nominal variable, with no natural order in the outcome values

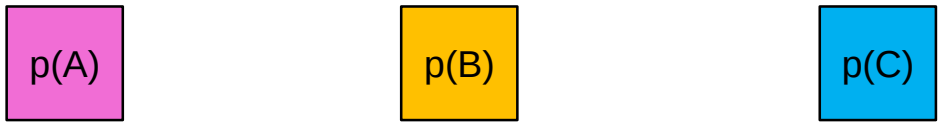
Convert multinomial to many binomial

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

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



2. Develop equation for three probabilities



3. 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

Logistic Regression Example: Donner Party

Donner Party - Background

- In 1846, the Donner and Reed families left Springfield, IL for California in a covered wagon
- In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming.
- There, its leaders decided to attempt a new and untested route to the Sacramento Valley.
- Having reached its full size of 87 people and 20 wagons, the party was delayed by a difficult crossing of the Wasatch Range and again in the crossing of the desert west of the Great Salt Lake.
- The group became stranded in the eastern Sierra Nevada mountains when the region was hit by heavy snows in late October.
- By the time the last survivor was rescued on April 21, 1847, 40 of the 87 members had died from famine and exposure to extreme cold.

From Ramsey, F.L. and Schafer, D.W. (2002). The Statistical Sleuth: A Course in Method of Data Analysis (2nd edition)

Donner Party – Survival Model

- **Goal**: Predict an individual's chance of survival in the Donner Party using their Age and Gender
- We can think of each person's survival as a Bernoulli trial with a success (or survival) parameter p
- Then, we calculate p using logistic regression with the features Age and Gender

$$p = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 \times Age + \beta_2 \times Gender))}$$

- Let $x_{i,1}$ represent the Age of, $x_{i,2}$ represent the Gender of, and p_i represents the probability of survival of person i

$$p_i = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2}))}$$

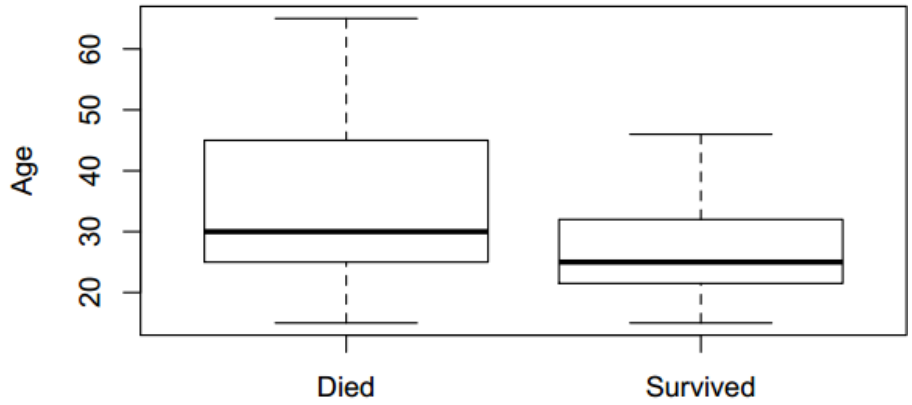
Donner Party - Data

	Age	Sex	Status
1	23.00	Male	Died
2	40.00	Female	Survived
3	40.00	Male	Survived
4	30.00	Male	Died
5	28.00	Male	Died
⋮	⋮	⋮	⋮
43	23.00	Male	Survived
44	24.00	Male	Died
45	25.00	Female	Survived

Data

	Male	Female
Died	20	5
Survived	10	10

Status vs Gender



Status vs Age

Donner Party Solution

- Get the following model after training

$$p = \frac{1}{1 + \exp(-(1.63 - 0.078 \times Age + 1.59 \times Gender))}$$

- Male model (Gender = 0)

$$p_{male} = \frac{1}{1 + \exp(-(1.63 - 0.078 \times Age))}$$

- Female model (Gender = 1)

$$p_{female} = \frac{1}{1 + \exp(-(1.63 - 0.078 \times Age + 1.59))}$$

$$= \frac{1}{1 + \exp(-(3.22 - 0.078 \times Age))}$$

Donner Party: Male and Female models