Machine Learning with R: Linear Regression

Humayun Kabir, BScN, MPH, MSc (Student)
MSc in Health Research Methodology

Department of Health Research Methods, Evidence, and Impact, McMaster University, Canada

DASH: Data Analysis Support Hub Workshop Series
Date: February 20, 2024

Lewis & Ruth Sherman Centre for Digital Scholarship
scds.ca

McMaster University | Library

McMaster University is located on the traditional Territories of the Mississauga and Haudenosaunee Nations, and within the lands protected by the “Dish With One Spoon” wampum agreement.

Laslovarga, “Webster Falls in Winter, Waterdown, Hamilton, Ontario, Canada - Spencer Gorge / Webster's Falls Conservation Area,” 23 January 2011, Wikimedia Commons - https://commons.wikimedia.org/wiki/File:Waterdawn_Webster_Falls_in_Winter8.jpg

Code of Conduct

The Sherman Centre and the McMaster University Library are committed to fostering a supportive and inclusive environment for its presenters and participants.

As a participant in this session, you agree to support and help cultivate an experience that is collaborative, respectful, and inclusive, as well as free of harassment, discrimination, and oppression. We reserve the right to remove participants who exhibit harassing, malicious, or persistently disruptive behaviour.

Please refer to our code of conduct webpage for more information:
scds.ca/events/code-of-conduct/

Certificate Program

The Sherman Centre offers a Certificate of Attendance that rewards synchronous participation at 7 workshops. We also offer concentrations in Data Analysis and Visualization, Digital Scholarship, and Research Data Management.

Learn more about the Certificate Program: https://scds.ca/certificate-program

Verify your participation at a session: https://u.mcmaster.ca/verification

At an unspecified point during the workshop, a code will be read aloud. This is the answer to the third question of the form.

DASH: Data Analysis Support Hub Workshops

Register for upcoming DASH events:

February 24: Introduction to Python – Vivek Jadon

February 27: Multivariable Analysis with R – Humayun Kabir

March 22: Machine Learning with R: Logistic Regression– Humayun Kabir

March 28: Intermediate Python Programming – Seyed Amirreza Mousavi

April 30: Survival Analysis with R – Humayun Kabir

Book an Appointment with the DASH Team

Receive help from a member of the DASH team! DASH can assist with the following topics:

- Creating data visualizations, including charts, graphs, and scatter plots
- Figuring out which statistical tests to run (e.g., t-test, chi-square, etc.).
- Analyzing data with software including SPSS, Python, R, SAS, ArcGIS, MATLAB, and Excel
- Choosing which software package to use, including free and open-source software
- Troubleshooting problems related to file formats, data retrieval, and download
- Selecting methodology and type of data analysis to use in a thesis project

Book an appointment: https://library.mcmaster.ca/services/dash

Session Recording and Privacy

This session is being recorded with the intention of being shared publicly via the web for future audiences. In respect of your privacy, participant lists will not be shared outside of this session, nor will question or chat transcripts.

Questions asked via the chat box will be read by the facilitator without identifying you. Note that you may be identifiable when asking a question during the session in an audio or visual format.

Machine Learning with R: Linear Regression

Objective

LEARNING THE BASICS OF LINEAR REGRESSION

MACHINE LEARNING WITH R

Covariance

$$\operatorname{cov}(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y})}{n-1}$$

Source: Dr. Dipak Kumar Mitra, North South University

Interpreting Covariance

$cov(X,Y) > 0 \longrightarrow$ X and Y are positively correlated

$cov(X,Y) < 0 \longrightarrow$ X and Y are inversely correlated

$cov(X,Y) = 0 \longrightarrow$ X and Y are independent

Source: Dr. Dipak Kumar Mitra, North South University

Correlation coefficient (standardized)

- Pearson's Correlation Coefficient is standardized covariance (unitless):

$$r = \frac{\operatorname{cov}ariance(x, y)}{\sqrt{\operatorname{var} x}\sqrt{\operatorname{var} y}}$$

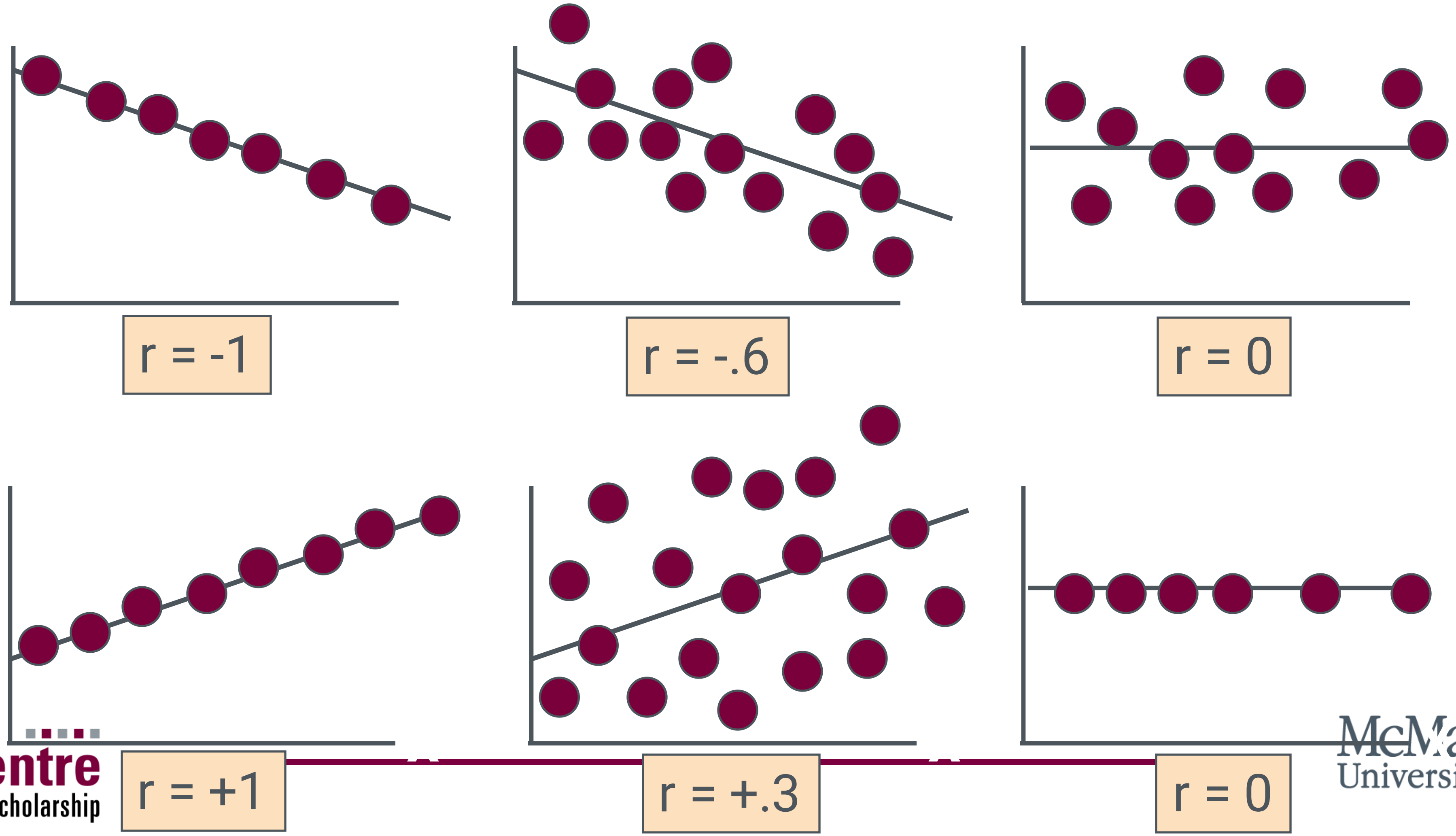
Source: Dr. Dipak Kumar Mitra, North South University

Correlation

- Measures the relative strength of the *linear* relationship between two variables
- Unit-less
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Source: Dr. Dipak Kumar Mitra, North South University

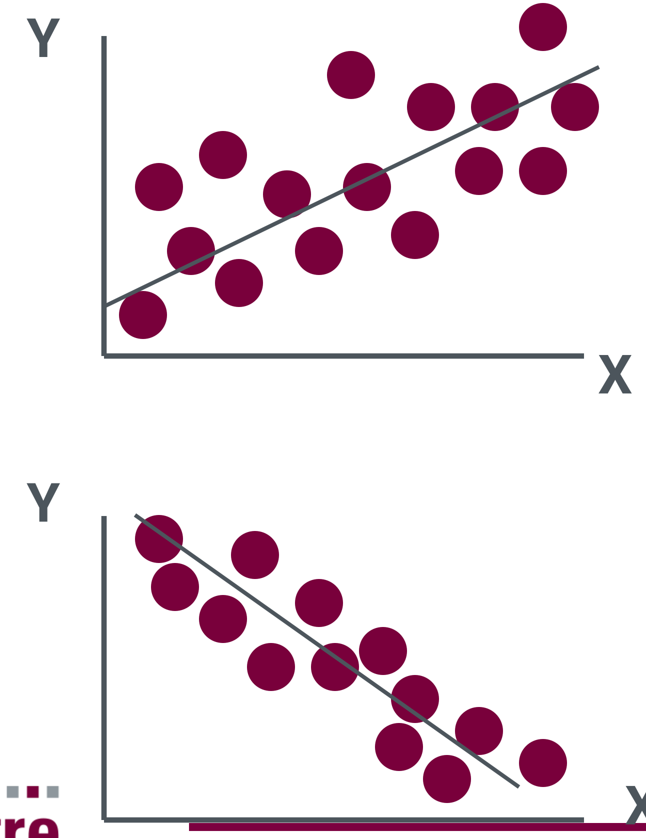
Scatter Plots of Data with Various Correlation Coefficients



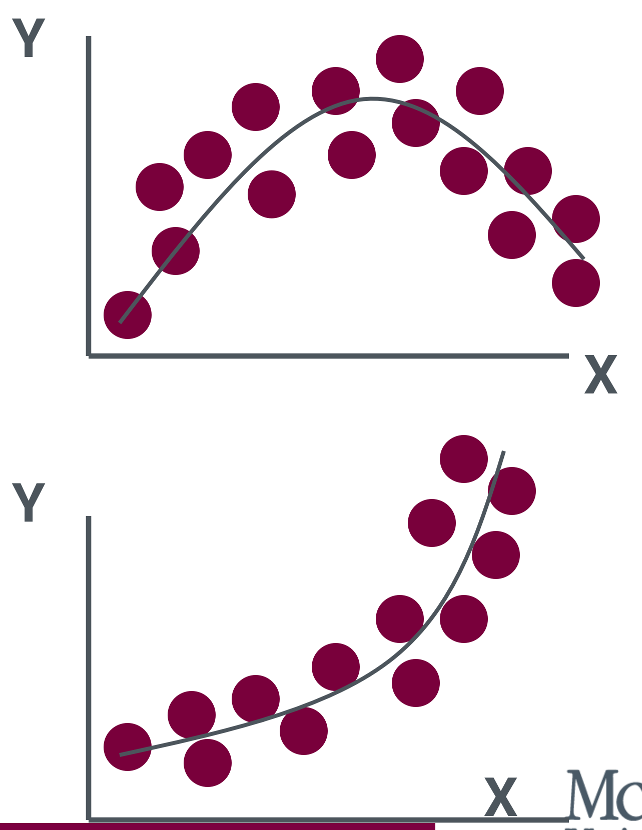
Source: Dr. Dipak Kumar Mitra, North South University

Linear Correlation

Linear relationships



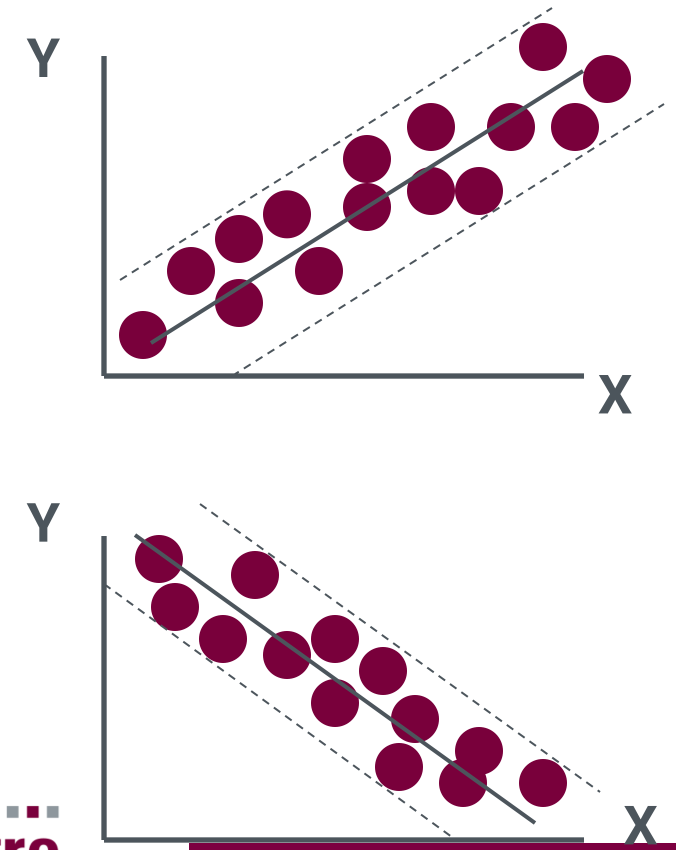
Curvilinear relationships



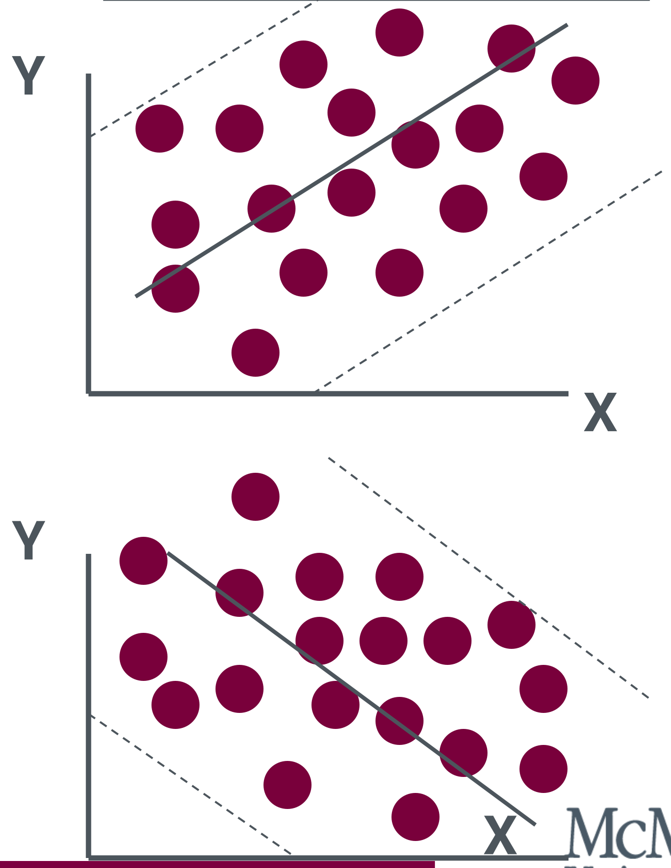
Source: Dr. Dipak Kumar Mitra, North South University

Linear Correlation

Strong relationships

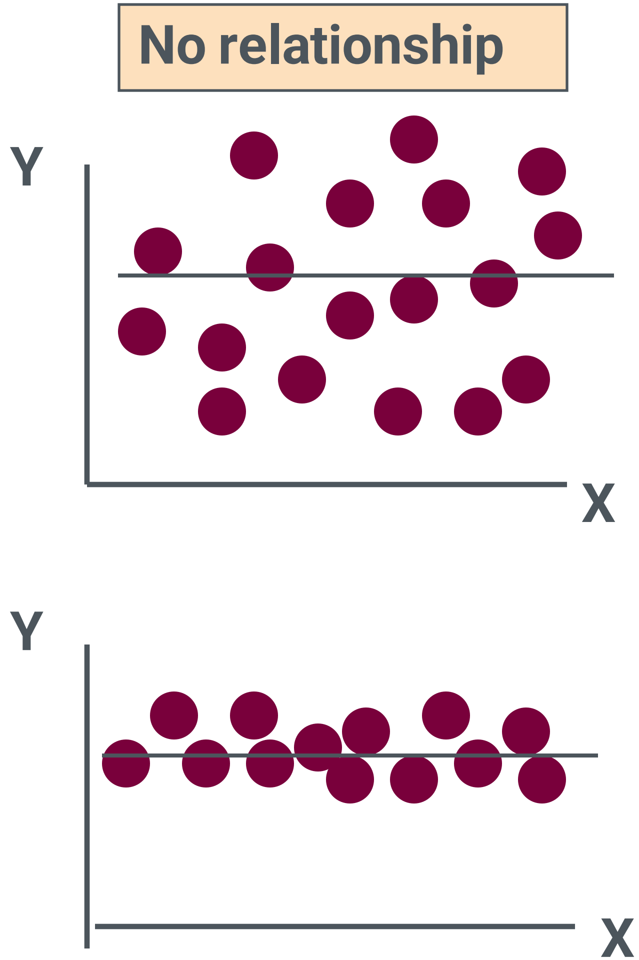


Weak relationships



Source: Dr. Dipak Kumar Mitra, North South University

Linear Correlation



Lewis & Ruth Sherman Centre for Digital Scholarship
scds.ca

Source: Dr. Dipak Kumar Mitra, North South University

McMaster University Library

Calculating by hand…

$$\hat{r} = \frac{\mathrm{cov}\,ariance(x,y)}{\sqrt{\mathrm{var}\ x}\sqrt{\mathrm{var}\ y}} = \frac{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}}$$

Source: Dr. Dipak Kumar Mitra, North South University

Simpler calculation formula…

$$\hat{r} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}} =$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerator of covariance

$$\hat{r} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerators of variance

Standard error of the correlation coefficient:

$$SE(\hat{r}) = \sqrt{\frac{1 - r^2}{n - 2}}$$

The sample correlation coefficient follows a t-distribution with n-2 degrees of freedom (since you have to estimate the standard error).

*note, like a proportion, the variance of the correlation coefficient depends on the correlation coefficient itself→substitute in estimated r

Linear regression

- 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*

Source: Dr. Dipak Kumar Mitra, North South University

Prediction

- If you know something about X, this knowledge helps you predict something about Y
- Sound familiar?…sound like conditional probabilities?

Source: Dr. Dipak Kumar Mitra, North South University

What’s Slope?

A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y.

Source: Dr. Dipak Kumar Mitra, North South University

Regression equation…

Expected value of y at a given level of *x*=

$$E(y_i / x_i) = \alpha + \beta x_i$$

Predicted value for an individual…

$$y_i = \underbrace{\alpha + \beta * x_i}_{\text{Fixed – exactly on the line}} + \boxed{\text{random error}_i}$$

Follows a normal distribution

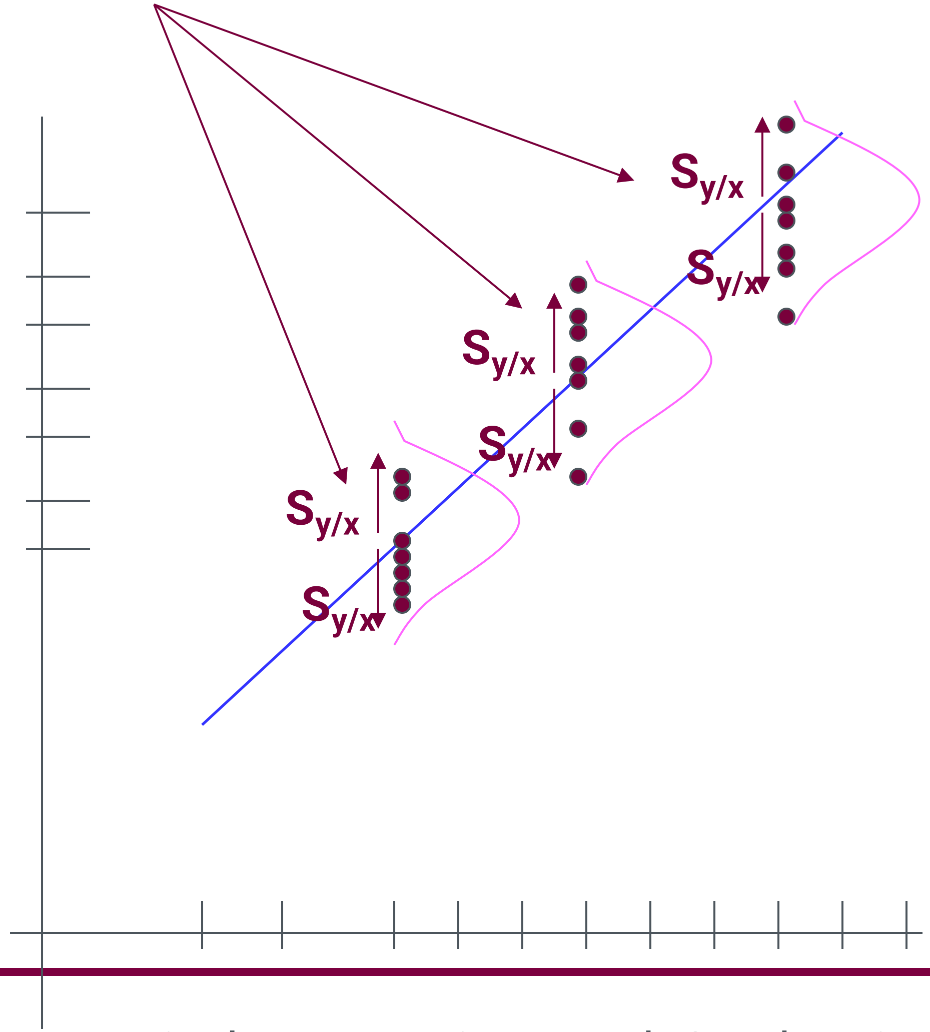
Source: Dr. Dipak Kumar Mitra, North South University

Assumptions (or the fine print)

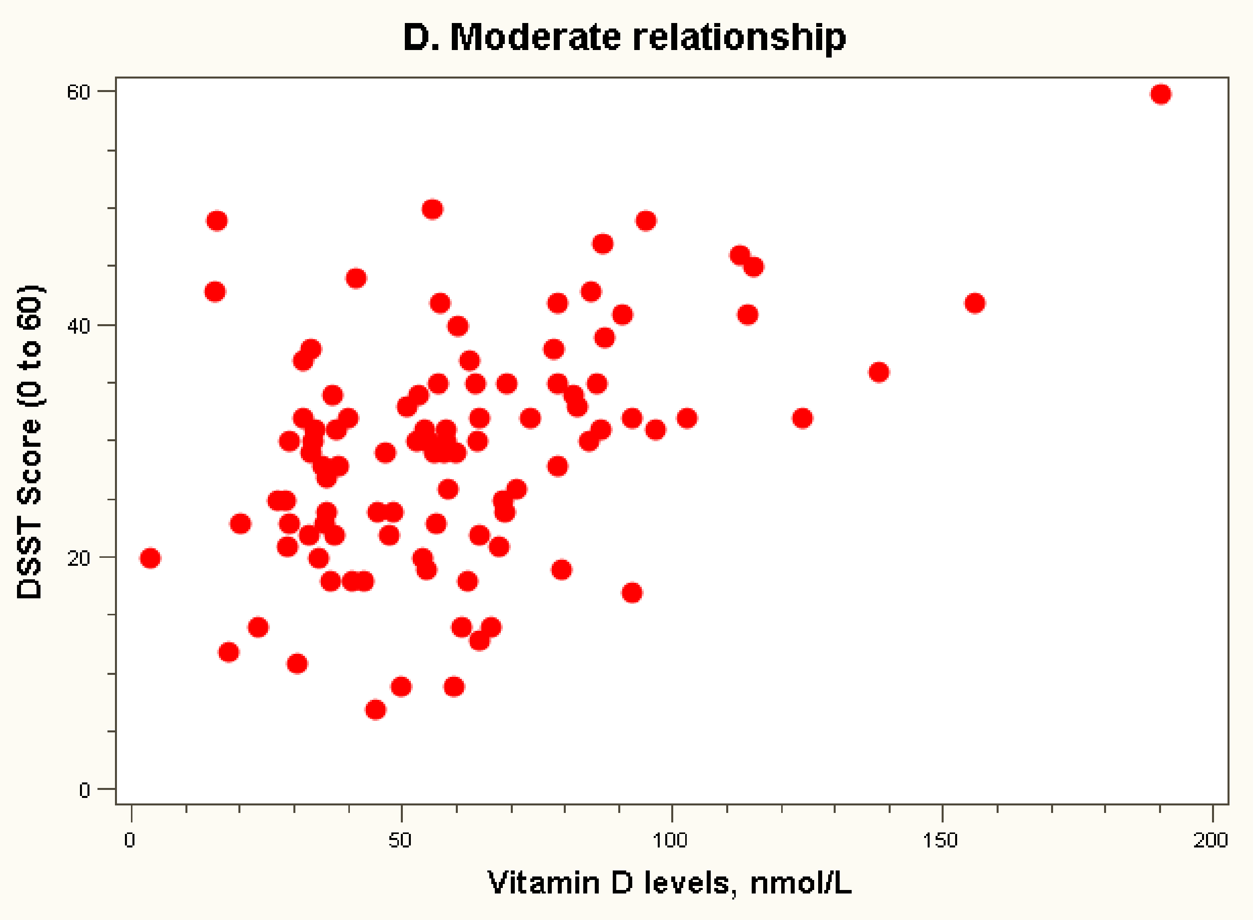
- Linear regression assumes that…
  1. Linearity: True mean of y is a linear function of x
  2. Y is distributed normally at each value of X
  3. The variance of Y at every value of X is constant (homogeneity of variances) (in next slide)
  4. The observations are independent

Source: Dr. Dipak Kumar Mitra, North South University

The standard error of Y given X is the average variability around the regression line. Variance is assumed to be equal at all values of X.

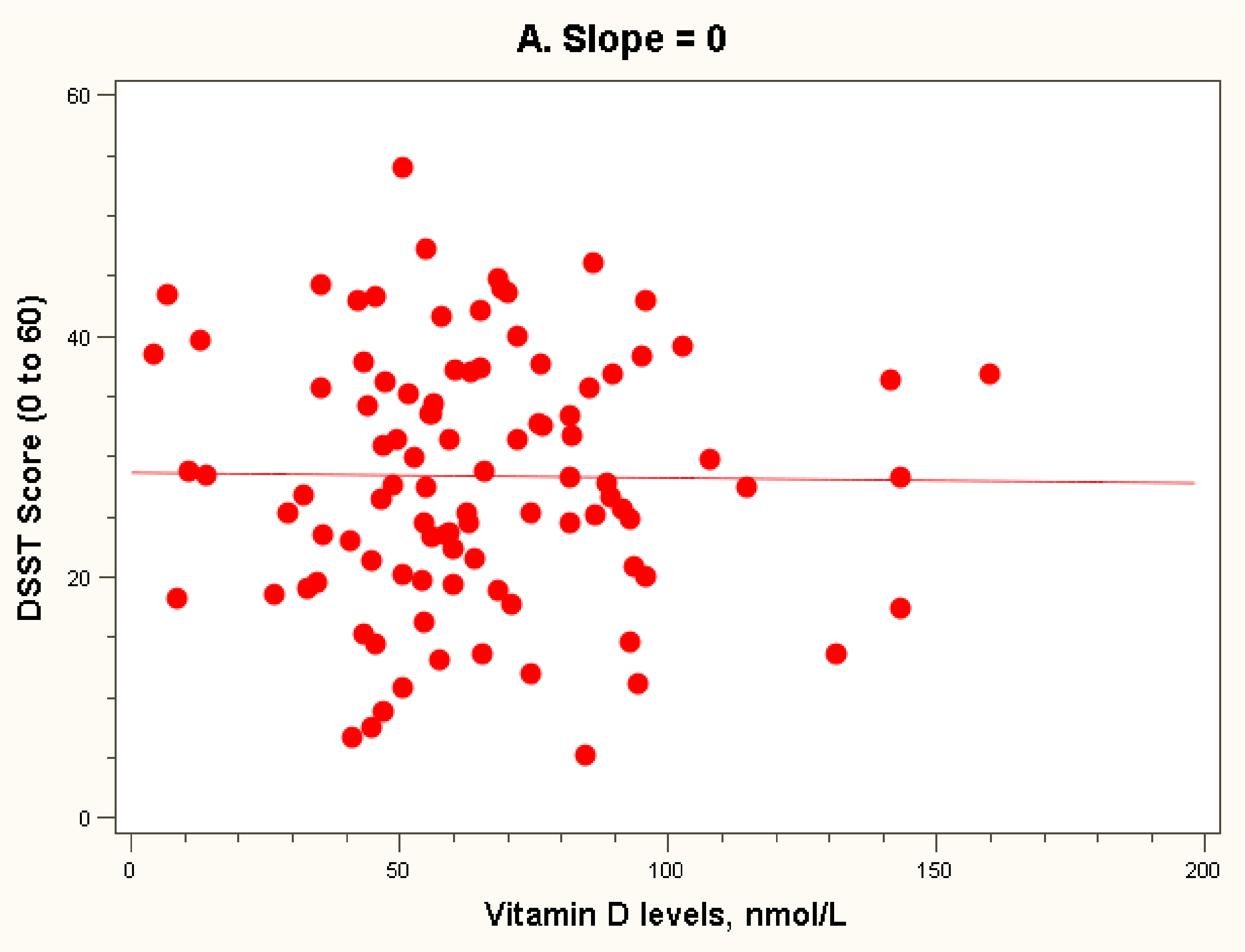


Moderate relationship



Source: Dr. Dipak Kumar Mitra, North South University

The “Best fit” line

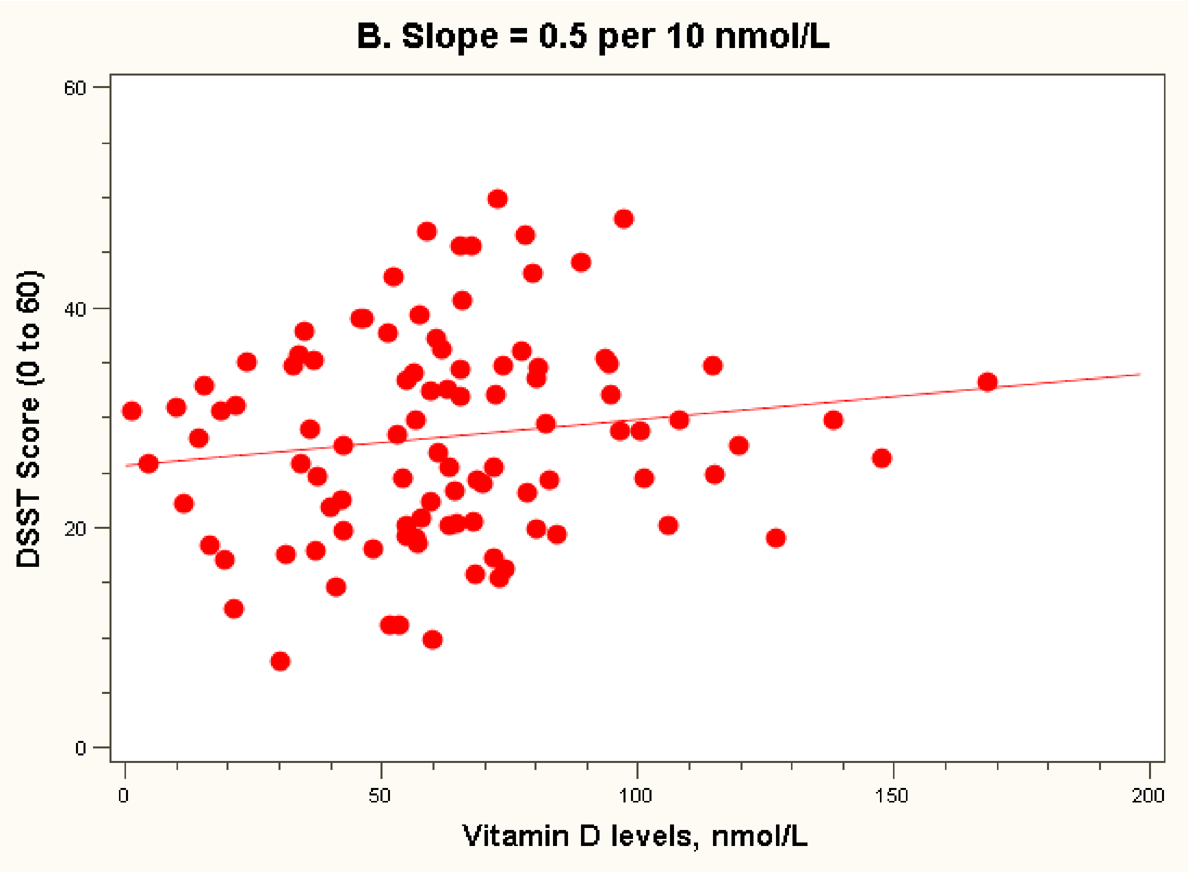


Regression equation:

$E(Y_i) = 28 + 0 * \text{vit } D_i$ (in 10 nmol/L)

Source: Dr. Dipak Kumar Mitra, North South University

The "Best fit" line



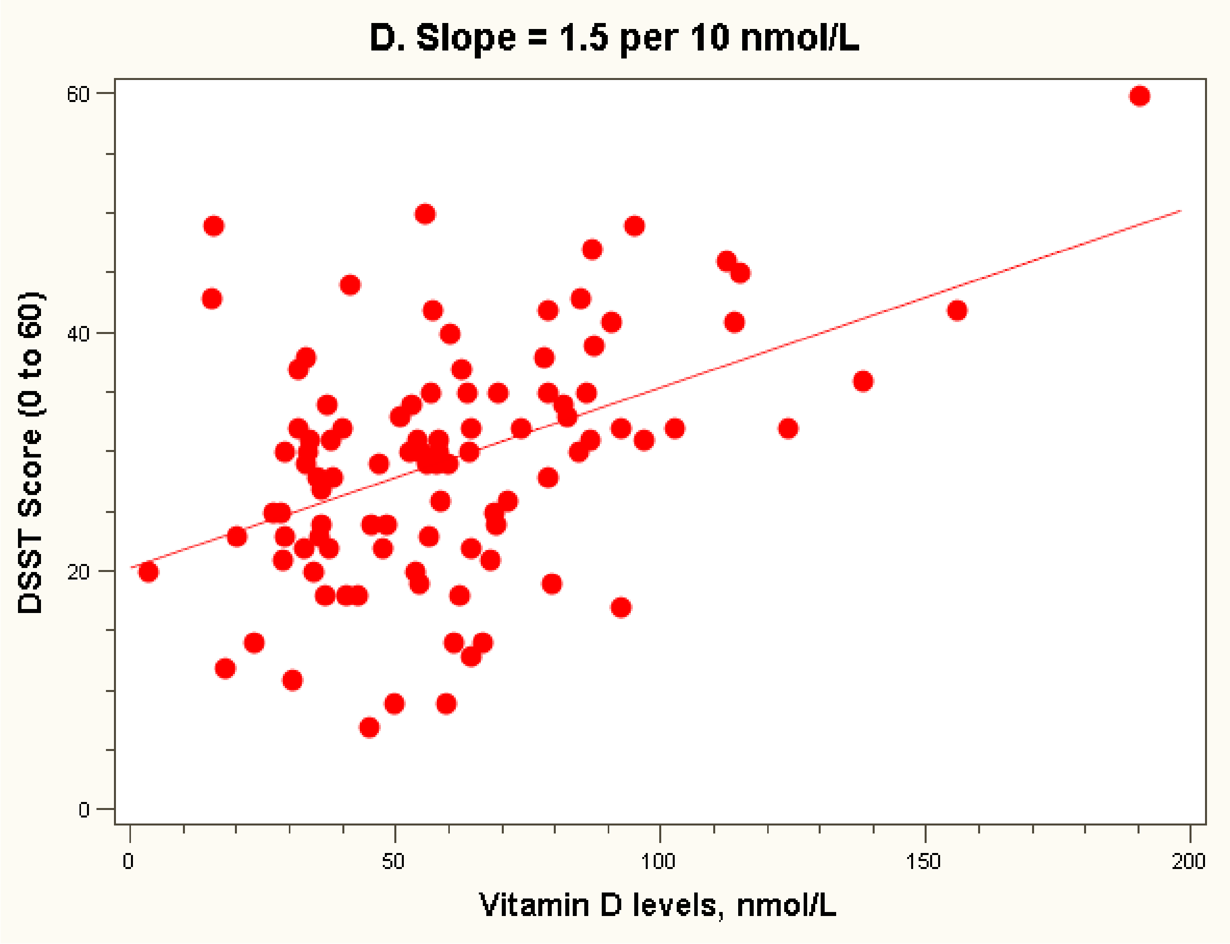
Note how the line is a little deceptive; it draws your eye, making the relationship appear stronger than it really is!

Regression equation:

$E(Y_i) = 26 + 0.5*\text{vitD}_i$ (in 10 nmol/L)

Source: Dr. Dipak Kumar Mitra, North South University

The "Best fit" line



Regression equation:

$E(Y_i) = 20 + 1.5*vit D_i$ (in 10 nmol/L)

Note: all the lines go through the point (63, 28)!

Statistical Concepts

- Based on ıtimate the model

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Since $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
- We can re-write the equation like $\hat{Y} = \bar{Y} + \hat{\beta}_1 (X - \bar{X})$
- So the regression line goes through the point $(\bar{X}, \bar{Y})$

Source: Dr. Shofiqul Islam, McMaster University

Correlation and Regression Slope

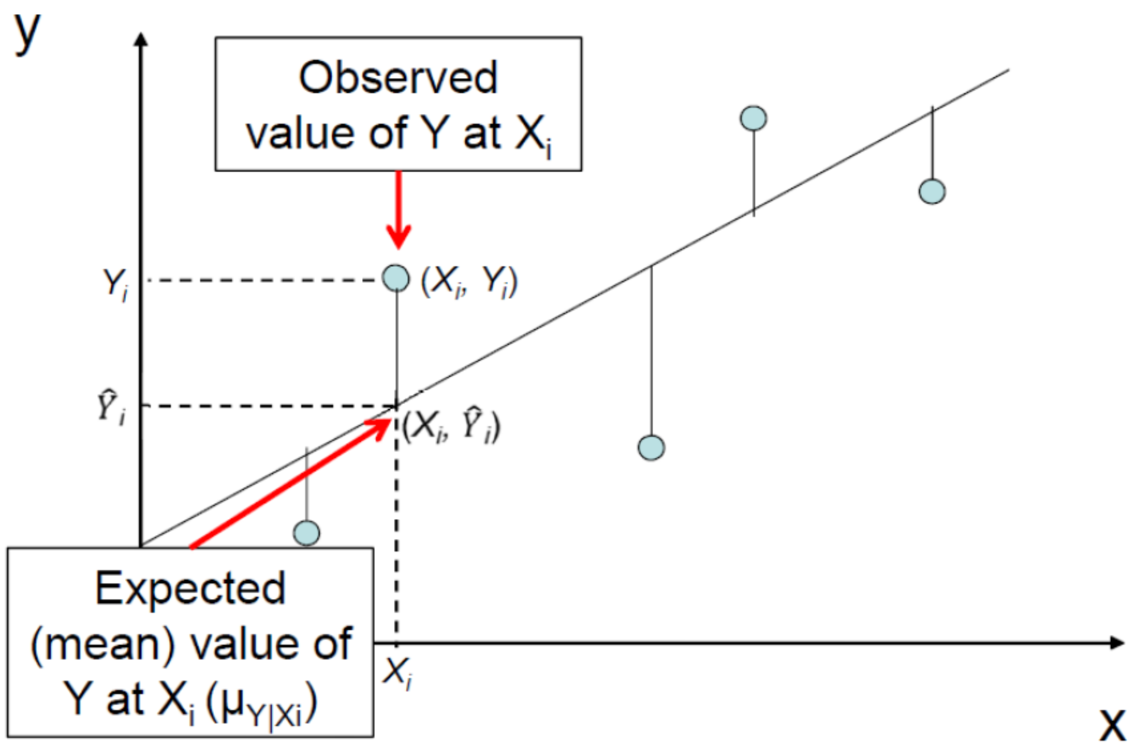
- Relationship between Pearson product-moment correlation and the slope of the simple linear regression line?
- Pearson's correlation coefficient measures the amount of linear association between Y and X
  - Non-directional –one 'is associated' with the other

$$r = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$$

- The slope of the simple linear regression line tells how much a change in X impacts a change in Y
  - Directional –one 'predicts' the other

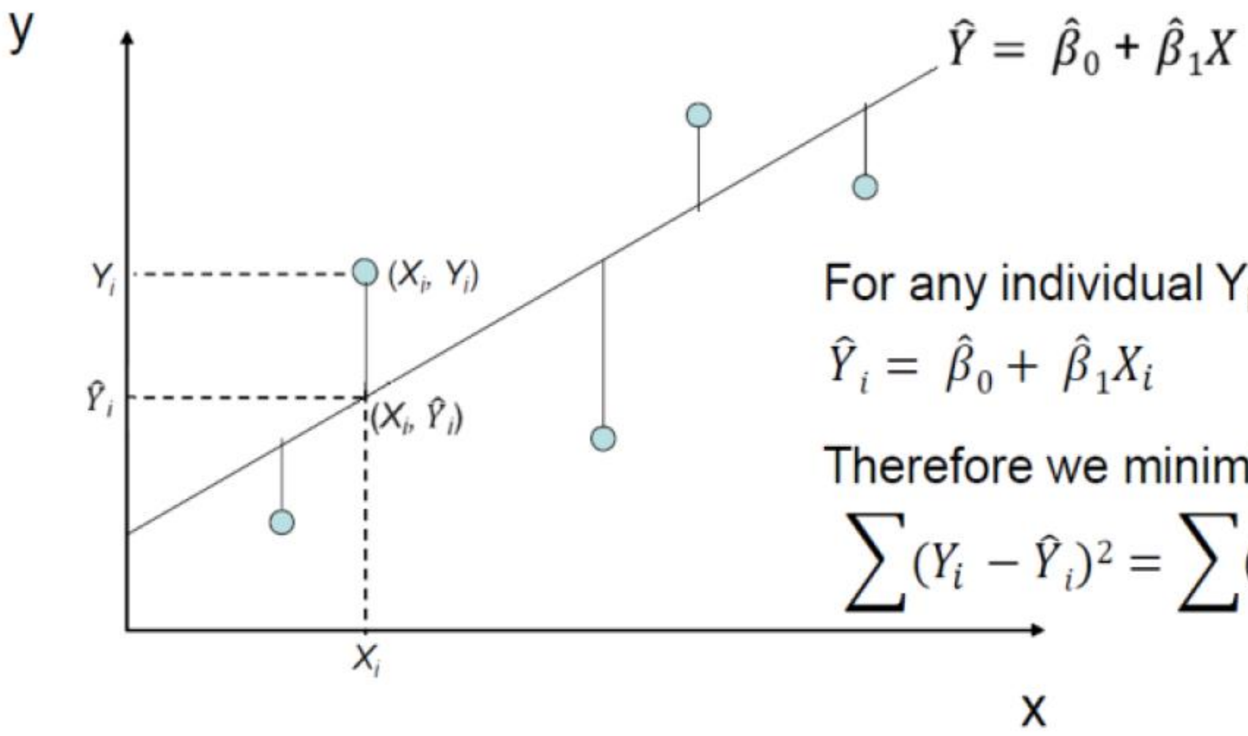
$$\hat{\beta}_1 = r\frac{S_Y}{S_X}$$

Geometry



Source: Dr. Shofiqul Islam, McMaster University

Geometry of Least Square



For any individual Y_i,

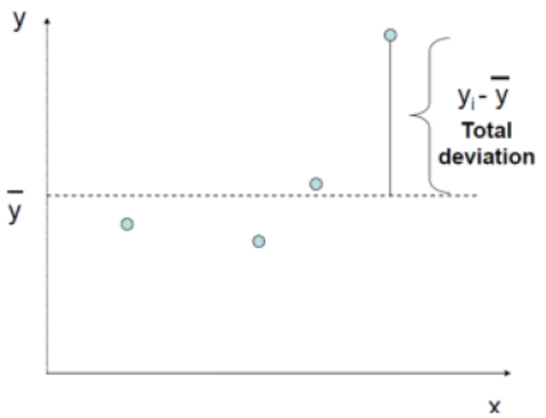
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Therefore we minimize:

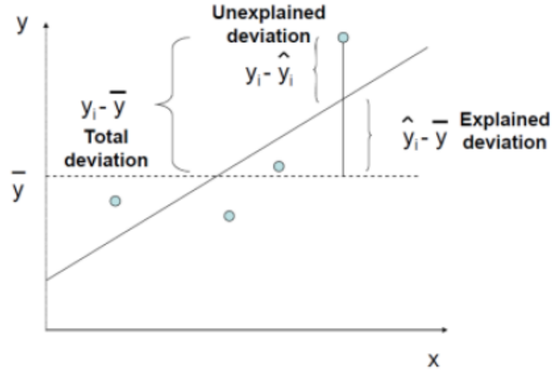
$$\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

Source: Dr. Shofiqul Islam, McMaster University

Variance decomposition



$$Var\ Y = \frac{\Sigma(y_i - \bar{y})^2}{n-1}$$

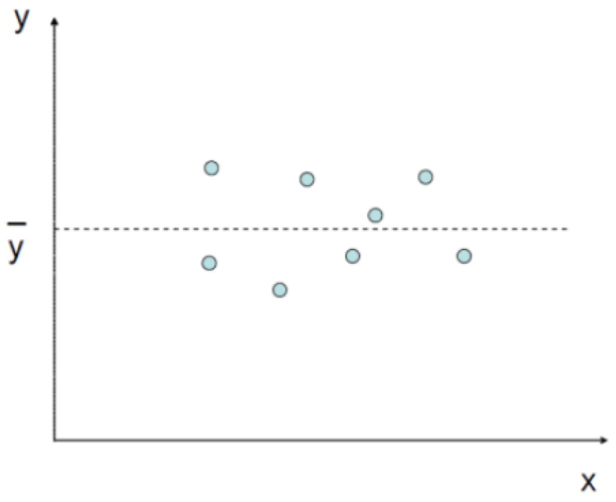


$$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2$$

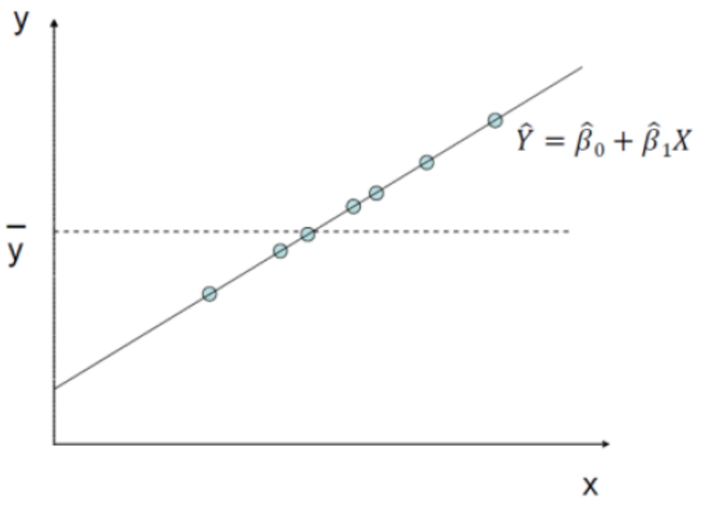
Total SS Residual SS Regression SS

Source: Dr. Shofiqul Islam, McMaster University

Extreme Results



$$SS_{Total} \cong SS_{Error}$$



$$SS_{Total} \cong SS_{Reg}$$

Source: Dr. Shofiqul Islam, McMaster University

ANOVA Table

of predictors (x variables)=1

	Sum of Squares SS	Degrees of Freedom df	Mean Square MS	F Value F
Regression	SS_{reg}	p	SS_{reg}/p	MS_{reg}/MS_{err}
Residual	SS_{err}	n-p-1	SS_{err} /(n-p-1)	
Total	SS_{tot}	n-1		

# of observations - # of parameters estimated

# of observation - 1

Where:

$SS_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

$SS_{err} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

$SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

$E(Ms_{err}) = \sigma^2$

$E(MS_{reg}) = \sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$

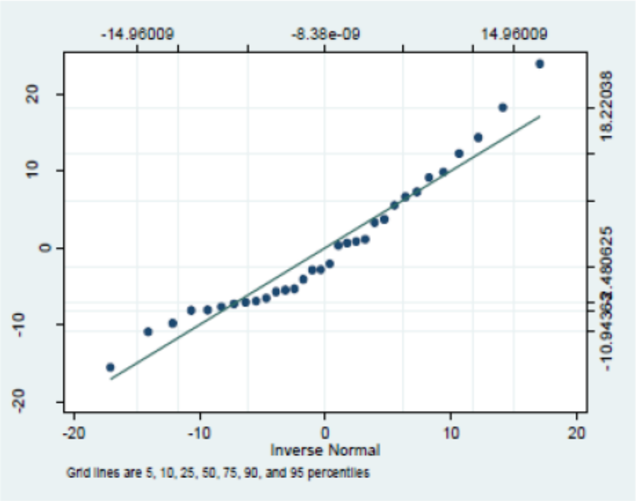
If H_o is true, i.e., $\beta_1 = 0$, then MS_{reg} is also an estimate of σ^2

Source: Dr. Shofiqul Islam, McMaster University

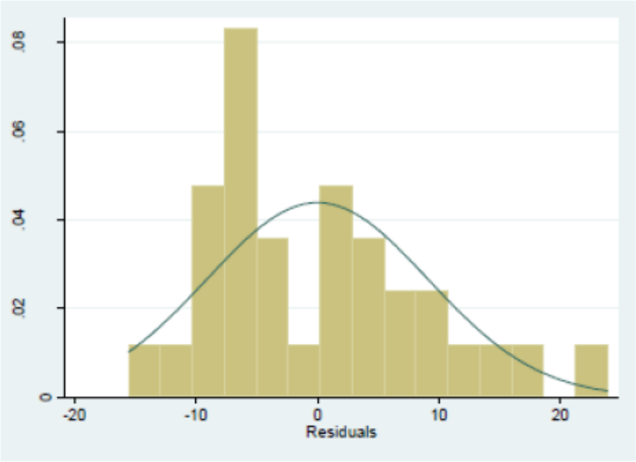
Assessing 'goodness of fit' assumptions

Normality – of residuals or of Y?

- Q-Q plot, histogram
- Shapiro-Wilk test



. qnorm residuals, grid



. histogram residuals, normal bin(15)

Source: Dr. Shofiqul Islam, McMaster University

Multiple linear regression…

- What if age is a confounder here?
  - Older men have lower vitamin D
  - Older men have poorer cognition
- “Adjust” for age by putting age in the model:
  - DSST score = intercept + $slope_1$ *x* vitamin D + $slope_2$ *x* age

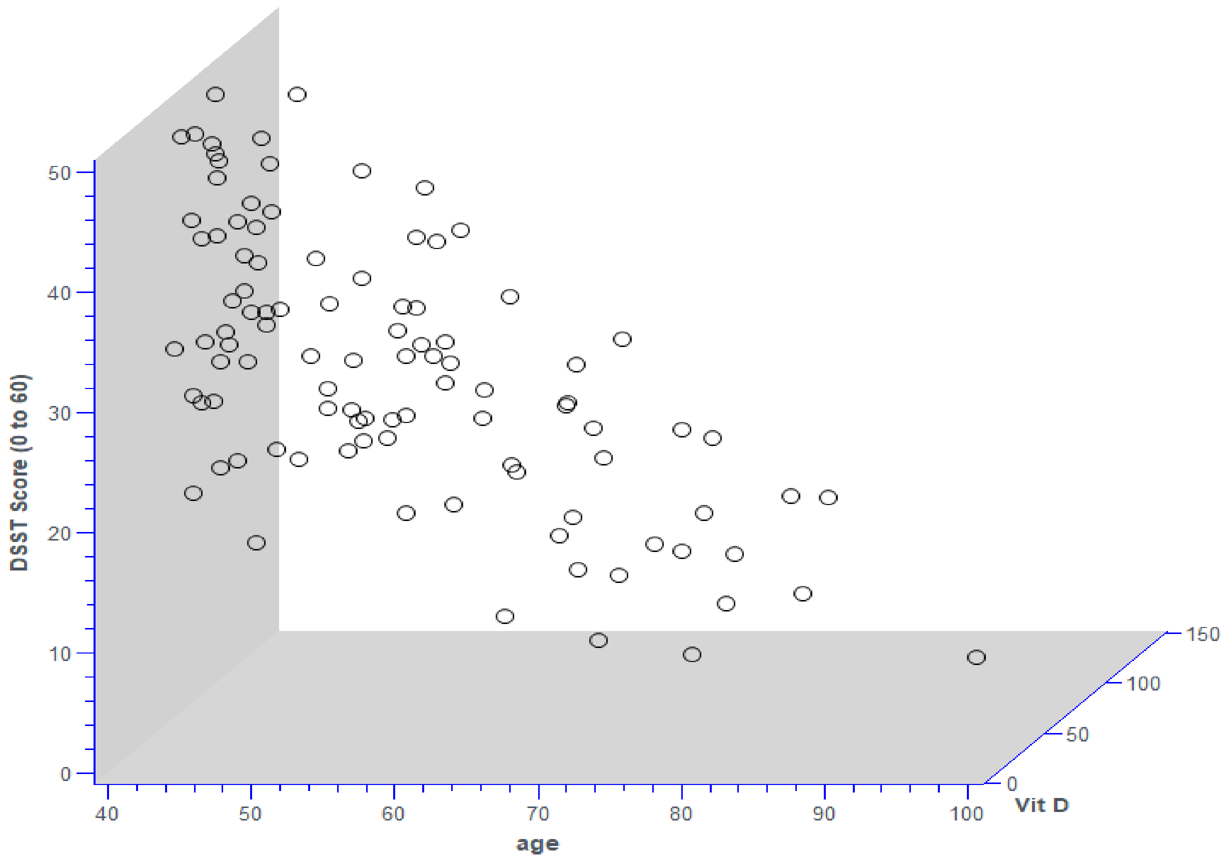
Multiple Linear Regression

- More than one predictor…

$$E(y) = \alpha + \beta_1 * X + \beta_2 * W + \beta_3 * Z \ldots$$

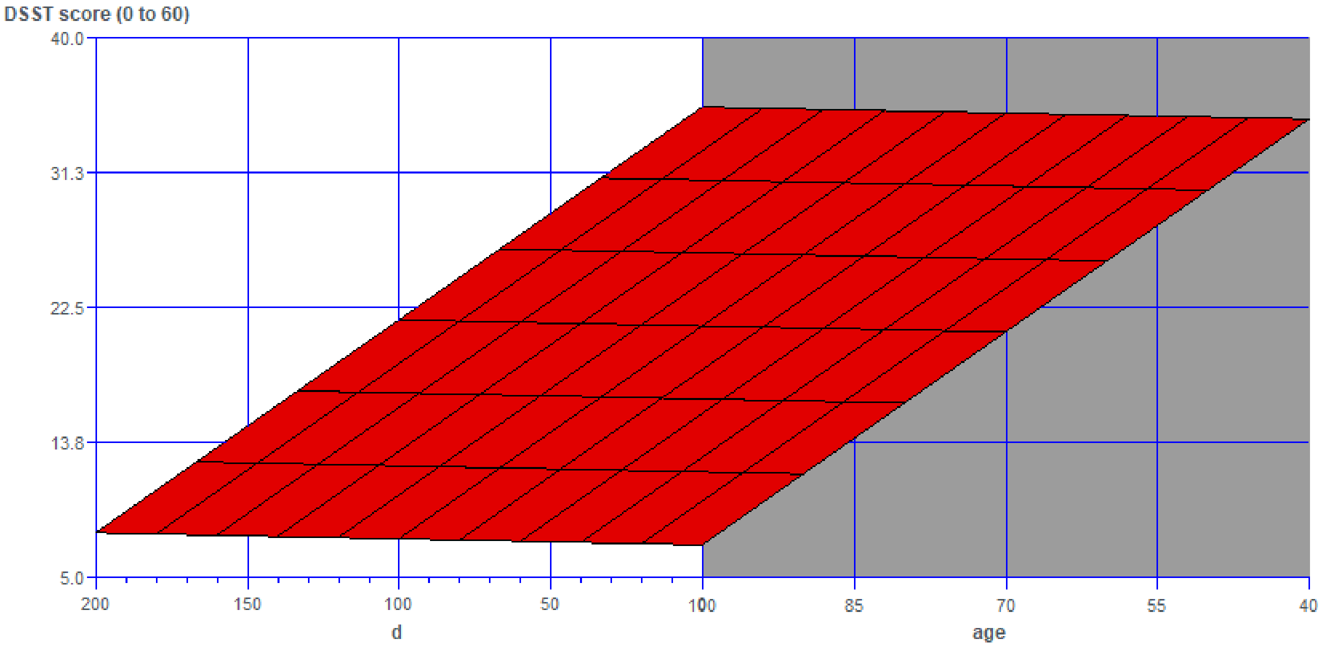
Each regression coefficient is the amount of change in the outcome variable that would be expected per one-unit change of the predictor, if all other variables in the model were held constant.

Different 3D view…



Source: Dr. Dipak Kumar Mitra, North South University

Fit a plane rather than a line…



On the plane, the slope for vitamin D is the same at every age; thus, the slope for vitamin D represents the effect of vitamin D when age is held constant.

Source: Dr. Dipak Kumar Mitra, North South University

Machine learning

Machine learning involves showing a large volume of data to a machine so that it can learn and make predictions, find patterns, or classify data.

Source: coursera.org

Types of machine learning

Basically, machine learning are three types.

Supervised

Unsupervised

Reinforcement learning.

Source: coursera.org

Supervised learning

*Machine learning **feeds historical input and output data** in machine learning algorithms, with processing in between each input/output pair that allows the algorithm to shift the model **to create outputs as closely aligned with the desired result as possible**.*

Common algorithms used during supervised learning include linear regression, neural networks, decision trees, and support vector machines.

Source: coursera.org

Unsupervised learning

*While supervised learning requires users to help the machine learn, unsupervised **learning algorithms don't use the same labeled training sets and data**. Instead, the machine looks for less obvious patterns in the data.*

*Unsupervised machine learning is very helpful when you need **to identify patterns** and use data to make decisions.*

Common algorithms used in unsupervised learning include k-means, hierarchical clustering, and Gaussian mixture models.

Source: coursera.org

Reinforcement learning

Reinforcement learning is the closest machine learning type to how humans learn.

*The algorithm used **learns by interacting with its environment and getting a positive or negative reward**.*

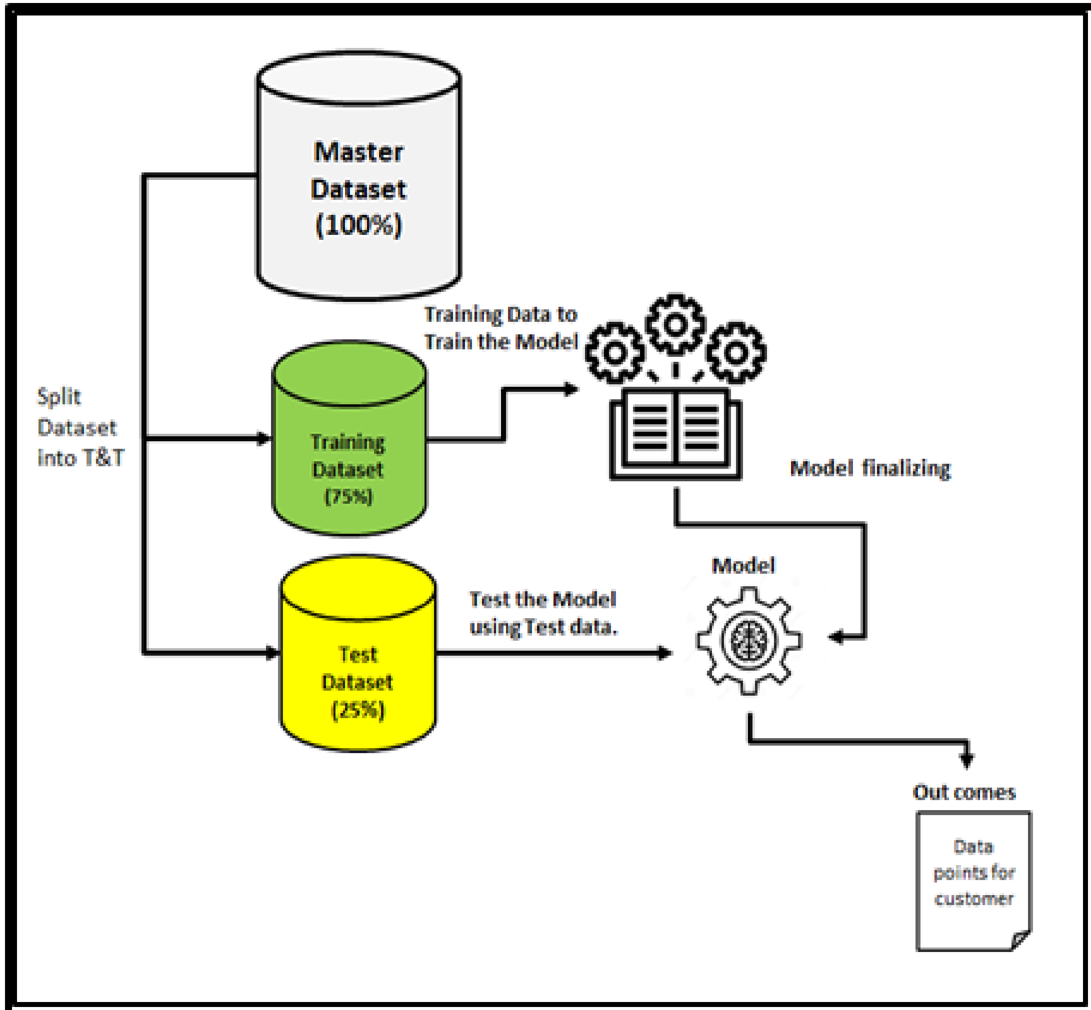
Common algorithms include temporal difference, deep adversarial networks, and Q-learning.

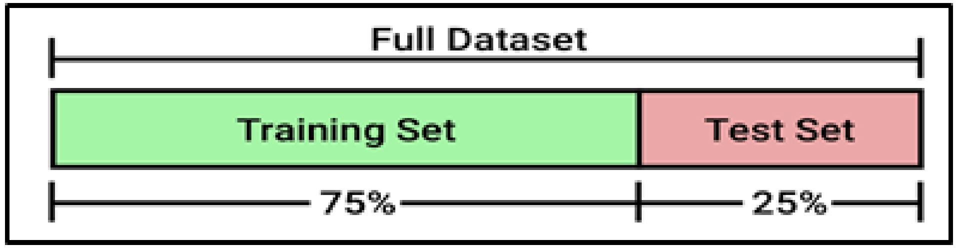
Source: coursera.org

Linear regression

- A supervised machine learning
- Learns from labeled data
- Make predictions on unseen data
- Goal of minimizing prediction errors

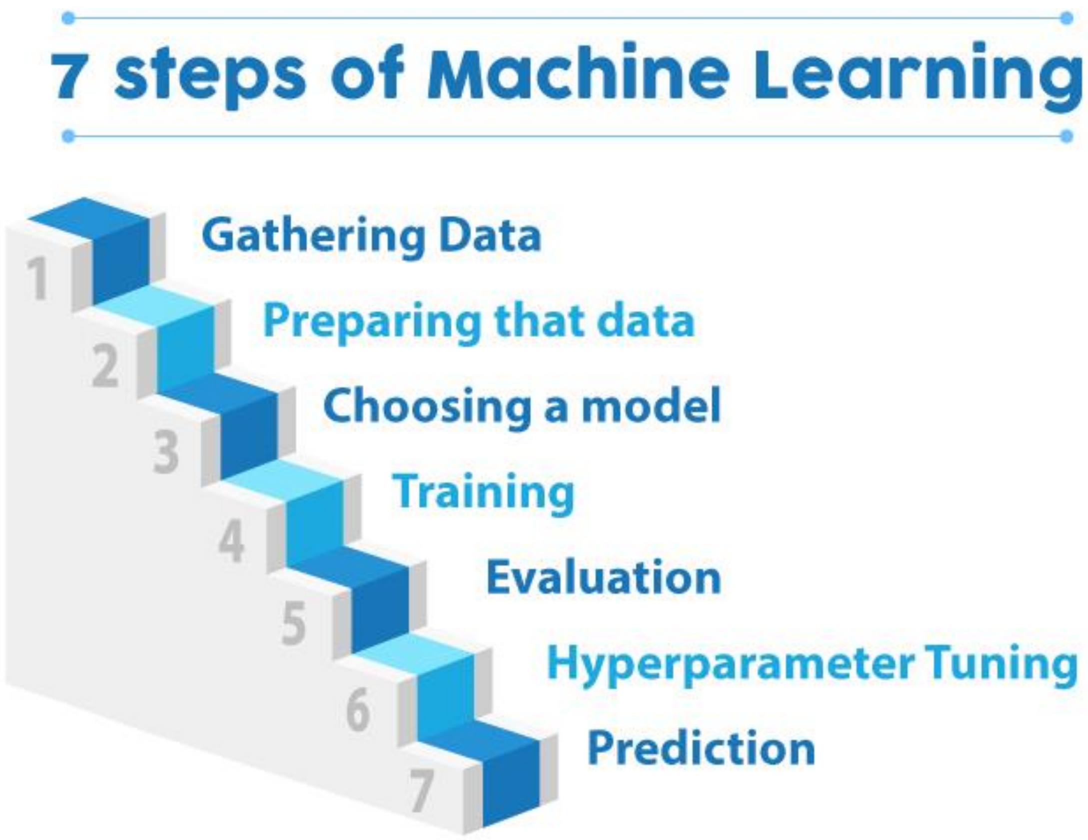
Steps of ML





https://www.analyticsvidhya.com/

Steps of ML including tuning



www.mygreatlearning.com

Contact

Book an appointment with DASH: https://library.mcmaster.ca/services/dash

Contact DASH: Data Analysis Support Hub: libdash@mcmaster.ca

Let move to the coding part

https://colab.research.google.com/drive/1PKxey0_YzdSrc_CcRPwh_Cd3JDglOi74#scrollTo=B_48FOMnQ5Yg