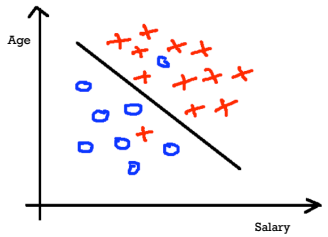
CHULA ENGINEERING
Foundation toward Innovation

COMPUTER



Linear Regression

2110498: AI for Engineers

Peerapon Vateekul, Ph.D.

Department of Computer Engineering,
Faculty of Engineering, Chulalongkorn University
Peerapon.v@chula.ac.th

Outlines

- Introduction
- Simple Linear Regression
- Multiple Linear Regression
- Other topics
- Demo

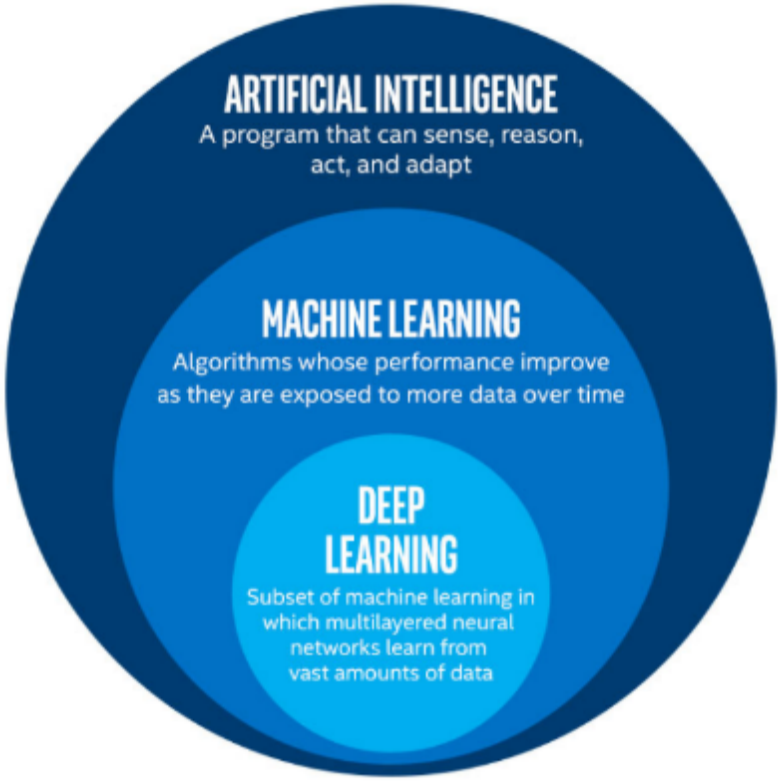
Introduction

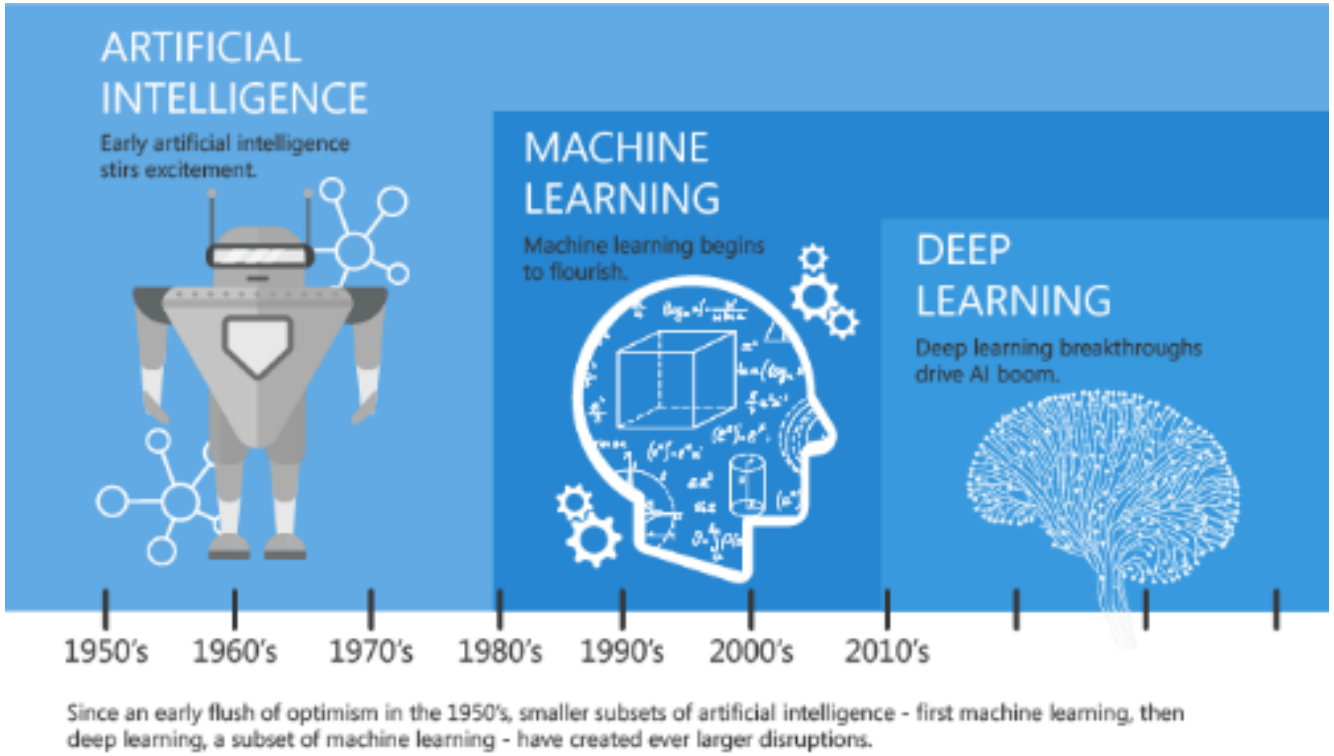
AI, Machine Learning, and Deep Learning

- **Machine Learning (ML)** is a subfield in AI focusing on making to learn by itself **without human intervention**.

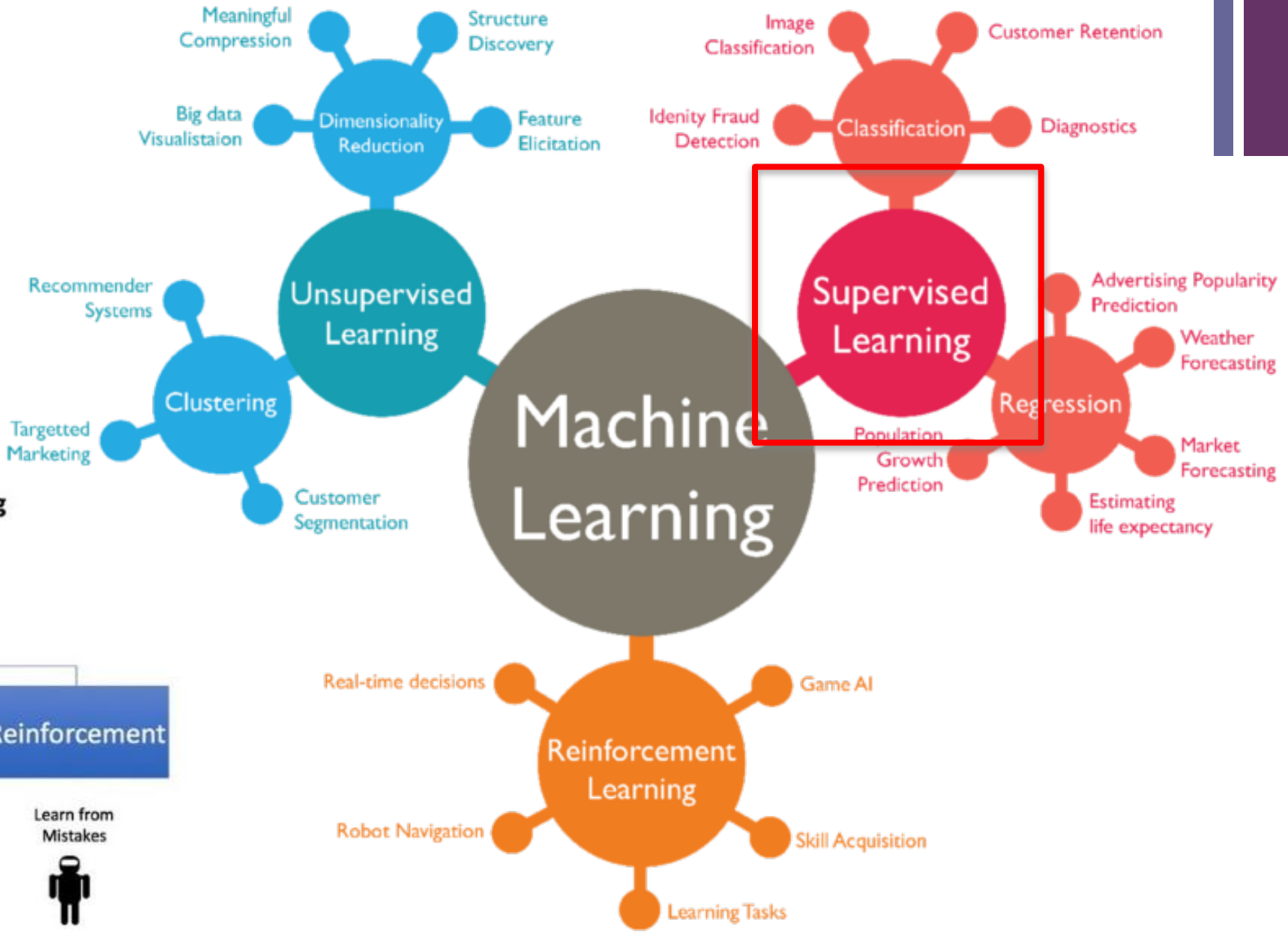
"Machine learning is the sc*ience of getting computers to act without being explicitly programmed*." — *Stanford University*

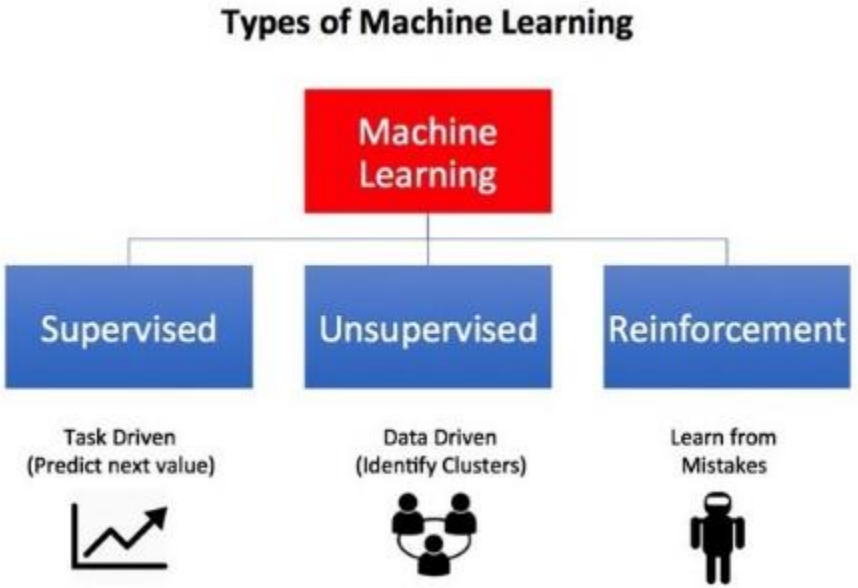
https://towardsdatascience.com/cousins-of-artificial-intelligence-dda4edc27b55





Machine Learning (cont.)



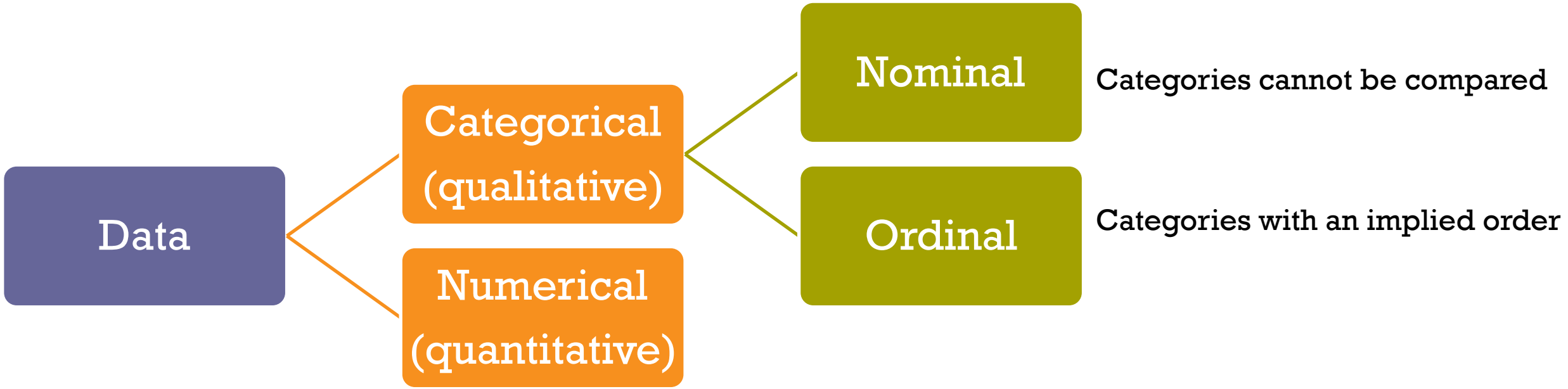


Terminology: Data table

		inputs		target
Age	**Income**	**Gender**	**Province**	**Purchase**
25	25,000	Female	Bangkok	Yes
35	50,000	Female	Nontaburi	Yes
32	35,000	Male	Bangkok	No

- Row
  - Example, instance, case, observation, subject
- Column
  - Feature, variable, attribute
- Input
  - Predictor, independent, explanatory variable
- Target
  - Output, outcome, response, dependent variable

Terminology: Kinds of data



Supervised Learning

Techniques for learning from target

Predictive model: a concise representation of the input and target association

SaleAmount = 0.7 + 0.9*Age + 0.5*Income

SaleAmount = **f**(Age, Income)

Training dataset:

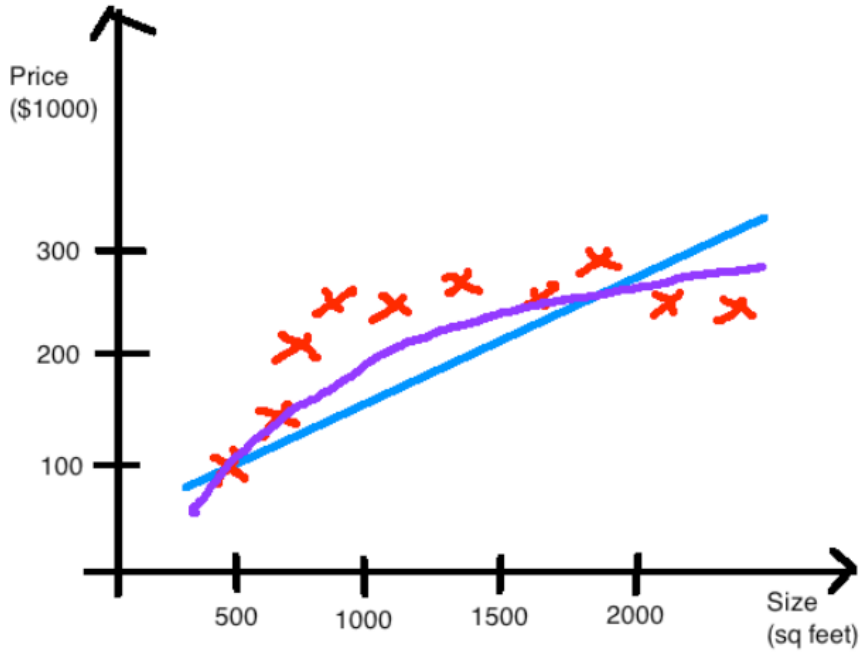
inputs	target
57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	1
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0	0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0	0
40,M,205,0,115,90,37,18,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1	0

Test dataset:

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0 ?

Regression: predict a continuous value

Linear Regression

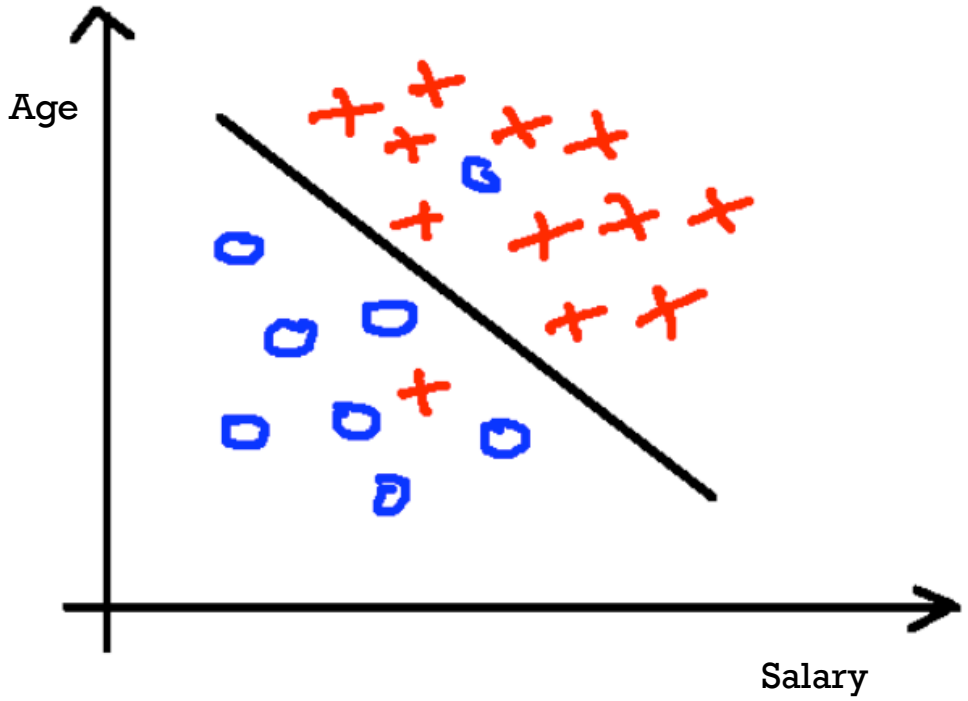


Predict a sale price of each house

- **Some techniques:**
  - Linear Regression / GLM
  - Decision Trees
  - Support vector regression
  - Neural Network
  - Ensembles
- **Sample Applications**
  - Financial risk management
  - Revenue forecasting

Classification: predicting a category

Logistic Regression



Predict targeted customers who tend to buy our product (yes/no)

- **Some techniques:**
  - Naïve Bayes
  - Decision Tree
  - Logistic Regression
  - Support Vector Machines
  - Neural Network
  - Ensembles
- **Sample Applications**
  - Database marketing
  - Fraud detection
  - Pattern detection
  - Churn customer detection

Prediction algorithms

- Decision Tree
- (Logistic) Regression
- kNN
- Support Vector Machine
- Neural Networks (NN)
- Deep Learning

Supervised learning (recap)

Training Data

	inputs			target
Age	**Income**	**Gender**	**Province**	**Churn**
25	25,000	Female	Bangkok	Yes
35	50,000	Female	Nontaburi	Yes
32	35,000	Male	Bangkok	No

Testing Data

Age	**Income**	**Gender**	**Province**	**Churn**
25	25,000	Female	Bangkok	?

Application: Direct Target Customer

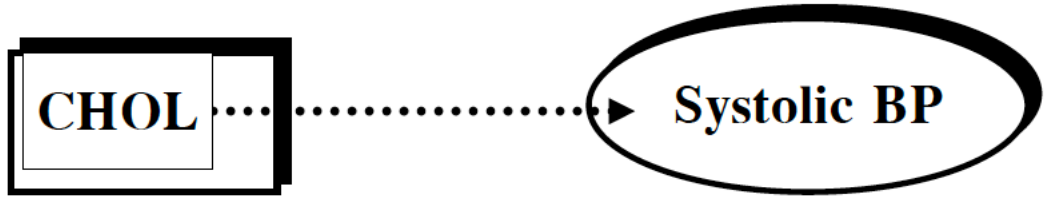
+

Simple Linear Regression

Problem: 1 input (predictor) & 1 output

- Collect data of 7 patients
- Systolic Blood Pressure (y) & Cholesterol (x)

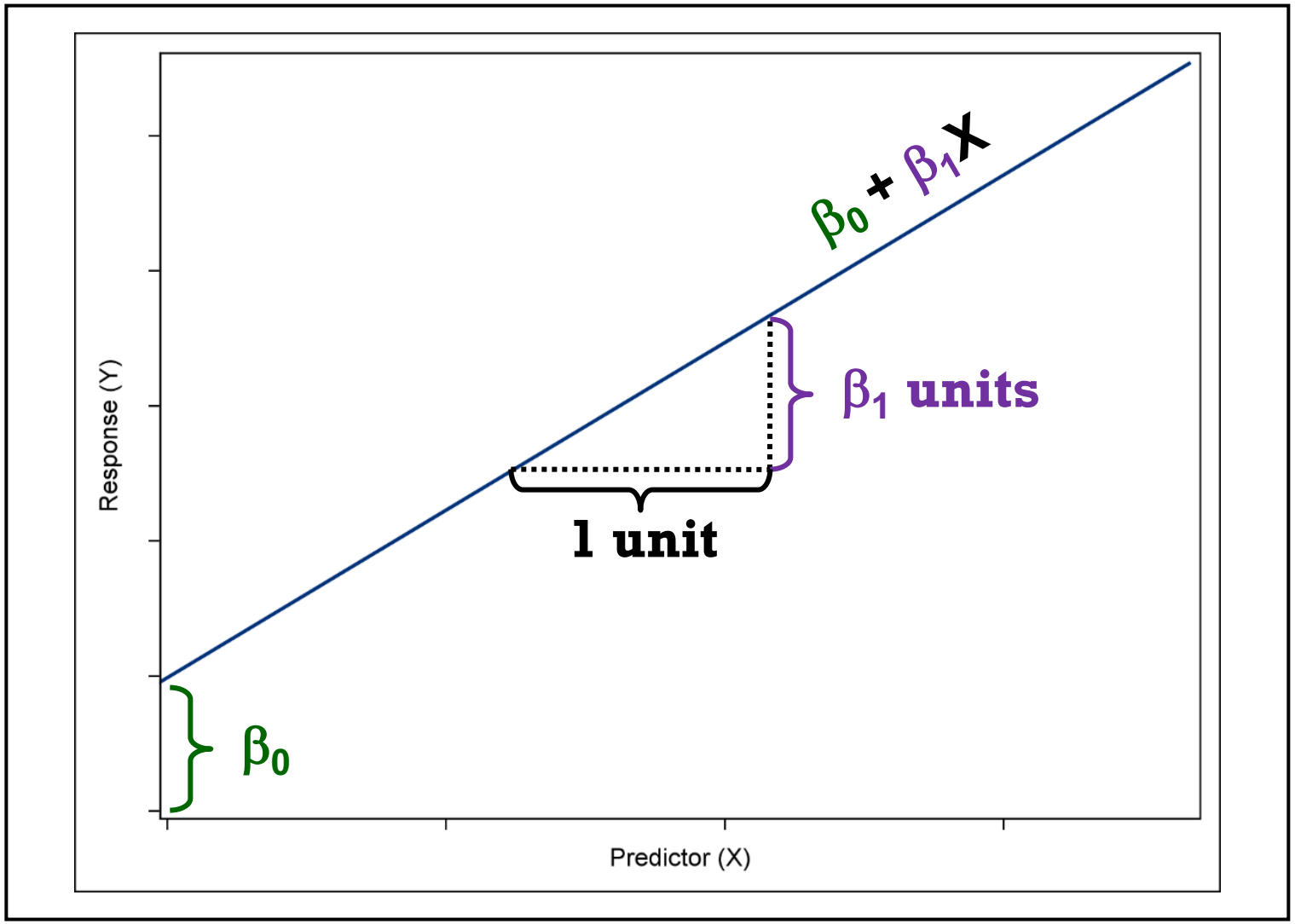
idno	chol (x)	sysbp (y)
1	437	194
2	264	121
3	249	131
4	297	159
5	243	123
6	272	161
7	161	115
รวม	1923	1004



$$\hat{y} = \beta_0 + \beta_1 x$$

$$\widehat{bp} = \beta_0 + \beta_1 chol$$

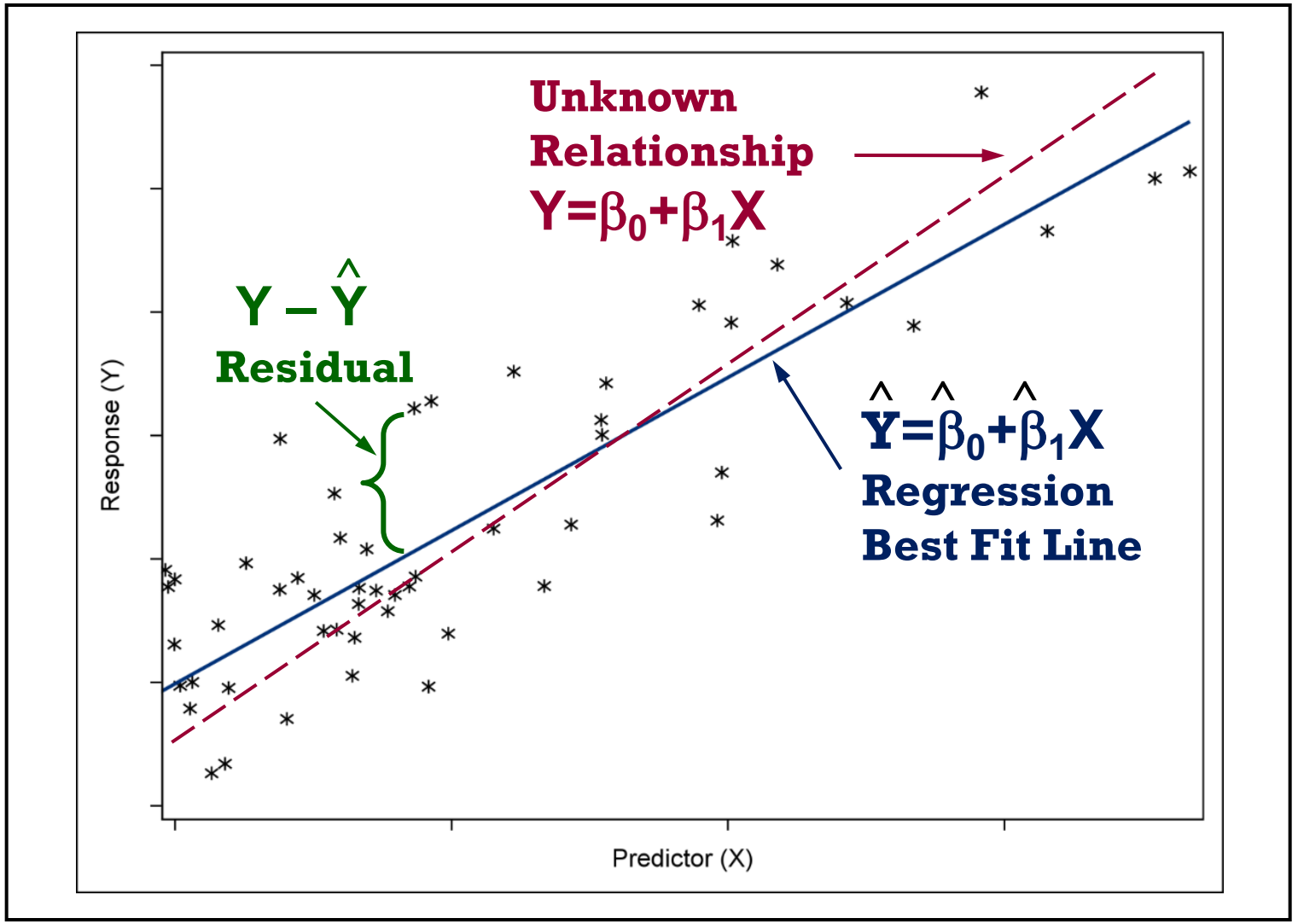
Simple Linear Regression Model



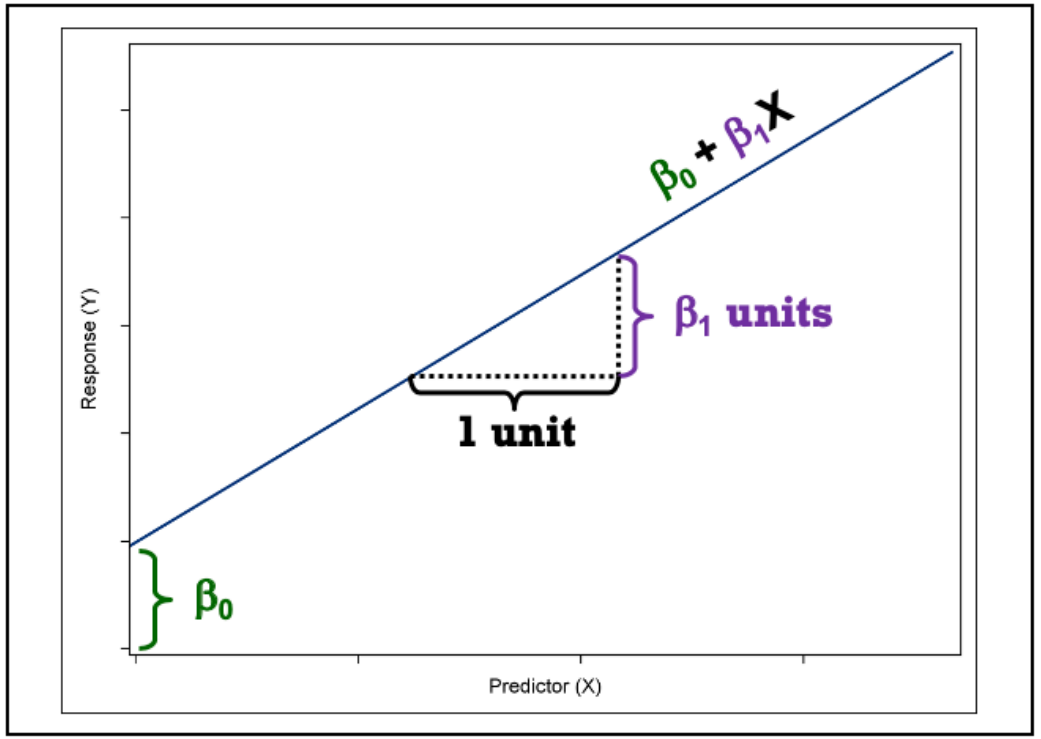
$$\hat{y} = \beta_0 + \beta_1 x$$

$$\widehat{bp} = \beta_0 + \beta_1 chol$$

Ordinary Least Squares (OLS) Regression



How to estimate parameters



$$\hat{y} = \beta_0 + \beta_1 x$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{s_{xy}}{s_{xx}} = \frac{\left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right)}{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)}$$

$$[Y] = [X][\beta]$$
$$[\beta] = [X]^{-1}[Y]$$

Example

- Systolic Blood Pressure (y)
- Cholesterol (x)

idno	chol(x)	sysbp(y)	x^2	xy	y^2
1	437	194	190969	84778	37636
2	264	121	69696	31944	14641
3	249	131	62001	32619	17161
4	297	159	88209	47223	25281
5	243	123	49049	29889	15129
6	272	161	73984	43792	25921
7	161	115	25921	18515	13225
รวม	1923	1004	569829	288760	148994

How to read an equation

$$\hat{y} = \beta_0 + \beta_1 x$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{s_{xy}}{s_{xx}} = \frac{\left(\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right)}{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)}$$

$$\bar{x} = \frac{1923}{7} = 247.7143, \bar{y} = \frac{1004}{7} = 143.4286$$

$$\beta_1 = \frac{s_{xy}}{s_{xx}} = \frac{\left(288760 - \frac{1923 \times 1004}{7}\right)}{\left(569829 - \frac{(1923)^2}{7}\right)} = 0.3116$$

$$\beta_0 = 143.4286 - (0.3116)(247.7143) = 57.8355$$

$$\hat{y} = 57.8355 + 0.3116x$$

$$\widehat{bp} = 57.8355 + 0.3116 \times chol$$

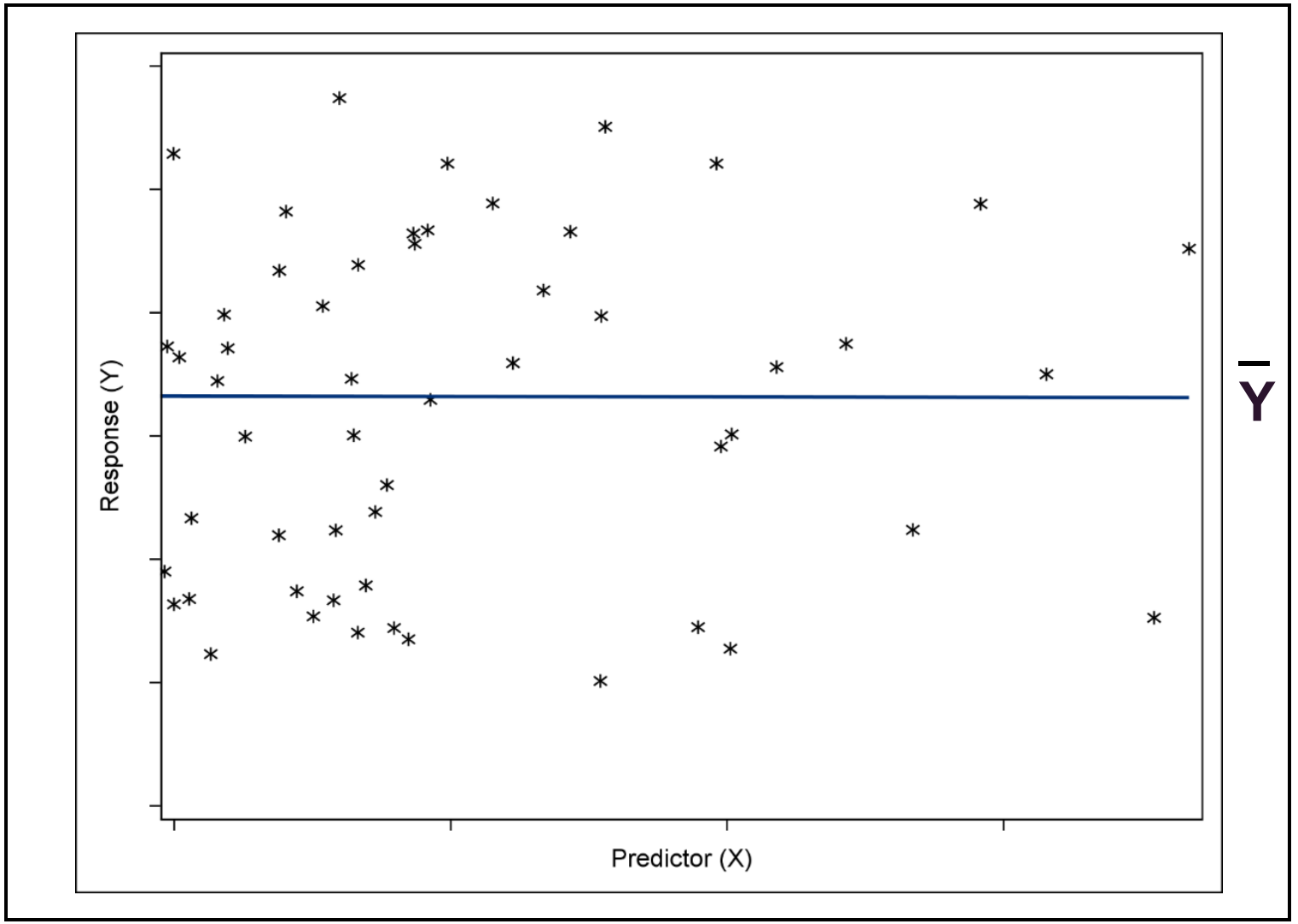
Example: Prediction

- Systolic Blood Pressure (y)
- Cholesterol (x)

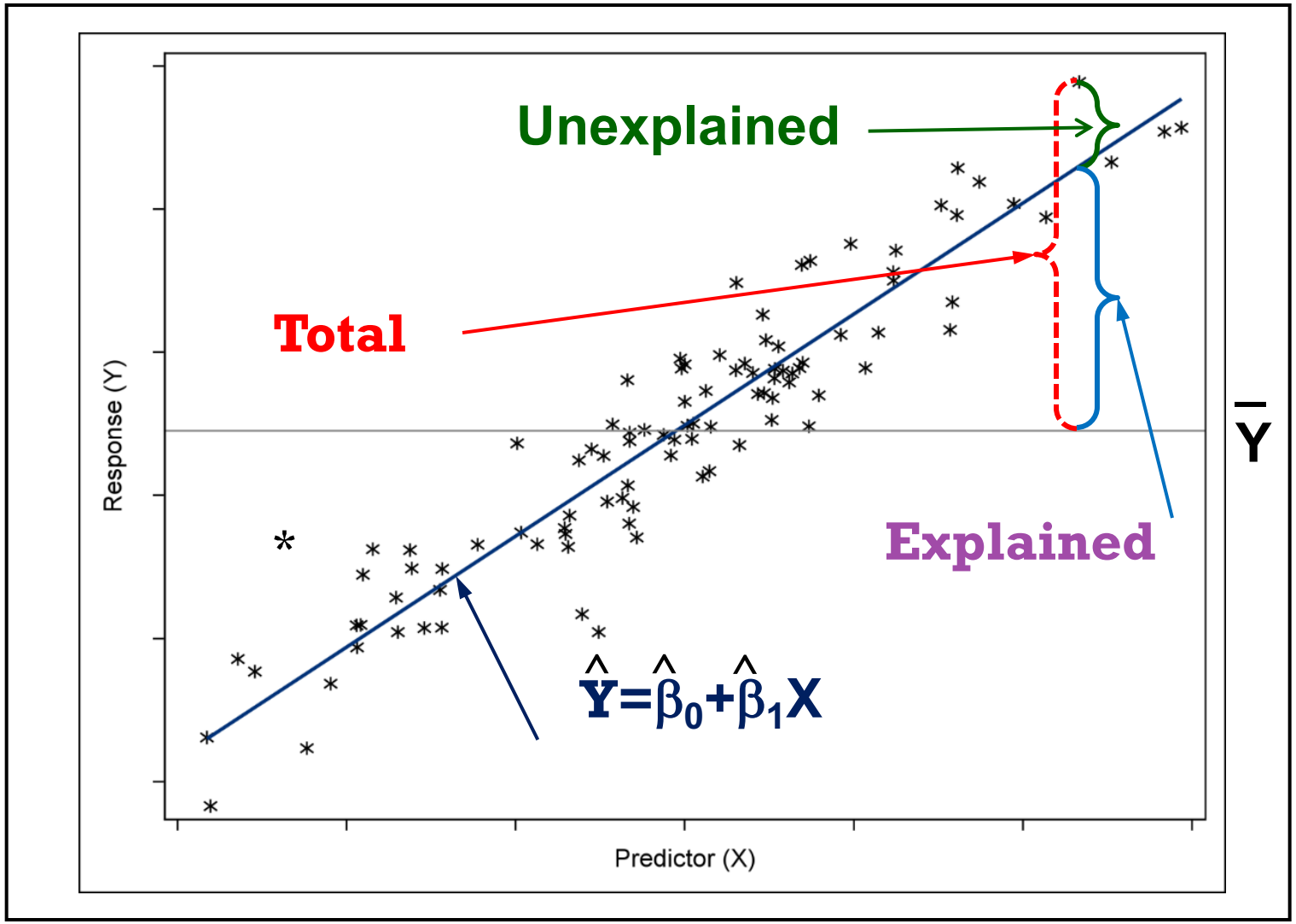
$$\widehat{bp} = 57.8355 + 0.3116 \times chol$$

idno	chol(x)	sysbp(y)
1	437	194
2	264	121
3	249	131
4	297	159
5	243	123
6	272	161
7	161	115
รวม	1923	1004

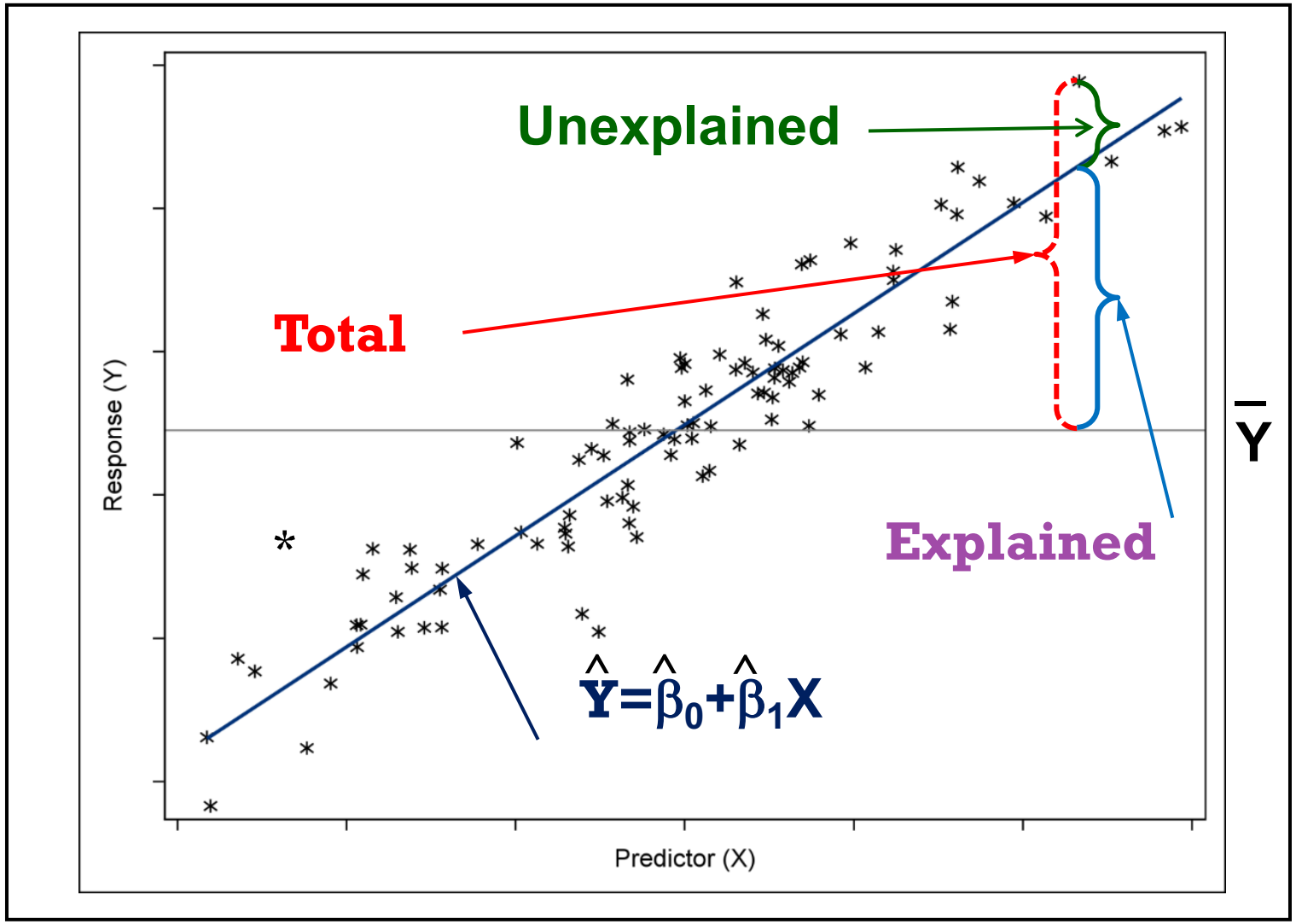
The Baseline Model (Null Hypothesis)



Explained versus Unexplained Variability



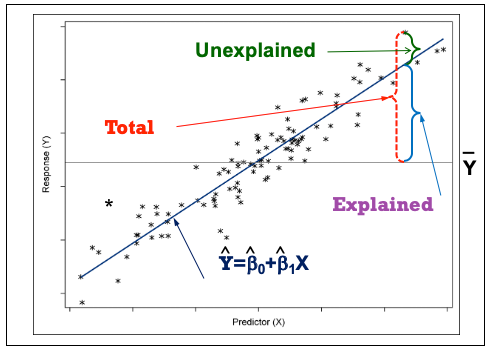
Coefficient of Determination



$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

- "Proportion of variance accounted for by the model"

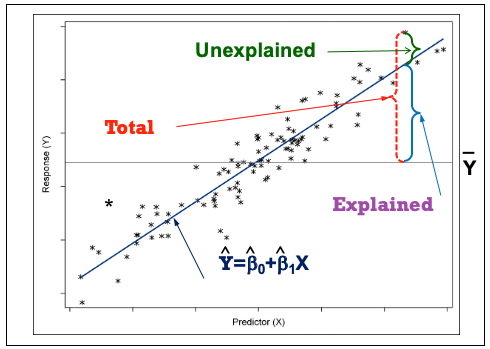
Coefficient of Determination (cont.)



$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

id	chol (x)	bp (y)	predict	error	squred error (SE)	guess	(y - y_bar)	squred total (ST)
1	437	194	196.1897	(2.1897)	4.7948	143.4286	50.5714	2,557.4694
2	264	121	141.4179	(20.4179)	416.8906	143.4286	(22.4286)	503.0408
3	249	131	136.6689	(5.6689)	32.1364	143.4286	(12.4286)	154.4694
4	297	159	151.8657	7.1343	50.8982	143.4286	15.5714	242.4694
5	243	123	134.7693	(11.7693)	138.5164	143.4286	(20.4286)	417.3265
6	272	161	143.9507	17.0493	290.6786	143.4286	17.5714	308.7551
7	161	115	108.8081	6.1919	38.3396	143.4286	(28.4286)	808.1837
average	274.7143	143.4286		SSE	972.2548		SST	4,991.7143
				MSE	138.8935			
				RMSE	**11.7853**			
	R^2	**1 - (SSE/SST)**	**0.8052**					

Coefficient of Determination (cont.)



- Train: R^2, RMSE
- Test: R^2, RMSE (honest estimate)

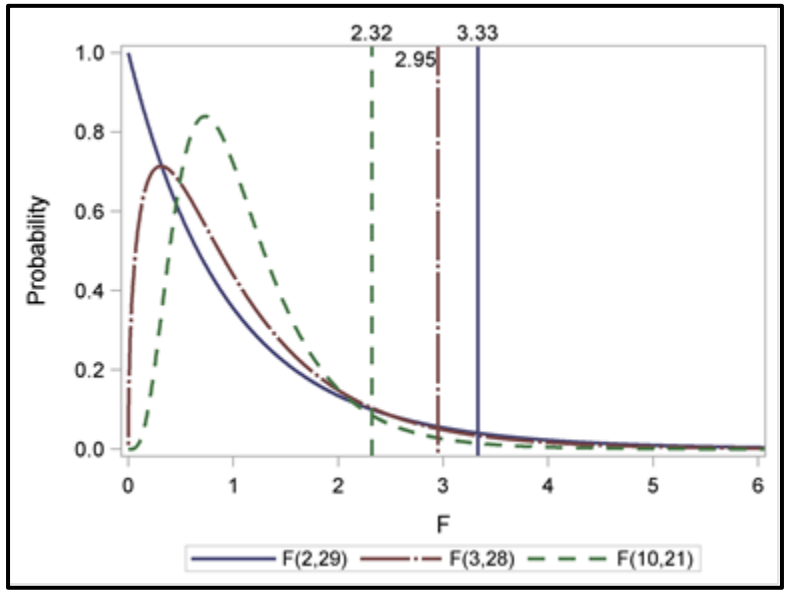
Training Data

Testing Data

id	chol (x)	bp (y)	predict	error	squred error (SE)	guess	(y - y_bar)	squred total (ST)
1	437	194	196.1897	(2.1897)	4.7948	143.4286	50.5714	2,557.4694
2	264	121	141.4179	(20.4179)	416.8906	143.4286	(22.4286)	503.0408
3	249	131	136.6689	(5.6689)	32.1364	143.4286	(12.4286)	154.4694
4	297	159	151.8657	7.1343	50.8982	143.4286	15.5714	242.4694
5	243	123	134.7693	(11.7693)	138.5164	143.4286	(20.4286)	417.3265
6	272	161	143.9507	17.0493	290.6786	143.4286	17.5714	308.7551
7	161	115	108.8081	6.1919	38.3396	143.4286	(28.4286)	808.1837
average	274.7143	143.4286		SSE	972.2548		SST	4,991.7143
				MSE	138.8935			
				RMSE	**11.7853**			
	R^2	**1 - (SSE/SST)**	**0.8052**					

Model Hypothesis Test
F Statistic and Critical Values at α=0.05

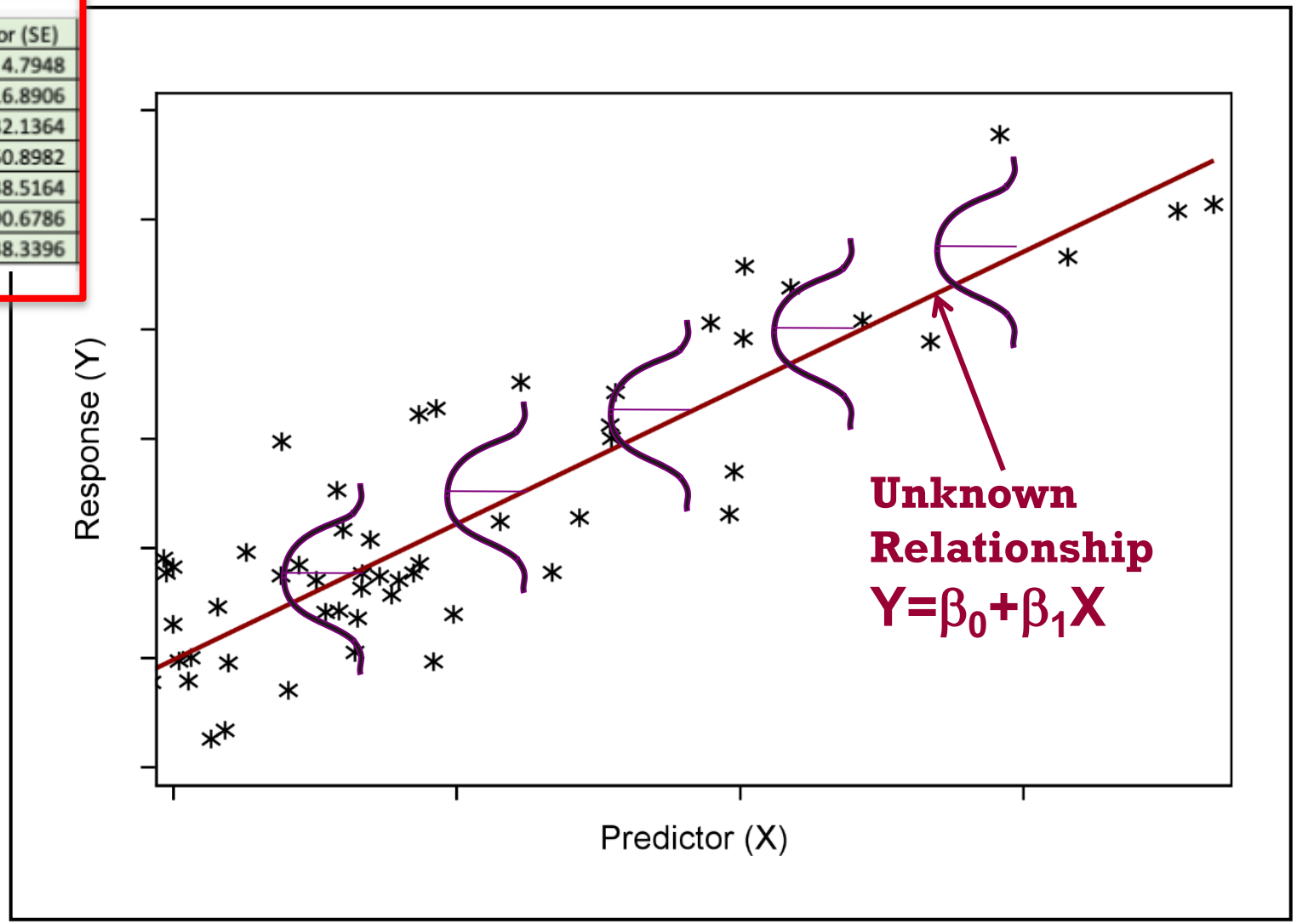
- Null Hypothesis:
  - The simple linear regression model does not fit the data better than the baseline model.
  - $\beta_1=0$
- Alternative Hypothesis:
  - The simple linear regression model does fit the data better than the baseline model.
  - $\beta_1\neq0$



F(Model df, Error df)=MS_M / MS_E

Assumptions of Simple Linear Regression

id	chol (x)	bp (y)	predict	error	squred error (SE)
1	437	194	196.1897	(2.1897	4.7948
2	264	121	141.4179	(20.4179	416.8906
3	249	131	136.6689	(5.6689	32.1364
4	297	159	151.8657	7.1343	50.8982
5	243	123	134.7693	(11.7693	138.5164
6	272	161	143.9507	17.0493	290.6786
7	161	115	108.8081	6.1919	38.3396



+

Multiple Linear Regression

Multiple Linear Regression with Two Variables

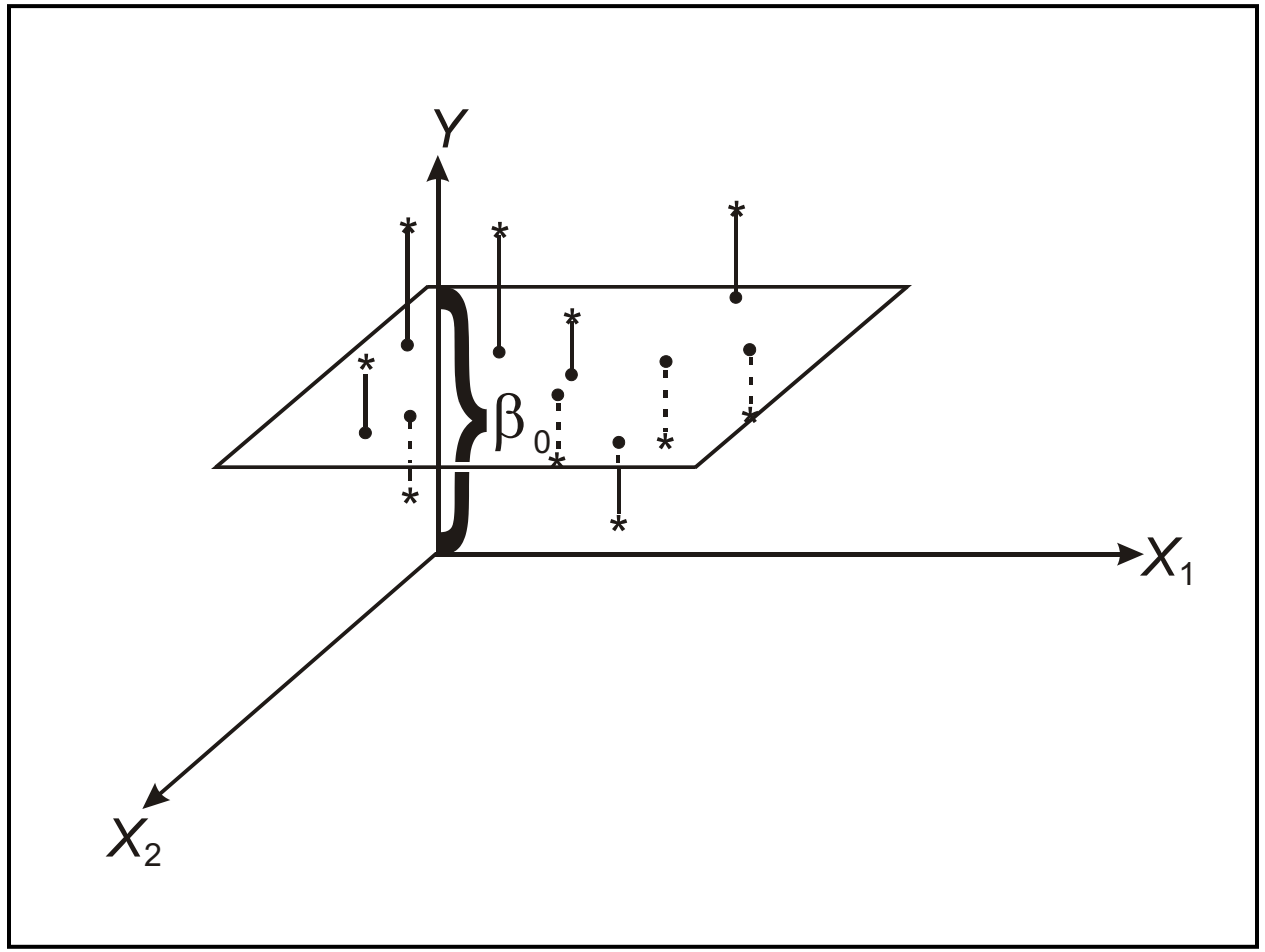
- Consider the two-variable model

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$$

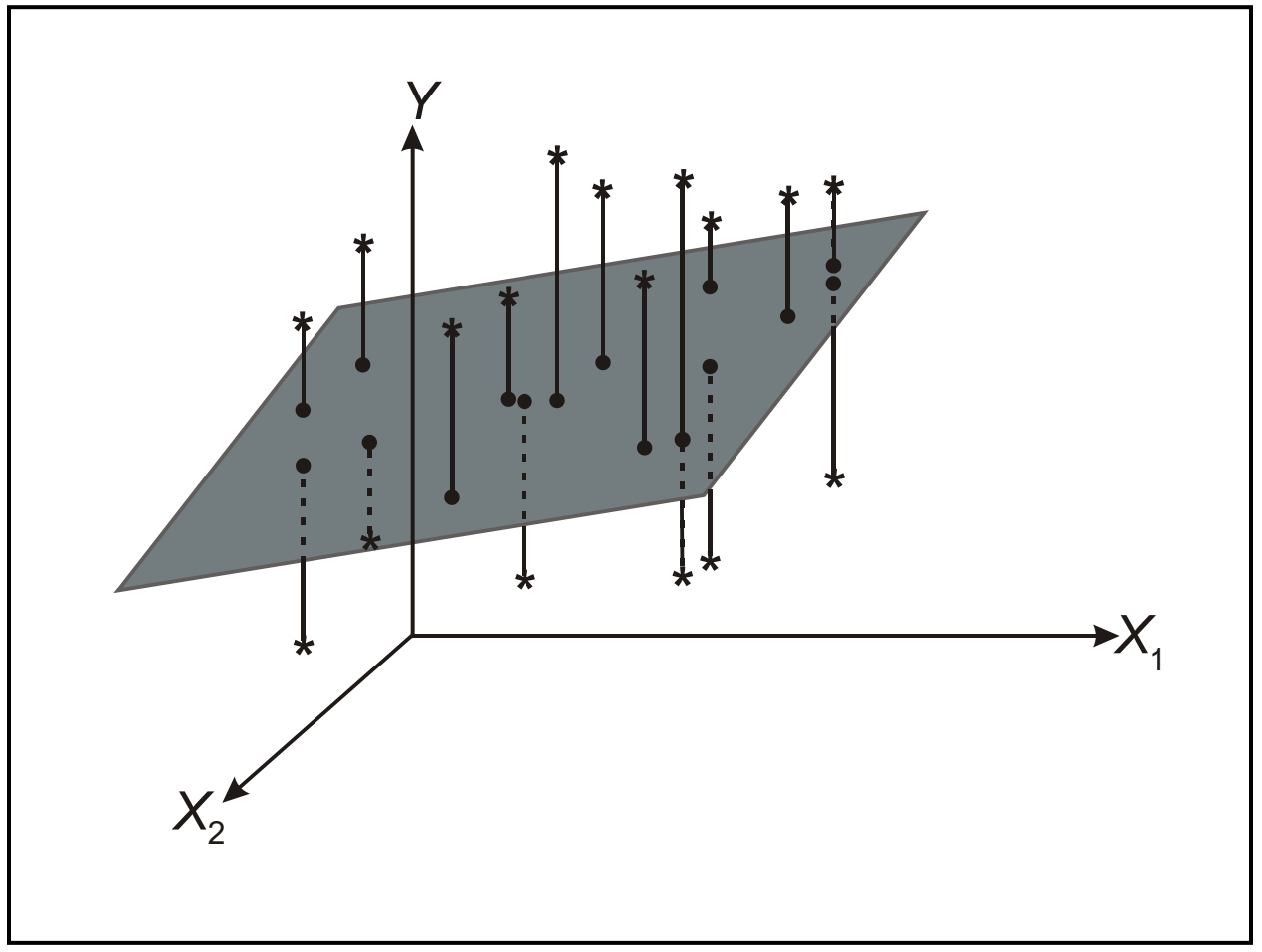
- where

Y	is the dependent variable.
X_1 and X_2	are the independent or predictor variables.
ε	is the error term.
β_0, β_1, and β_2	are unknown parameters.

Picturing the Model: No Relationship



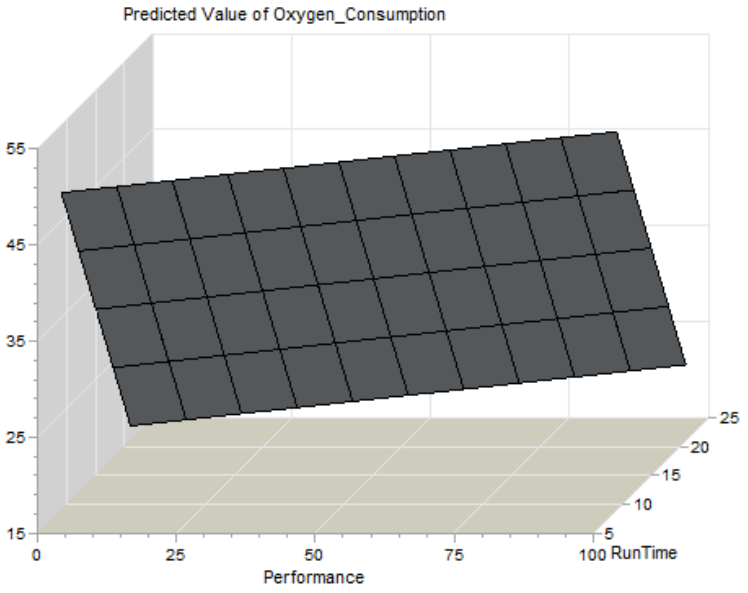
Picturing the Model: A Relationship



The Multiple Linear Regression Model

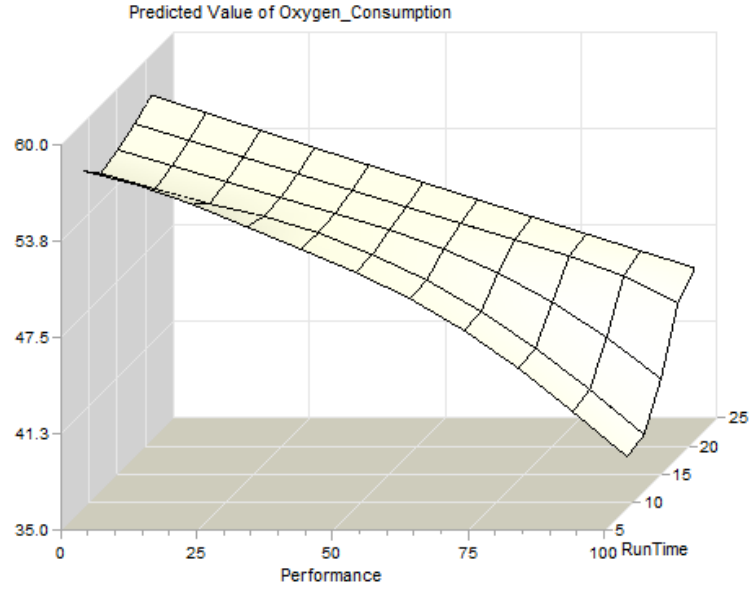
- In general, you model the dependent variable, Y, as a linear function of *k* independent variables, X_1 through X_k:

$$Y=\beta_0+\beta_1X_1+\ldots+\beta_kX_k+\varepsilon$$



$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$$

Linear Model with only Linear Effects



$$Y=\beta_0+\beta_1X_1+\beta_2X_1^2+\beta_3X_2+\beta_4X_2^2+\varepsilon$$

Linear Model with Nonlinear Effects

The Multiple Linear Regression Model (cont.) Matrix Multiplication Approach

inputs		target
Age	**Income**	**Spending**
25	25,000	400
35	50,000	500
32	35,000	550

$$Y=\beta_0(1)+\beta_1X_1+\beta_2X_2+\varepsilon$$

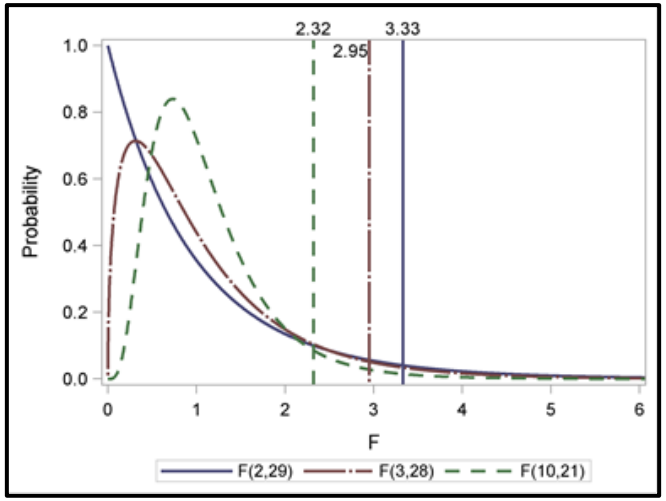
$$[Y] = [X][\beta]$$
$$[\beta]=[X]^{-1}[Y]$$

https://sites.google.com/site/mystatistics01/regression-correlation-analysis/multiple-linear-regression-step

Model Hypothesis Test

- Null Hypothesis:
  - The regression model does not fit the data better than the baseline model.
  - $\beta_1=\beta_2=\ldots=\beta_k=0$

- Alternative Hypothesis:
  - The regression model does fit the data better than the baseline model.
  - Not all β_is equal zero.



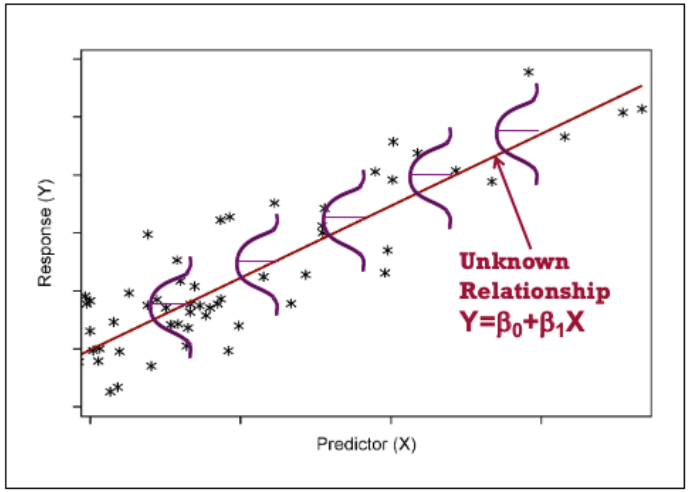
F(Model df, Error df)=MS_M / MS_E

Assumptions for Linear Regression

- The mean of the Ys is accurately modeled by a linear function of the X_i.
  - (y, x_i) = linear relationship (correlation)

- The random error term, ε, is assumed to have a normal distribution with a mean of zero.
- The random error term, ε, is assumed to have a constant variance, σ^2.
  - Not skew

- The errors are independent.



Multiple Linear Regression versus Simple Linear Regression

- **Main Advantage**
- Multiple linear regression enables you to investigate the relationship among Y and several independent variables simultaneously.
- **Main Disadvantages**
- Increased complexity makes it more difficult to do the following:
  - ascertain which model is "best"
  - interpret the models

Common Applications of Multiple Regression

- Multiple linear regression is a powerful tool for the following tasks:
  - Prediction – to develop a model to predict future values of a response variable (Y) based on its relationships with other predictor variables (Xs)
  - Analytical or Explanatory Analysis – to develop an understanding of the relationships between the response variable and predictor variables

Prediction

- The terms in the model, the values of their coefficients, and their statistical significance are of secondary importance.
- The focus is on producing a model that is the best at predicting future values of Y as a function of the Xs. The predicted value of Y is given by this formula:

$$\underline{\hat{Y}} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$$

Analytical or Explanatory Analysis

- The focus is on understanding the relationship between the dependent variable and the independent variables.
- Consequently, the statistical significance of the coefficients is important as well as the magnitudes and signs of the coefficients.

$$\hat{Y} = \underline{\hat{\beta}_0} + \underline{\hat{\beta}_1} X_1 + \ldots + \underline{\hat{\beta}_k} X_k$$

Adjusted R Square

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSE}{SST}$$

$$R^2_{ADJ} = 1 - \frac{(n-i)(1-R^2)}{n-p}$$

- i=1 if there is an intercept and 0 otherwise
- n=the number of observations used to fit the model
- p=the number of parameters in the model

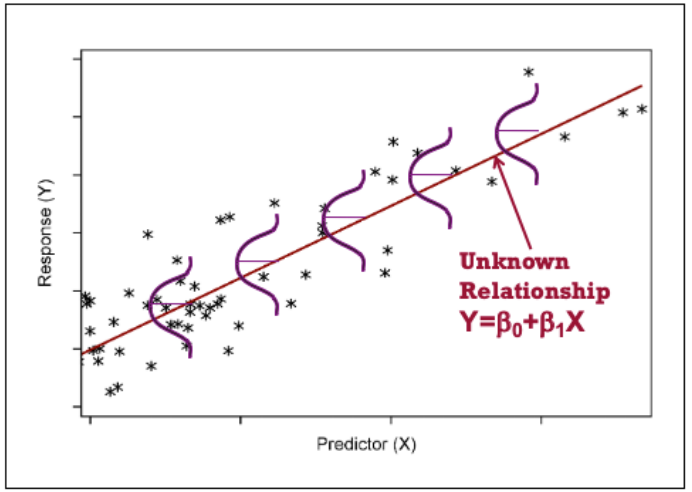
Other topics

Assumptions for Linear Regression

- The mean of the Ys is accurately modeled by a linear function of the X_i.
  - (y, x_i) = linear relationship (correlation)

- The random error term, ε, is assumed to have a normal distribution with a mean of zero.
- The random error term, ε, is assumed to have a constant variance, σ^2.
  - Not skew

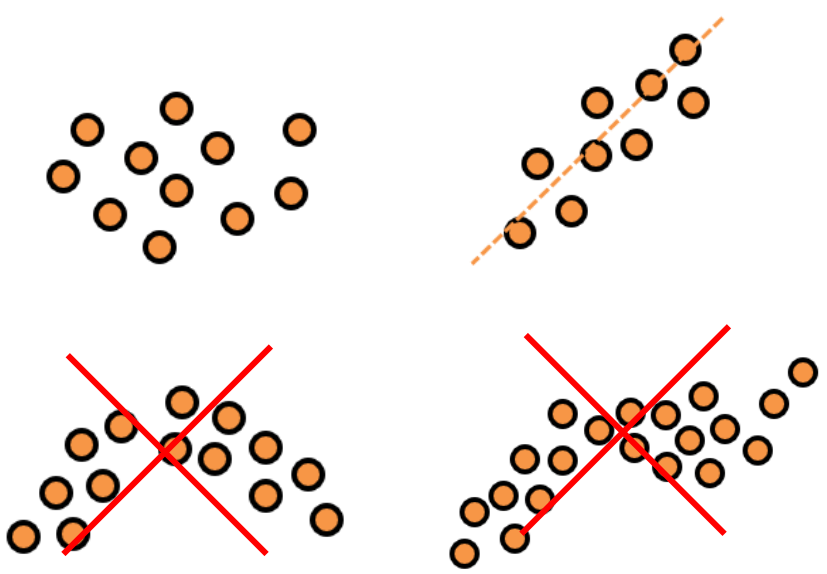
- The errors are independent.



Pearson Correlation

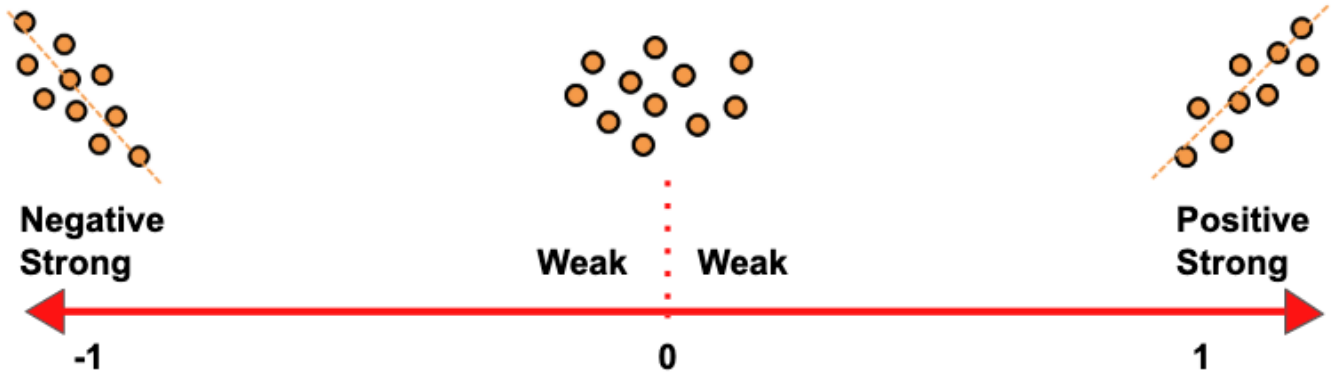
- Pearson Correlation, which is the Pearson Product Moment Correlation (PPMC), is used to evaluate **linear** **relationships** between two **continuous variable**
- Here's the most commonly used formula to find the Pearson correlation coefficient, which can be called Pearson's R:

$$r = \frac{\sum(x_i - x_{\text{average}})\,(y_i - y_{\text{average}})}{\sqrt{\sum(x_i - x_{\text{average}})^2 * \sum\,(y_i - y_{\text{average}})^2}}$$



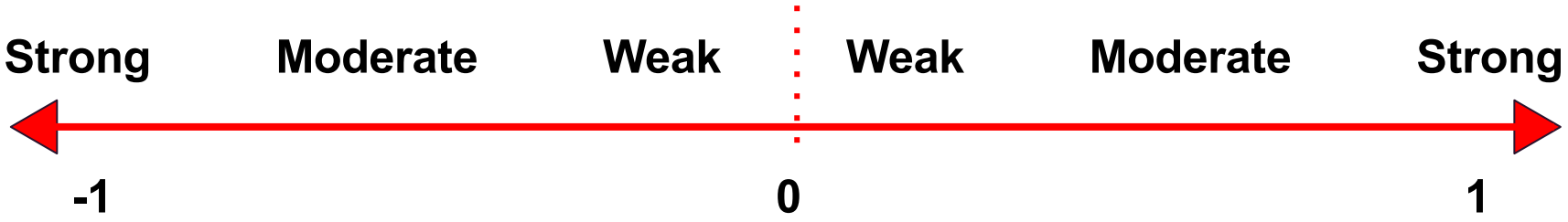
Correlation Coefficient

- The numerical measure of the degree of association between two continuous variables is called the **correlation coefficient (r)**.
- The coefficient value is always between -1 and 1 and it measures both the **strength** and **direction** of the linear relationship between the variables.



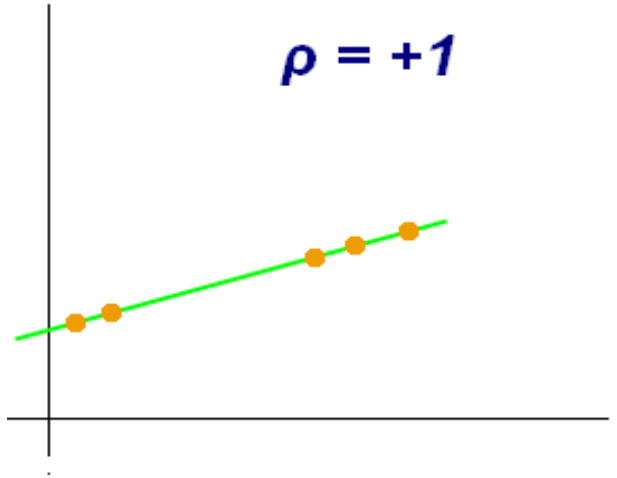
Correlation Coefficient (cont.): Strength

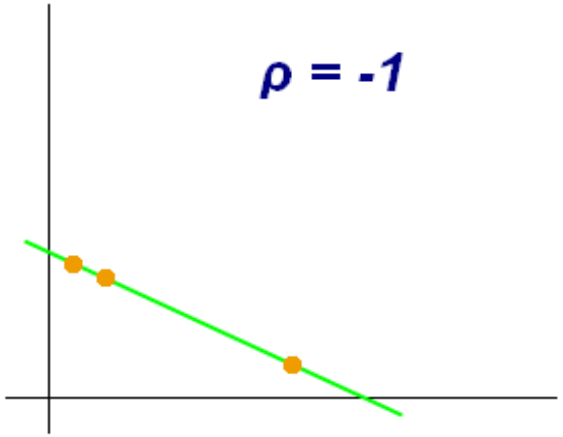
- **Strength**
  - The values of **-1 and 1** indicate a perfect **linear relationship** when all the data points fall on a line. Normally, either positive or negative, is **rarely** found.
  - A coefficient of **0** indicates no linear relationship between the variables. This is what you are likely to get with two sets of random numbers.
  - Values **between 0 and +1/-1** represent a scale of weak, moderate and strong relationships. As the coefficient gets closer to either -1 or 1, the strength of the relationship increases.



Correlation Coefficient (cont.): Direction

- **Direction**
  - **Positive coefficients** represent **direct** linear association (upward-sloping)
  - **Negative coefficients** represent **inverse** linear association (downward-sloping)



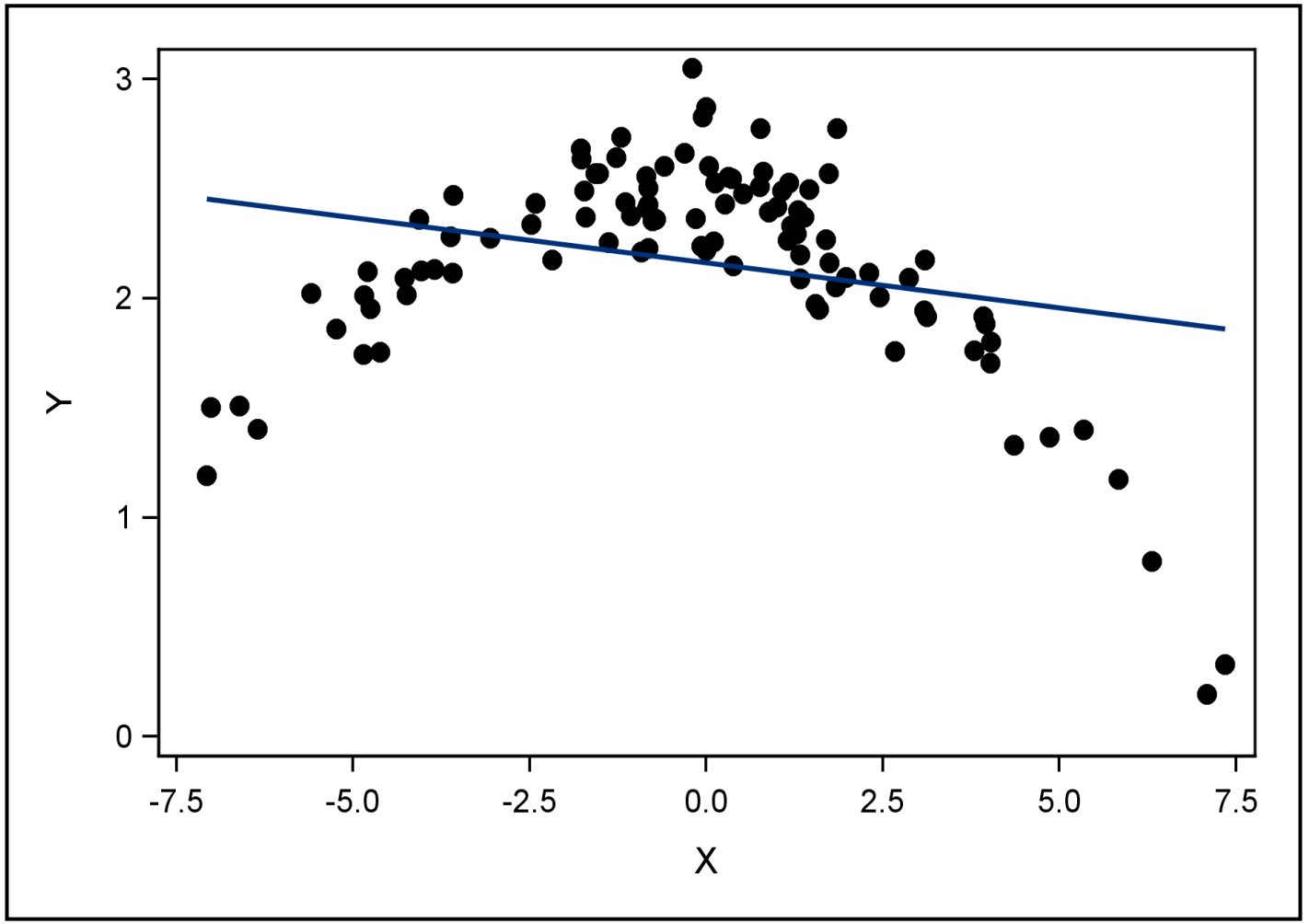


Hypothesis Test for a Correlation

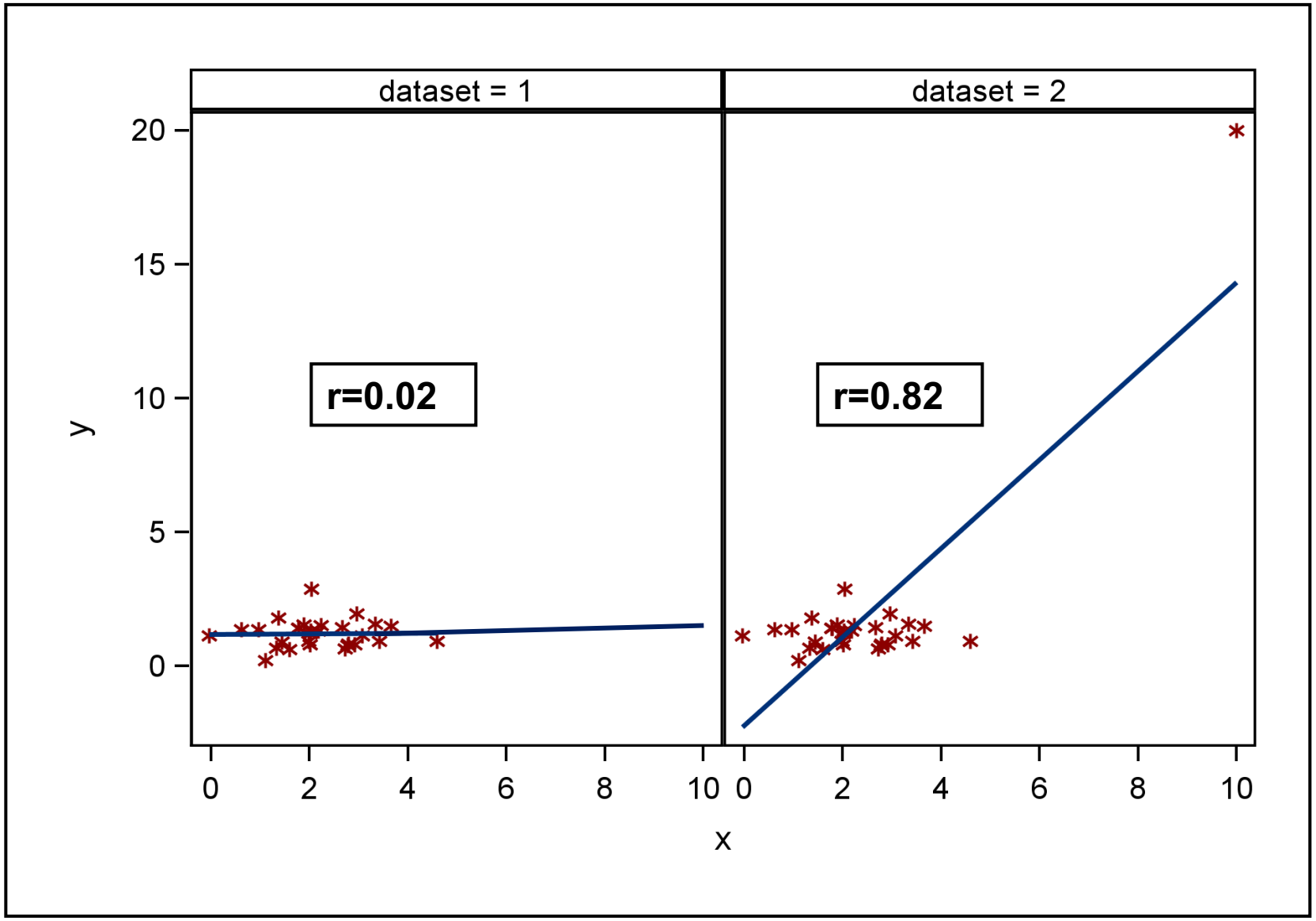
- The parameter representing correlation is ρ.
- ρ is estimated by the sample statistic r.

- H_0: $\rho=0$
- Rejecting H0 indicates only great confidence that ρ is not exactly zero.
- A *p*-value does not measure the magnitude of the association.
- Sample size affects the *p*-value.

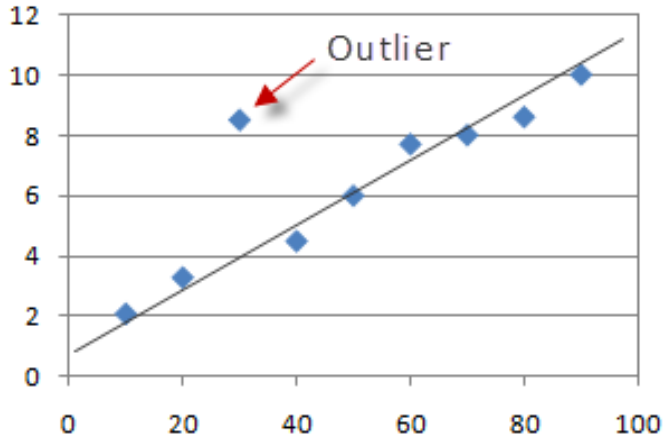
Remark1: Missing Another Type of Relationship



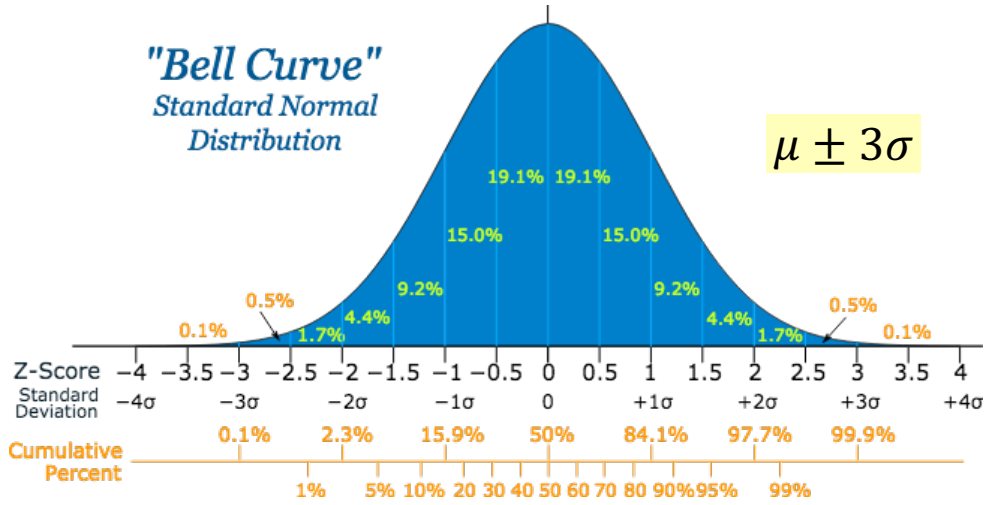
Remark2: Extreme Data Values

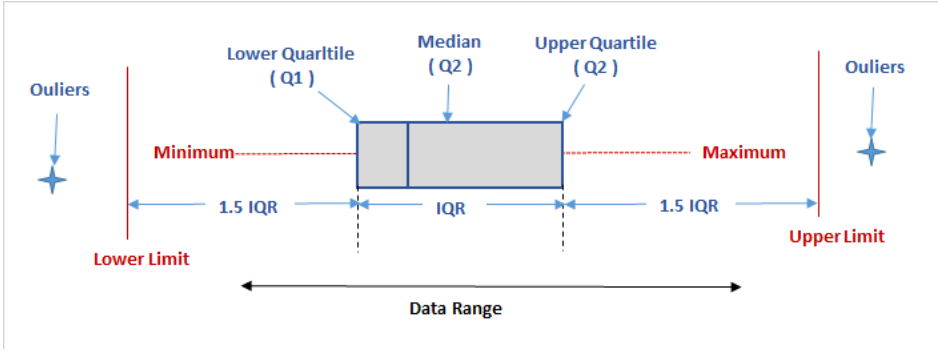


Truncate outliers



- Outlier, leverage points, extreme values

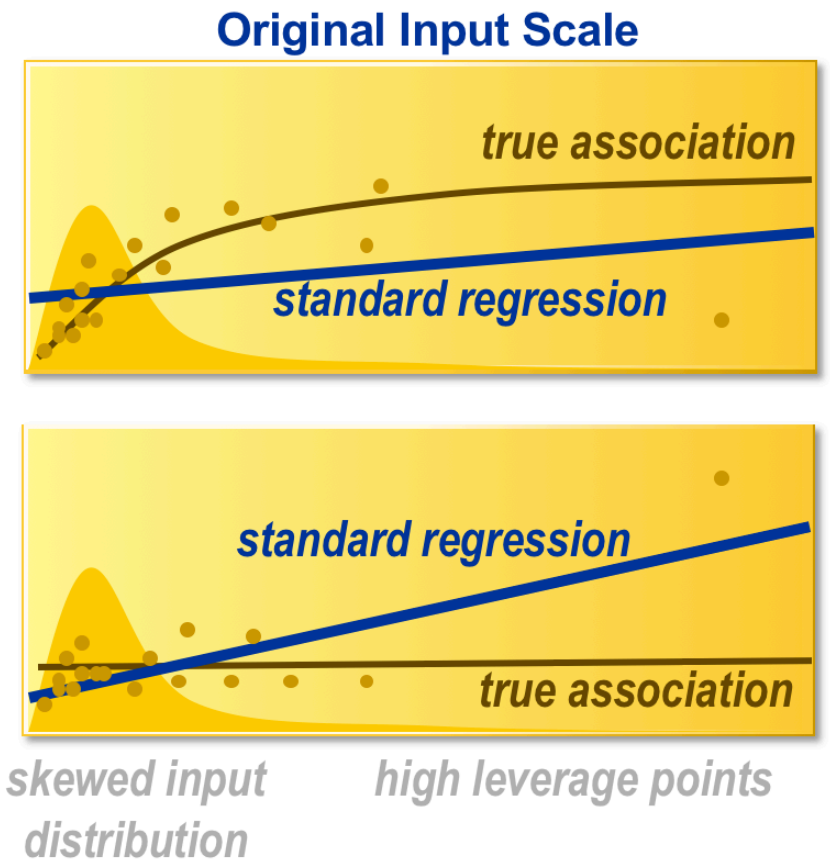




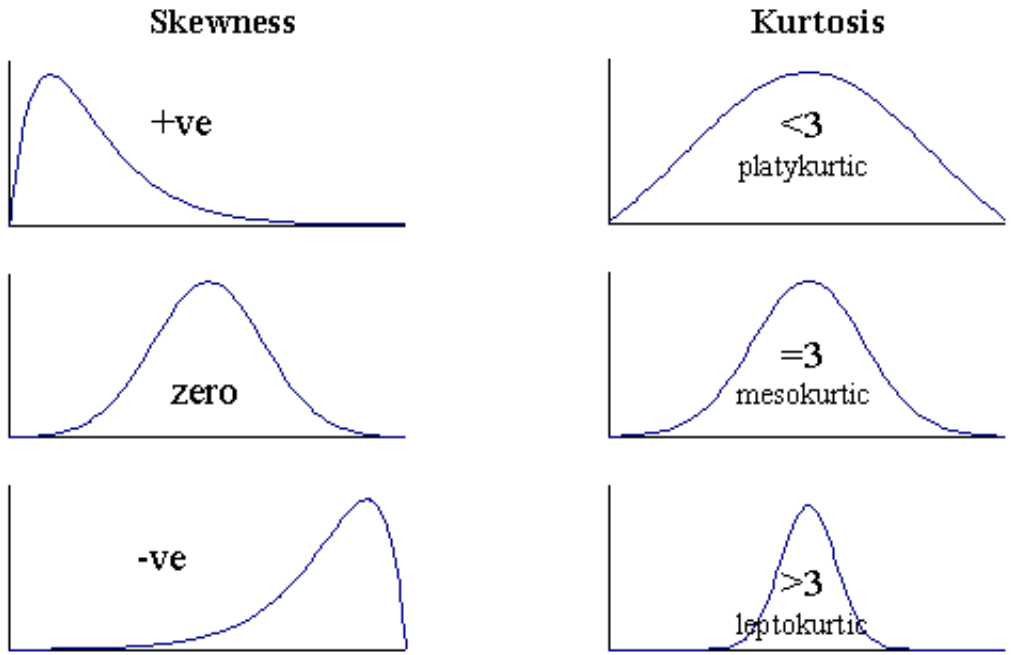
Percentile
1st
2.5th
5th
10th
25th
50th
75th
90th
95th
97.5th
99th

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x_1 + \hat{w}_2 x_2$$

Outliers & Feature transformation



- Skewness
- Example: Salary, Balance in bank account
- **Solutions: Log, Binning**



Outliers & Feature transformation (cont.)

	Spending	Spending with Outliers	LOG10(Spending)	LOG10(Spending with Outliers)
	3,000.00	3,000.00	3.48	3.48
	3,200.00	3,200.00	3.51	3.51
	4,000.00	4,000.00	3.60	3.60
	4,500.00	4,500.00	3.65	3.65
	5,000.00	5,000.00	3.70	3.70
		1,000,000.00		6.00
mean	3,940.00	169,950.00	3.59	3.99

Demo

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

sklearn.linear_model.LinearRegression

class sklearn.linear_model.**LinearRegression**(*fit_intercept=True*, *normalize=False*, *copy_X=True*, *n_jobs=None*) [source]

Ordinary least squares Linear Regression.

LinearRegression fits a linear model with coefficients w = (w1, ..., wp) to minimize the residual sum of squares between the observed targets in the dataset, and the targets predicted by the linear approximation.

Parameters:

fit_intercept : *bool, optional, default True*
Whether to calculate the intercept for this model. If set to False, no intercept will be used in calculations (i.e. data is expected to be centered).

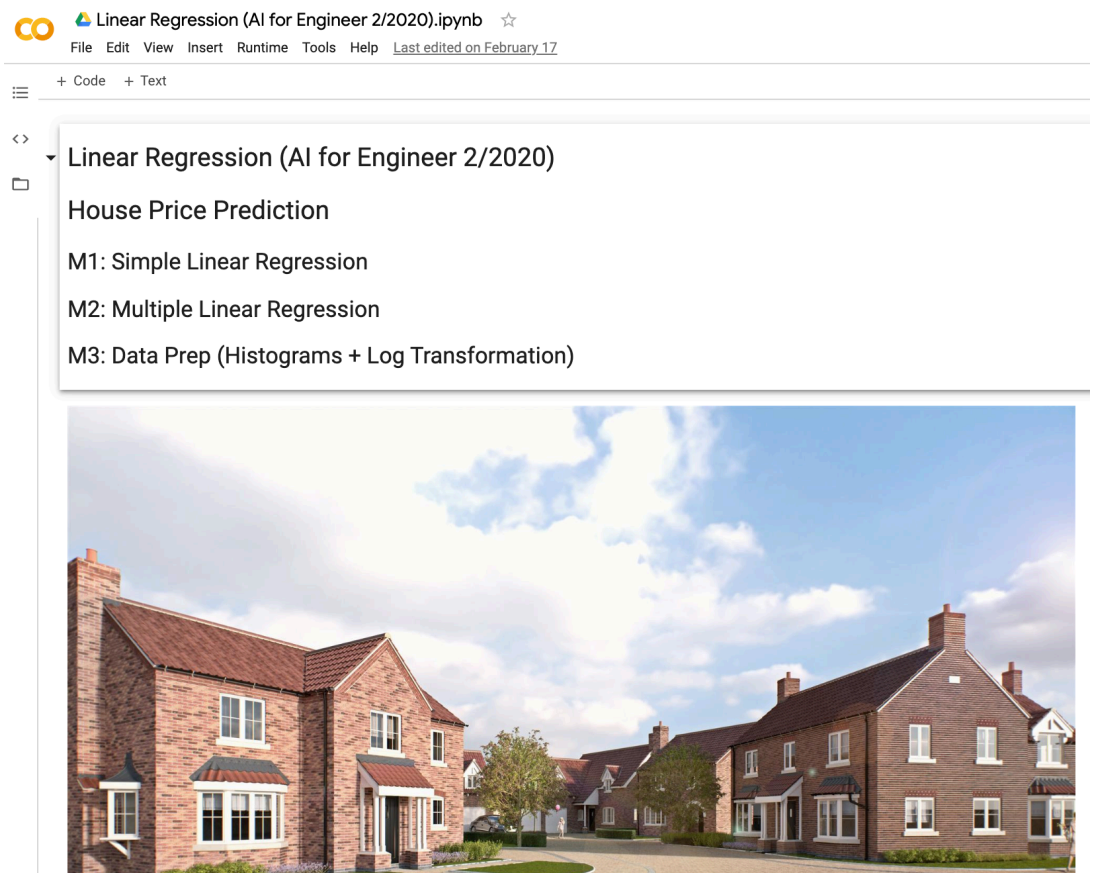
normalize : *bool, optional, default False*
This parameter is ignored when `fit_intercept` is set to False. If True, the regressors X will be normalized before regression by subtracting the mean and dividing by the l2-norm. If you wish to standardize, please use

Methods

fit(self, X, y[, sample_weight])	Fit linear model.
get_params(self[, deep])	Get parameters for this estimator.
predict(self, X)	Predict using the linear model.
score(self, X, y[, sample_weight])	Return the coefficient of determination R^2 of the prediction.
set_params(self, **params)	Set the parameters of this estimator.

House Price Prediction: Target=MEDV

http://bit.ly/AIENG2020-Reg2



Homework

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

About Citation Policy Donate a Data Set Contact

UCI Machine Learning Repository

Search

Repository Web Google

View ALL Data Sets

Center for Machine Learning and Intelligent Systems

Wine Quality Data Set

Download: Data Folder, Data Set Description

Abstract: Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests (see [Cortez et al., 2009], [Web Link]).

Data Set Characteristics:	Multivariate	**Number of Instances:**	4898	**Area:**	Business
Attribute Characteristics:	Real	**Number of Attributes:**	12	**Date Donated**	2009-10-07
Associated Tasks:	Classification, Regression	**Missing Values?**	N/A	**Number of Web Hits:**	1141924

Source:

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal
@2009

Attribute Information:

For more information, read [Cortez et al., 2009].
Input variables (based on physicochemical tests):
1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

train_test_split

- test = 0.25
- random_state = 101