Data Analytics

Yong Zheng

Illinois Institute of Technology
Chicago, IL, 60616, USA

School of Applied Technology
ILLINOIS INSTITUTE OF TECHNOLOGY

Week 4 - 6

- Review on Statistical Basics
- Supervised Learning
- Predictive Models
- Simple Linear Regression
- Multiple Linear Regression
- Advanced Topics in Regression Models

Week 4 - 6

- Review on Statistical Basics
- Supervised Learning
- Predictive Models
- Simple Linear Regression
- Multiple Linear Regression
- Advanced Topics in Regression Models

Statistics Basics

- Descriptive Statistics
  - Data: Quantitative and Qualitative
  - Describe data numerically or by visualizations
  - Interpret data distributions
- Inferential Statistics
  - Estimate population by sample statistics
  - Hypothesis Testing
    - One sample
    - Two Independent samples
    - Two paired samples
  - Predictive Models

Learning Styles

- Statistical Basics
  - Theories
  - Calculations
  - Understandings
  - R Practice
- Predictive Models
  - No manual calculations
  - Less theories
  - Focus on practical skills
    - Understandings
    - R Practice to build models
    - Be able to read and interpret the outputs
    - Be able to identify and fix issues in the models
    - Be able to compare different models

Week 4 - 6

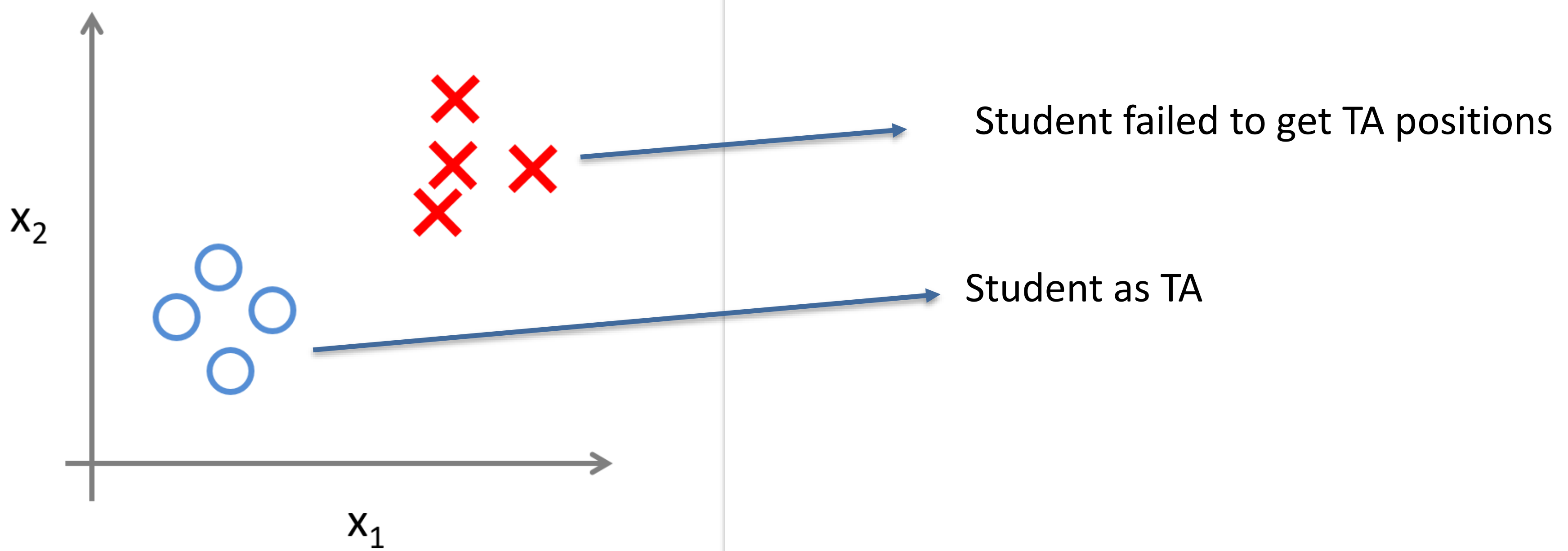
- Review on Statistical Basics
- Supervised Learning
- Predictive Models
- Simple Linear Regression
- Multiple Linear Regression
- Advanced Topics in Regression Models

Supervised v.s. Unsupervised Learning

- Supervised Learning: infer a (predictive) function from data associated with pre-defined targets/classes/labels or known values
  - We have the truth/knowledge
    - Regression models, if it is a numerical variable
    - Classification models, if it is a nominal variable
  - We learn from training data to validate on test set
  - We have metrics to evaluate the models
- Unsupervised Learning: discover or describe underlying structure from unlabelled data
  - We do not have the truth/knowledge
  - It is used to discover unknown structure or patterns
  - There are no metrics to evaluate the results

Supervised v.s. Unsupervised Learning

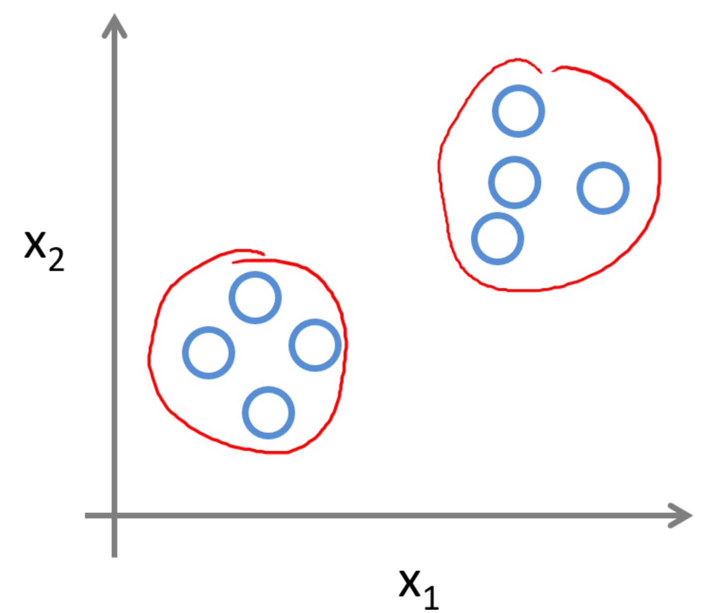
Supervised Learning



Supervised v.s. Unsupervised Learning

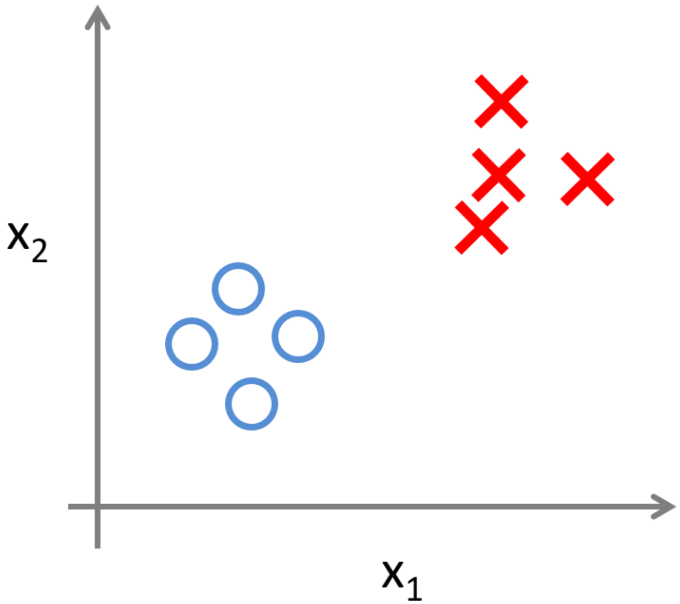
Just categorize students into N groups

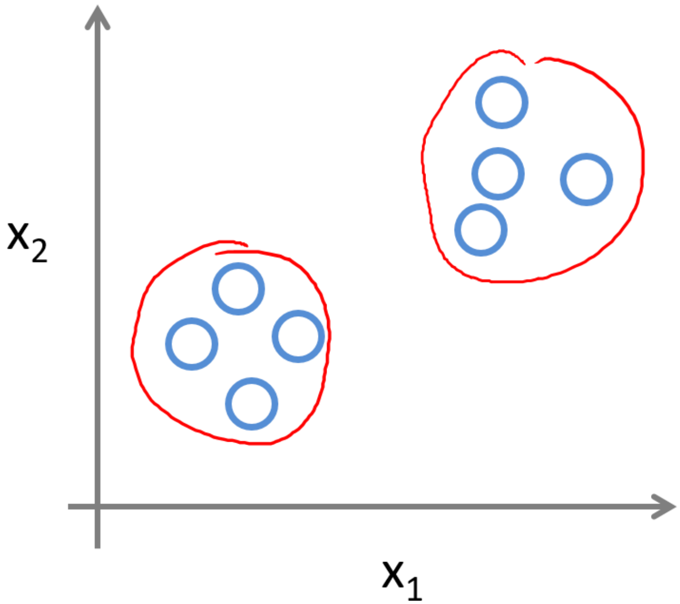
Unsupervised Learning



Supervised v.s. Unsupervised Learning

Supervised Learning	Unsupervised Learning





Examples

- We have undergraduate, graduate, PhD students in IIT, given the information of a student, such as age, gender, home country, living address, department, I want to predict he or she is an undergraduate, graduate or PhD student

Is it a supervised or unsupervised learning process?

Examples

- We analyze customers' purchasing behaviors in order to discover their shopping patterns. For example, we may find out that a customer who bought milk is highly possible to purchase bread on the same transaction.

Is it a supervised or unsupervised learning process?

Supervised v.s. Unsupervised Learning

Machine Learning Algorithms *(sample)*

	Unsupervised	Supervised
Continuous	• Clustering & Dimensionality Reduction ○ SVD ○ PCA ○ K-means	• Regression ○ Linear ○ Polynomial • Decision Trees • Random Forests
Categorical	• Association Analysis ○ Apriori ○ FP-Growth • Hidden Markov Model	• Classification ○ KNN ○ Trees ○ Logistic Regression ○ Naive-Bayes ○ SVM

Supervised Learning

- We know the truth
- We build predictive models to predict the values in the response variable (either numerical or nominal)
- We can compare the predicted values and the truth based on pre-defined metrics to say how accurate our prediction model is

Supervised Learning

Two examples:

- Linear regression
  Given some factors (hrs in studying, hrs in sleeping, hrs in games), try to <u>predict a numerical variable</u> (e.g., student grade)
- Classification
  Given some factors (age, height, weight, eye color, hair color), try to <u>predict a categorical variable</u> (e.g., gender)

Standard Process In Supervised Learning

- We use a linear regression as an example

Age	Years of exp	GPA	Salary
23	3	3.5	6K
25	4	4	7K
21	2	3.9	5K
20	2	3.1	4K
24	2	3.6	5K
27	2	3.7	6K

Standard Process In Supervised Learning

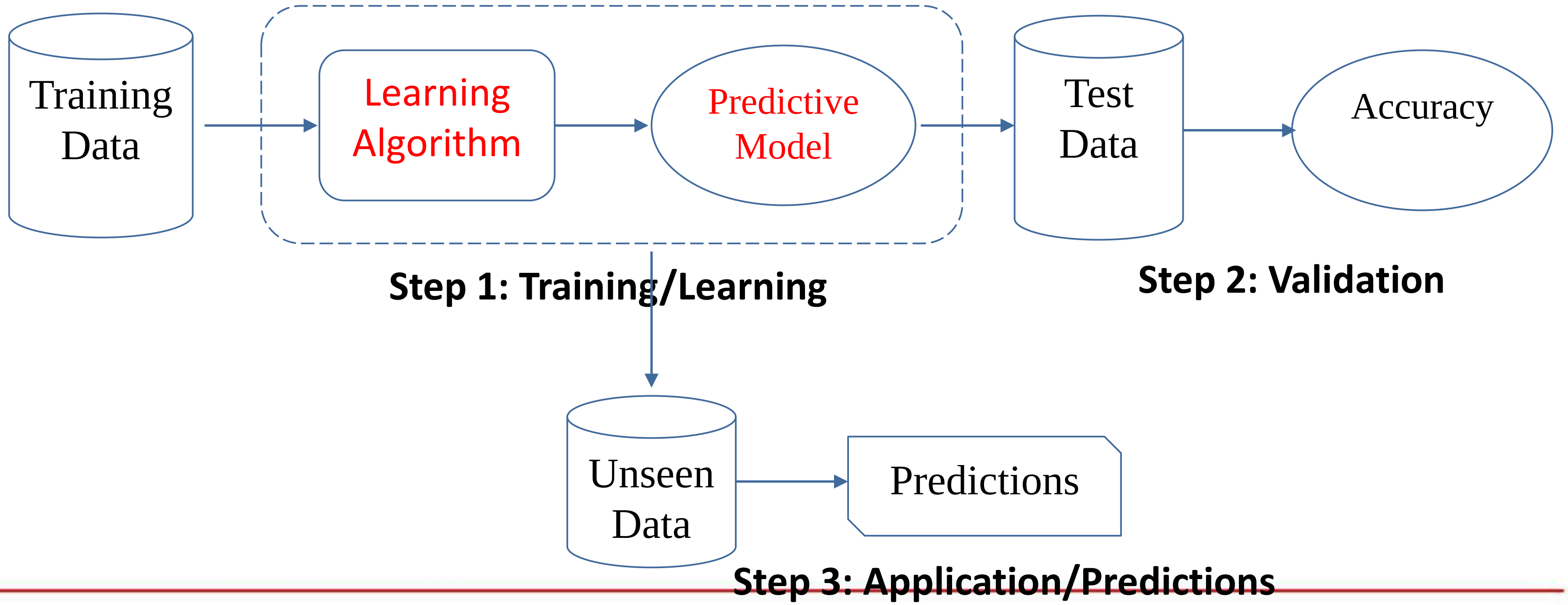
- We use a classification problem as an example

Color	Weight	Stripes	Tiger?
Orange	300 lbs	no	no
White	50 lbs	yes	no
Orange	490 lbs	yes	yes
White	510 lbs	yes	yes
Orange	490 lbs	no	no
White	450 lbs	no	no

Standard Process In Supervised Learning

- Train: Learn a model using the training data
- Validation/Test: Test using test data to assess accuracy
- Application: Apply the selected model to unseen data

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}}$$



Data Splits for Evaluations

1). Hold-out Evaluation → If your data is large enough

Color	Weight (lbs)	Stripes	Tiger?
Orange	300	no	no
White	50	yes	no
Orange	490	yes	yes
White	510	yes	yes
Orange	490	no	no
White	450	no	no
Orange	40	no	no
Orange	200	yes	no
White	500	yes	yes
Green	560	yes	no
Orange	500	yes	?
White	50	yes	?

Training Data Set

Validation Data Set

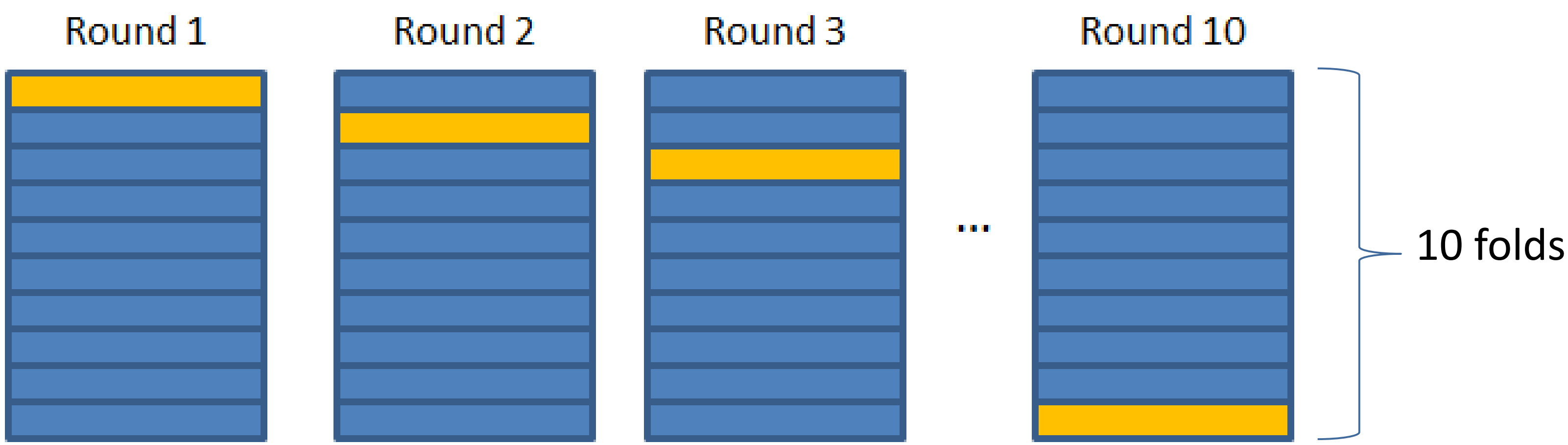
Unseen data set

Data Splits for Evaluations

2). N-folds Cross Evaluation → If your data is relatively small

Validation Set
Training Set

Usually we choose N as 5 or 10



Validation Accuracy:	93%	90%	91%	95%

Final Accuracy = Average(Round 1, Round 2, ...)

Predictive Models

- Common misunderstanding/mistakes
  - Which one we should choose? ➔ it depends on how large our data is
  - Some students use both of them, and produce the metrics, such as accuracy. They found that they could get a higher accuracy by using hold-out evaluation. Then they simply believe the hold-out evaluation is better ➔ wrong!

How it works: Build a Model

Instances

Features + classes

Training Data

Classification Algorithms

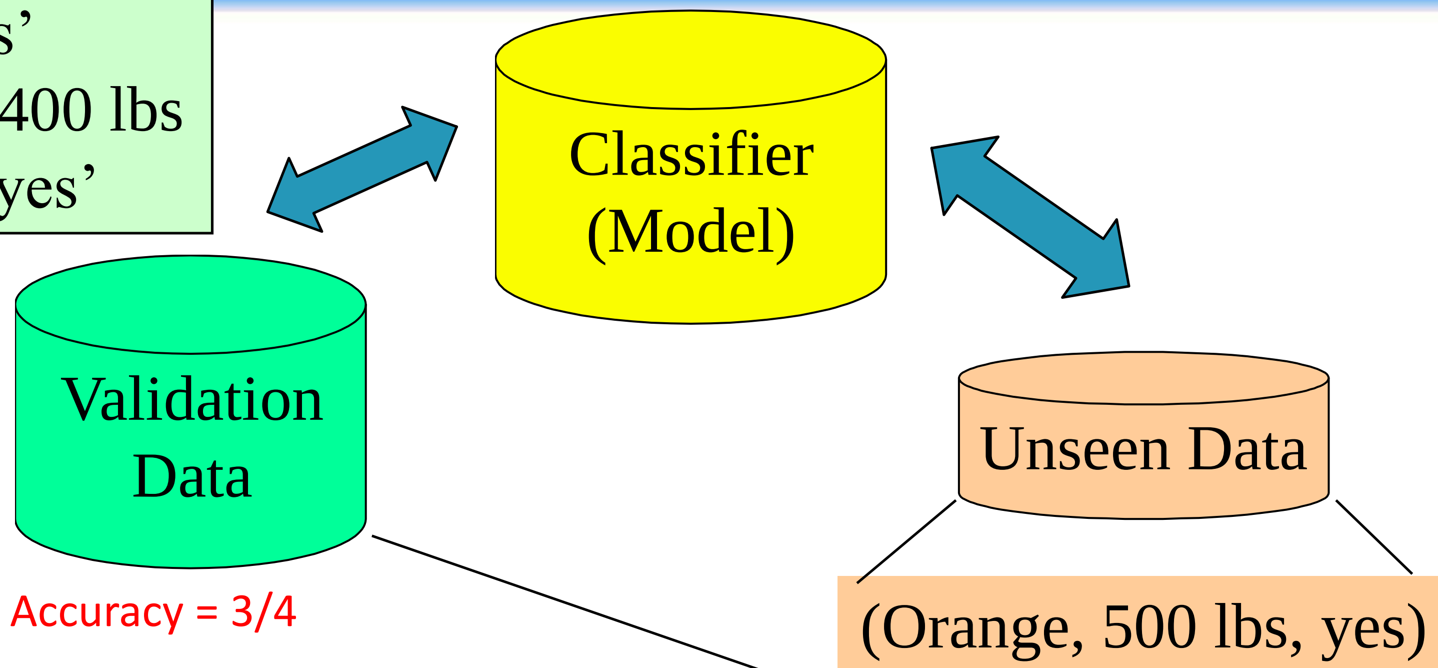
Classifier (Model)

Color	Weight	Stripes	Tiger?
Orange	300 lbs	no	no
White	50 lbs	yes	no
Orange	490 lbs	yes	yes
White	510 lbs	yes	yes
Orange	490 lbs	no	no
White	450 lbs	no	no

IF stripes = ‘yes’
AND weight > 400 lbs
THEN tiger = ‘yes’

How it works: Predictions

IF stripes = ‘yes’
AND weight > 400 lbs
THEN tiger = ‘yes’



Color	Weight	Stripes	Pred	Truth
Orange	40 lbs	no	no	no
Orange	200 lbs	yes	no	no
White	500 lbs	yes	yes	yes
Green	560 lbs	yes	yes	no

Tiger?

Yes

Predictive Models

- Evaluation Metrics
  - We have the truth and predictions, therefore we can always use specific metrics to evaluate the supervised learning
  - For linear regression, the variable to be predicted is numerical. We usually use error-based metrics
  - For classification, the variable to be predicted is categorical. We usually use accuracy-based metrics

Week 4 - 6

- Review on Statistical Basics
- Supervised Learning
- Predictive Models
- Simple Linear Regression
- Multiple Linear Regression
- Advanced Topics in Regression Models

Predictive Models

- Predictive Models
  - Predictive modeling is a process that uses related techniques (e.g., statistics, data mining and machine learning) to forecast outcomes.
  - Each model is made up of a number of predictors or factors, which are variables that are likely to influence future results.
  - You are able to exploit the impacts on the target/goals by different predictors, and try to capture the internal patterns for prediction purpose

Regression Analysis

Powerful method used to compute predictive analytics involving the analysis of several variables to predict or explain variations in another variable of interest.

- Examples:
  - **Real estate app**: Sale price of a property assessed through land value, improvement value and living space
  - **IS application**: Productivity rates as a measure of project effort, project duration, levels of experience with equipment and in project management, numbers of basic transactions and data entities.
  - **CAPM model in finance** – used to estimate asset's systematic risk. (Assets with higher betas are more sensitive to the market.)

Types of Regression Analysis

- Simple Linear Regression Analysis
  - Exploits relations between one dependent variable y and one independent variable x
  - $y = f(x) = \beta_0 + \beta_1 x + e$
- Multiple Linear Regression Analysis
  - Exploits relations between one dependent variable y and multiple independent variable $x_1, x_2, x_3, \ldots$
  - $y = f(x_1, x_2, x_3, \ldots, x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + e$

Types of Regression Analysis

- Multivariate Regression Analysis
  - Exploits relations between multiple dependent variable y and one independent variable x
  - $y_1, y_2, y_3, \ldots, y_m = f(x_1, x_2, x_3, \ldots, x_n)$
- In our class, we only focus on simple and multiple linear regression models. Multivariate regression analysis may be introduced in ITMD 529

Week 4 - 6

- Review on Statistical Basics
- Supervised Learning
- Predictive Models
- Simple Linear Regression
- Multiple Linear Regression
- Advanced Topics in Regression Models

Simple Linear Regression

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable
- Dependent variable: the variable we wish to predict or explain
- Independent variable: the variable used to predict or explain the dependent variable

Simple Linear Regression

- Data are pairs on (y_i, x_i), i.e., two columns y and x (paired!!)
- Data show a linear association between Y and X → same rate of change in Y for any one-unit change in X.
- The observations on y satisfy the following model

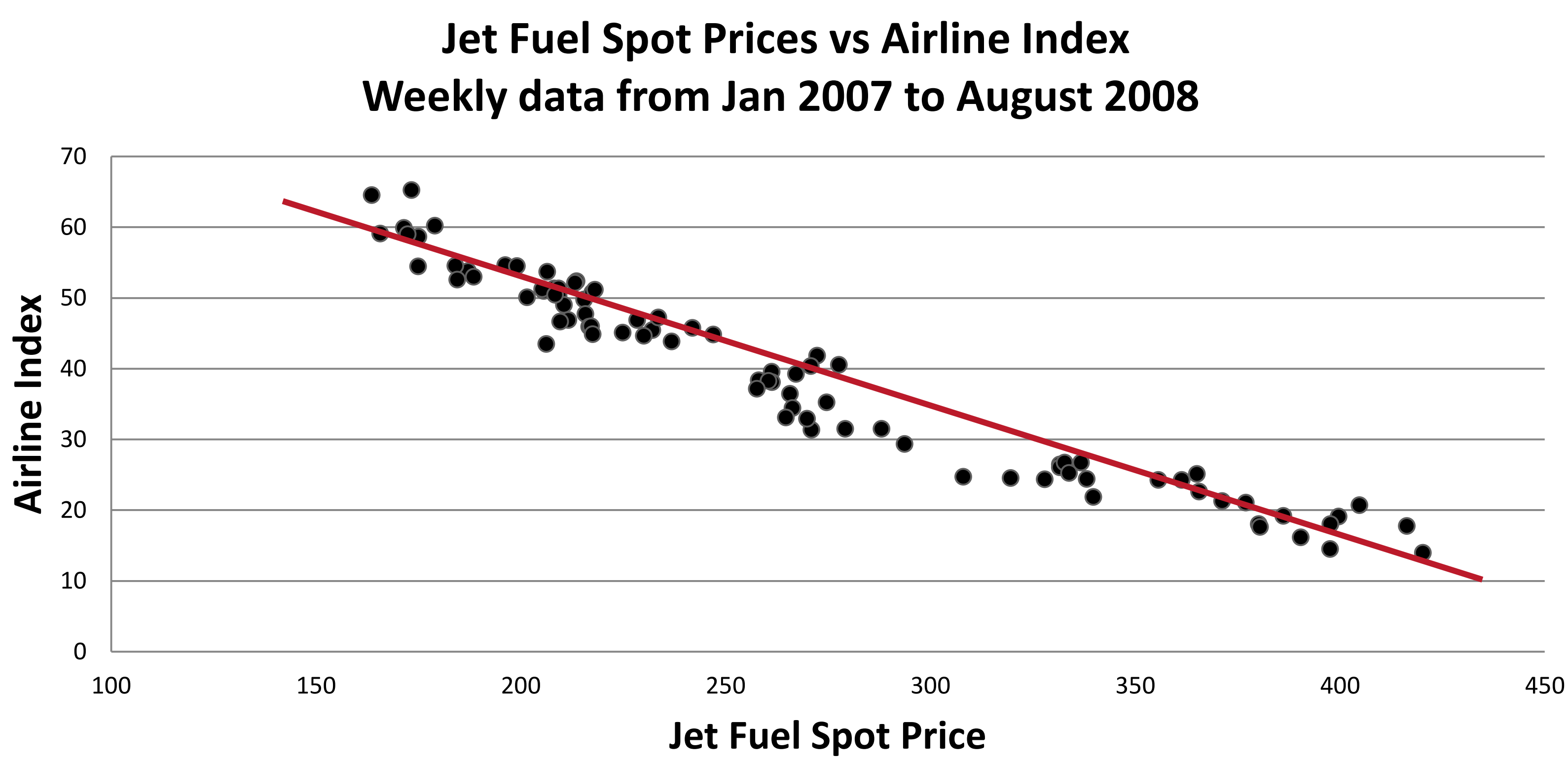
Data = prediction + error

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- The terms e_i are the model errors – they are measured by the residuals!
  - Error = diff between observed and (unobserved) true value
  - Residual = diff between observed and predicted value

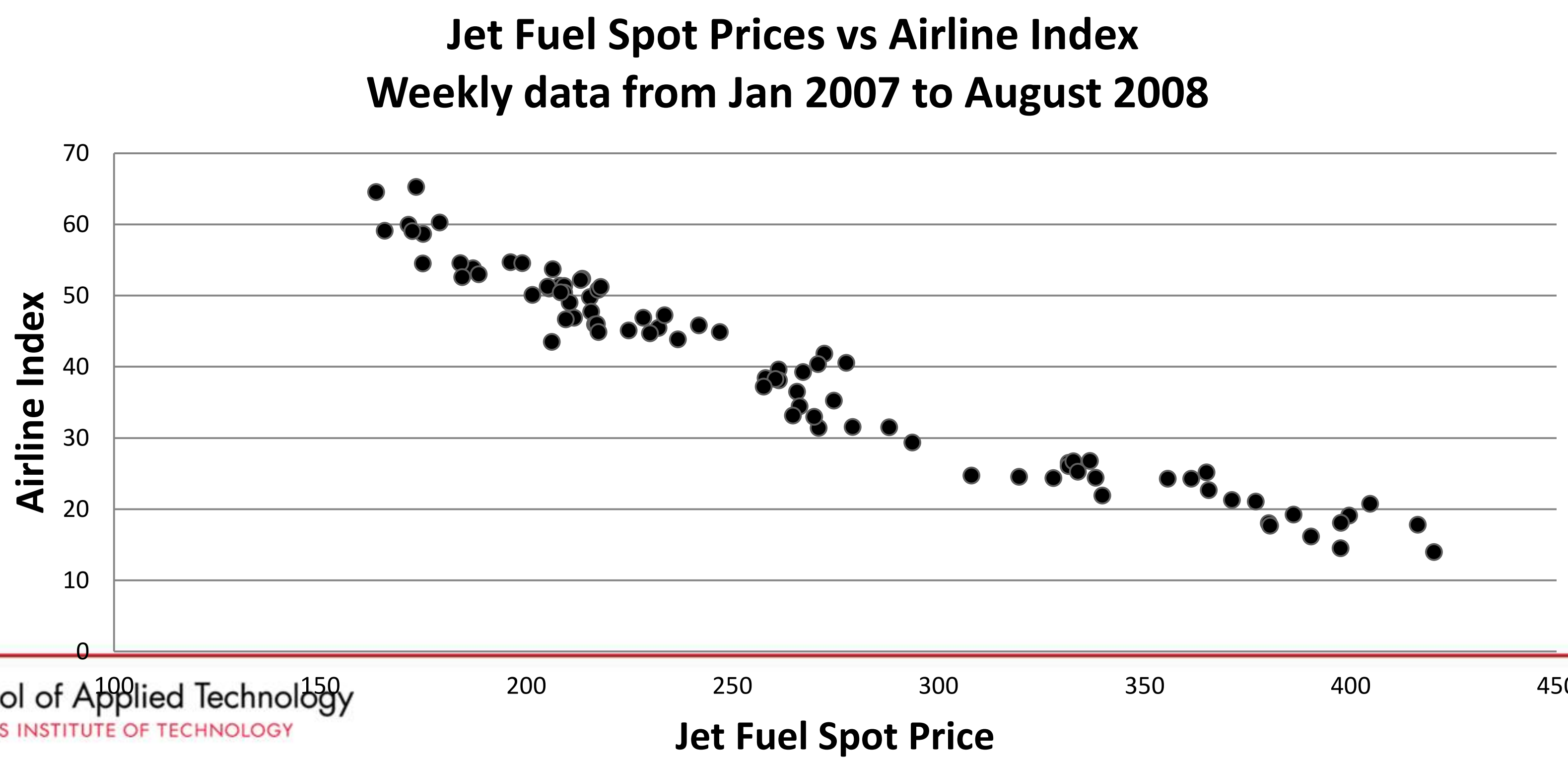
Simple Linear Regression

- Simple Linear Regression Model = Straight Line Model
- A straight line can be used to model data points x_i and y_i



Simple Linear Regression

- A scatter plot can be used to show the relationship between two variables
- Correlation analysis is used to measure the strength of the association (linear relationship) between two variables



Simple Linear Regression

- The first-order simple linear regression model

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

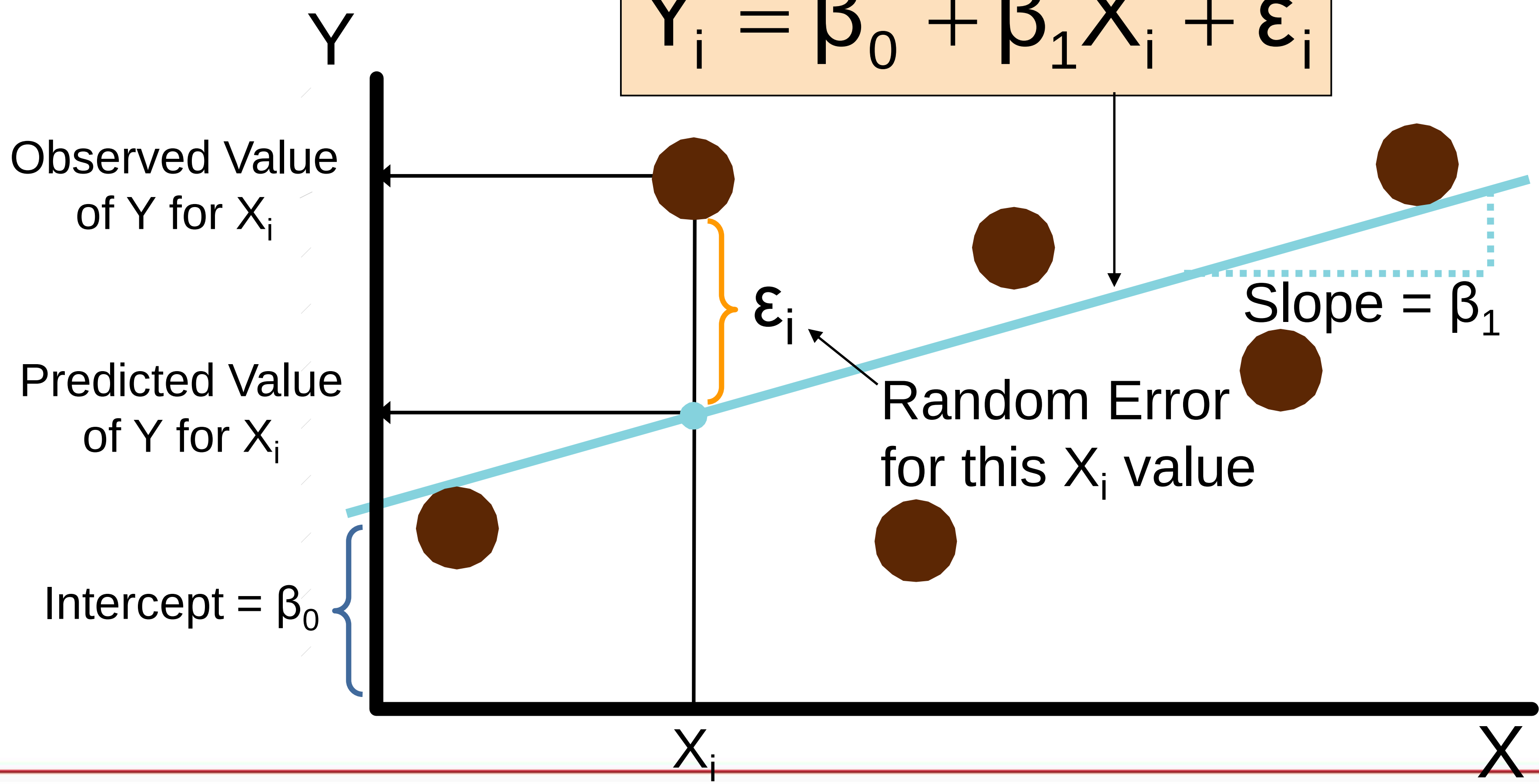
Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Simple Linear Regression

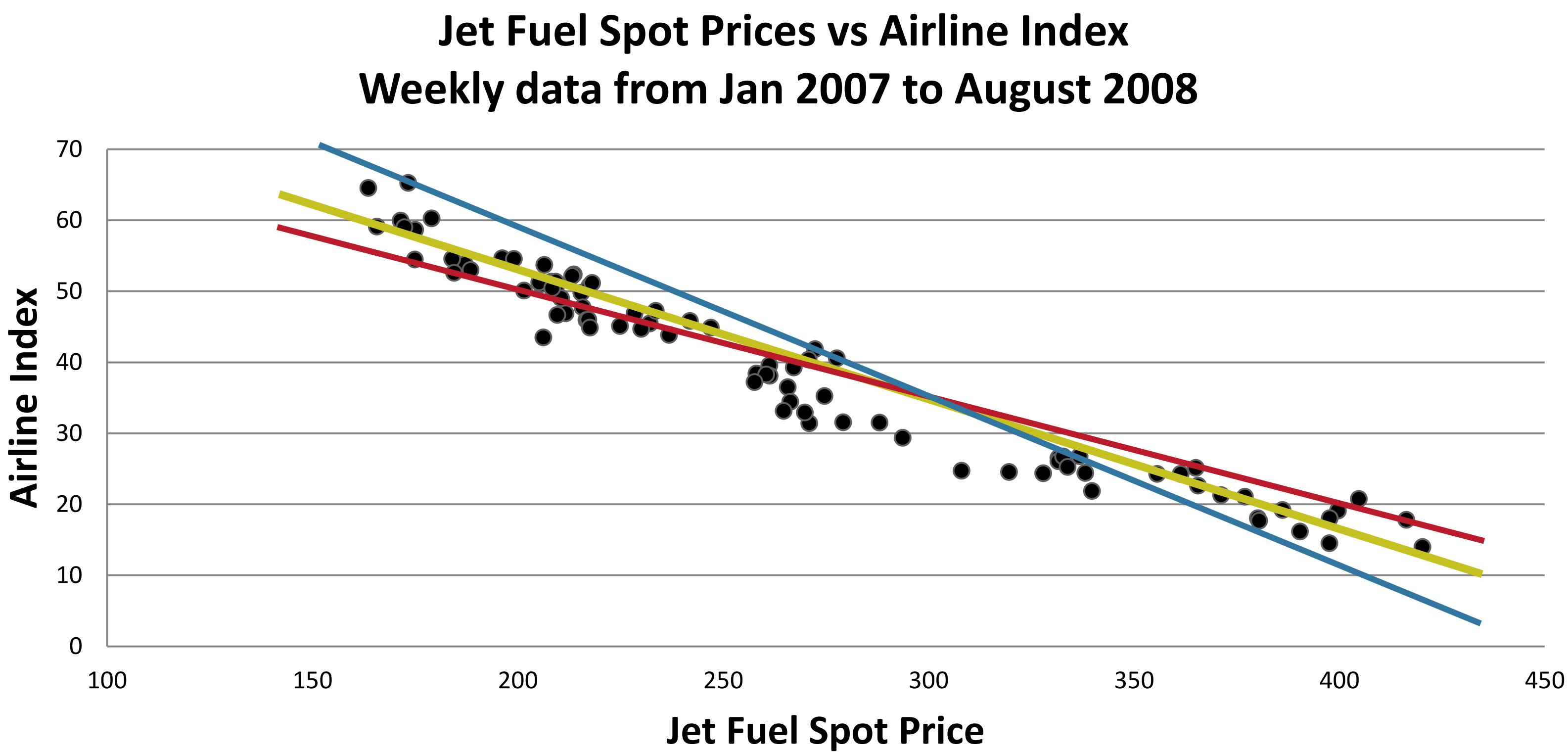
- The first-order simple linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Simple Linear Regression

- There could be multiple lines, which one is the best?



Simple Linear Regression

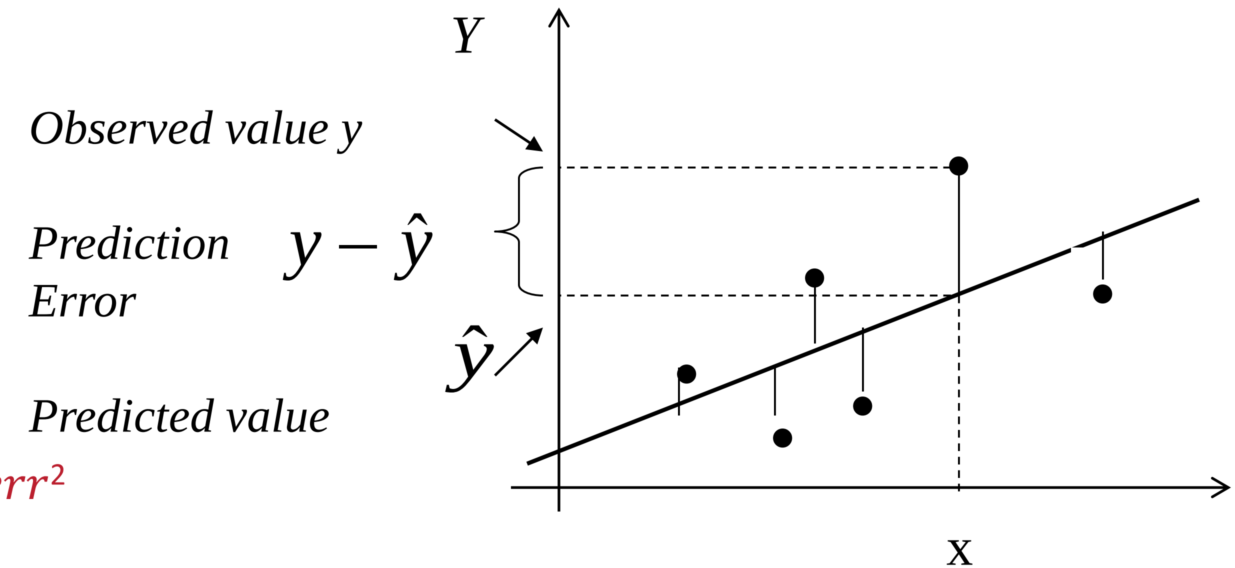
The regression line is used to predict the response $\hat{y}$ at any given x. Regression line minimizes the vertical distances between observed y and the point on the line. The accuracy of the prediction depends on how much spread out the observations are around the line.

Error of prediction, err = $y - \hat{y}$

Squared error = err^2

Sum of error (SE) = $\sum err$

Sum of squared errors (SSE) = $\sum err^2$



Simple Linear Regression

Regression line is also known as Least Square Regression Line, because we use Least Squares as the optimization goal. **The parameter estimates** are the values for β's that minimize the sum of the prediction square errors:

$$\min_{\beta_0,\beta_1}\sum_i (y_i-\hat{y}_i)^2=\min_{\beta_0,\beta_1}\sum_i [y_i-(\beta_0+\beta_1 x_i)]^2$$

The optimal values for $\hat{\beta}_0$ *and* $\hat{\beta}_1$ are found using standard optimization theory (~ solving first derivatives equal to zero)

Simple Linear Regression: Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

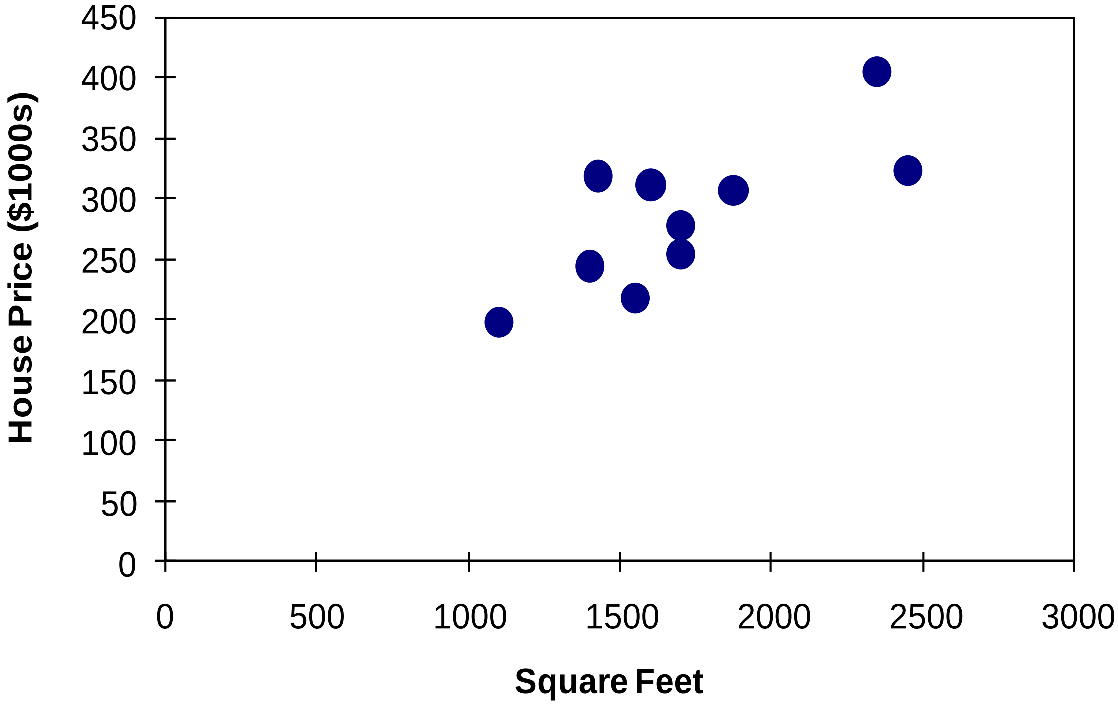
- A random sample of 10 houses is selected
  - Dependent variable (Y) = house price in $1000s
  - Independent variable (X) = square feet

Simple Linear Regression: Example

House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Simple Linear Regression: Example

House price model: Scatter Plot



$$\hat{Y}_i = b_0 + b_1 X_i$$

Simple Linear Regression: Example

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{houseprice} = 98.24833 + 0.10977(squarefeet)$$

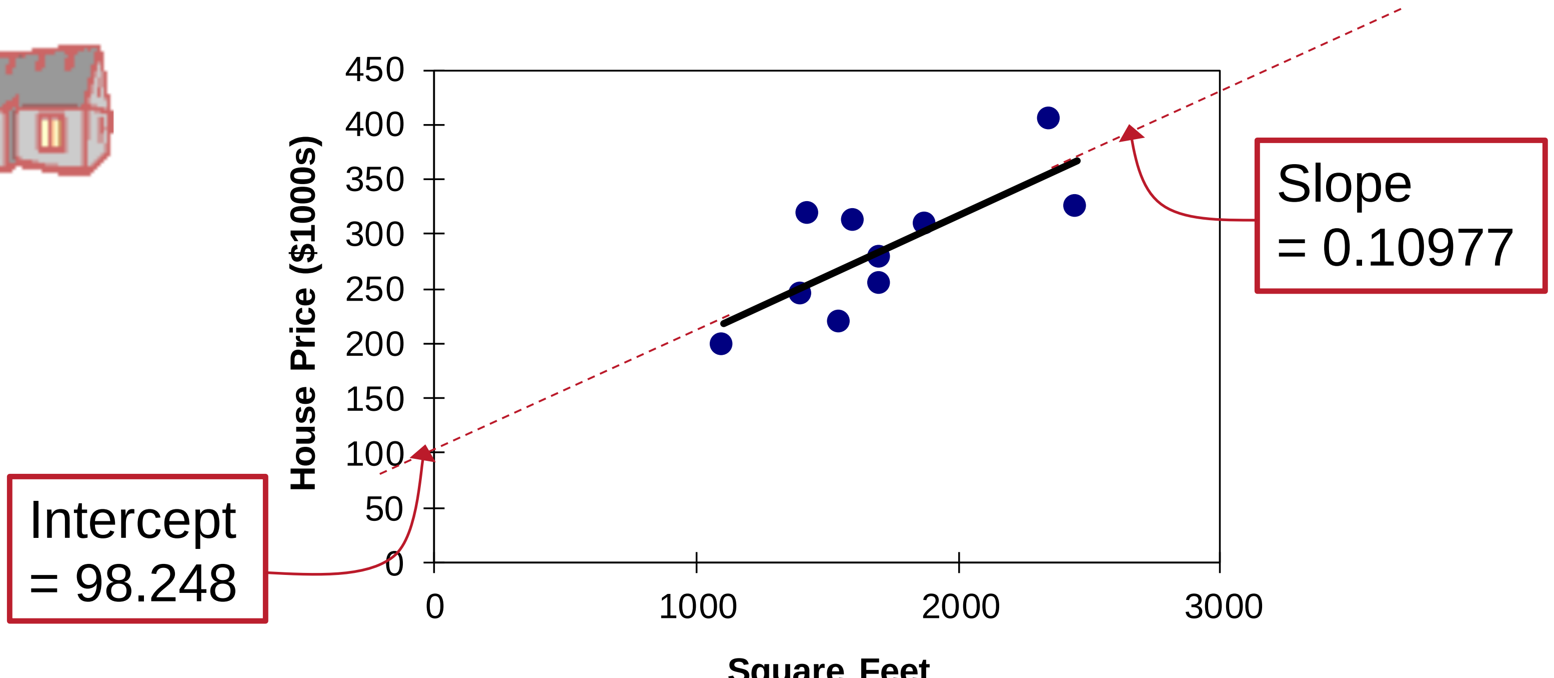
$$\hat{Y}_i = b_0 + b_1 X_i$$

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Simple Linear Regression: Example

House price model: Scatter Plot and Prediction Line



$$\widehat{\text{houseprice}} = 98.24833 + 0.10977(\text{squarefeet})$$

Simple Linear Regression: Example

How to interpret a simple linear regression model?

$$\widehat{\text{houseprice}} = 98.24833 + \boxed{0.10977}(\text{squarefeet})$$

- b_1 estimates the change in the mean value of Y as a result of a one-unit increase in X
  - Here, $\boxed{b_1 = 0.10977}$ tells us that for every increase of 100 square feet, the average house price will be increased by ?????

Simple Linear Regression: Example

How to interpret a simple linear regression model?

$$\widehat{\text{houseprice}} = \boxed{98.24833} + 0.10977(\text{squarefeet})$$

- b_0 is the estimated mean value of Y when the value of X is zero (if X = 0 is in the range of observed X values)
- In this application, due to that a house cannot have a square footage of 0, b_0 has no practical application

Simple Linear Regression: Example

Use your model for prediction purpose

$$\widehat{\text{house price}} = \boxed{98.24833} + 0.10977(\text{square feet})$$

Predict the price for a house with 2000 square feet:

$$\text{house price} = 98.25 + 0.1098\,(\text{sq.ft.}) = 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Other Steps

- Goodness of Fit Test
  - F-test is used to evaluate whether all of the x variables are useful and which x variables are influential
- Residual Analysis
  - To validate whether the residuals meet the requirements, whether there is problems in the model
- Evaluations and Predictions
  - To calculate metrics and evaluate multiple models based on the test set

Week 4 - 6

- Review on Statistical Basics
- Supervised Learning
- Predictive Models
- Simple Linear Regression
- Multiple Linear Regression
- Advanced Topics in Regression Models

More general regression model

Consider one Y variable and **k** independent variables X_i, e.g. X_1, X_2, X_3.

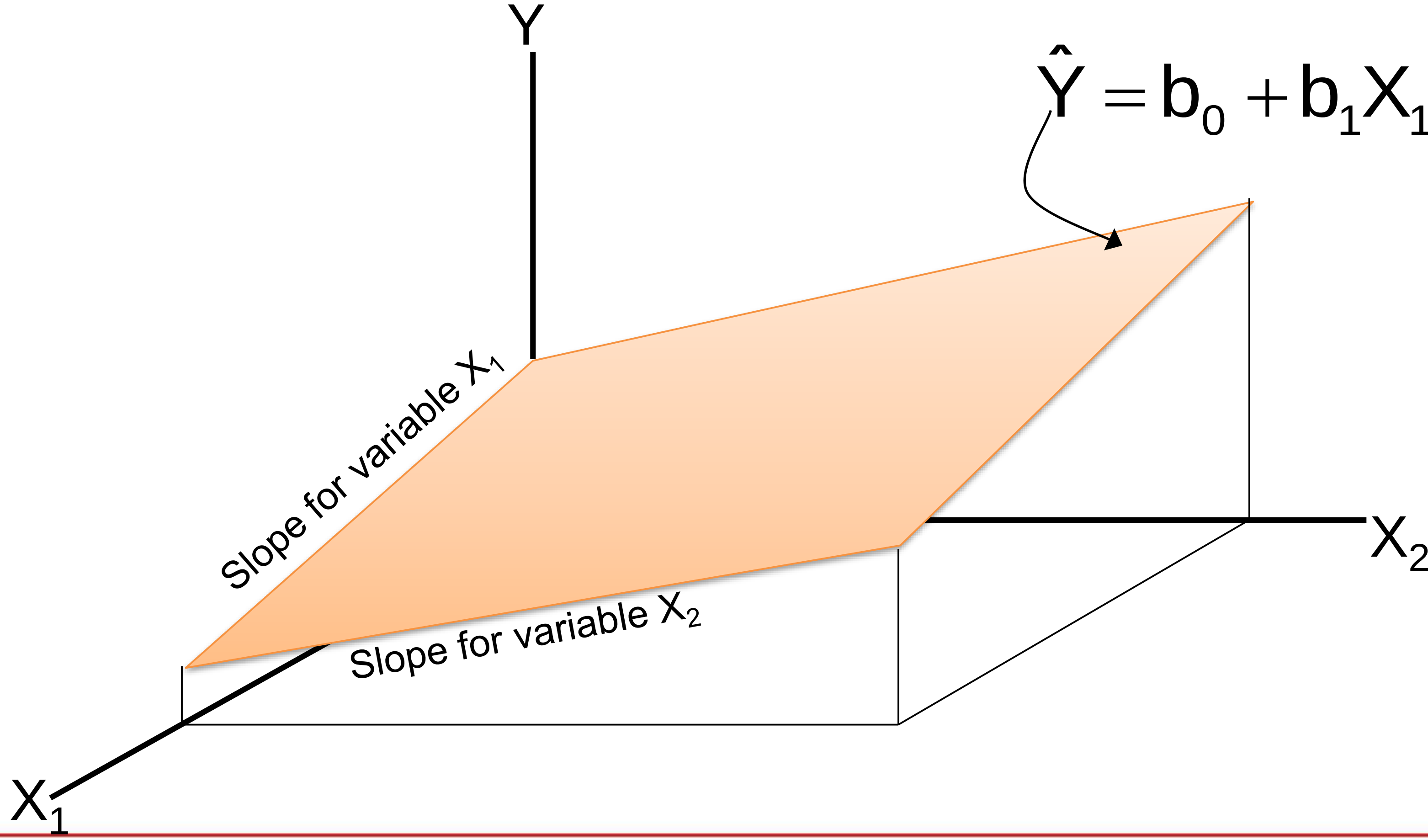
- Data on n tuples **(y_i, x_{i1},x_{i2},x_{i3}).**
- Scatter plots show linear association between Y and the X-variables
- The observations on y can be assumed to satisfy the following model

$$\underbrace{y_i}_{\boxed{\text{Data}}} = \underbrace{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}}_{\boxed{\text{Prediction}}} + \underbrace{e_i}_{\boxed{\text{error}}} \quad for\ i = 1, \ldots, n$$

A Multiple Regression Model with X1 and X2

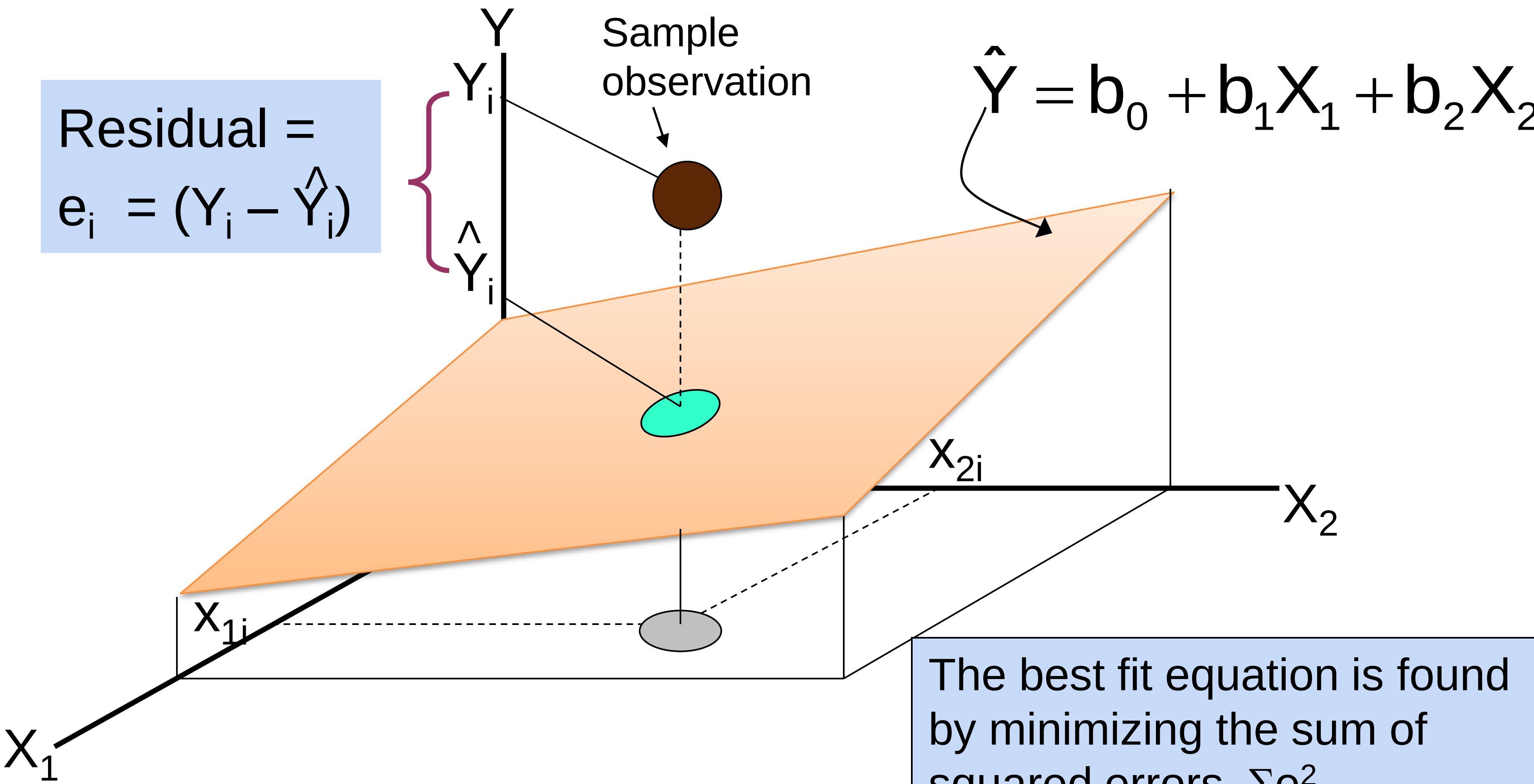
Two variable model

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$



A Multiple Regression Model with X1 and X2

Two variable model



The best fit equation is found by minimizing the sum of squared errors, Σe^2

Multiple Linear Regression

Important Steps in Multiple Linear Regression

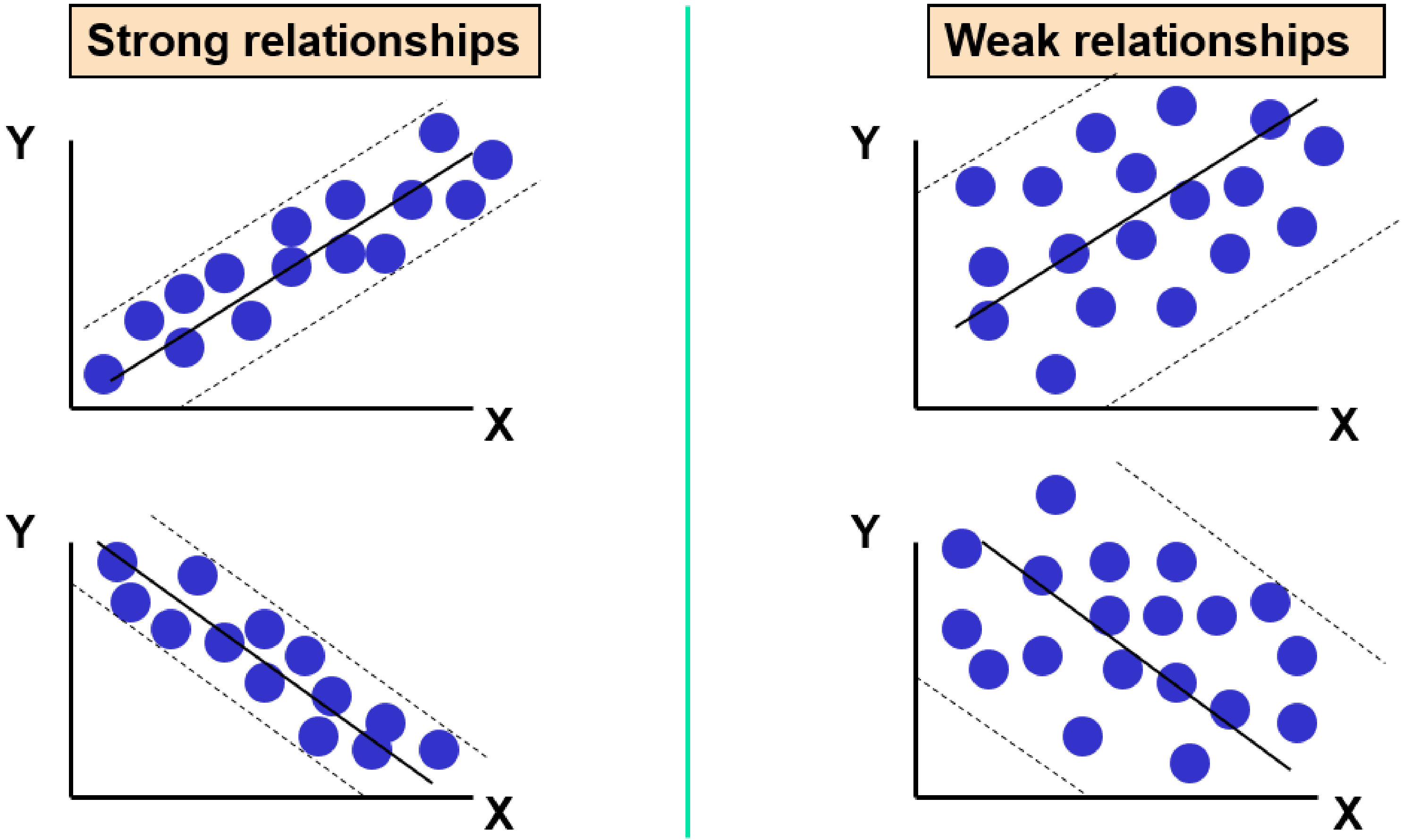
- Data Splits – build a model based on train set, and evaluate it based on the test set
- Determine linear relationship between y and x variables
- Build a multiple linear regression model by parameter estimates
- Goodness of fit test
- Residual analysis – the last step to tell your model is qualified
- Interpret the performance of the training process
- Evaluations and predictions – evaluate it based on test set

Multiple Linear Regression

How to determine x and y have a linear relationship?

- Draw a scatter plot Y and X-variables to observe the straight line pattern
- Calculate the correlations

Scatter Plot



Correlation Coefficient

- It implies a strong correlation between X and Y.
- The correlation coefficient *r* is the measure of the **linear** association between two variables.

The correlation coefficient is defined as

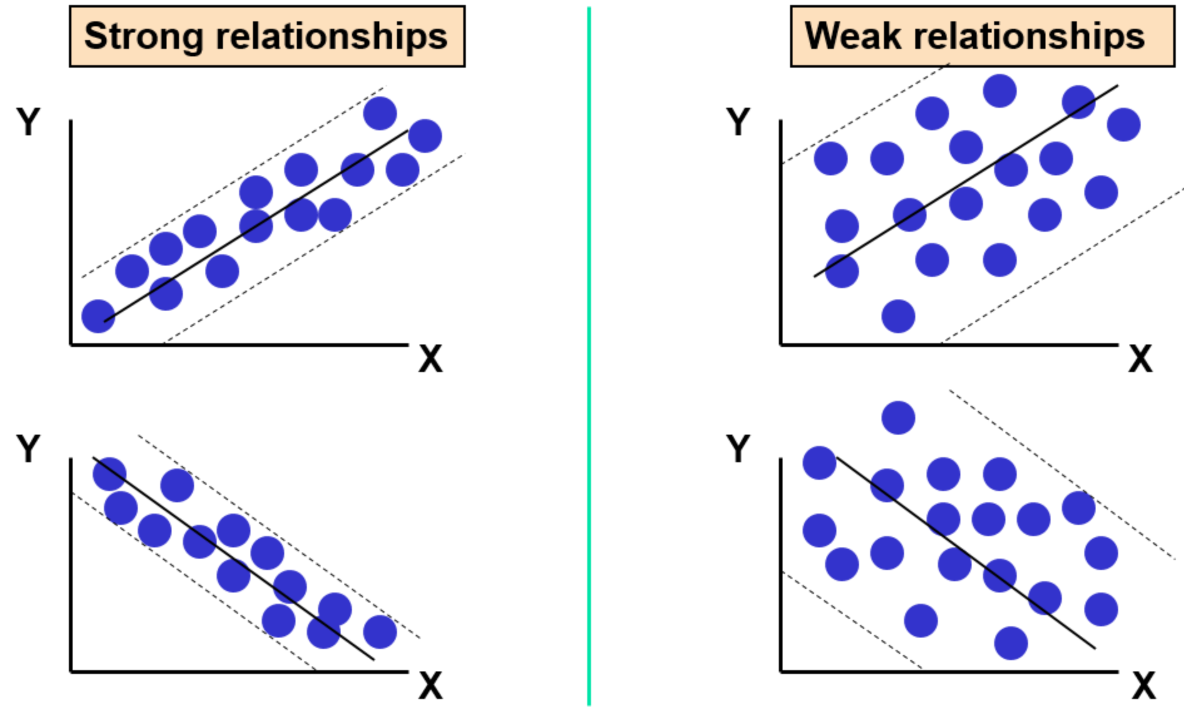
$$r(X,Y) = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

- Where X has average x-bar and standard deviation s_x, and Y has average y-bar and standard deviation s_y.

Multiple Linear Regression

- It is much more obvious from the visualization

1. r lies between -1 and +1.
   r > 0 indicates a positive linear relationship.
   r < 0 indicates a negative linear relationship.
   r = 0 indicates no linear relationship.
   r = ±1 indicates perfect linear relationship.
2. The larger the absolute value of r, the stronger the linear relationship.



What if there are no linear relationship?

What if there are no linear relations or not clear linear relations between Y and X?

Solutions:

- You may need to perform transformations on either or both of the Y and X variables. Usually, we try X first, then Y.
  - Square transformation: X' = X * X
  - Log transformation: X' = logX
  - Inversion transformation: X' = 1/X
- If transformation does NOT work, you may need to ignore the variable X. But you may consider polynomial regressions or non-linear analytics models.

Multiple Linear Regression

Important Steps in Multiple Linear Regression

- Data Splits – build a model based on train set, and evaluate it based on the test set
- Determine linear relationship between y and x variables
- Build a multiple linear regression model by parameter estimates
- Goodness of fit test
- Residual analysis – the last step to tell your model is qualified
- Interpret the performance of the training process
- Evaluations and predictions – evaluate it based on test set