Chapter 6

BUILDING MODELS FROM DATA

Classification and Prediction.

Building models. Why?

- **Relationship exists between variables.**
- Formal **ways to describe**, **encode**, and **test** if and how **one or more variables relate to others** is to **build and evaluate models** from the data.
- **Models**:
  - **describe important relationships** in the data:
    - **Strength and direction**—positive or negative—of the relation.
    - Encode **linear and nonlinear relationsh**ips in the data
  - Can also be used to **confirm a hypothesis** about relationships.
  - Help to **summarize and understand the data**.
  - Making **prediction**

Building models. Why? example

A **data set** of **historical purchases** along with **customer geographical** and **demographic data** (such as the customer's age, location, salary, and so on) could be collected and used to generate a model that **encodes what type of products clients' purchase**.

Once the model is built, it could be used to **identify** from a list of **potential clients** those **most likely to make a purchase**, and **customers** on this **prioritized list** could be **targeted** with marketing material or other promotions.

Building models. How?

- How can we build models from data sets?
- **A model** is usually built to **predict values** for a **specific variable**.
  - For example:
    - A data set composed of *historical data* containing **attributes of pharmaceuticals** and **their observed side effects** is collected.
    - **Model:** *Predict the side effects from the pharmaceuticals' attributes*.
- A **variable that a model is to predict** is often referred to as **y-variable or response variable or dependent variable.**
- **Variables that will be encoded** in the model and **used in predicting** this response are referred to as the **x-variables or the independent variables.**

Example.

TABLE 4.1 Example of Telecommunications Data Used to Build a Model

ID	Month	Age	Income	Customer length	Gender	Monthly calls	Service requests	Churn
A	January	45	$72k	36	Female	46	1	0
B	March	27	$44k	24	Male	3	5	1
C	July	51	$37k	47	Male	52	0	0
D	February	17	0	16	Female	62	1	1
E	December	45	$63k	63	Female	52	0	0
F	October	24	$36k	24	Male	72	1	0
G	March	39	$48k	5	Male	36	0	0
H	June	46	$62k	17	Male	1	0	1
...	...	...	...	...	...	...	...	...

Example.

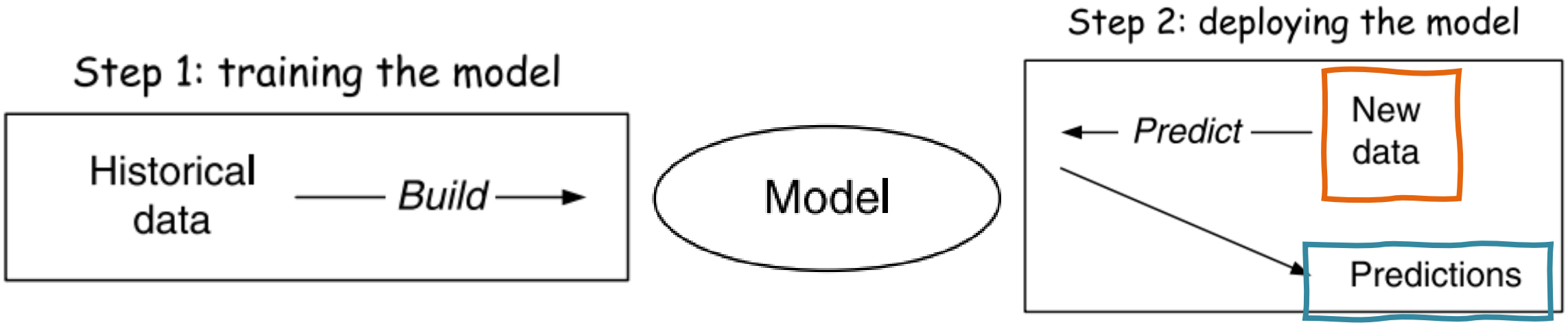
TABLE 4.2 Data Collected on Customers for the Current Month

ID	Month	Age	Income	Customer length	Gender	Monthly calls	Service requests	Churn
a	May	52	$84k	52	Female	52	0	?
b	May	26	$28k	14	Male	12	2	?
c	May	64	$59k	4	Male	31	1	?
...	...	...	...	...	...	...	...	...

TABLE 4.3 Customers Predicted to Change Services this Month, and a Measure of the Likelihood of Switching

ID	Month	Age	Income	Customer length	Gender	Monthly calls	Service requests	Predicted churn	Churn probability
a	May	52	$84k	52	Female	52	0	0	0.33
b	May	26	$28k	14	Male	12	2	1	0.74
c	May	64	$59k	4	Male	31	1	1	0.88
...	...	...	...	...	...	...	...	...	...

Process of generating and using a prediction model.



Variables.

A **generalized format** for the **model** is shown where some **function** of the **independent variables (x_i)** is **used to predict** the **response (y),** which in this case is MPG.

Response

Independent variables

Name	MPG (y)	Cyclinders (x1)	Displacement (x2)	Horsepower (x3)	Weight (x4)	Acceleration (x5)
Chevrolet Chevelle Malibu	18	8	307	130	3504	12
Buick Skylark 320	15	8	350	165	3693	11.5
Plymouth Satellit	18	8	318	150	3436	11
AMC Rebel SST	16	8	304	150	3433	12
Ford Torino	17	8	302	140	3449	10.5

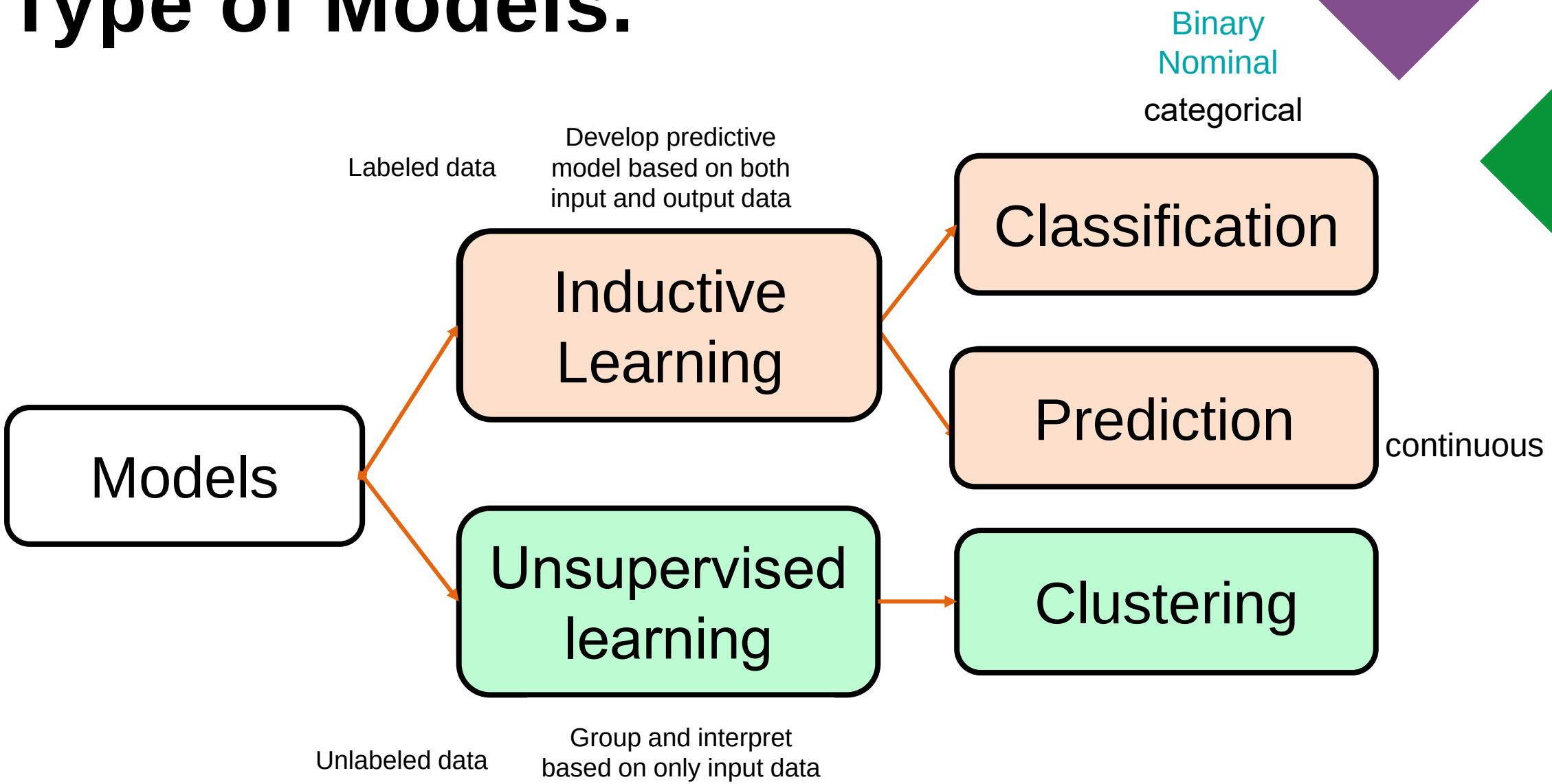
$$y = f(x_i)$$

Model

Model what?

Predicting car's fuel efficiency.

Type of Models.



Adapted from: https://www.sciencedirect.com/science/article/pii/S2210832717301485

Type of Models.

Models built to predict **categorical variables** (such as a binary variable or a nominal variable) are referred to as **classification models**

Classification trees
k-Nearest Neighbors
Logistic regression
Naïve Bayes classifiers
Neural networks
Rule-based classifiers
Support vector machines

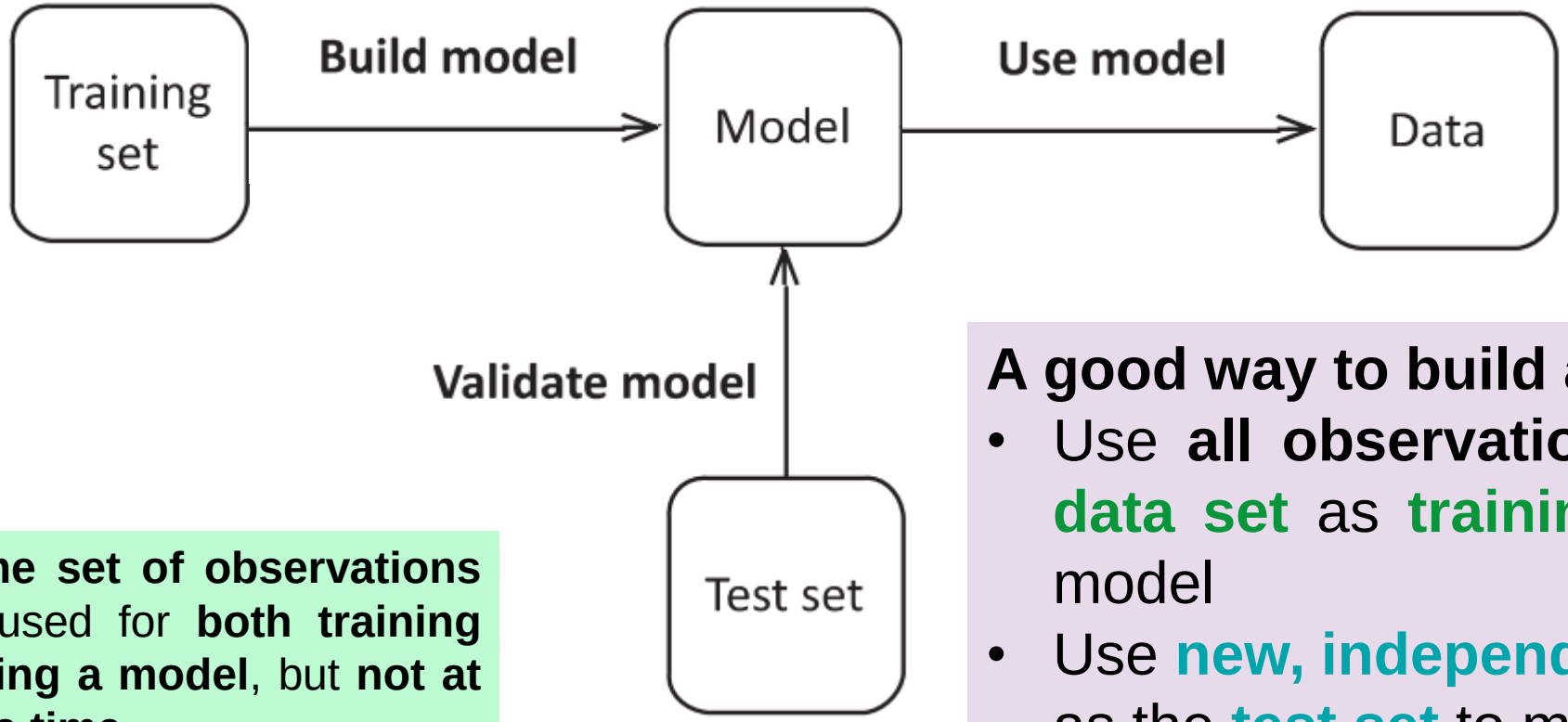
Models that predict **continuous variables** are called **regression models or prediction models**

Regression trees
k-Nearest Neighbors
Linear regressions
Neural networks
Nonlinear regression
Partial least squares
Support vector machines

Selection of independent/response variables for a good model.

- Use as **few independent variables** as possible.
- Use **different visualizations** and metrics in understanding relationships in the data, such as scatterplots, contingency tables, t-tests, Chi-Square tests, for **prioritizing candidate independent variables** when building a model, especially where there are many variables to consider.
- Consider **relationship between the independent variables**:
  - Combinations of variables that have ***strong relationships*** to each other should be **avoided** since they will be essentially encoding the **same relationship** to the response.
- Use **derived variables**, that is, a new variable that is a function of one or more variables.

Generating models from data sets



The **same set of observations** can be used for **both training** and **testing a model**, but **not at the same time**.

A good way to build and test a model:

- Use **all observations** in the **original data set** as **training set** to build the model
- Use **new, independent observations** as the **test set** to measure accuracy.

FIGURE 6.2 Use of training sets to build models and test sets to assess their performance.


AFTER THE BREAK

#EXAMREMINDER

ACADEMIC CALENDAR: 2021-22 – II SEM M Tech/MCA

Events/Activities	Start Date	End Date	Mode
End Semester Examination (First semester)	**February 07, 2022**	**February 18, 2022**	**ONLINE**
Semester Break	February 19, 2022	February 27, 2022	-----
Commencement of Second semester classes	**February 28, 2022**		**OFFLINE**
Last Instructional day	June 18, 2022		-----
End Semester Examination (Second semester)	**June 23, 2022**	**July 04, 2022**	**OFFLINE**
Commencement of, Project work for second year M Tech / III semester Classes for MCA*	**July 11, 2022 (M Tech)**		**OFFLINE***
	July 25, 2022* (MCA)		

Classification and Prediction: Prediction Problems

(Numeric)

- **Classification** is about determining a (categorial) class (or label) for an element in a dataset.
- Model used to classify **unknown value called a classifier**.
- A **classifier** is constructed from a training set composed of the records of databases and their corresponding class names

- Prediction is about predicting a **missing/unknown** element (continuous value) of a dataset.
- Model used to predict **unknown value called a predictor.**
- A **predictor** is also constructed from a training set and its accuracy refers to how well it can estimate the value of new data.

Classification and Prediction: Prediction Problems

Typical applications

- Credit/loan approval
- Medical diagnosis: if a tumor is cancerous or benign
- Fraud detection: if a transaction is fraudulent
- Target marketing

Back again to supervised vs unsupervised.

- **Supervised learning (classification)**
  - **Supervision:** The training data (observations, measurements, etc.) are accompanied by **labels (known)** indicating the class of the observations
  - New data is classified based on the training set
- **Unsupervised learning (clustering)**
  - The **class labels** of training data is **unknown**
  - Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Testing Model Accuracy

- Using different data ensures that the model has not **overfitted** the training data.
- **Test data:**
  - **Before** data set is used to **train any models**, a set of data (say one-third) selected randomly is set aside for the sole purpose of testing the quality of the results.
  - These **observations will not be used in building the model**, but they will be used with any built model to test the model's predictive performance.
  - **Ideal case: Test set is only used for model assessment**.
  - However, in practical situations, **not enough data** available.

Testing Model Accuracy: Test set is small (1)

Cross-validation: k-fold partitioning method

- Original data set is divided into **k equally sized partitions**.
- Model is measured **k times.**
- **First iteration:**
  - One of the partitions is selected as the test set, and
  - Remaining partitions comprise the training set.
  - Model is tested and an **accuracy score** is generated.
- **Each subsequent iteration:**
  - A partition different from any already used as test set selected as the test set
  - Remaining partitions become the training set.
  - **Another score** is calculated.
- At the end of this process, the **accuracy of the model** is based on the **average of the k scores.**

Testing Model Accuracy: Test set is small (2)

Cross-validation: k-fold partitioning method example:

- Suppose data set portioned into **10 partitions** where each partition consists of observations **randomly selected** from the data set.
- In **each of the 10 iterations**, designate one partition (**10% of data set**) as the **test set** and the **other 9 partitions** (**90% of data set**) as the **training set.**
- At the **end of the 10 iterations**, an **average of 10 scores** is used to assess the model's accuracy.

- **Extreme k-fold partitioning: k** number of observations in data set and **each partition** contains a **single observation**. This is a **cross-validation method** known as **leave-one-out**.

Testing Model Accuracy: BIAS

- In cross-validation, **each partition** will have been **used as a test set**
  - Or, **every observation** in data set will have been **tested once.**
- ***Why?***
  - To **avoid introducing bias** into a model and to ensure that **a prediction** will be calculated for **every observation** in the data set.

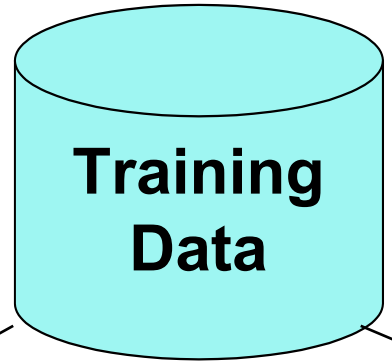
Bias is a **measure** of the **model's accuracy** and indicates **how close** the **predictions of the response value** made by the model are to the **actual response value** of new observations

Testing Model Accuracy: BIAS?

- Can be introduced when models become overly complex by optimizing the model for just the training set used to build the model.
- When **performance** is tested for these overtrained models against either a separate test set or through cross-validation: **poor performance**
- In cross-validation methods, bias can be introduced when:
  - The training sets overlap (some observations are used more than once), or
  - The combined training sets do not cover the data set (some observations are never used).

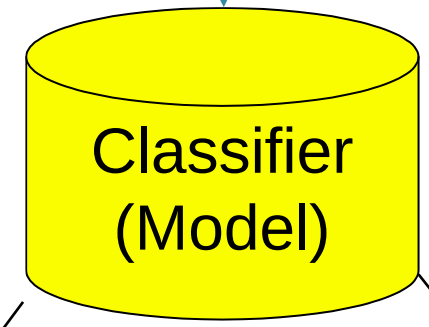
CLASSIFICATION MODELS:

Phase 1: Model Construction or Training



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

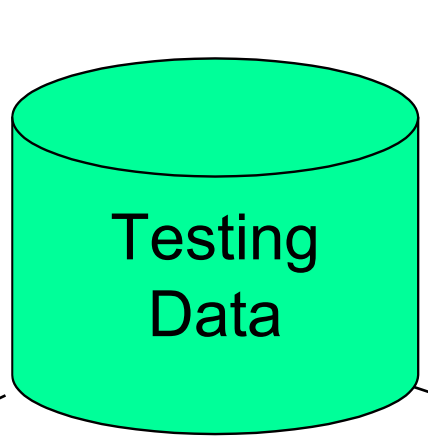
Classification Algorithms

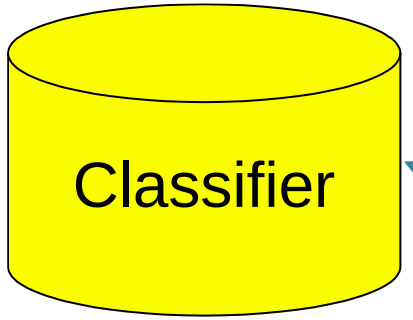


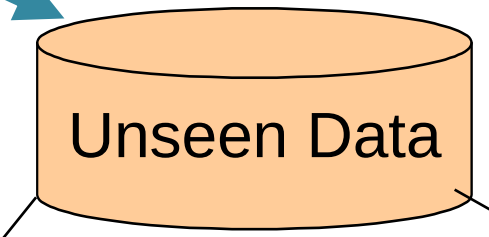
IF rank = ‘professor’
OR years > 6
THEN tenured = ‘yes’

CLASSIFICATION MODELS.

Phase 2: Use model in Prediction, Applying







NAME	RANK	YEARS	TENURED	Predicted
Tom	Assistant Prof	2	no	no
Merlisa	Associate Prof	7	no	**yes**
George	Professor	5	yes	yes
Joseph	Assistant Prof	7	yes	yes

(Jeff, Professor, 4)

Tenured?

Classification Models: Assess Performance?

Look at the **results of applying the models** (such as the results from a test set or the cross-validation results) and **determine** how many observations are **correctly or incorrectly classified**.

Accuracy or concordance of a model is based on the proportion or percentage of correctly predicted observations in comparison to the whole set.

For example:
If test set contained **100 observations** and the model predicted **78 correctly**, that is, it predicted **22 incorrectly**, then the **concordance** would be **78/100 or 78%.**

Classification Models: Type?

Model to **predict a binary response**, where a **true response** is coded as **1** and a **false response** is coded as **0.**

True response encoded as **1** when **there is evidence** for an oil deposit

False response as **0** When **there is no evidence** for an oil deposit

After Break Time

Selecting the best method.

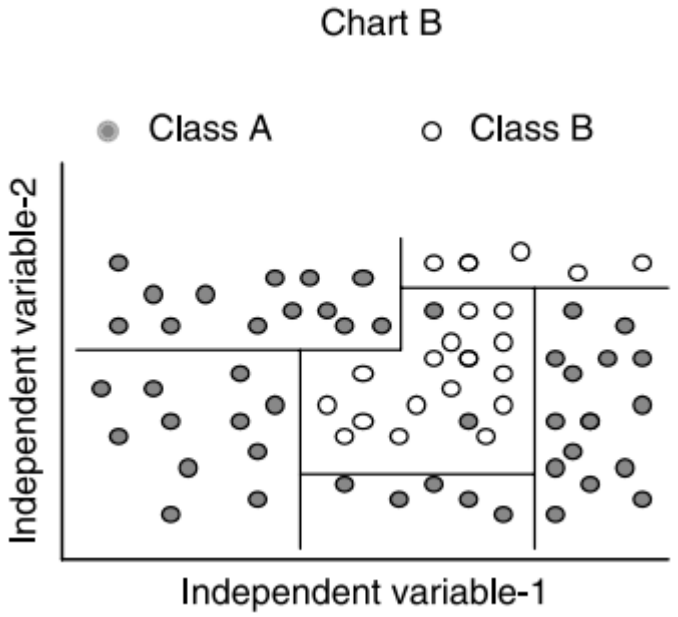
Understanding of the data

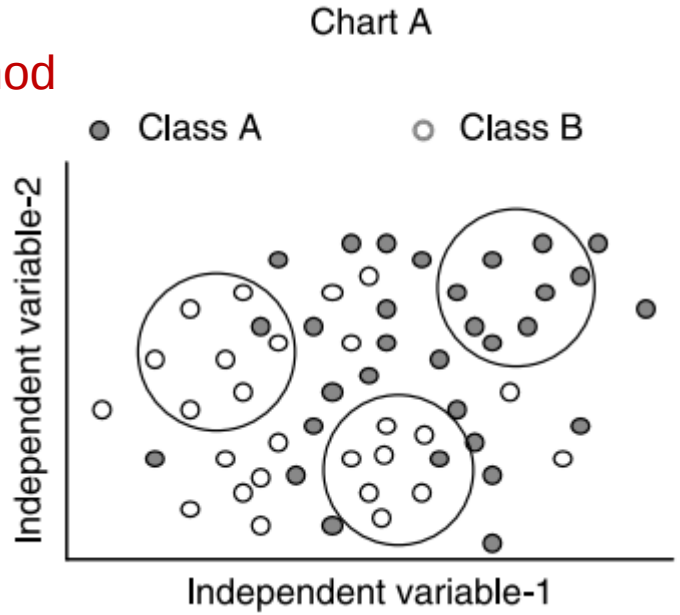
How different methods operate?

kNN method

Classification trees

Discriminant analysis





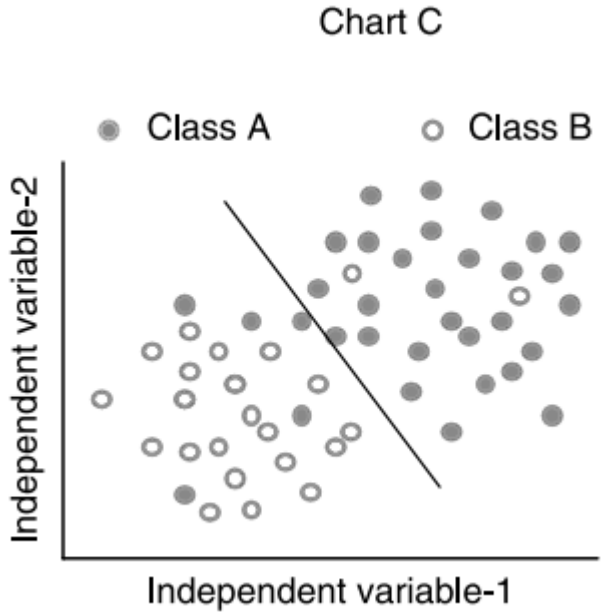


Figure 4.3 Illustration of different classification modeling approaches

Selecting the best method.

Understanding of the data

How different methods operate?

Linear relationship

nonlinear relationship

regression trees

kNN

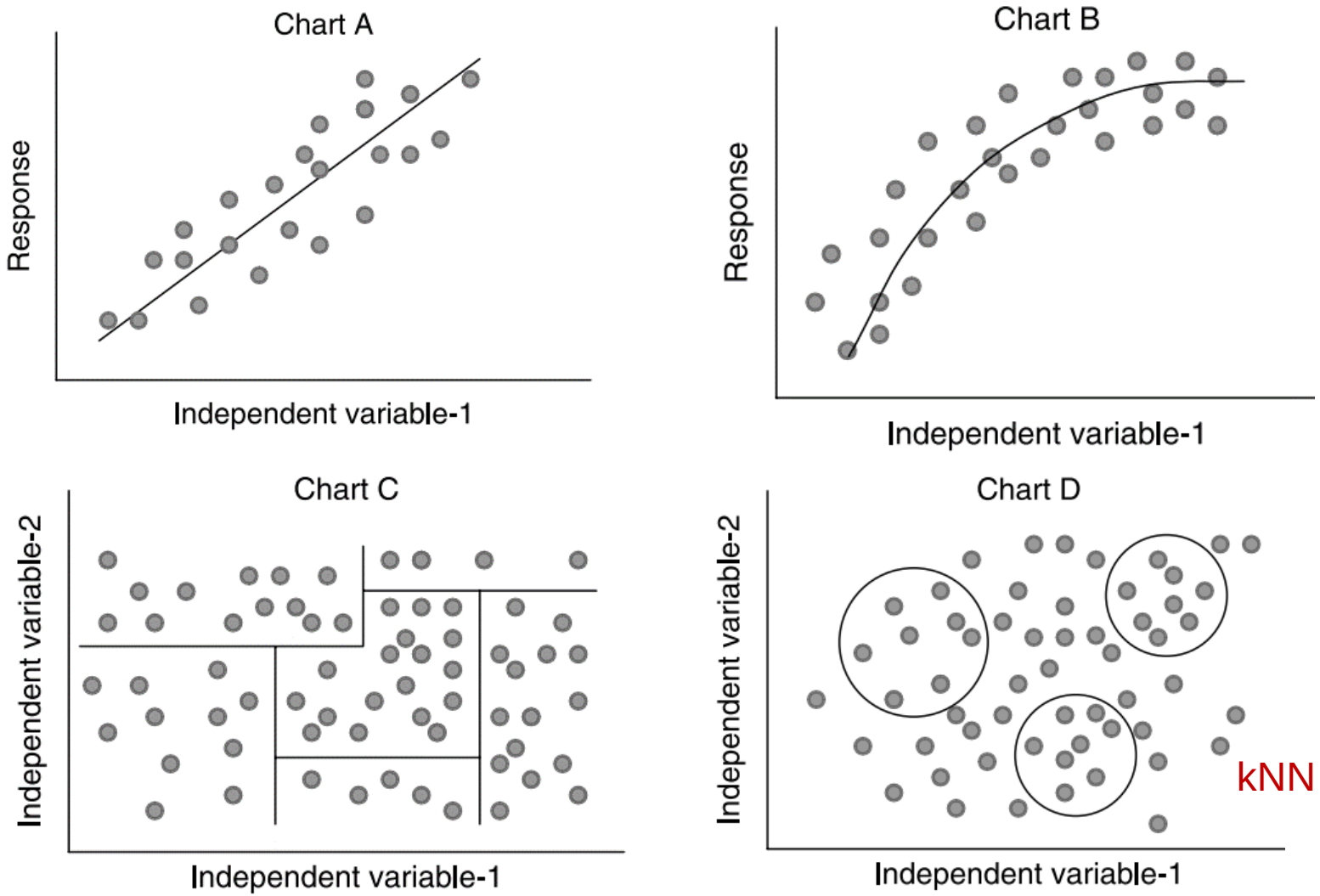


Figure 4.2 Illustrations of different regression modeling approaches

Evaluating Classification Models' Predictive Accuracy

Response variable: cylinders
Categorical (5 values): 3, 4, 5, 6, and 8

Actual (cylinders)

Predicted (cylinders)	3	4	5	6	8	Totals
3	3	1	0	0	0	4
4	1	170	1	32	0	204
5	0	1	0	0	0	1
6	0	1	0	45	0	46
8	0	26	2	6	103	137
Totals	4	199	3	83	103	392

Total number of **correctly classified** observations = **Sum of counts on the diagonal**

Figure 4.6 Contingency table showing predicted values against actual values (results of a classification model)

Evaluating Classification Models' Predictive Accuracy.

Total number of correctly classified observations:
Sum the counts on diagonal:3, 170, 0, 45, and 103 = 321

Overall accuracy: correct 321 observations divided by 392 (total number of observations) = 0.82.

Error rate: One minus overall accuracy level = 0.18.

Good classification models have high values along the diagonals in the contingency table.

	3	4	5	6	8	Totals
3	3	1	0	0	0	4
4	1	170	1	32	0	204
5	0	1	0	0	0	1
6	0	1	0	45	0	46
8	0	26	2	6	103	137
Totals	4	199	3	83	103	392

Error rate: The error rate, or misclassification rate, is 1 minus the accuracy value,

$$1 - \frac{TP + TN}{TP + FP + FN + TN}$$

Evaluating Binary Models' Predictive Accuracy.

4 properties

1. **True positive (TP):** The number of observations predicted to be true (1) that are in fact true (1).
2. **True negative (TN):** The number of observations predicted to be false (0) that are in fact false (0).
3. **False positive (FP):** The number of observations that are incorrectly predicted to be positive (1), but which are in fact negative (0).
4. **False negative (FN):** The number of observations that are incorrectly predicted to be negative (0), but which are in fact positive (1).

Evaluating Binary Models' Predictive Accuracy.

TABLE 6.1 Contingency Table Summarizing the Correct and Incorrect Predictions from a Binary Classification Model

		Actual		
		True (1)	False (0)	
Prediction	True (1)	True positives (TP)	False positive (FP)	*Number of observations predicted true* (1)
	False (0)	False negatives (FN)	True negatives (TN)	*Number of observations predicted false* (0)
		Number of actual true (1) *values*	*Number of actual false* (0) *values*	*Total observations*

Overall Accuracy: calculated based on the number of correctly classified divided by the total number of observations,

Contingency table or Confusion matrix

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Contingency Table and Performance Metrics.

Model 1

		Actual		
		1	0	Total
Prediction	1	36	12	48
	0	13	39	52
	Total	49	51	100
Concordance	75.0%	$\frac{TP + TN}{TP + FP + FN + TN}$		
Sensitivity	73.5%	$\frac{TP}{TP + FN}$		
Specificity	76.5%	$\frac{TN}{TN + FP}$		

Model 2

		Actual		
		1	0	Total
Prediction	1	31	17	48
	0	3	49	52
	Total	34	66	100
Concordance	80.0%	$\frac{TP + TN}{TP + FP + FN + TN}$		
Sensitivity	91.2%	$\frac{TP}{TP + FN}$		
Specificity	74.2%	$\frac{TN}{TN + FP}$		

FIGURE 6.3 Contingency tables and performance metrics for three models.

Contingency Table and Performance Metrics.

Model 3				
		Actual		
		1	0	Total
Prediction	1	40	9	49
	0	22	39	61
	Total	62	48	110
Concordance	71.8%	$\frac{TP + TN}{TP + FP + FN + TN}$		
Sensitivity	64.5%	$\frac{TP}{TP + FN}$		
Specificity	81.3%	$\frac{TN}{TN + FP}$		

To better assess the overall performance of a binary classification model

Necessary to calculate additional metrics.

Two commonly used calculations are:

Sensitivity

Specificity

Evaluating Binary Models' Predictive Accuracy.

Sensitivity: This is the *true positive rate*, also referred to as the *hit rate*, or *recall*. It is calculated using the number of observations identified as true positives, divided by the actual number of positive observations (TP + FN),

Sensitivity

$$\frac{\text{TP}}{\text{TP} + \text{FN}}$$

Specificity: This is the number of negative observations that are correctly predicted to be negative, or the *true negative rate*. It is calculated using the number of correctly predicted negative observations, divided by the total number of actual negative observations (TN + FP),

Specificity

$$\frac{\text{TN}}{\text{TN} + \text{FP}}$$

Evaluating Binary Models' Predictive Accuracy.

False positive rate: This value is the same as 1 minus the sensitivity and is calculated using the number of incorrectly predicted negative observations divided by the actual number of negative observations (FP + TN),

$$\frac{FP}{FP + TN}$$

Positive predictive value: This value is also called *precision*, and it is the number of correctly predicted positive observations divided by the total number of predicted positive observations (TP + FP),

$$\frac{TP}{TP + FP}$$

Evaluating Binary Models' Predictive Accuracy.

Negative predictive value: This value is the total number of correctly predicted negative observations divided by the number of negative predictions (TN + FN),

$$\frac{\text{TN}}{\text{TN} + \text{FN}}$$

False discovery rate: This value is the number of incorrectly predicted positive observations divided by the number observations predicted positive (FP + TP),

$$\frac{\text{FP}}{\text{FP} + \text{TP}}$$

Homework

Model A
Actual

Predicted	1	0	Totals
1	79	28	107
0	72	213	285
Totals	151	241	392

Model B
Actual

Predicted	1	0	Totals
1	140	38	178
0	11	203	214
Totals	151	241	392

Model C
Actual

Predicted	1	0	Totals
1	129	18	147
0	22	223	245
Totals	151	241	392

Figure 4.7
Summary of three different models

TABLE 4.7 Comparison of Different Metrics Across Three Models

	Model A	Model B	Model C
Accuracy			
Error			
Sensitivity			
Specificity			
False positive rate			
Positive predictive value			
Negative predictive value			
False discovery rate			

Answer these questions based on performance analysis:

1. Model **A/B/C** is **most accurate**? (which one)
2. Model **A/B/C** performed **best** based on **sensitivity** score.
3. Model **A/B/C** has the **highest specificity** score.

REGRESSION MODELS.

Models that predict **continuous variables** are called **regression models.**

How to generate linear models to describe a relationship between one or more independent variables and a single response variable?

Linear model using a **single independent** variable is referred to as **simple linear regression**

Linear model using **more than one independent variable** is referred to as **multiple linear regression**

Fitting a Simple Linear Regression Model

- A **straight line** representing a model can be drawn through the center of the points.
- A model that would **predict values along this line** would provide a good model.
- A straight line can be described using the formula:

$$y = a + bx$$

$$y = \beta_0 + \beta_1 x$$

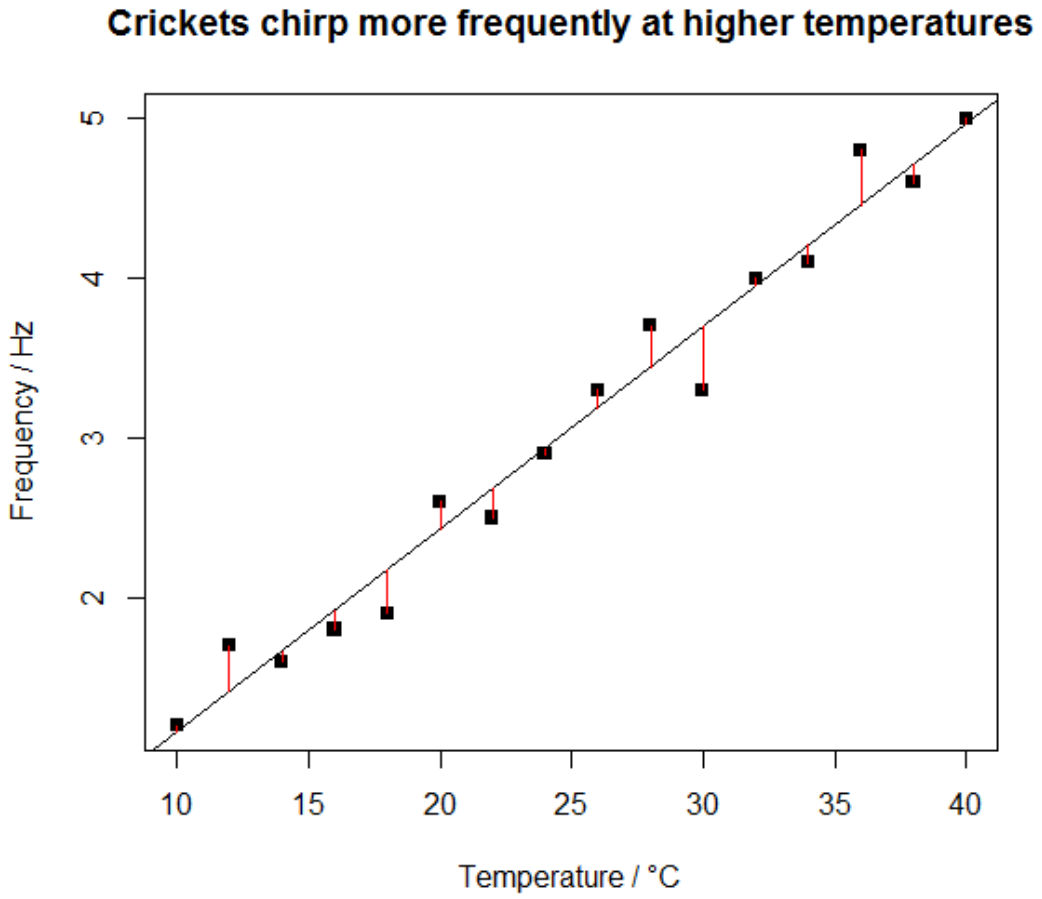
$$y = b_0 + b_1 x$$

where

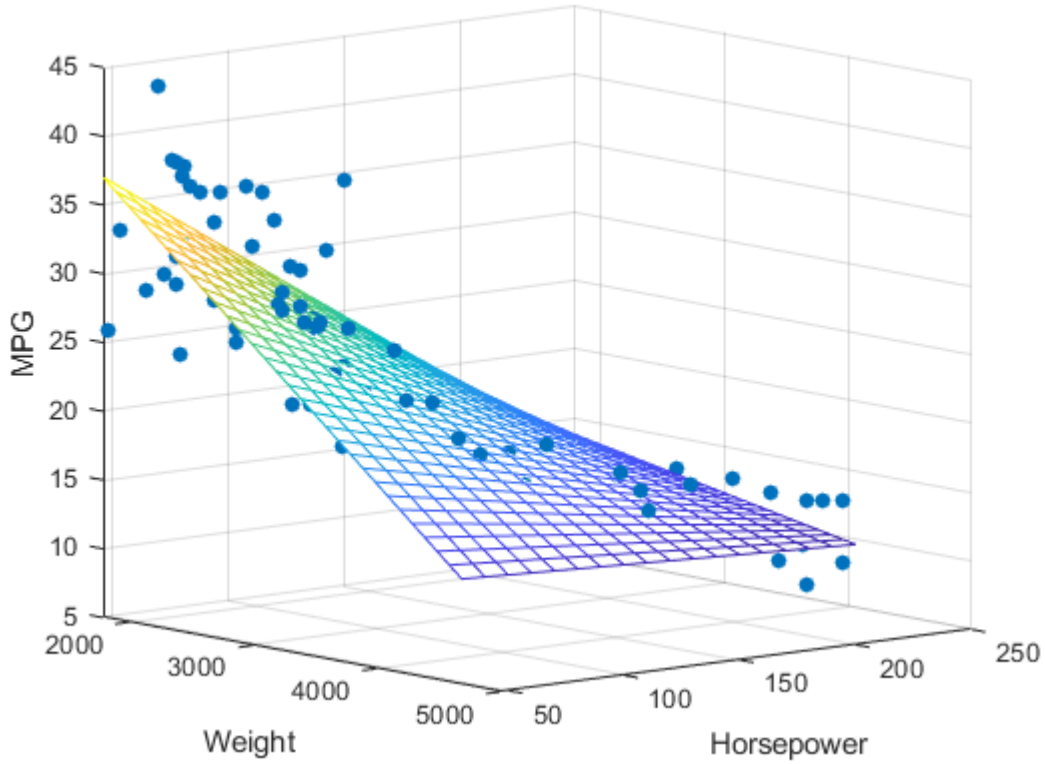
- **a** is the point of intersection with the y-axis, and
- **b** is the slope of the line

where $b_0(\beta_0)$ is the point of intersection with the y-axis and b_1 (β_1) is the slope of the line

Simple Linear Regression vs Multiple Linear Regression.



$$y = a + bx$$



$$y = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{pi} + e_i$$

Fitting a Simple Linear Regression Model

A **simple linear regression model** can be generated where there is a **linear relationship** between **two variables.**

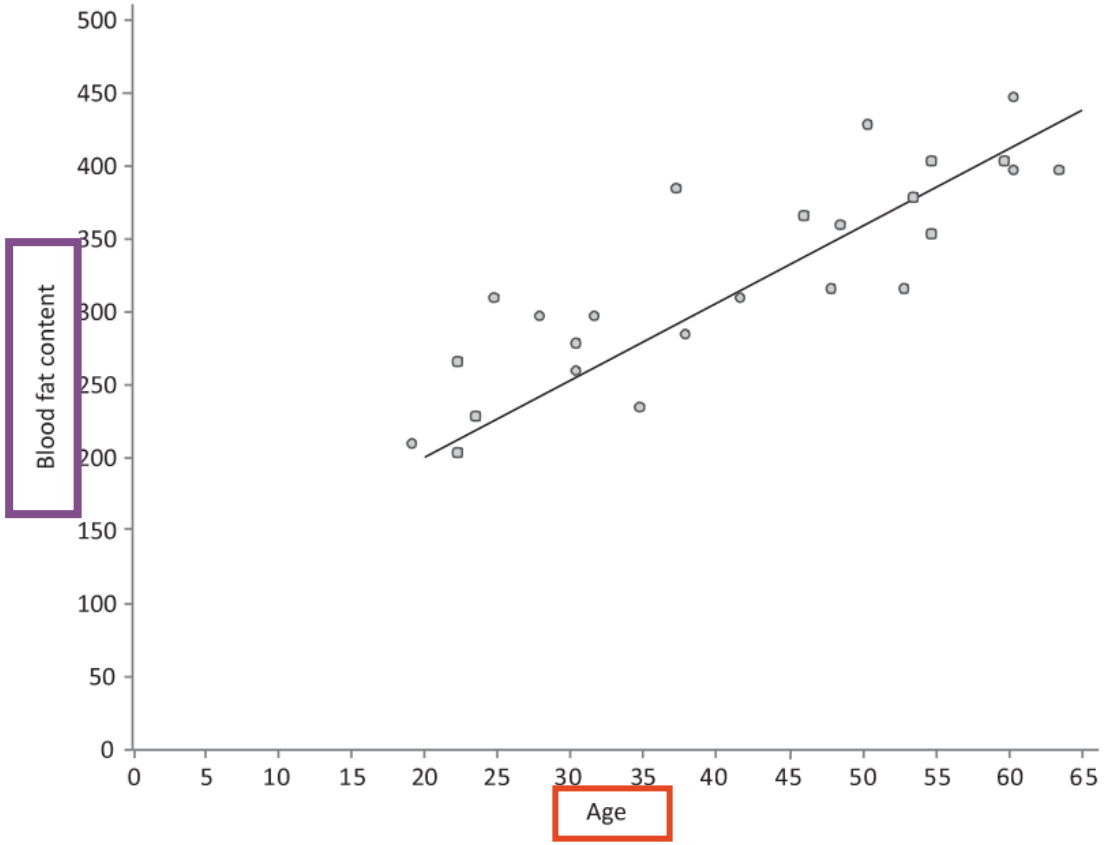


FIGURE 6.5 A straight line drawn through the relationship between variables *Age* and *Blood fat content*.

one descriptor/independent variable
one response/dependent variable

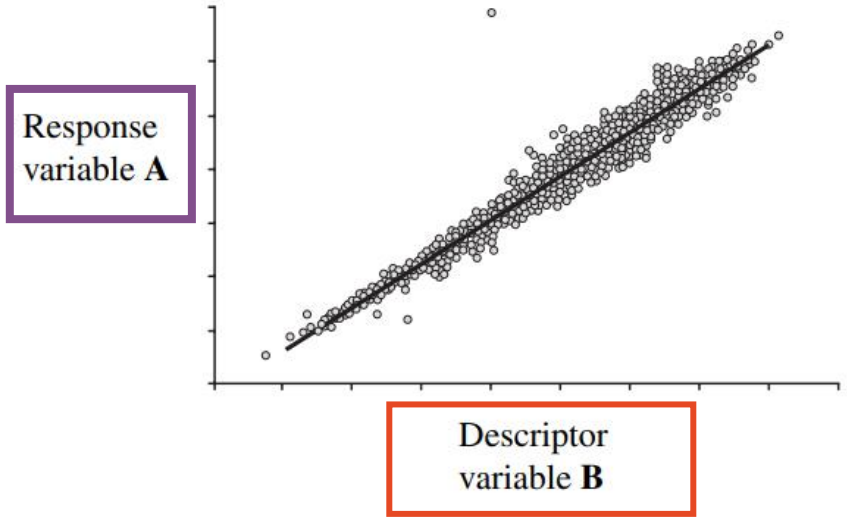


Figure 7.9. Scatterplot illustrating a simple linear model

Fitting a Simple Linear Regression Model

High degree of correlation between the two variables.

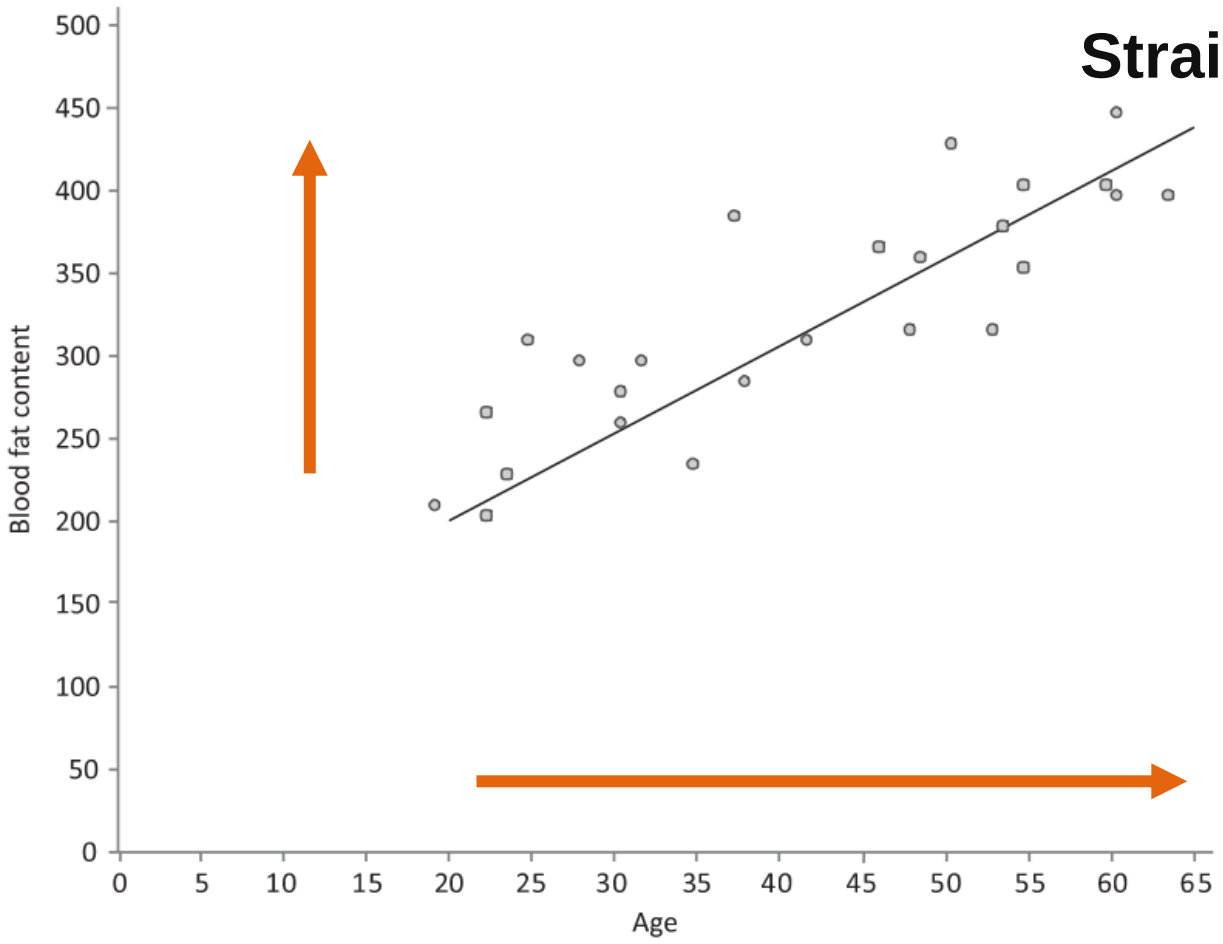


FIGURE 6.5 A straight line drawn through the relationship between variables *Age* and *Blood fat content*.

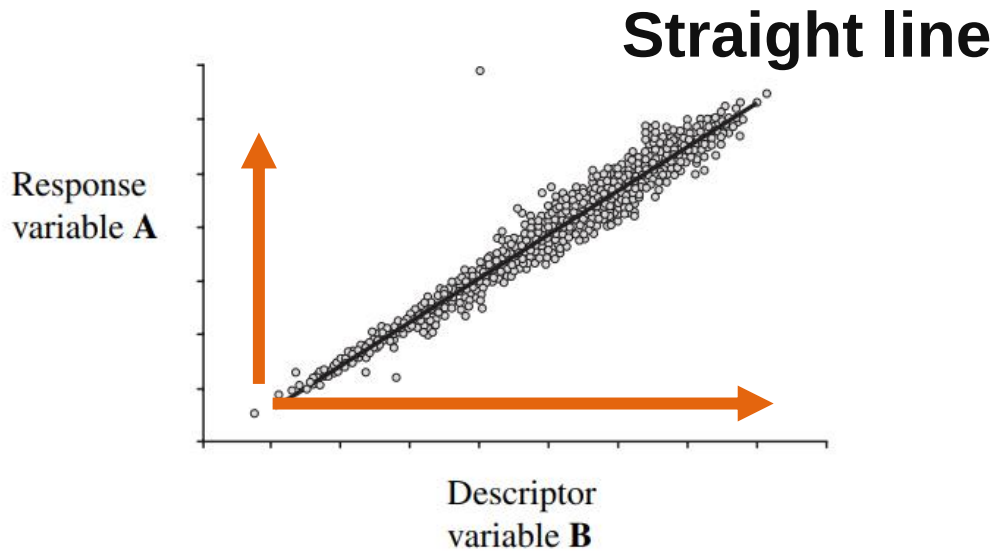


Figure 7.9. Scatterplot illustrating a simple linear model

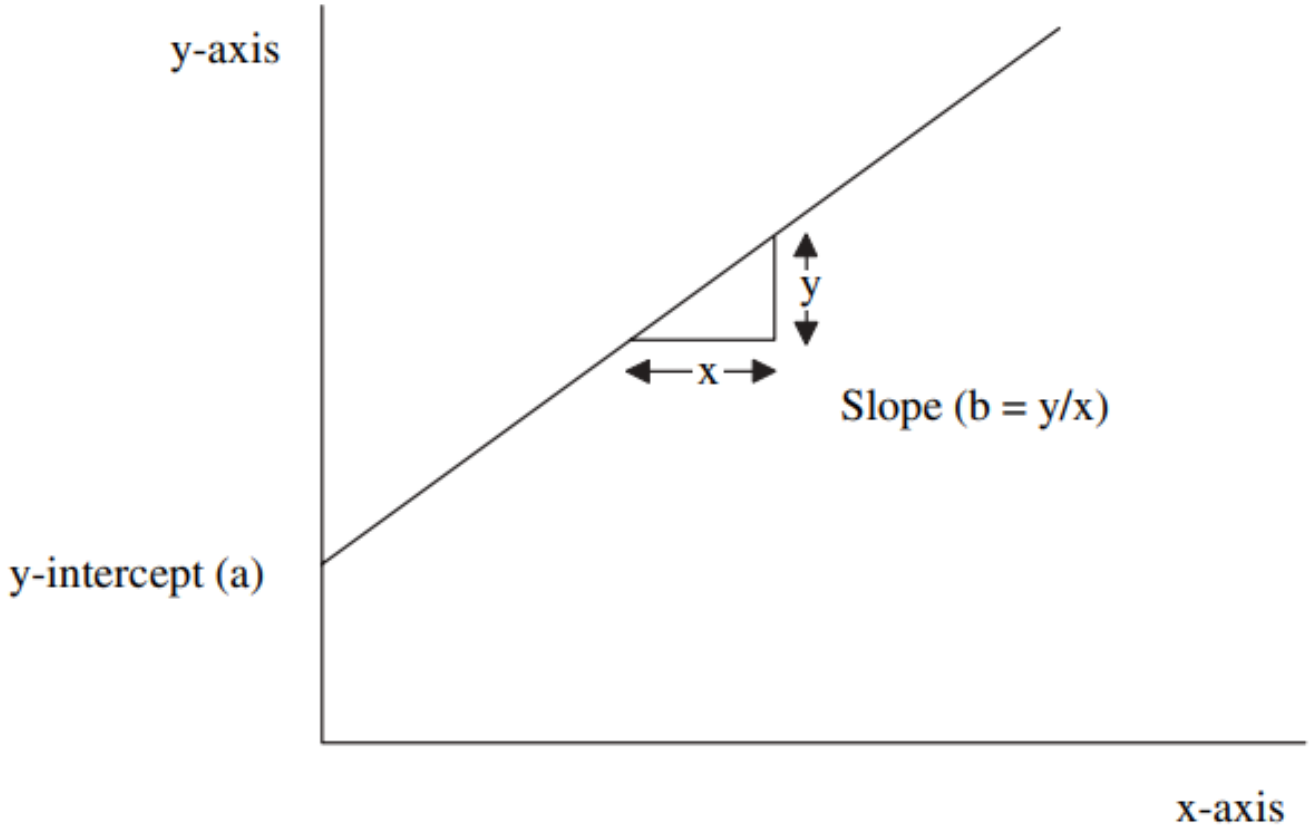
Example.

In Table 7.8, a data set of observations from a grocery store contains variables Income and Monthly sales. The variable Income refers to the yearly income for a customer and the Monthly sales represent the amount that particular customer purchases per month

Income (x)	Monthly Sales (y)
$15,000.00	$54.00
$16,000.00	$61.00
$17,000.00	$70.00
$18,000.00	$65.00
$19,000.00	$68.00
$20,000.00	$84.00
$23,000.00	$85.00
$26,000.00	$90.00
$29,000.00	$87.00
$33,000.00	$112.00
$35,000.00	$115.00
$36,000.00	$118.00
$38,000.00	$120.00
$39,000.00	$118.00
$41,000.00	$131.00
$43,000.00	$150.00
$44,000.00	$148.00
$46,000.00	$151.00
$49,000.00	$157.00
$52,000.00	$168.00
$54,000.00	$156.00
$52,000.00	$158.00
$55,000.00	$161.00
$59,000.00	$183.00
$62,000.00	$167.00
$65,000.00	$186.00
$66,000.00	$191.00

Table 7.8. Table of the customer's Income and Monthly Sales

Figure 7.10. Calculation of the slope of a straight line



Manually generation Example

Income (x)	Monthly Sales (y)
$15,000.00	$54.00
$16,000.00	$61.00
$17,000.00	$70.00
$18,000.00	$65.00
$19,000.00	$68.00
$20,000.00	$84.00
$23,000.00	$85.00
$26,000.00	$90.00
$29,000.00	$87.00
$33,000.00	$112.00
$35,000.00	$115.00
$36,000.00	$118.00
$38,000.00	$120.00
$39,000.00	$118.00
$41,000.00	$131.00
$43,000.00	$150.00
$44,000.00	$148.00
$46,000.00	$151.00
$49,000.00	$157.00
$52,000.00	$168.00
$54,000.00	$156.00
$52,000.00	$158.00
$55,000.00	$161.00
$59,000.00	$183.00
$62,000.00	$167.00
$65,000.00	$186.00
$66,000.00	$191.00

$$y = a + bx$$

For this data set an approximate formula for the relationship between Income and Monthly sales is:

Monthly sales = 20 + 0:0025 x Income

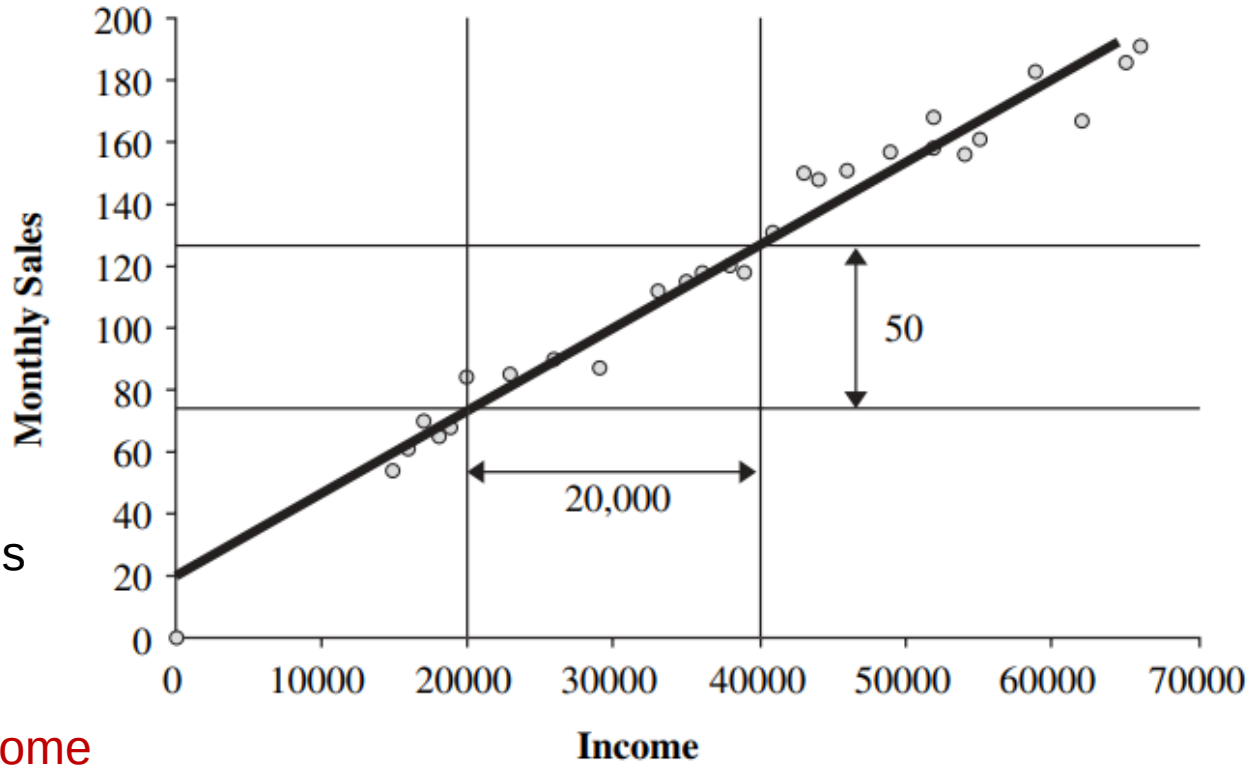


Figure 7.12. Calculating the slope of the line

1. Once a formula for the straight line has been established, predicting values for the y response variable based on the x descriptor variable can be easily calculated.
2. NOT: Formula should only be used for values **of x-variable within range** from which the formula was derived. Monthly sales should only be predicted based on Income between ***$15,000 and $66,000***

For a customer with an **Income of $31,000**, **Monthly sales** would be **predicted**:
Monthly sales (predicted) = 20 + 0.0025 x $31,000
Monthly sales (predicted) = $ 97.50

Least Squares Method.

- ***Parameters a and b can be derived manually*** by drawing a best guess line through the points in the scatterplot and then visually inspecting where the line crosses the y-axis (a) and measuring the slope (b)
- The **least squares method** is able to calculate these parameters **automatically.**
- Formula for calculating a slope **b** is:

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

where x_i and y_i are the individual values for the descriptor variable (x_i) and the response (y_i). x is the mean of the descriptor variable x and y is the mean of the response variable y.

Calculate the scope.

The formula for calculating the intercept with the y-axis is:

$$a = \bar{y} - b\bar{x}$$

Example.

Table 7.9. Calculation of linear regression with least squares method

Income (x)	Monthly Sales (y)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
$15,000.00	$54.00				
$16,000.00	$61.00				
$17,000.00	$70.00				
$18,000.00	$65.00				
$19,000.00	$68.00				
$20,000.00	$84.00				
$23,000.00	$85.00				
$26,000.00	$90.00				
$29,000.00	$87.00				
$33,000.00	$112.00				
$35,000.00	$115.00				
$36,000.00	$118.00				
$38,000.00	$120.00				
$39,000.00	$118.00				
$41,000.00	$131.00				
$43,000.00	$150.00				
$44,000.00	$148.00				
$46,000.00	$151.00				
$49,000.00	$157.00				
$52,000.00	$168.00				
$54,000.00	$156.00				
$52,000.00	$158.00				
$55,000.00	$161.00				
$59,000.00	$183.00				
$62,000.00	$167.00				
$65,000.00	$186.00				
$66,000.00	$191.00				
Totals					

Table 7.9. Calculation of linear regression with least squares method

Income (x)	Monthly Sales (y)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
$15,000.00	$54.00	−23,963	−70.22	1,682,733	574,223,594
$16,000.00	$61.00	−22,963	−63.22	1,451,770	527,297,668
$17,000.00	$70.00	−21,963	−54.22	11,908,801	482,371,742
$18,000.00	$65.00	−20,963	−59.22	1,241,473	439,445,816
$19,000.00	$68.00	−19,963	−56.22	1,122,362	398,519,890
$20,000.00	$84.00	−18,963	−40.22	762,733	359,593,964
$23,000.00	$85.00	−15,963	−39.22	626,103	254,816,187
$26,000.00	$90.00	−12,963	−34.22	443,621	168,038,409
$29,000.00	$87.00	−9,963	−37.22	370,844	99,260,631
$33,000.00	$112.00	−5,963	−12.22	72,881	35,556,927
$35,000.00	$115.00	−3,963	−9.22	36,547	15,705,075
$36,000.00	$118.00	−2,963	−6.22	18,436	8,779,150
$38,000.00	$120.00	−963	−4.22	4,066	927,298
$39,000.00	$118.00	37	−6.22	−230	1,372
$41,000.00	$131.00	2,037	6.78	13,807	4,149,520
$43,000.00	$150.00	4,037	25.78	104,066	16,297,668
$44,000.00	$148.00	5,037	23.78	119,770	25,371,742
$46,000.00	$151.00	7,037	26.78	188,436	49,519,890
$49,000.00	$157.00	10,037	32.78	328,992	100,742,112
$52,000.00	$168.00	13,037	43.78	570,733	169,964,335
$54,000.00	$156.00	15,037	31.78	477,844	226,112,483
$52,000.00	$158.00	13,037	33.78	440,362	169,964,335
$55,000.00	$161.00	16,037	36.78	589,807	257,186,557
$59,000.00	$183.00	20,037	58.78	1,177,733	401,482,853
$62,000.00	$167.00	23,037	42.78	985,473	530,705,075
$65,000.00	$186.00	26,037	61.78	1,608,510	677,927,298
$66,000.00	$191.00	27,037	66.78	1,805,473	731,001,372
Totals				17,435,222	6,724,962,963

Using the data from Table 7.8, the slope and intercept are calculated using Table 7.9.

The average Income is $38,963
The average Monthly sales is $124.22.

$$\text{Slope } (b) = 17,435,222/6,724,962,963$$

$$\text{Slope } (b) = 0.00259$$

$$\text{Intercept } (a) = 124.22 - (0.00259 \times 38,963)$$

$$\text{Intercept } (a) = 23.31$$

Hence the formula is:

Monthly sales = 23:31+ 0:00259 x Income
These values are close to the values calculated using the manual approach

AFTER THE BREAK.

Other evaluation measures.

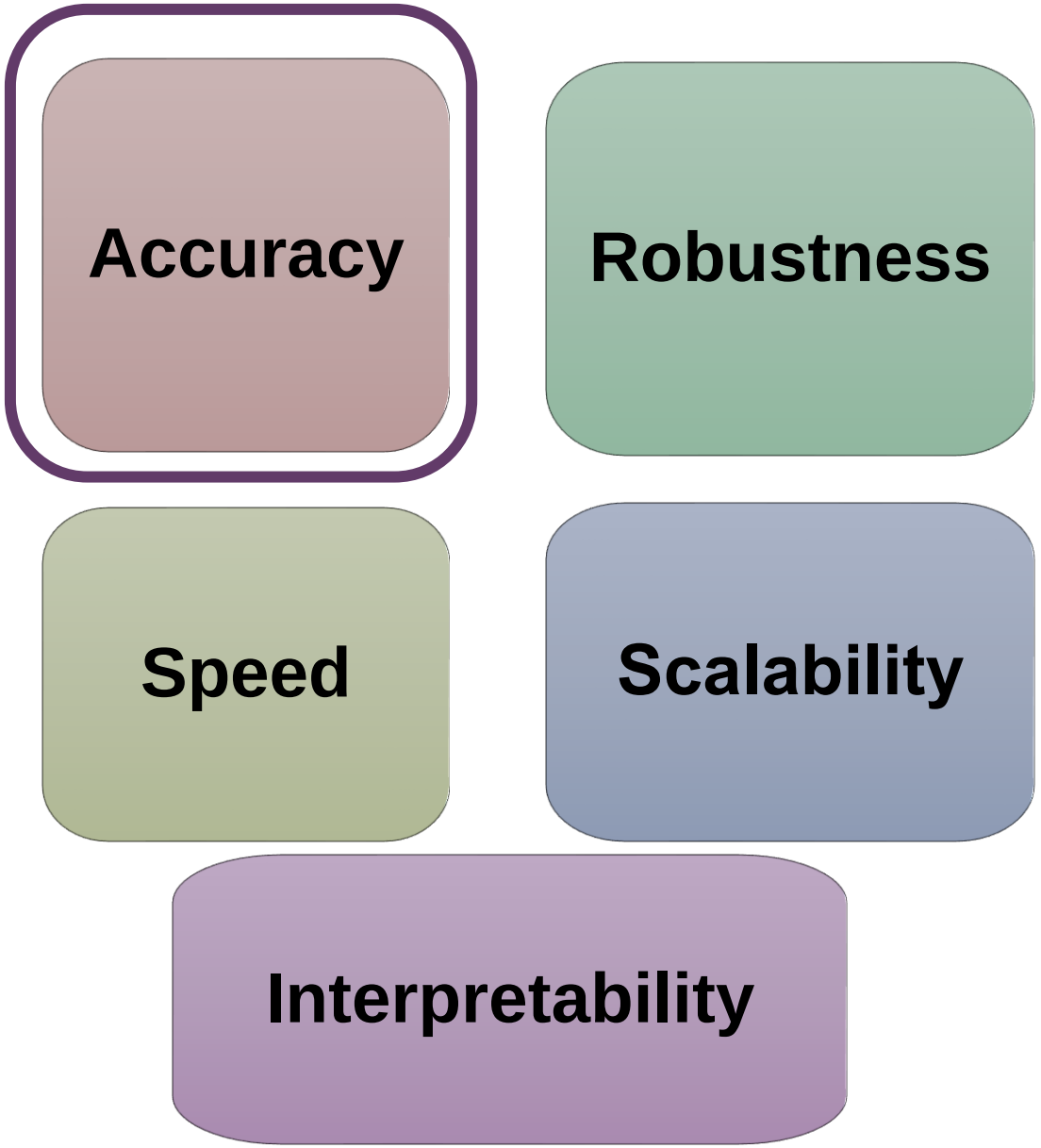


Table 7.5. Table showing an example of actual and predicted values

Actual response	Predicted response
True (1)	True (1)
False (0)	False (0)
False (0)	False (0)
True (1)	True (1)
True (1)	False (0)
False (0)	True (1)
True (1)	True (1)
False (0)	False (0)
True (1)	True (1)
False (0)	False (0)
False (0)	False (0)
True (1)	True (1)
True (1)	False (0)
True (1)	True (1)
False (0)	False (0)
False (0)	False (0)
True (1)	True (1)
True (1)	True (1)

Other evaluation measures.

	Actual response		
Predicted response	(TP) 8	(FP) 2	10
	(FN) 1	(TN) 7	8
	9	9	18

Accuracy rate, error rate and others

Table 7.5. Table showing an example of actual and predicted values

Actual response	Predicted response
True (1)	True (1)
False (0)	False (0)
False (0)	False (0)
True (1)	True (1)
True (1)	False (0)
False (0)	True (1)
True (1)	True (1)
False (0)	False (0)
True (1)	True (1)
False (0)	False (0)
False (0)	False (0)
True (1)	True (1)
True (1)	False (0)
True (1)	True (1)
False (0)	False (0)
False (0)	False (0)
True (1)	True (1)
True (1)	True (1)

Model A
Actual

Predicted	1	0	Totals
1	79	28	107
0	72	213	285
Totals	151	241	392

Model B
Actual

Predicted	1	0	Totals
1	140	38	178
0	11	203	214
Totals	151	241	392

Model C
Actual

Predicted	1	0	Totals
1	129	18	147
0	22	223	245
Totals	151	241	392

Figure 4.7
Summary of three different models

Homework

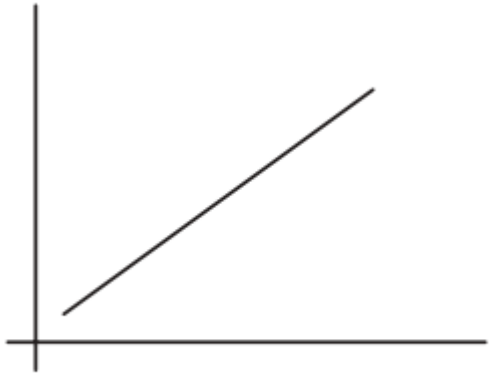
TABLE 4.7 Comparison of Different Metrics Across Three Models

	Model A	Model B	Model C
Accuracy	0.75	0.88	0.90
Error	0.26	0.13	0.10
Sensitivity	0.52	0.93	0.86
Specificity	0.88	0.84	0.93
False positive rate	0.12	0.16	0.07
Positive predictive value	0.74	0.79	0.88
Negative predictive value	0.75	0.95	0.91
False discovery rate	0.26	0.21	0.12

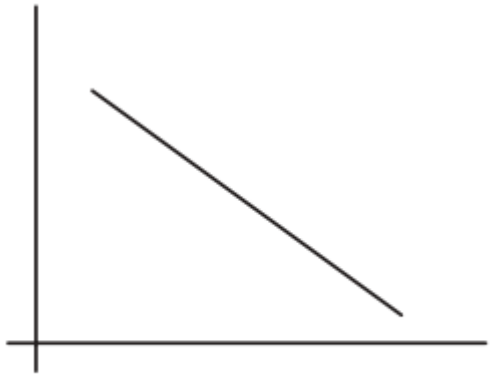
Answer these questions based on performance analysis:

1. Model is **most accurate?** **C** most accurate, **B then A**
2. Model performed **best** based on **sensitivity** score? **B**
3. Model **A/B/C** has the **highest specificity** score. **C**

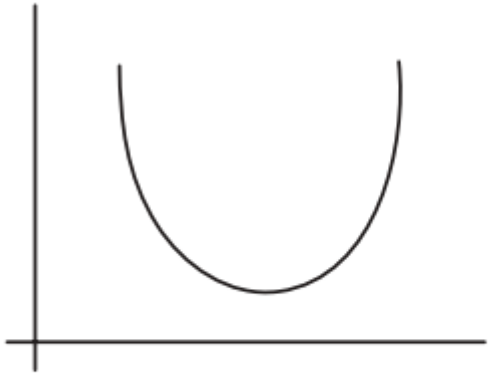
VISUALIZING RELATIONSHIPS BETWEEN VARIABLES



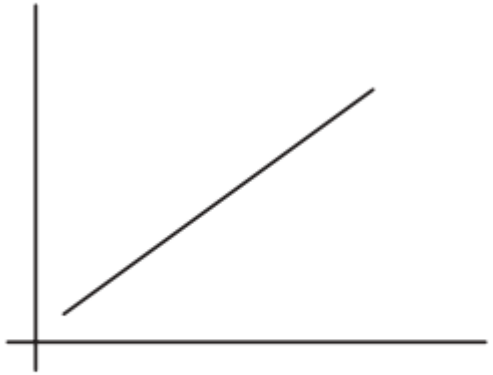
Positive



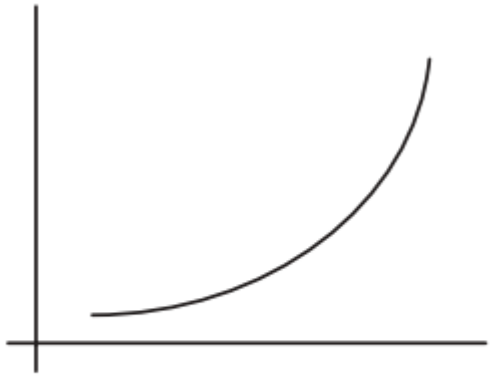
Negative



Both positive and negative at various points



Linear



Nonlinear

From **Chapter 4:** Understanding relationships

Simple Linear Regression model, why?

- Establish if **there is a relationship** between two variables
  - Statistically significant, linear (one increases, other increases)
  - Examples: age and blood fat count, income and monthly expenditure, salary and gender, student height and exam scores
- **Predict new observations**
  - Use what we know to predict outcomes for unseen values.
  - Examples: Given an age, can we predict blood fat content, what will be sales in the next month or quarter.

Simple Regression Model.

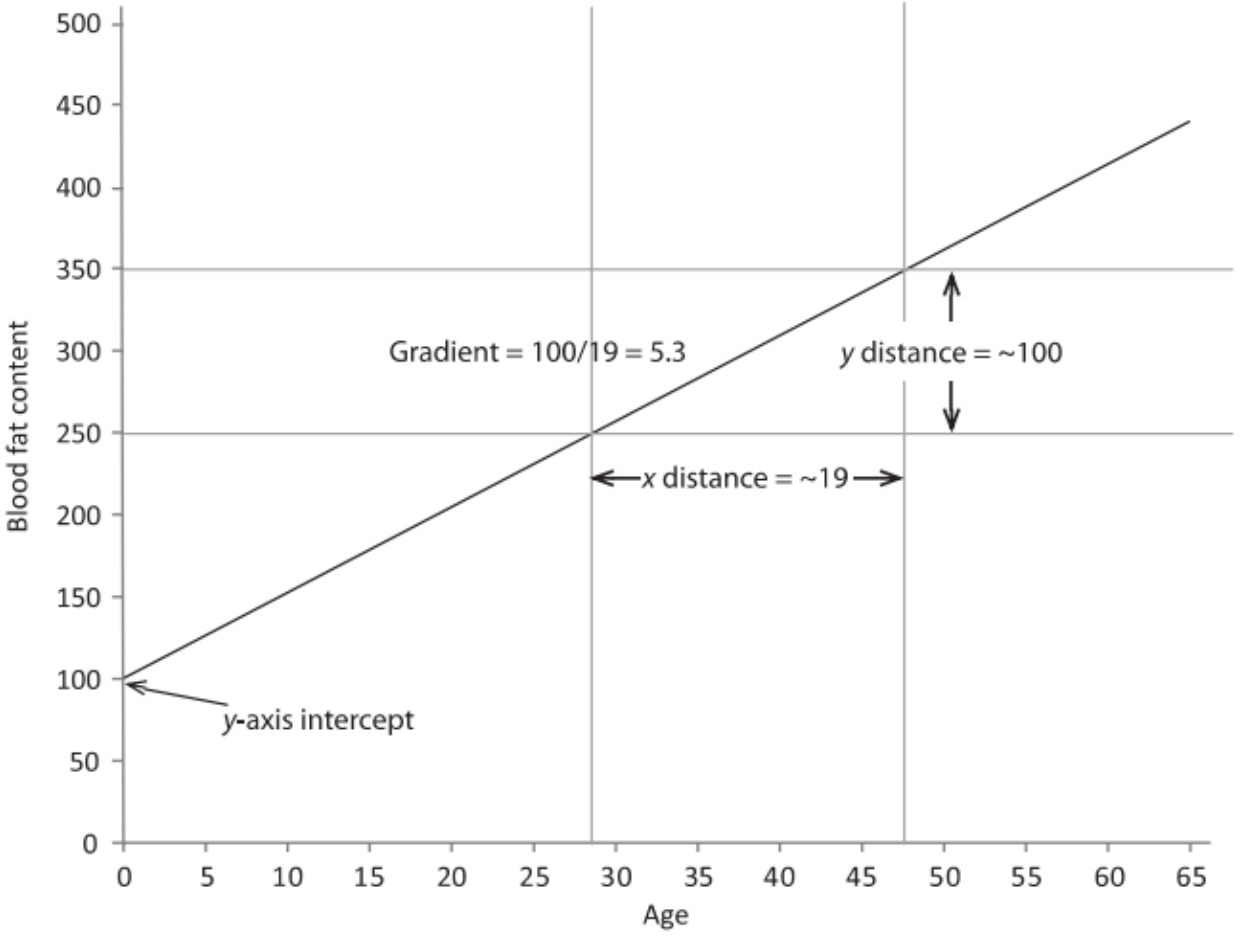


FIGURE 6.6 Deriving the straight line formula from the graph.

The diagram shows a **high degree of correlation** between the two variables. As variable **Age increases**, response variable **Blood fat content increases** proportionally, thereby linear.

$$y = b_0 + b_1 x$$

For this data set an approx. formula for the relationship between Age and blood fat content:

Blood fat content = 100 + 5.3 × Age

Example: Relationship between the independent variable Age and the response variable Blood fat content.

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad b_0 = \bar{y} - b_1\bar{x}$$

$$Slope\,(b_1) = 19{,}157.84/3{,}600.64$$

$$Slope\,(b_1) = 5.32$$

$$Intercept\,(b_0) = 310.72 - (5.32 \times 39.12)$$

$$Intercept\,(b_0) = 102.6$$

Hence the equation is

$$Blood\ fat\ content = 102.6 + 5.32 \times Age$$

age is 33.

$$Blood\ fat\ content = 102.6 + 5.32 \times 33 = 278.16$$

TABLE 6.2 Calculation of Linear Regression with Least Square Method

X	Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
46	354	6.88	43.28	297.7664	47.3344
20	190	−19.12	−120.72	2,308.1664	365.5744
52	405	12.88	94.28	1,214.3264	165.8944
30	263	−9.12	−47.72	435.2064	83.1744
57	451	17.88	140.28	2,508.2064	319.6944
25	302	−14.12	−8.72	123.1264	199.3744
28	288	−11.12	−22.72	252.6464	123.6544
36	385	−3.12	74.28	−231.7536	9.7344
57	402	17.88	91.28	1,632.0864	319.6944
44	365	4.88	54.28	264.8864	23.8144
24	209	−15.12	−101.72	1,538.0064	228.6144
31	290	−8.12	−20.72	168.2464	65.9344
52	346	12.88	35.28	454.4064	165.8944
23	254	−16.12	−56.72	914.3264	259.8544
60	395	20.88	84.28	1,759.7664	435.9744
48	434	8.88	123.28	1,094.7264	78.8544
34	220	−5.12	−90.72	464.4864	26.2144
51	374	11.88	63.28	751.7664	141.1344
50	308	10.88	−2.72	−29.5936	118.3744
34	220	−5.12	−90.72	464.4864	26.2144
46	311	6.88	0.28	1.9264	47.3344
23	181	−16.12	−129.72	2,091.0864	259.8544
37	274	−2.12	−36.72	77.8464	4.4944
40	303	0.88	−7.72	−6.7936	0.7744
30	244	−9.12	−66.72	608.4864	83.1744
			Sum	19,157.84	3,600.64

Assessing the Model Fit.

TABLE 6.4 Cruise Ship Data Annotated with Predicted Values and Calculated Residuals

Order	Ship Name	Cruise Line	Cabins	Passenger Density	Crew	Predicted ($\hat{y}$)	Residual ($\hat{e}$)
1	Journey	Azamara	3.55	42.64	3.55	3.85	−0.3
2	Quest	Azamara	3.55	42.64	3.55	3.85	−0.3
3	Celebration	Carnival	7.43	31.8	6.7	6.35	0.35
4	Destiny	Carnival	13.21	38.36	10	10.9	−0.9
5	Ecstasy	Carnival	10.2	34.29	9.2	8.52	0.68

- A set of statistics **are usually generated that help** to understand the **overall accuracy of the model**.
- The residual (ê) is an error term representing the difference between the observed value (y) and the predicted value (ŷ):

$$\hat{e} = y - \hat{y}$$

Assessing the Model Fit.

In the example of the cruise ship linear model SSR is 1,484 SSE is 88.3 and SST is 1,573.

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

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

$$SST = SSR + SSE$$

The coefficient of determination R^2

$$R^2 = \frac{SSR}{SST}$$

$$R^2 = \frac{1,484}{1,573} = 0.94$$

TABLE 6.4 Cruise Ship Data Annotated with Predicted Values and Calculated Residuals

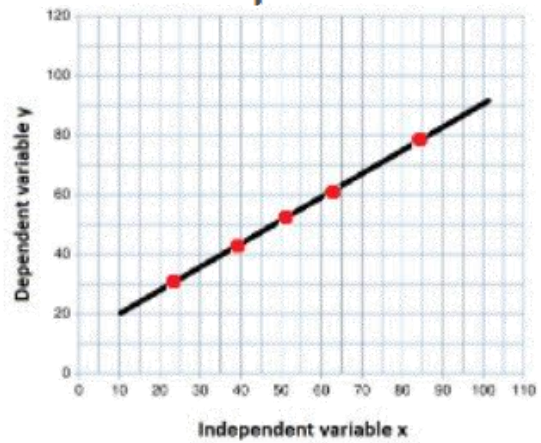
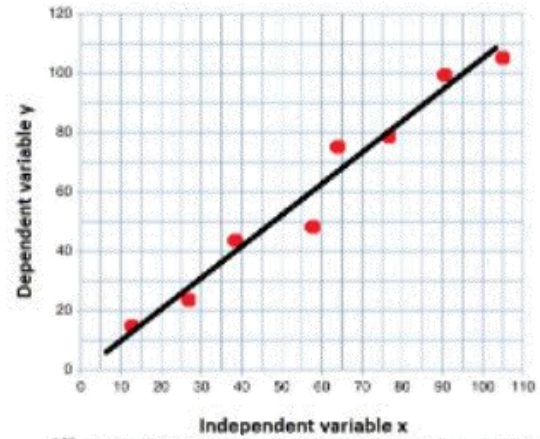
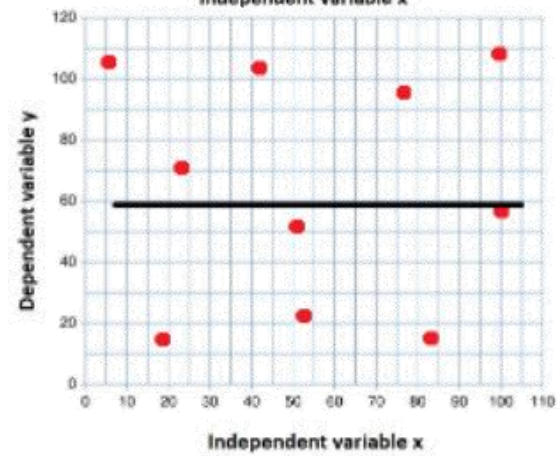
Order	Ship Name	Cruise Line	Cabins	Passenger Density	Crew	Predicted ($\hat{y}$)	Residual ($\hat{e}$)
1	Journey	Azamara	3.55	42.64	3.55	3.85	−0.3
2	Quest	Azamara	3.55	42.64	3.55	3.85	−0.3
3	Celebration	Carnival	7.43	31.8	6.7	6.35	0.35
4	Destiny	Carnival	13.21	38.36	10	10.9	−0.9
5	Ecstasy	Carnival	10.2	34.29	9.2	8.52	0.68

The *sum of squares total* (*SST*) is a measure of the variation of the *y*-values about their mean:

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

Assessing the Model Fit: Conclusion

- Looking at the residual, ***difference between the prediction and the actual value,*** helps to better understand how well the model is performing.
- Previous example: 94% of the variable Crew can be explained by the variability in the independent variables (Cabins and Passenger density) with rest attributable to something else.
- **R^2 values vary between 0 and 1.**
- **The closer the values are to 1, the more accurate are the predictions of the model; we say these models have a closer fit**

R^2 Values	Interpretation	Graph
$R^2 = 1$	All the variation in the y values is accounted for by the x values	
$R^2 = 0.83$	83% of the variation in the y values is accounted for by the x values	
$R^2 = 0$	None of the variation in the y values is accounted for by the x values	

Assessing the Model Fit: Part 2

- It is also a typical practice to calculate the **standard error of the estimate ($S_{y,x}$),** which is a measure of the variation of the y-values about the regression line.
- This value is interpreted in a similar manner to standard deviation and has the formula:

$$s_{y.x} = \sqrt{\frac{SSE}{n-2}}$$

In the cruise ship example, s would be

$$s_{y.x} = \sqrt{\frac{88.3}{154-2}} = 0.76$$

- The value given by $S_{y,x}$ indicates the **model's accuracy**: the ***larger the value for the standard error of estimate***, the ***lower the precision***.

Assumptions of simple linear regression.

1. **Homogeneity of variance (homoscedasticity):** When size of the error in prediction doesn't change significantly across the values of the independent variable.
2. **Independence of observations:** Observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3. **Normality:** The data follows a normal distribution.
4. **Relationship between the independent and dependent variable is linear**: the line of best fit through the data points is a straight line

https://www.scribbr.com/statistics/simple-linear-regression/

TAKE A BREAK

Relationship was non-linear?

- **Quadratic (U shape)**

(total working hours/week vs. overall happiness)

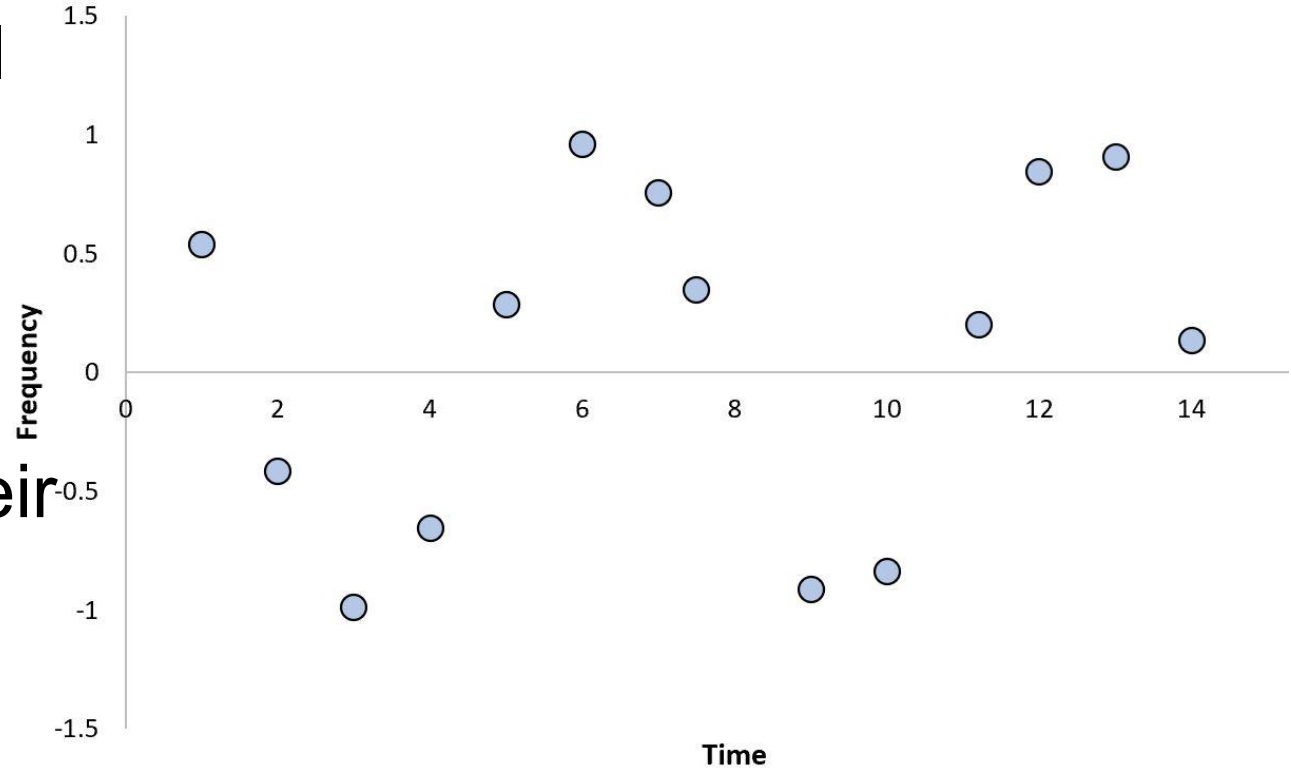
- **Cubic Relationships (sigmoid)**
- **Exponential Relationships**

(Life span of bamboo plants and their yearly growth)

- **Logarithmic Relationships**

(efficiency of smart-home technologies and time)

Cosine (wave)



Simple Non-linear Regression

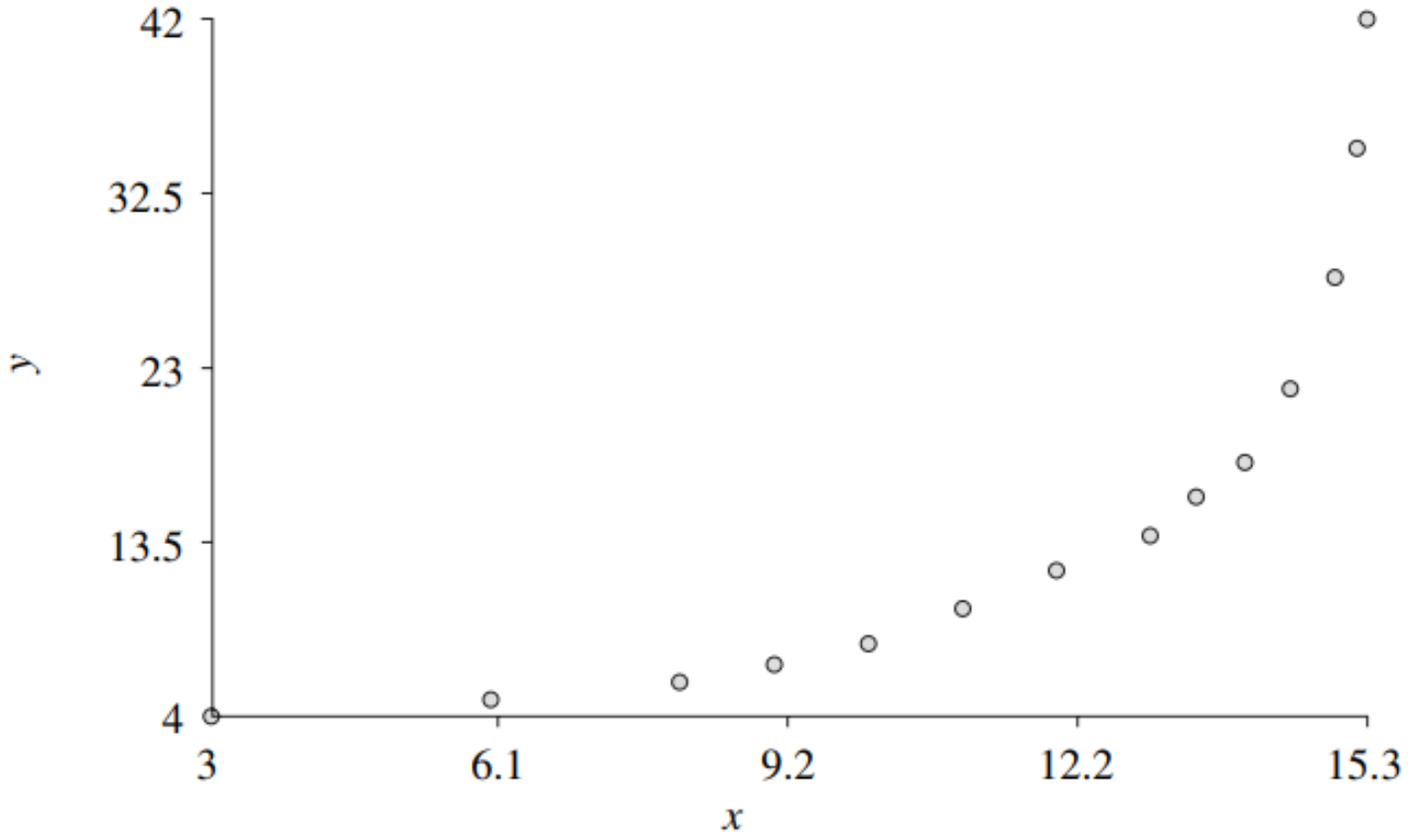
- Simple way of generating a regression equation is to **transform** the **nonlinear relationship to a linear relationship** using a **mathematical transformation.**
- A linear model can then be generated.
- Once a prediction has been made, the **predicted value is transformed back to the original scale.**
- To generate a model, **transform x or y** or **both** to create a linear relationship.
- Transform the y variable using the following formula:

$$y' = \frac{-1}{y}$$

Example.: How do we generate regression model?

Table 7.10. Table of observations for variables x and y

x	y
3	4
6	5
9	7
8	6
10	8
11	10
12	12
13	14
13.5	16
14	18
14.5	22
15	28
15.2	35
15.3	42



Scenarios for non-linear relationships.

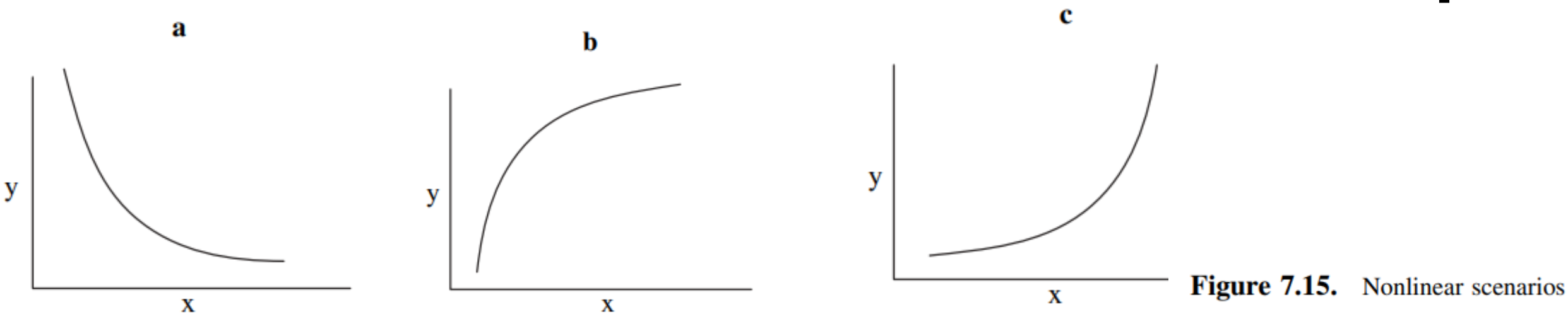


Figure 7.15. Nonlinear scenarios

Following transformation may create a linear relationship for the charts shown:

- **Situation a:** Transformations on the x, y or both x and y variables such as **log or square root.**
- **Situation b:** Transformation on the x variable such as **square root, log or -1/x**
- **Situation c:** Transformation on the y variable such as **square root, log or -1/y**.

This approach to creating simple nonlinear models **can only be used** when there is a **clear transformation of the data to a linear relationship**

Example.: Transformation to a linear regression model.

Table 7.10. Table of observations for variables x and y

x	$y' = -1/y$	y
3	−0.25	4
6	−0.2	5
9	−0.14286	7
8	−0.16667	6
10	−0.125	8
11	−0.1	10
12	−0.08333	12
13	−0.07143	14
13.5	−0.0625	16
14	−0.05556	18
14.5	−0.04545	22
15	−0.03571	28
15.2	−0.02857	35
15.3	−0.02381	42

1. Transform the y variable using the following formula: $y' = \frac{-1}{y}$

2. Using the least squares method described previously, an equation for the linear relationship between x and y' can be calculated. The equation is:

$$y' = -0.307 + 0.018 \times x$$

3. Using x calculate predicted value for the transformed value of y (y')

4.To map this new prediction of y' perform **an inverse transformation**, that is, **y= 1/y'**

Example.: How do we generate regression model?

Table 7.12. Prediction of y using a nonlinear model

x	y	$y' = -1/y$	Predicted y'	Predicted y
3	4	−0.25	−0.252	3.96
6	5	−0.2	−0.198	5.06
9	7	−0.143	−0.143	6.99
8	6	−0.167	−0.161	6.20
10	8	−0.125	−0.125	8.02
11	10	−0.1	−0.107	9.39
12	12	−0.083	−0.088	11.33
13	14	−0.071	−0.070	14.28
13.5	16	−0.062	−0.061	16.42
14	18	−0.056	−0.052	19.31
14.5	22	−0.045	−0.043	23.44
15	28	−0.036	−0.033	29.81
15.2	35	−0.029	−0.023	33.45
15.3	42	−0.024	−0.028	35.63

The **Predicted y** values are close to the **actual y** values


After a break.

OVERFITTING vs. UNDERFITTING.

Not interested in learning

A

Class test ~50%

Test ~47%

Under-fit/ biased learning

Memorizing the lessons

B

Class test ~98%

Test ~69%

Over-fit/ Memorizing

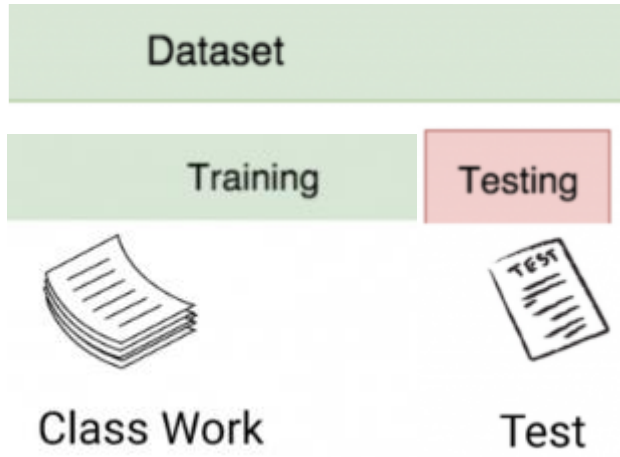
Conceptual Learning

C

Class test ~92%

Test ~89%

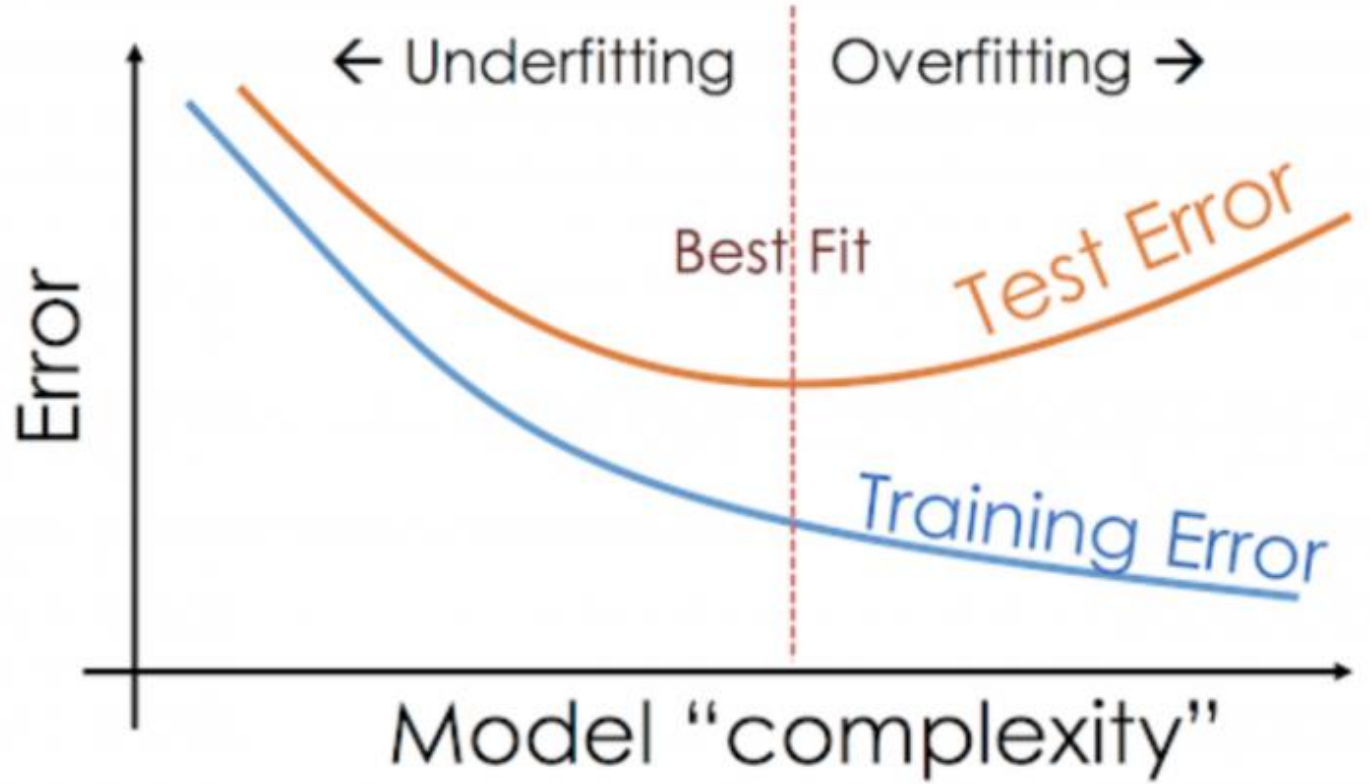
Best-fit



https://www.analyticsvidhya.com/blog/2020/02/underfitting-overfitting-best-fitting-machine-learning/

OVERFITTING vs. UNDERFITTING.

Any model is **performing poorly** on **training data** and **test set** is called an **underfitting model**.



Any model is **performing very well** on **training data** but performance **drops significantly** over **test set** is called an **overfitting model**.

Underfitting – High bias and low variance

Overfitting – High variance and low bias

- **Bias:** Assumptions made by a model to make a function easier to learn.
- **Variance:** Data trained on training data obtaining a very low error. When data is changed and then training the same previous model experiences a high error.

https://www.analyticsvidhya.com/blog/2020/02/underfitting-overfitting-best-fitting-machine-learning/

OVERFITTING vs. UNDERFITTING.

Any model is **performing poorly** on **training data** and **test set** is called an **underfitting model**.

Any model is **performing very well** on **training data** but performance **drops significantly** over **test set** is called an **overfitting model**.

Underfitting – High bias and low variance

Overfitting – High variance and low bias

- **Bias:** Assumptions made by a model to make a function easier to learn.
- **Variance:** Data trained on training data obtaining a very low error. When data is changed and then training the same previous model experiences a high error.

https://www.analyticsvidhya.com/blog/2020/02/underfitting-overfitting-best-fitting-machine-learning/

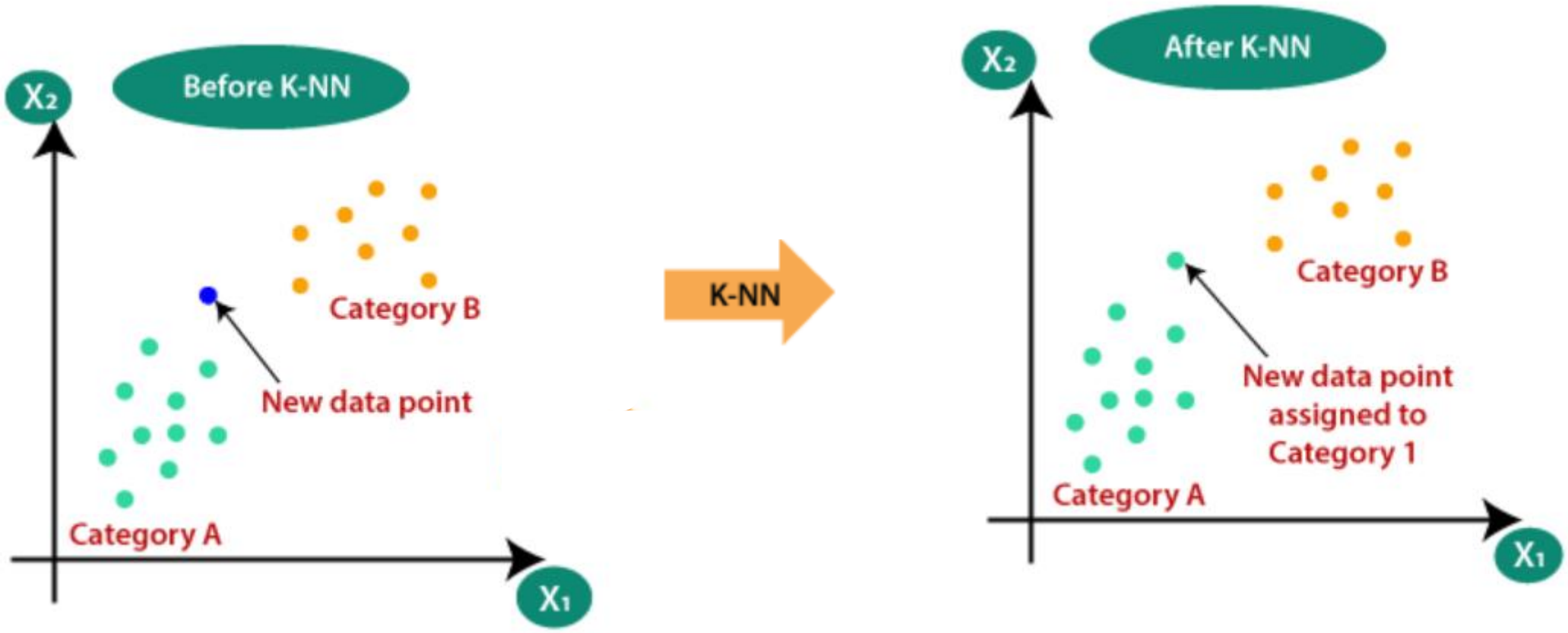
EAGER vs LAZY LEARNERS

When given a set of training data construct a classification model before receiving test data to classify.
Example:
Regression Models

- Simply stores it and waits until it is given a test data.
- Only when it sees the test tuple does it perform generalization to classify the observation based on its similarity to the stored training observations are also referred to as **instance-based learners**.
- Lazy learners naturally support incremental learning.
- Example : k-nearest neighbor, Bayesian Classifiers

KNN: K Nearest Neighbour

KNN: K Nearest Neighbour



KNN: K Nearest Neighbour

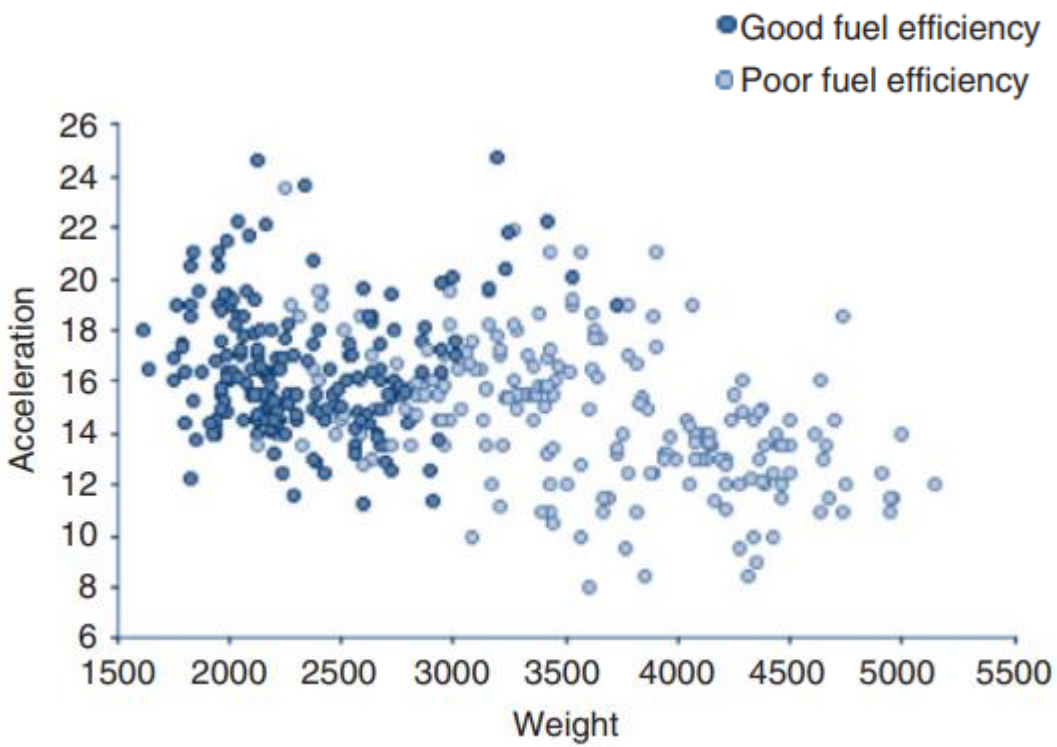


FIGURE 6.11 Scatterplot showing fuel efficiency classifications.

Response variable is a dichotomous variable

- Simple approach to calculating predictions for unknown observations.
- Calculates a prediction by looking at **similar observations** in the training set and uses some function of their response values, such as an average, to calculate the prediction.
- Like all prediction methods, it starts with a training set but it differs from other methods by determining the optimal number of similar observations to use in making the prediction rather than producing a mathematical model.

KNN: K Nearest Neighbour

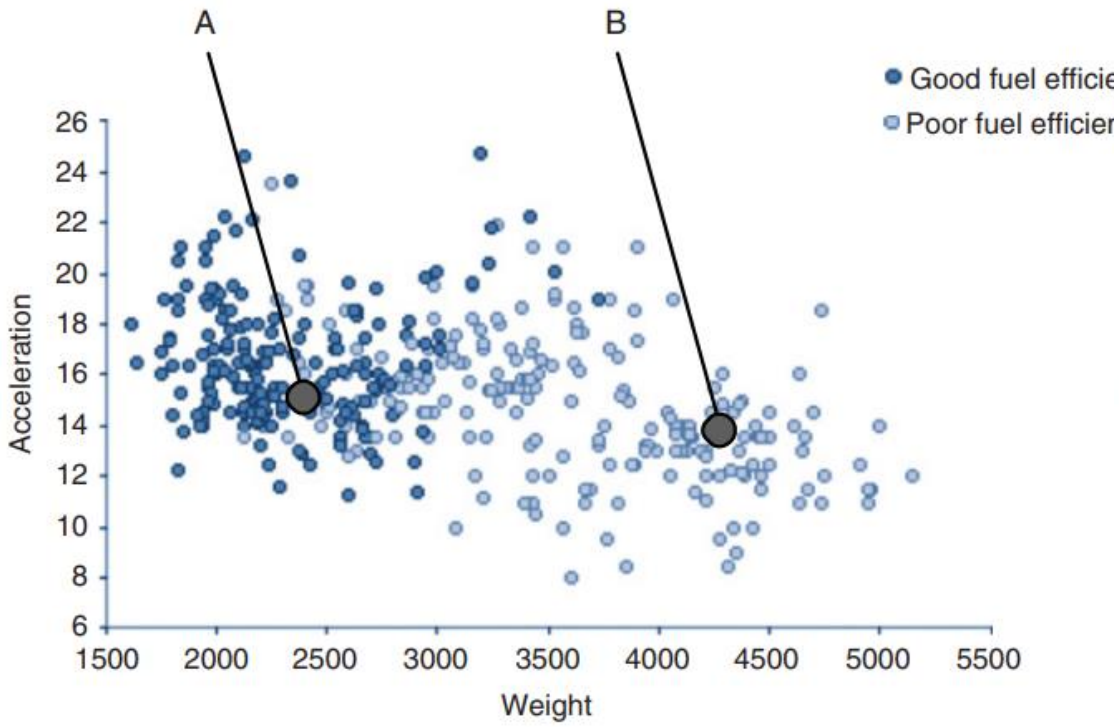


FIGURE 6.12 Two test set observations (A and B).

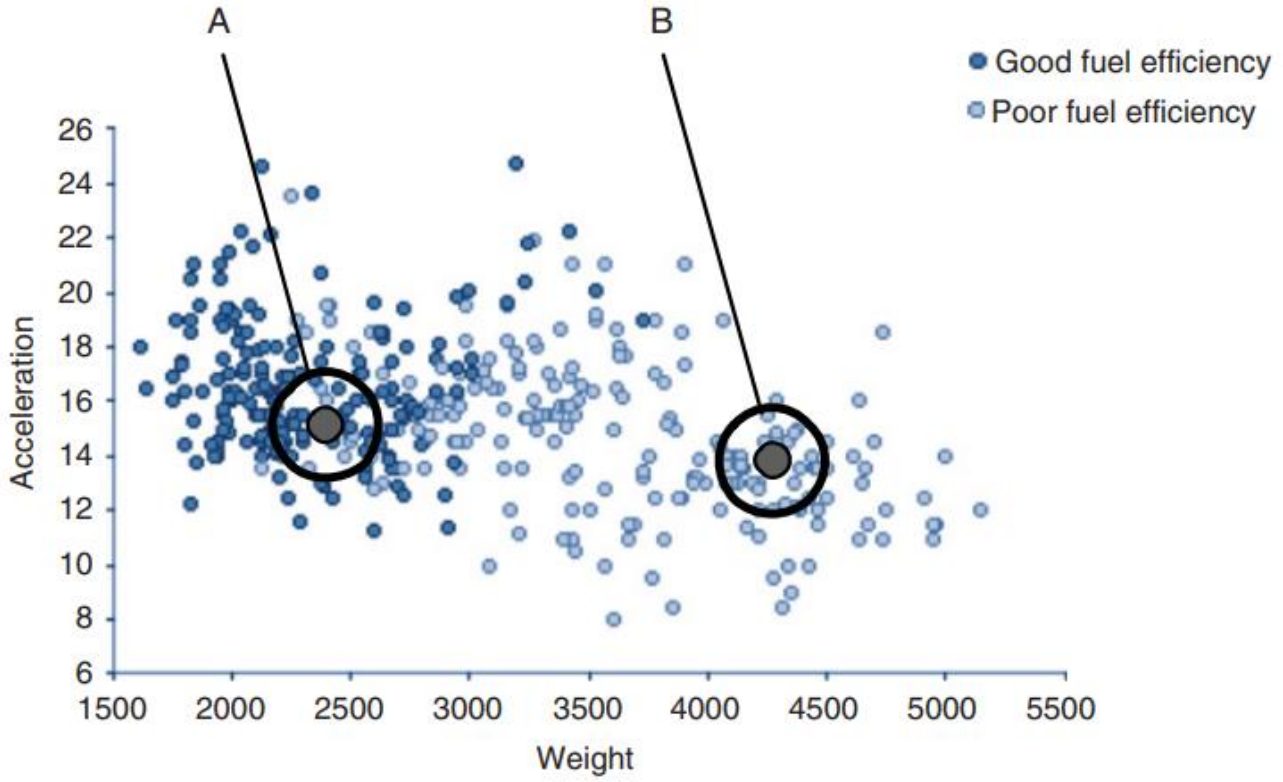


FIGURE 6.13 Looking at similar observation to support making predictions for A and B.

KNN: K Nearest Neighbour: LEARNING.

- A kNN model **uses the k most similar neighbors** to the observation to calculate a prediction.
- Where a **response variable** is **continuous**, the prediction is the **mean of these nearest neighbours**.
- Where a **response variable** is **categorical,** the prediction can be presented as a **mean or a particular classification scheme** that selects the most common classification term.

In the **learning phase**, three items should be determined:

1. **Best similarity method**
2. **k**
3. **Combination of descriptors**

KNN: K Nearest Neighbour: LEARNING.

- **Best Similarity Method**:
  - Euclidean or Jacard distance
  - Prior to calculating the similarity, Normalize variables to a common range so that no variables are considered to be more important
- **k:**
  - This is the number of similar observations that produces the best predictions.
  - If this value is too high, then the kNN model will overgeneralize.
  - If the value is too small, it will lead to a large variation in the prediction.
- **Combination of descriptors:**
  - It is important to understand which combination of descriptors results in the best prediction.

KNN: K Nearest Neighbour: LEARNING.

- **Selection of k** is performed by **adjusting the values of k** within a **range** and **selecting the value** that gives the **best prediction**.
- To ensure that models generated using different values of k are not overfitting, a **separate training and test set** should be used.
- To **assess different values for k,** the **sum of squares of error (SSE) evaluation** criteria will be used:

$$SSE = \sum_{i=1}^{k} (\hat{y}_i - \bar{y})^2$$

- Smaller SSE values indicate that the predictions are closer to the actual values.

Table 7.13. Table for detecting the best values for *k*

K	*SSE*
2	3,533
3	3,414
4	3,465
5	3,297
6	3,218
7	3,355
8	3,383
9	3,445
10	3,577
11	3,653
12	3,772
13	3,827
14	3,906
15	3,940
16	3,976
17	4,058
18	4,175
19	4,239
20	4,280

KNN: K Nearest Neighbour: PREDICTING.

1. Once a **value for k** has been **set** in the **training phase**, the **model** can now be used to **make predictions.**
2. For example, an **observation x** has **values for the descriptor** variables but **not for the response**.
3. Using the **same technique for determining similarity** as used in the model building phase, **observation x** is **compared** against all **observations in the training set**.
4. A **distance is comp**uted between **x** and **each training set observation.**
5. The **closest k** observations are **selected** and **a prediction is made**, for example, using the ***average value.***

EXAMPLE.

Sl . No.	Height (in cms)	Weight (in kgs)	T Shirt Size	Distance	Rank
1	158	58	M		
2	158	59	M		
3	158	63	M		
4	160	59	M		
5	160	60	M		
6	163	60	M		
7	163	61	M		
8	160	64	L		
9	163	64	L		
10	165	61	L		
11	165	62	L		
12	165	65	L		
13	168	62	L		
14	168	63	L		
15	168	66	L		
16	170	63	L		
17	170	64	L		
18	170	68	L		

$$d_{p,q} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

$$d_{p,q} = \sum_{i=1}^{n} |p_i - q_i|$$

New customer
Height= 161cm
Weight= 61kg

KNN: Good or bad?

- **ADVANTAGES:**
  - **Noise:** kNN is relatively insensitive to errors or outliers in the data.
  - **Large sets:** kNN can be used with large training sets.

- **DISADVANTAGE:**
  - **Speed:** kNN can be computationally slow when it is applied to a new data set since a similar score must be generated between the observations presented to the model and every member of the training set

NEXT.

Multilinear Regression.

- ***Practical situations***: A ***simple linear regression*** is **not sufficient** because **models** will need **more than one independent variable**.
- General form for a multiple linear regression equation is a linear function of the independent variables:

$$y = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{pi} + e_i$$

- where the response variable
  - **y** is shown with p independent variables (x-variables),
  - b_0 is a constant value
  - **k** is the number of coefficients of the independent variables, and
  - e_i refers to an error term measuring the unexplained variation or noise in the linear relationship.

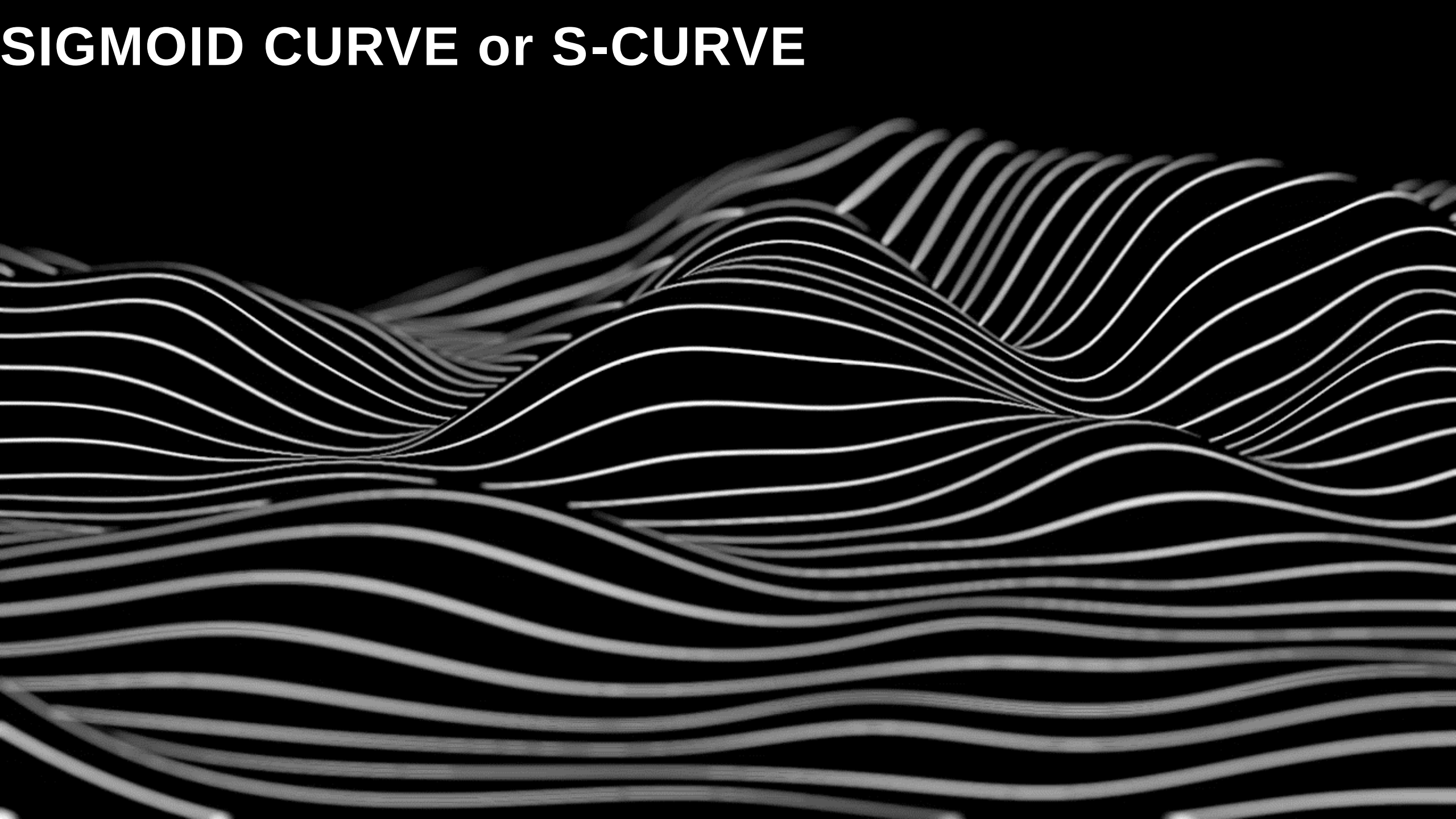
Multilinear Regression.

- **Set of coefficients are calculated** as part of the model building process to **minimize the overall differences** between the **observed** and the **predicted response values**.
- Since the **mathematics for computing** all but the simplest models make it i**mpossible** to **compute by hand**, **software tools** are typically used to perform the computation.
- This results in an equation where the coefficients are estimated:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_k$$

- In this equation, the coefficients are shown with a **"hat"** to represent that they are **estimated.**

SIGMOID CURVE or S-CURVE



LOGISTIC REGRESSION

- **Linear(simple and multilinear) regression** approach makes predictions when **response variable is continuous**.
- **Logistic regression** is a popular approach to building models where the **response variable is usually binary (dichotomous).**
- For example, response variable could indicate whether:
  - A consumer purchases a product (1 for purchase and 0 if do not)
  - A candidate drug is potent (1 if the candidate drug is potent and 0 if it is not).
- Logistic regression provides a flexible and easy-to-interpret method for building models from binary data. The following section outlines how to build, use, and assess logistic regression models

Fitting a Simple Logistic Regression Model

- **Linear(simple and multilinear) regression** approach makes predictions when **response variable is continuous**.
- **Logistic regression** is a popular approach to building models where the **response variable is usually binary (dichotomous).**
- For example, response variable could indicate whether:
  - A consumer purchases a product (1 for purchase and 0 if do not)
  - A candidate drug is potent (1 if the candidate drug is potent and 0 if it is not).
- Logistic regression provides a flexible and easy-to-interpret method for building models from binary data. The following section outlines how to build, use, and assess logistic regression models

LOGISTIC REGRESSION

Whether or not a person smokes

Binary Response

$$Y = \begin{cases} \text{Non} - \text{smoker} \\ \text{Smoker} \end{cases}$$

Success of a medical treatment

$$Y = \begin{cases} \text{Survives} \\ \text{Dies} \end{cases}$$

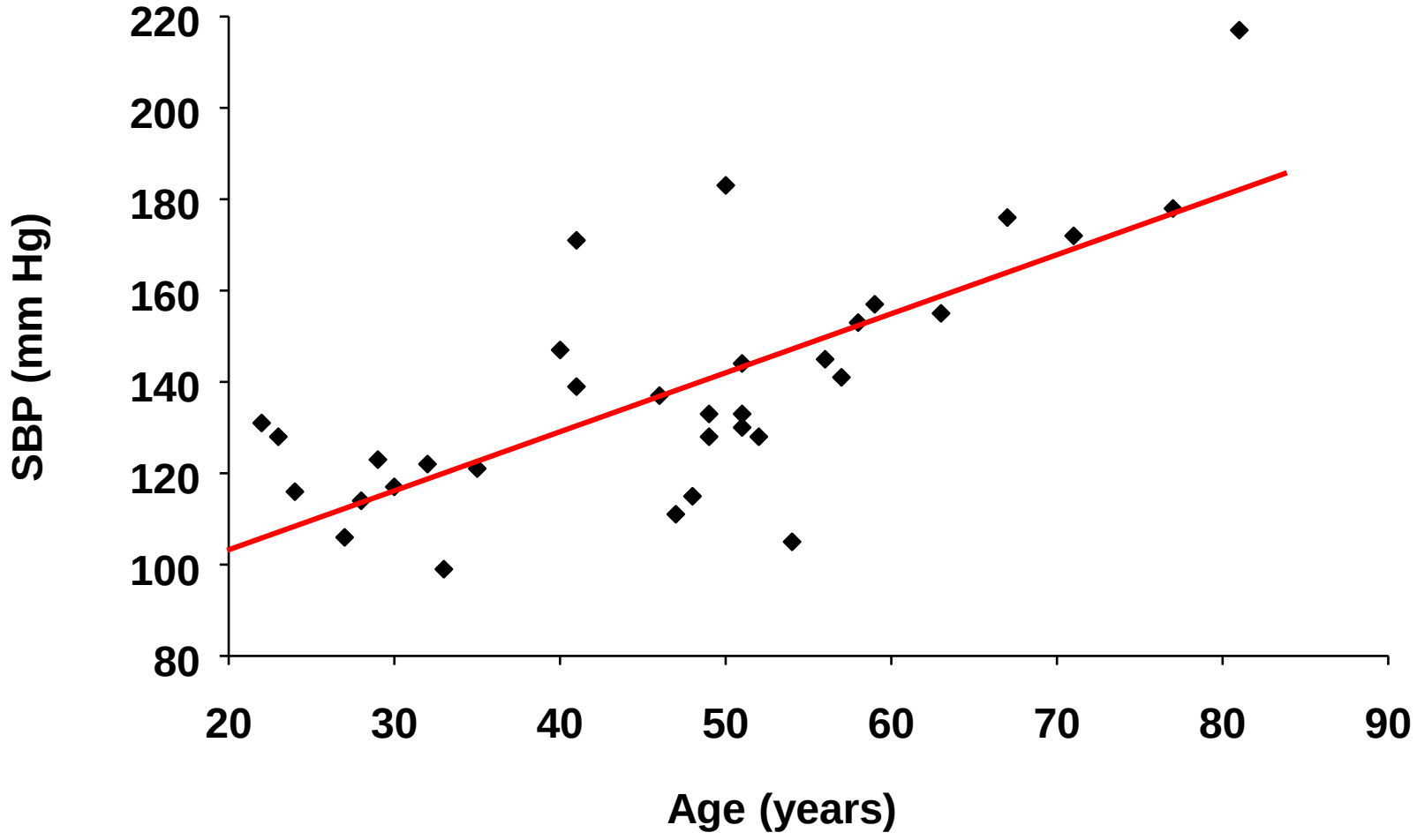
Opinion poll responses

Ordinal Response

$$Y = \begin{cases} \text{Agree} \\ \text{Neutral} \\ \text{Disagree} \end{cases}$$

LINEAR REGRESSION.

$$SBP = 81.54 + 1.222 \cdot Age$$



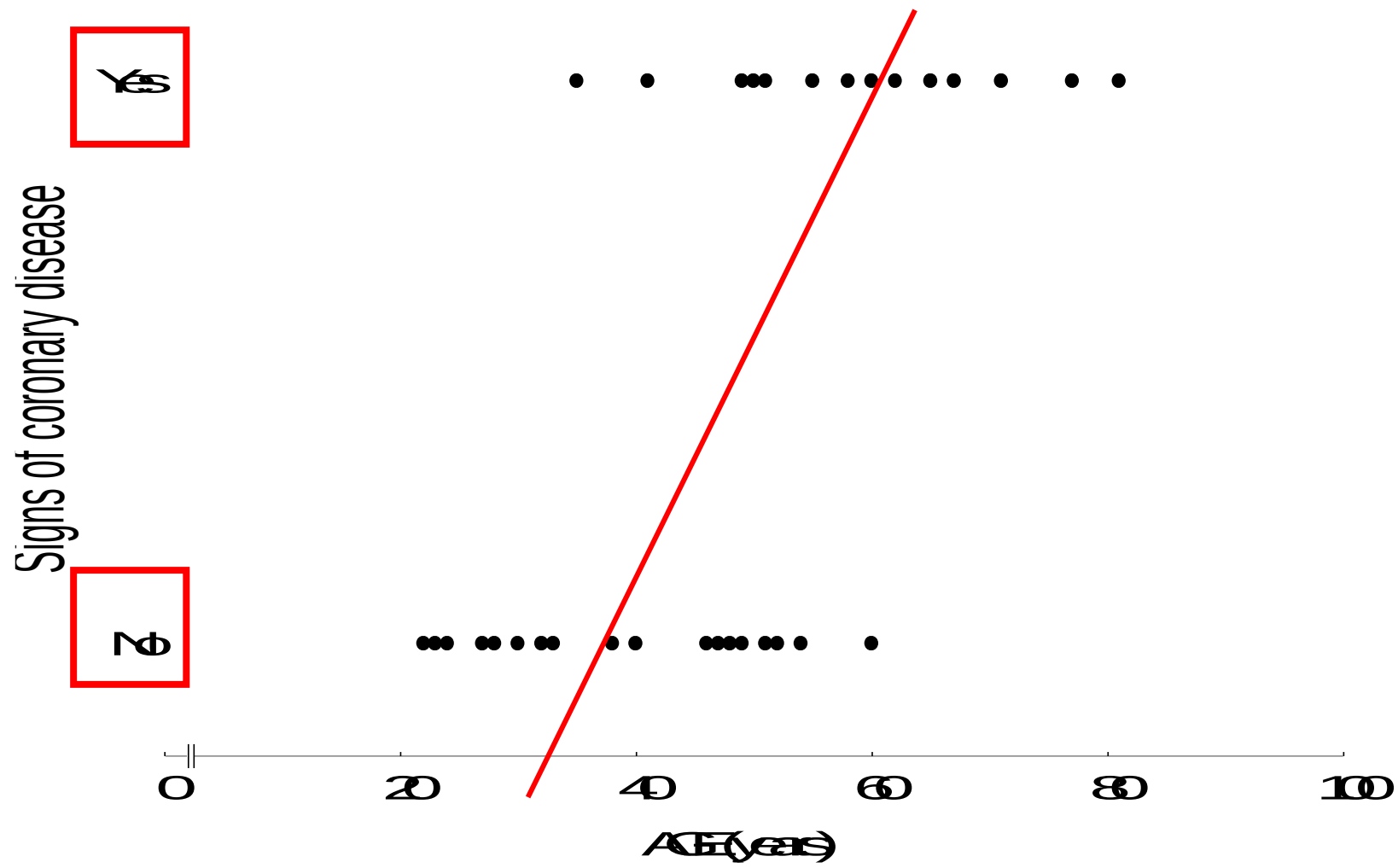
Adapted from Colton T. Statistics in Medicine. Boston: Little Brown, 1974

LINEAR to LOGISTIC REGRESSION.

Table 2 Age and signs of coronary heart disease (CD)

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

LOGISTIC REGRESSION.



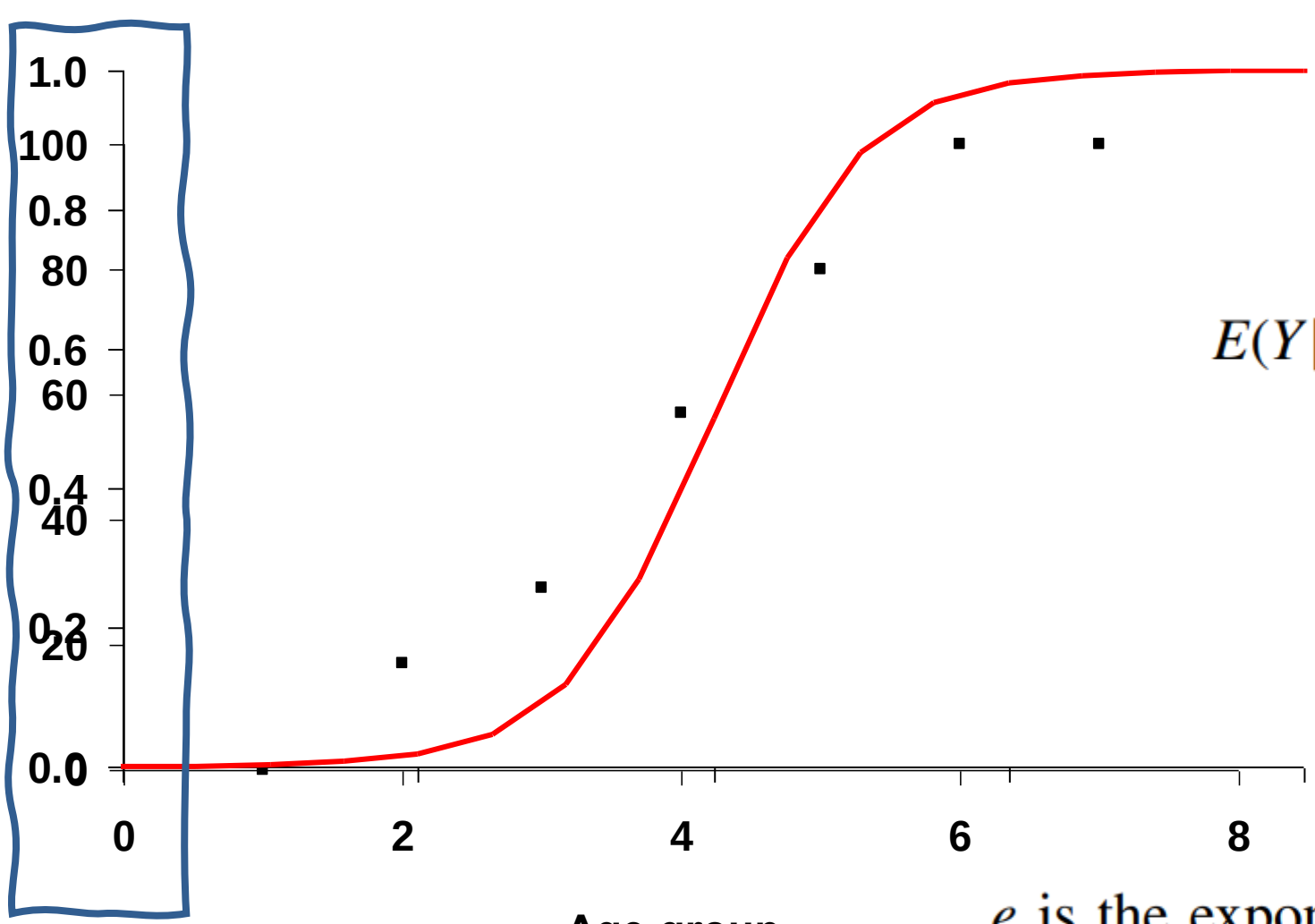
Dot-plot: Data from Table 2

LOGISTIC REGRESSION.

Table 3 Prevalence (%) of signs of CD according to age group

		Diseased	
Age group	# in group	#	%
20 - 29	5	0	0
30 - 39	6	1	17
40 - 49	7	2	29
50 - 59	7	4	57
60 - 69	5	4	80
70 - 79	2	2	100
80 - 89	1	1	100

LOGISTIC REGRESSION.



$$E(Y|x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_i x}}$$

e is the exponential function

β_0, β_1 are constant values.

Example: Looking for El Dorado.

x-variable (*log(Sb level)*)

response or *y*-variable

(*Gold deposit proximity*),

TABLE 6.6 The Mean Value for the Variable Gold Deposit Proximity for Different Ranges of Log(Sb Level)

Log(Sb Level) Ranges	Mean Gold Deposit Proximity
−1.1 to −0.7	0
−0.7 to −0.3	0.04
−0.3 to 0.1	0.43
0.1 to 0.5	0.89
0.5 to 0.9	0.9
0.9 to 1.3	1

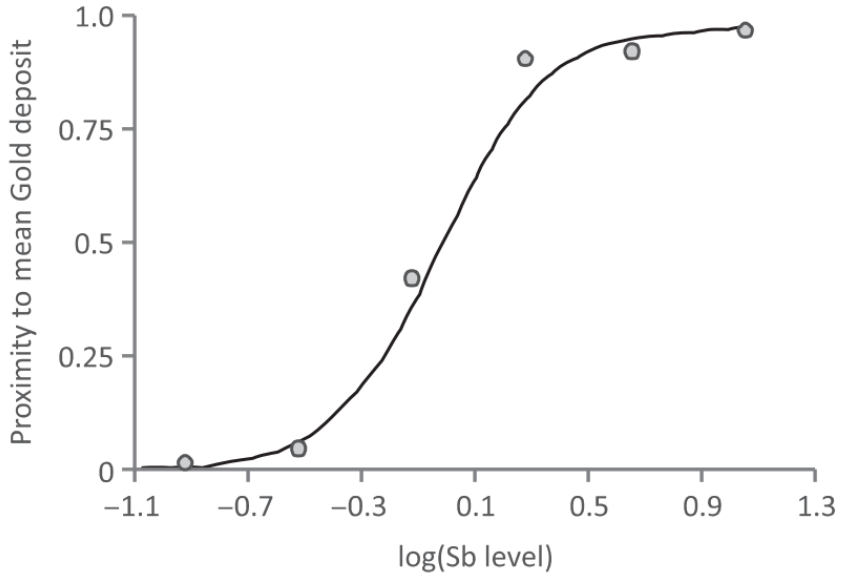


FIGURE 6.10 Shape of the graph showing how the mean values for variable Gold deposit proximity increase as values for the log-transformed Sb level increase.

Is there a gold deposit within 0.5 km?

- Expected value of y given x (E(Y|x)) is calculated where **e is the exponential function** and $\beta 0$, $\beta 1$ are constant values.
- Formula calculatea values for the E(Y|x) along the “S”-shaped curve.
- Ensures that values do not exceed 1 or go below 0.

Source: Sahoo and Pandalai (1999)

Fitting logistic regression to the data

- **Logistic regression: Maximum likelihood method**
- **Iterative computing**
  - Reiteration until maximisation (plateau)
- Results
  - Maximum Likelihood Estimates (MLE) for α and β
  - Estimates of E(y) for a given value of x

If *log*(*Sb level*) was 0.4, then

$$E(\textit{Gold deposit proximity}|\textit{log}(\textit{Sb level}) = 0.4) = \frac{e^{-0.0728+5.82\times0.4}}{1+e^{-0.0728+5.85\times0.4}} = 0.905$$

Since the expected value is close to 1, we could conclude that it is likely there will be a gold deposit within 0.5 km.

Thank you

Next up Bayesian Classifier