IFN509

Data Exploration and Mining

Week 10

Algorithms of Predictive Data Mining
Regression Mining

Prof Richi Nayak
r.nayak@qut.edu.au

School of Computer Science
Centre for Data Science
Faculty of Science and Engineering
https://research.qut.edu.au/adm

Learning Objectives: Week 10

- Predictive Modelling Algorithms
  - Regression Modelling: Classification and Regression
    - Liner Regression
    - Logistic Regression
    - Nonlinear regression

What Should You Do in Week 10?

- Listen to the lecture recording and review the lecture slides (Regression Mining)
- Tutorial: Attempt the exercise questions related to the lecture on Decision Tree mining.
- Practical: Complete practical tasks on Decision Trees
- Consult the Lecturer or Tutor if you have any questions related to the subject.
- Assessment Item 2
  - Association mining: Should have finished
  - Clustering: Should have finished
  - Decision Tree: Should start attempting

Regression Modelling for Regression

Linear Regression

Regression Algorithms

- Regression algorithms project the attribute space into a continuous function
  - Linear Regression
    - builds a predictive model that attempts to fit a straight line through a plot of the data.
  - Nonlinear Regression
    - builds a predictive model that attempts to fit a non-linear function through a plot of the data.
  - Radial basis function
    - builds a predictive model that attempts to fit a weighted sum of a set of nonlinear functions through a plot of the data.
  - Logistic Regression, Poisson regression
    - builds a predictive model that can **make the classification**.

Linear Regression

- Works most naturally with numeric attributes
- Simple Case: Involves a target attribute *y* and a single input attribute *x*

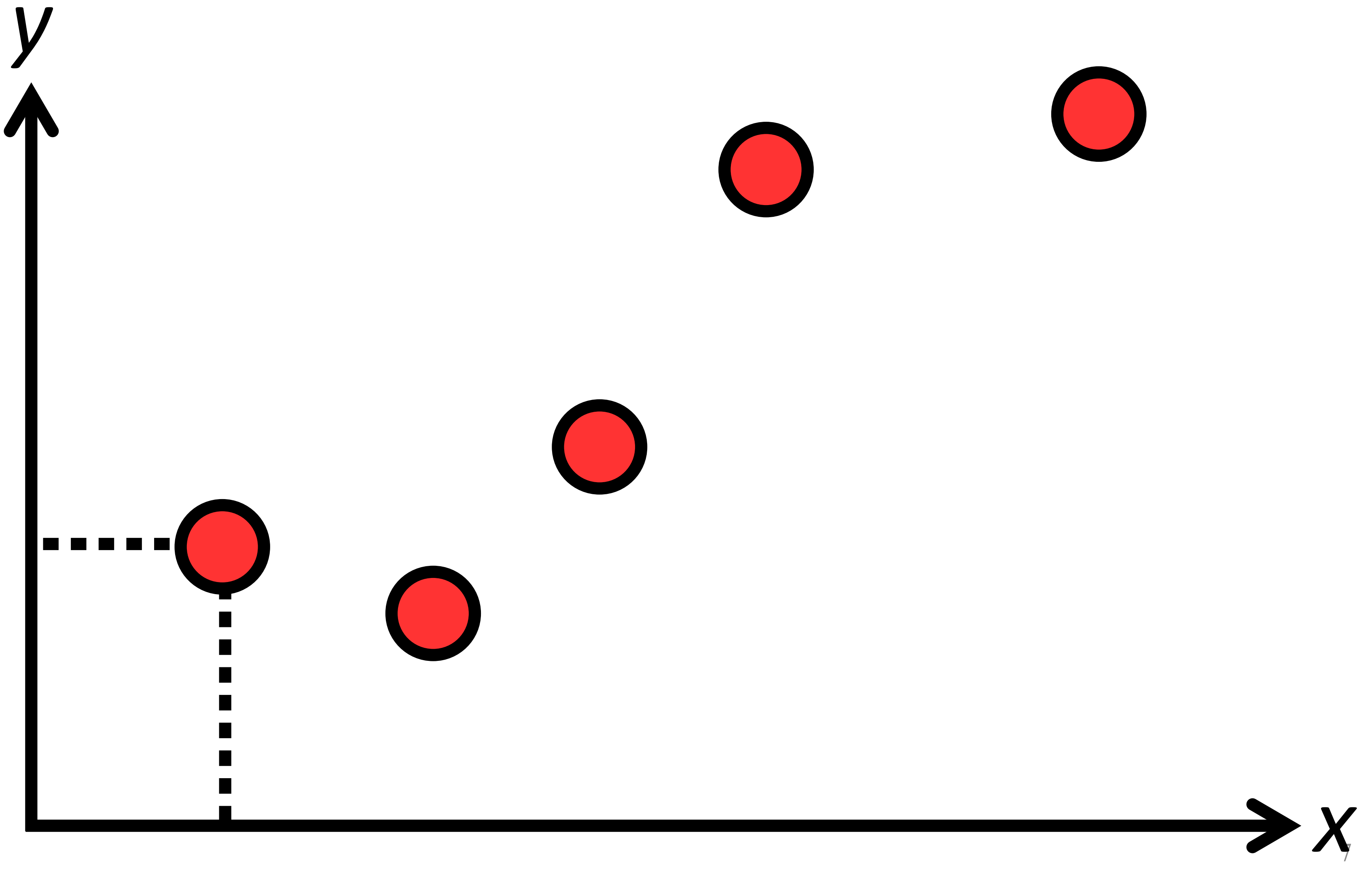
  $y = w_0 + w_1 x$

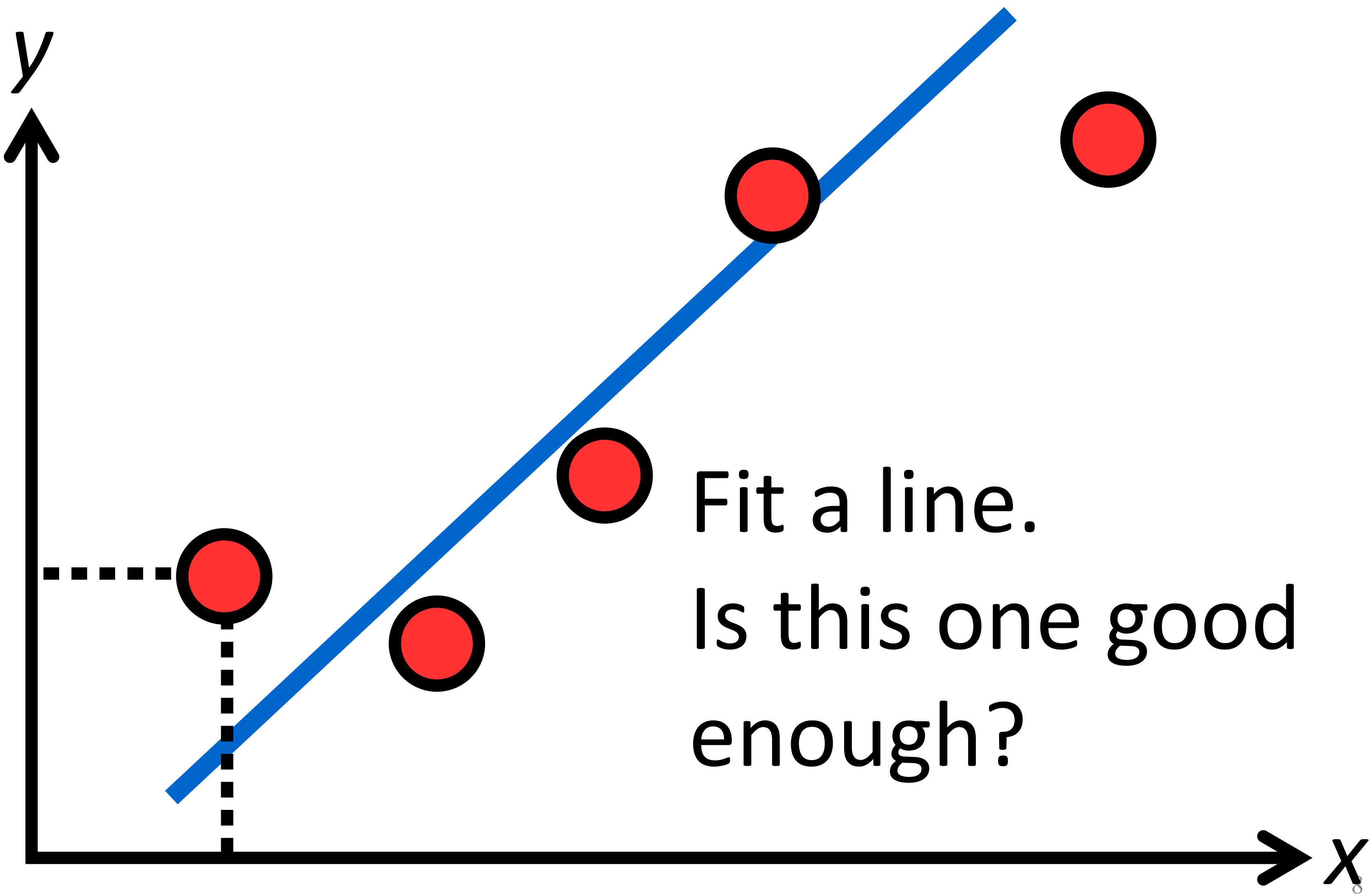
  where $\underline{w_0 \text{ (}y\text{-intercept)}}$ and $\underline{w_1 \text{ (slope)}}$ are regression coefficients

- Coefficients are calculated from the training data
- Method of least squares estimates the best-fitting straight line

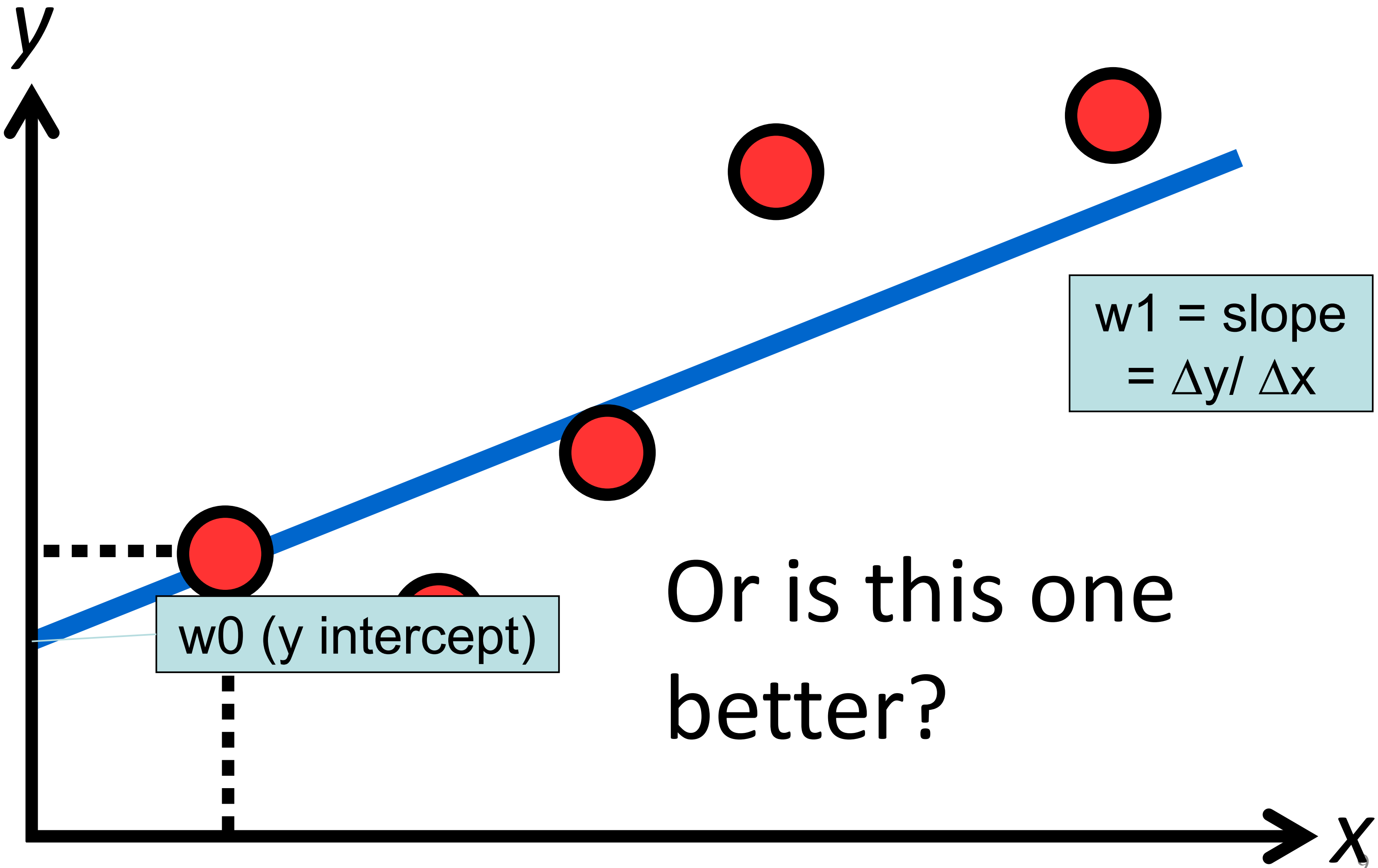
$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2} \qquad w_0 = \bar{y} - w_1 \bar{x}$$

Linear Regression: Process

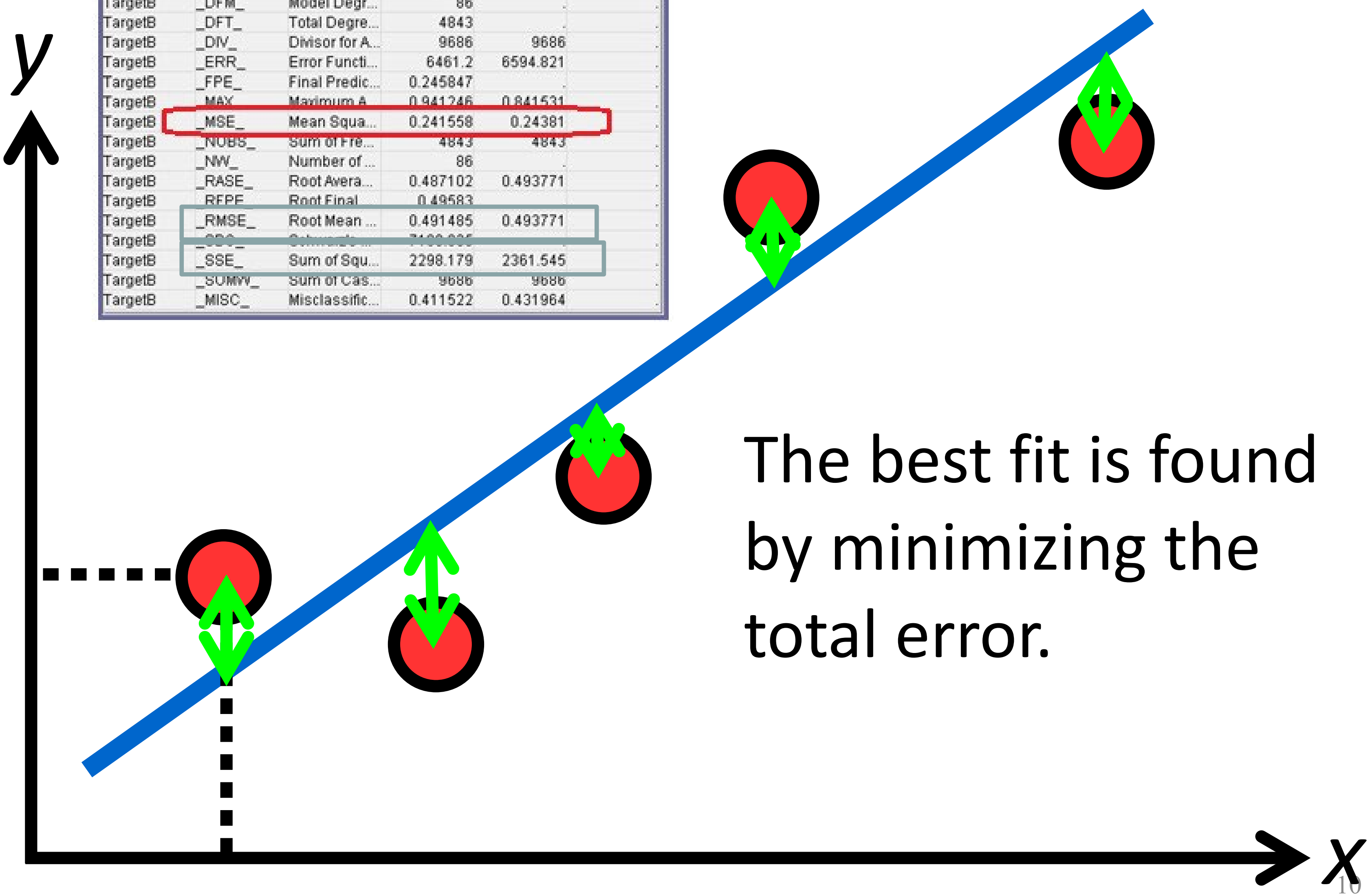



y
x
Fit a line.
Is this one good enough?

$$y = w_0 + w_1 x$$



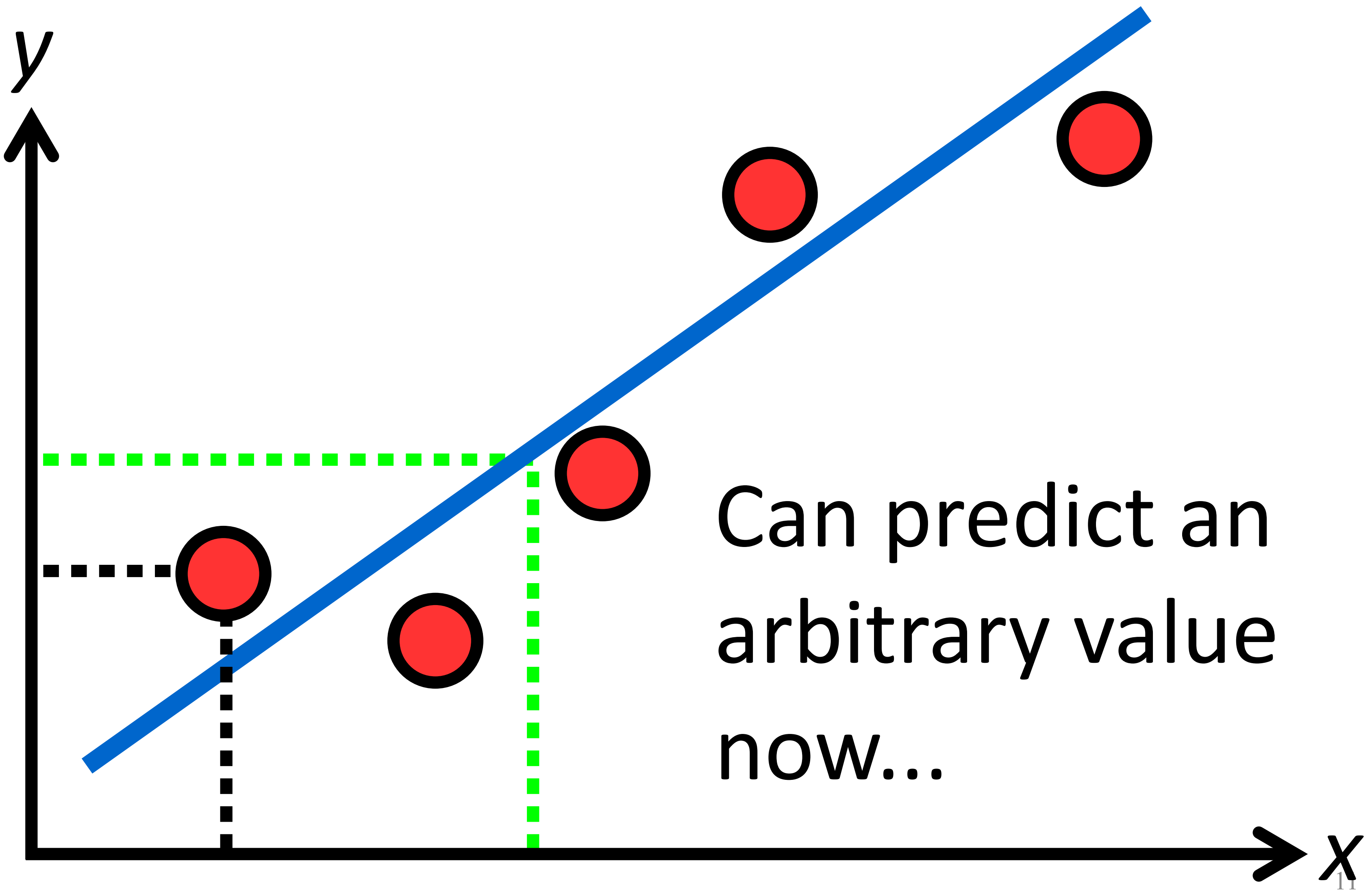
Or is this one better?



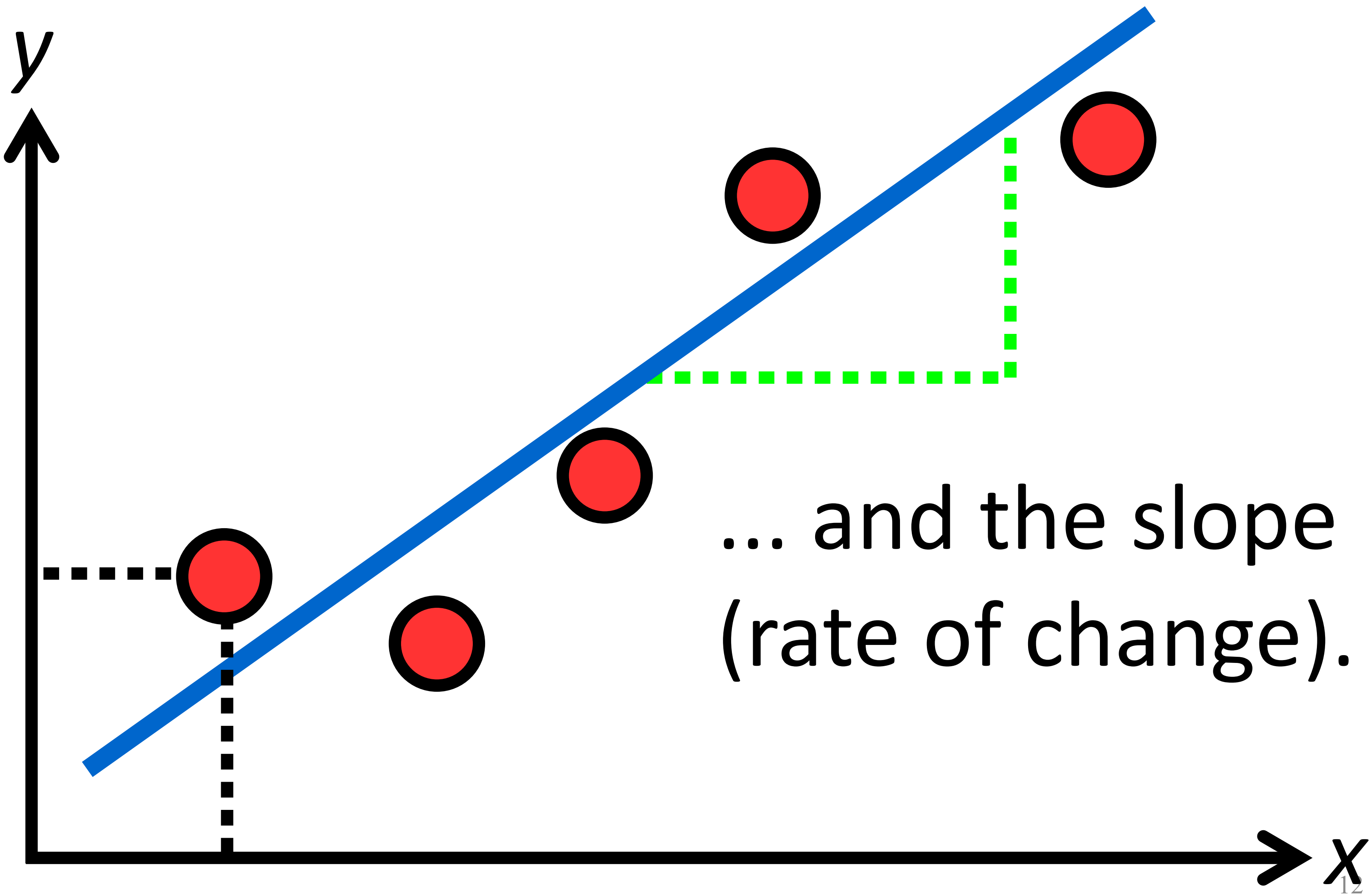
Fit Statistics

Target	Fit Statistics	Statistics Label	Train	Validation	Test
TargetB	_AIC_	Akaike's Inf...	6633.2	.	.
TargetB	_ASE_	Average Sq...	0.237268	0.24381	.
TargetB	_AVERR_	Average Err...	0.667066	0.680861	.
TargetB	_DFE_	Degrees of ...	4757	.	.
TargetB	_DFM_	Model Degr...	86	.	.
TargetB	_DFT_	Total Degre...	4843	.	.
TargetB	_DIV_	Divisor for A...	9686	9686	.
TargetB	_ERR_	Error Functi...	6461.2	6594.821	.
TargetB	_FPE_	Final Predic...	0.245847	.	.
TargetB	_MAX_	Maximum A...	0.941246	0.841531	.
TargetB	_MSE_	Mean Squa...	0.241558	0.24381	.
TargetB	_NOBS_	Sum of Fre...	4843	4843	.
TargetB	_NW_	Number of ...	86	.	.
TargetB	_RASE_	Root Avera...	0.487102	0.493771	.
TargetB	_RFPE_	Root Final ...	0.49583	.	.
TargetB	_RMSE_	Root Mean ...	0.491485	0.493771	.
TargetB	_SBC_	Schwarz's ...	[illegible]	.	.
TargetB	_SSE_	Sum of Squ...	2298.179	2361.545	.
TargetB	_SUMW_	Sum of Cas...	9686	9686	.
TargetB	_MISC_	Misclassific...	0.411522	0.431964	.

The best fit is found by minimizing the total error.



Can predict an arbitrary value now...



... and the slope (rate of change).

A Simple 2-D Data

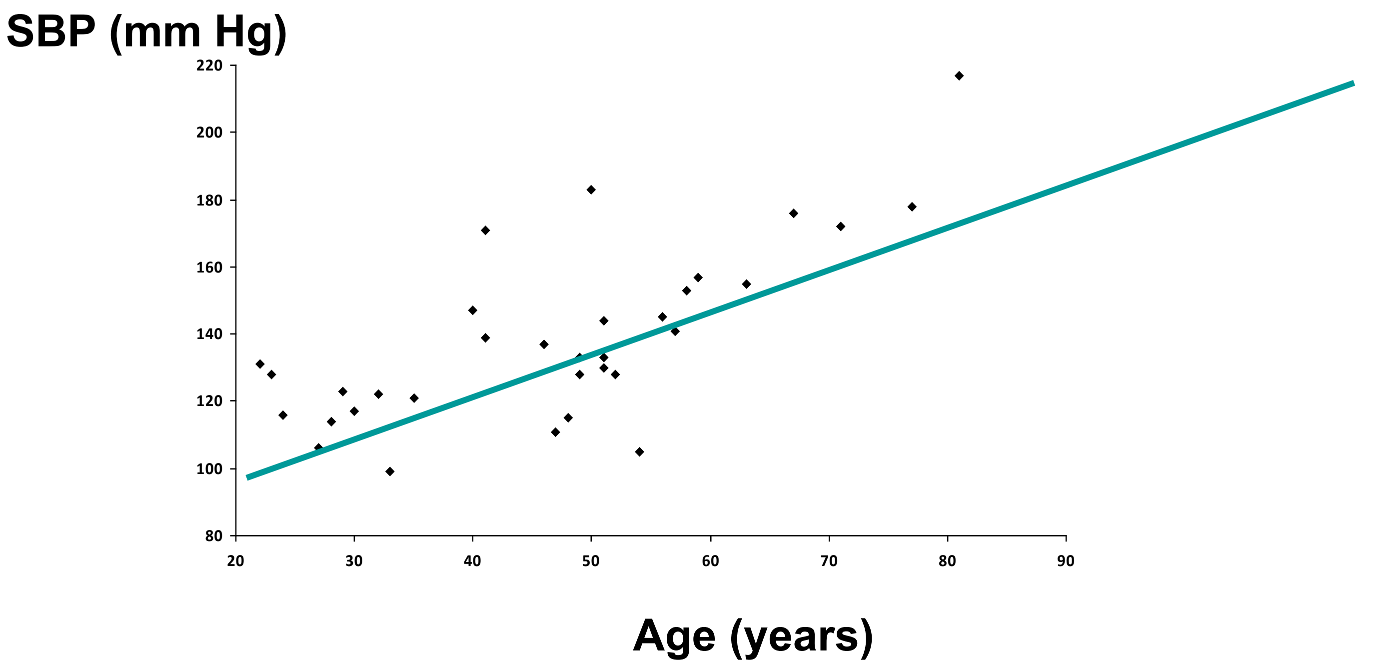
Table: Age and systolic blood pressure (SBP) among 33 adult women

Age	SBP	Age	SBP	Age	SBP
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217

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

A linear regression model

$$\text{SBP} = 95.54 + 1.222 \cdot \text{Age}$$



Another Example: Regression

- **Research question:** How fast does Coronary Heart Disease (CHD) mortality rise with a one unit increase in smoking? (Source: Howell, 2004)
- **Input attribute** = Av. # of cigs per adult per day
- **Target attribute** =

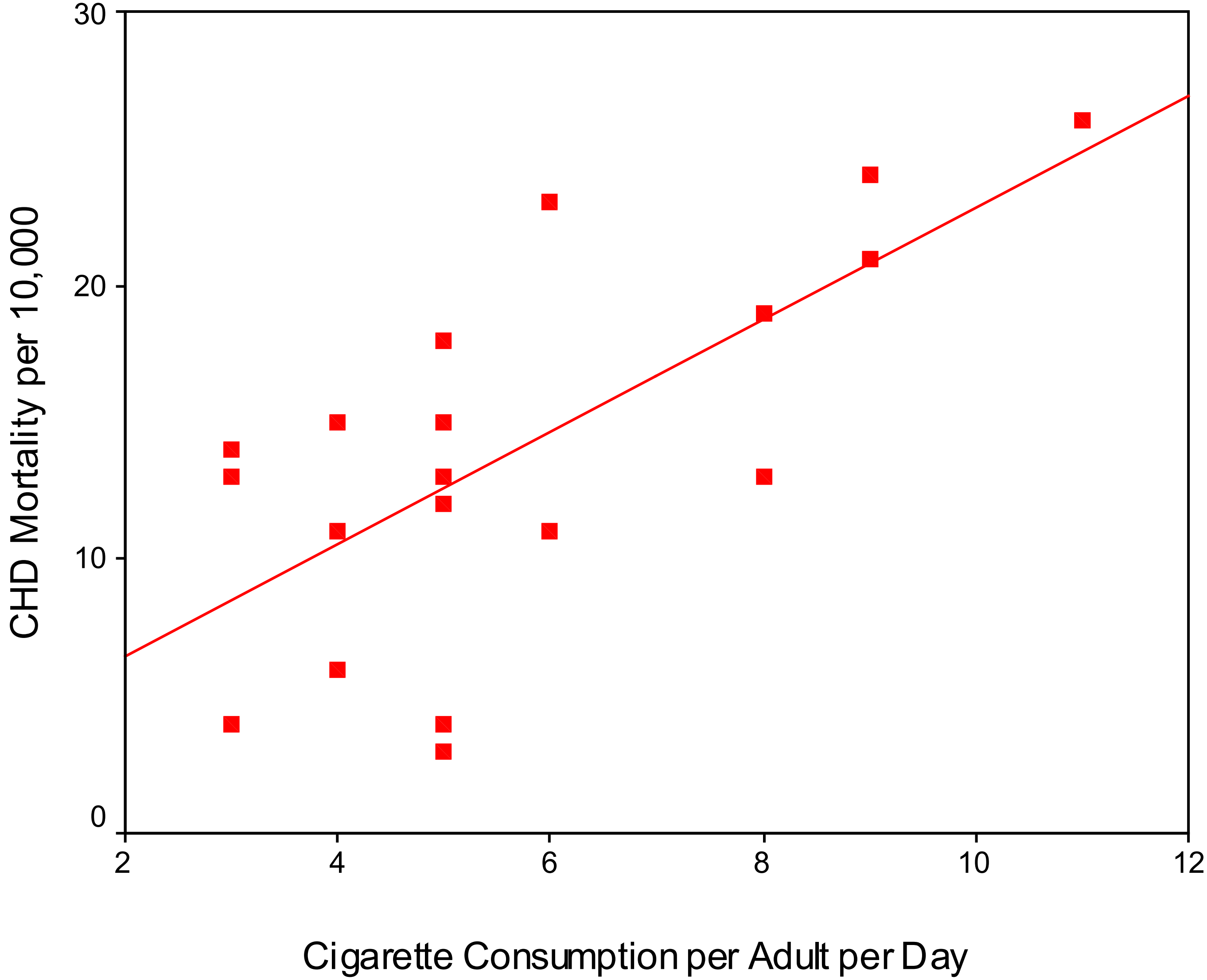
Cigarette Consumption and Coronary Heart Disease Mortality for 21 Countries

Cig.	11	9	9	9	8	8	8	6	6	5	5
CHD	26	21	24	21	19	13	19	11	23	15	13

Cig.	5	5	5	5	4	4	4	3	3	3
CHD	4	18	12	3	11	15	6	13	4	14

Cig. = Cigarettes per adult per day
CHD = Cornary Heart Disease Mortality per 10,000 population

Linear regression - Scatterplot with Line of Best Fit

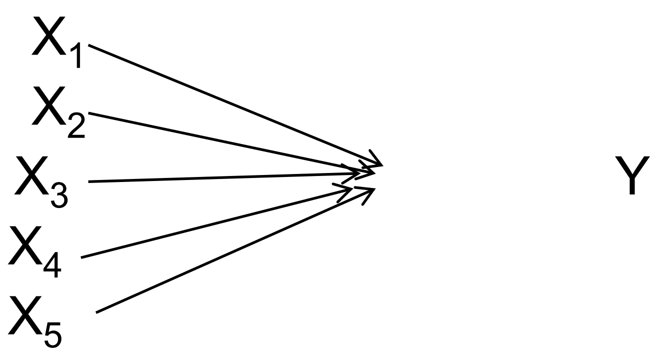


Multiple Linear Regression

Linear Regression

$$X \longrightarrow Y$$

Multiple Linear Regression



- For *2*-D data (i.e. with two input attributes), we may have:
  - $y = w_0 + w_1 x_1 + w_2 x_2$
- For *k*-D data (i.e. a data set with *k* number of attributes)

$$y = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_k x_k$$

- Solvable by extension of the least square method

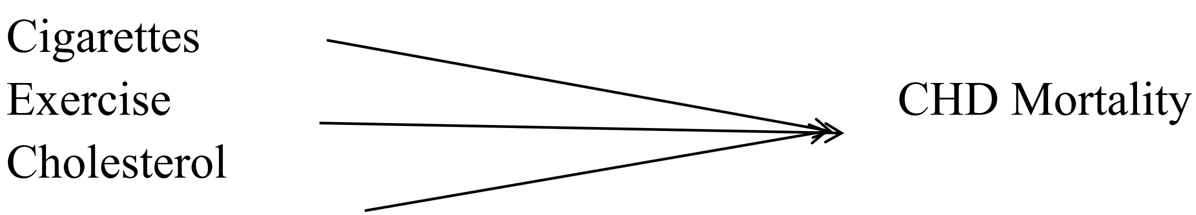
Example: Multiple Regression

Input attributes:

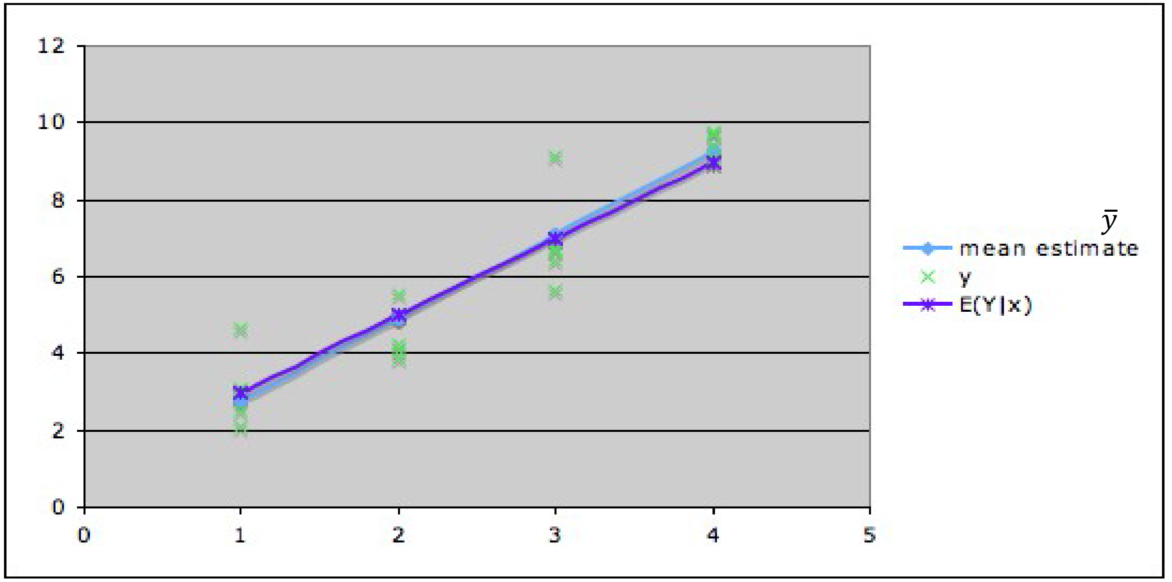
# of cigarettes per day; exercise; and cholesterol

Predict

CHD mortality



Interpretation of Regression Equation



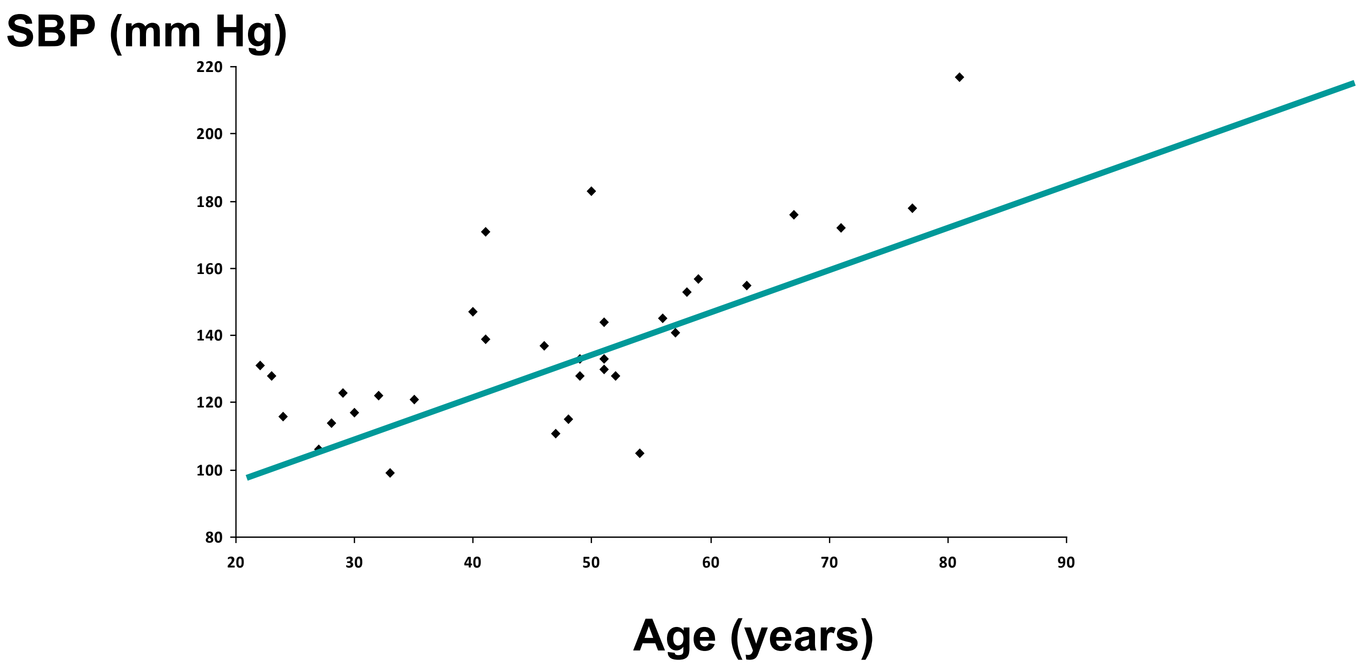
$$y = 0.56 + 2.18\,x$$

Interpreting the slope (w_1): For each change of one unit in x, the *average* change in the mean of Y is about 2.18 units.

Interpreting the Intercept (w_0): If x=0, then we predict y is 0.56.

Interpretation: Another example

$$\mathrm{SBP} = 95.54 + 1.222 \cdot \mathrm{Age}$$

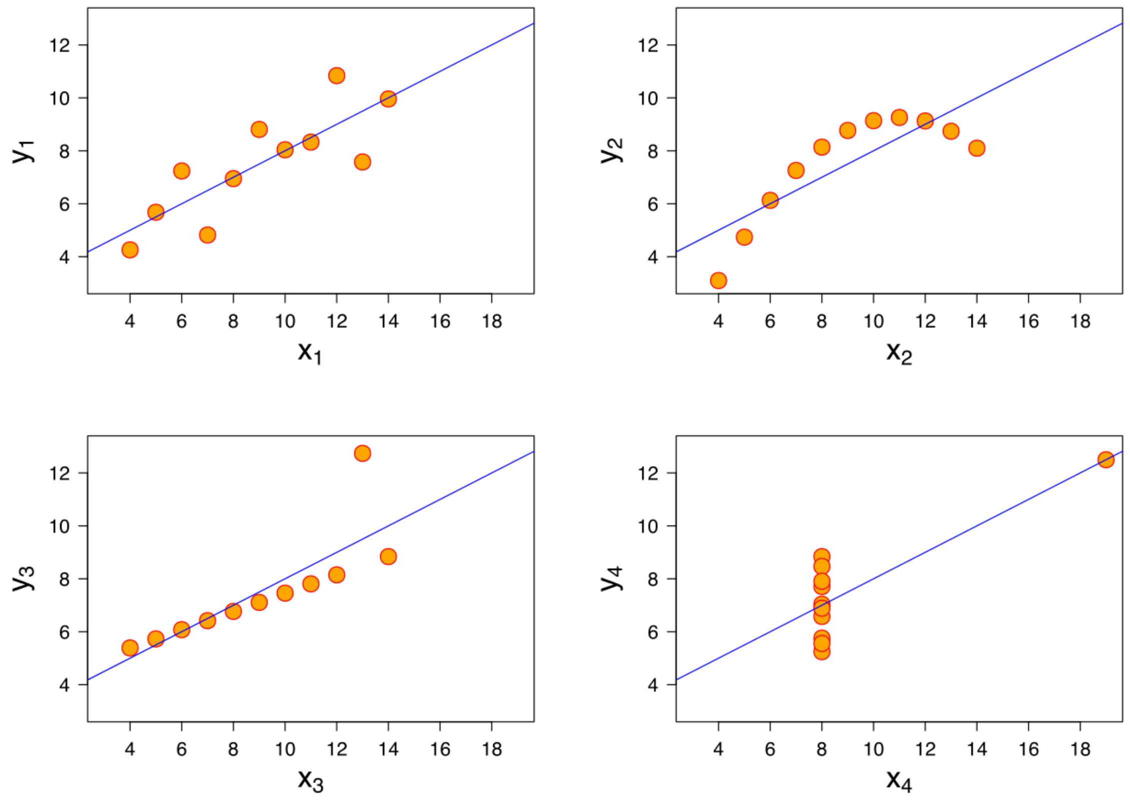


Interpreting the slope (w_1): For each change of one unit in Age, the *average* change in the mean of SBP is about 1.222 units.

Interpreting the Intercept (w_0): If Age=0, then we predict SBP is 95.54.

Model Evaluation: R^2-values

We can **always** fit a linear model to any dataset, but how do we know if there is a **real linear relationship**?



R^2-values

Approach: Measure how much the total "noise" (variance) is reduced when we include the line as an offset.

R-squared: a suitable measure. Let $\hat{y} = X\,\hat{w}$ be a predicted value, and $\bar{y}$ be the sample mean. Then the R-squared value is

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

And can be described as the fraction of the total variance not explained by the model.

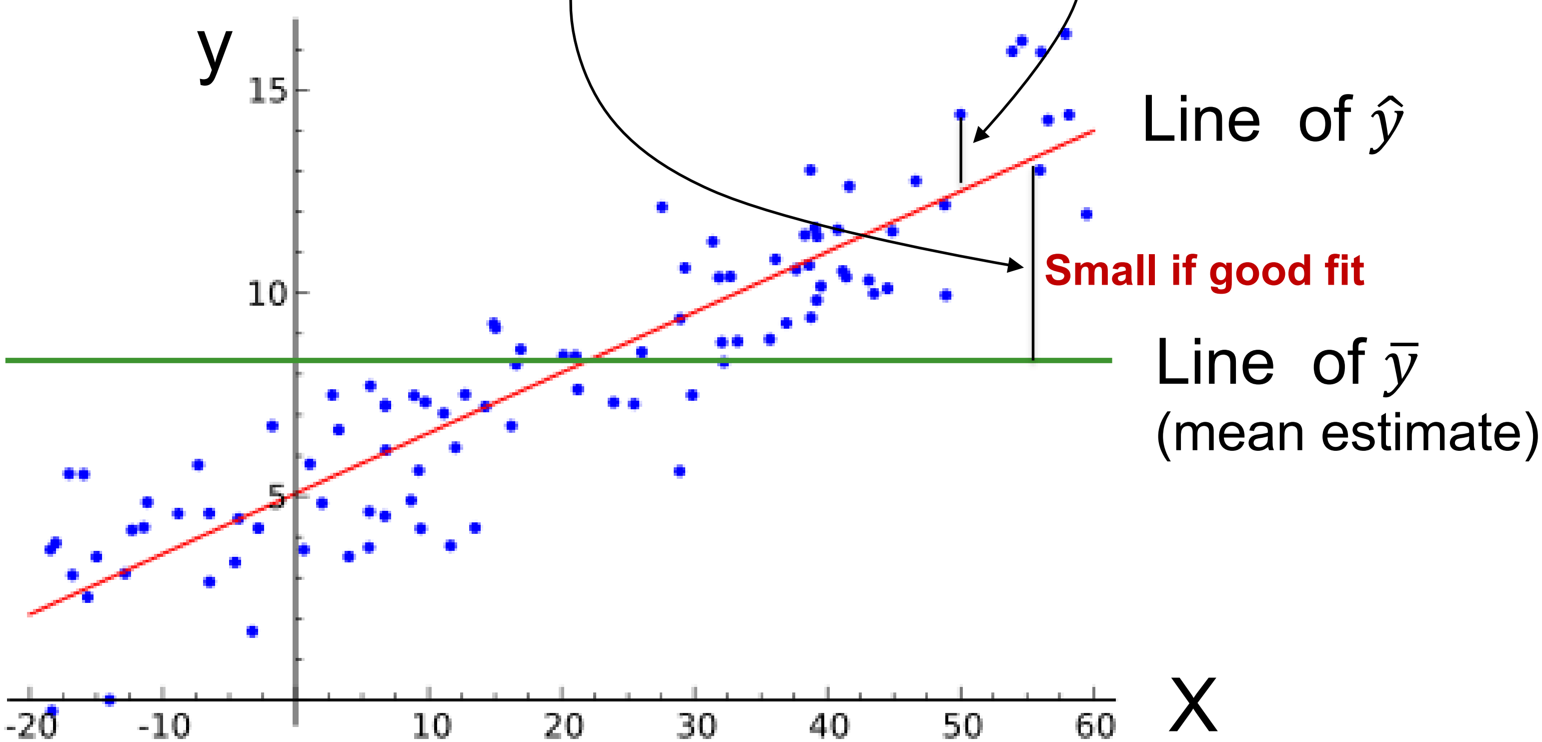
$R^2 = 0$: bad model. No evidence of a linear relationship.

$R^2 = 1$: good model. The line perfectly fits the data.

R-squared Coefficient

Large if good fit

$$R^2 = 1 - \frac{\sum(y_i - \widehat{y_i})^2}{\sum(y_i - \bar{y})^2}$$



Regression Modelling for Classification

Logistic Regression

Why use logistic regression?

- There are many data problems for which the target (or dependent) variable is "limited."
- For example, voting, morbidity or mortality, and participation data are not continuous or distributed normally.
- Binary logistic regression is a type of regression analysis where the target variable is a dummy variable: coded 0 (did not vote) or 1(did vote)

Categorical Target Variables

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

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

Proportion of “Success”: π

- In ordinary regression, the model predicts the *mean* Y for any combination of input variables.

- What’s the “mean” of a 0/1 indicator variable?

- Goal of logistic regression: Predict the “true” proportion of success, π, at any value of the varaible.

$$\bar{y} = \frac{\sum y_i}{n} = \frac{\#\text{ of } 1's}{\#\text{ of trials}} = \text{Proportion of "success"}$$

Logistic Regression Model

Y = Binary response

X = Quantitative variable

π = proportion of 1's (yes, success) at any X or P(Y|X)

- Equivalent forms of the logistic regression model:

Logit form

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \quad X$$

Probability form

$$\Pi \text{ or } P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

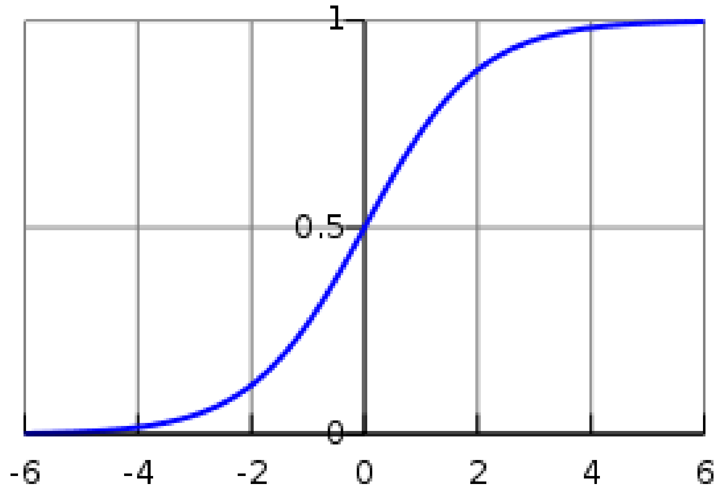
N.B.: This is natural log (aka "ln")

- π or p is the probability that the event Y occurs, p(Y=1)
- p/(1-p) is the "odds ratio"
- ln[p/(1-p)] is the log odds ratio, or "logit"

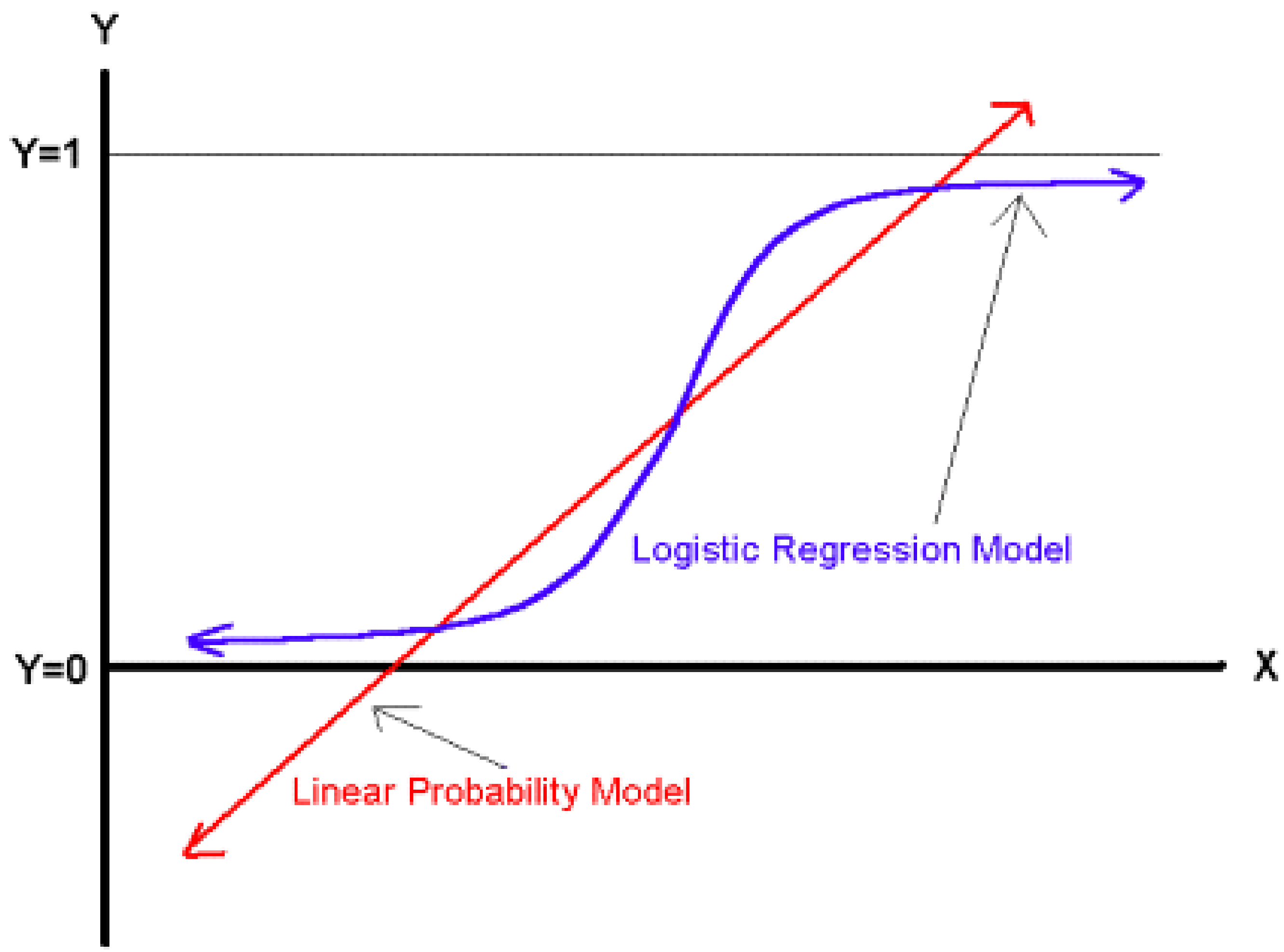
What does this function look like?

Logit Function or Logistic Regression

- The logistic distribution constrains the estimated probabilities to lie between 0 and 1.
- The estimated probability is: $p(X) = \frac{1}{1+\exp(-X\beta)}$
  - if βX =0, then p = .50
  - if βX gets really big, p approaches 1
  - if βX gets really small, p approaches 0



Comparing the LP and Logit Models



Maximum Likelihood Estimation (MLE)

- MLE is a statistical method for estimating the model coefficients (α, β).
- The likelihood function (L) measures the probability of observing the particular set of input variable values ($v_1, v_2, \ldots, v_n$) that occur in the sample:

  $$L = \text{Prob}\,(v_1 * v_2 * * * v_n)$$

- The higher the L, the higher the probability of observing the v's in the sample.

Maximum Likelihood Estimation (MLE)

- MLE finds the coefficients (α, β)
  - that make the log of the likelihood function (LL < 0) as large as possible.
  - Or that make -2 times the log of the likelihood function (-2LL) as small as possible.
- MLE solves the following condition:

$$\{Y - p(Y=1)\}X_i = 0$$

summed over all observations, i = 1,…,n

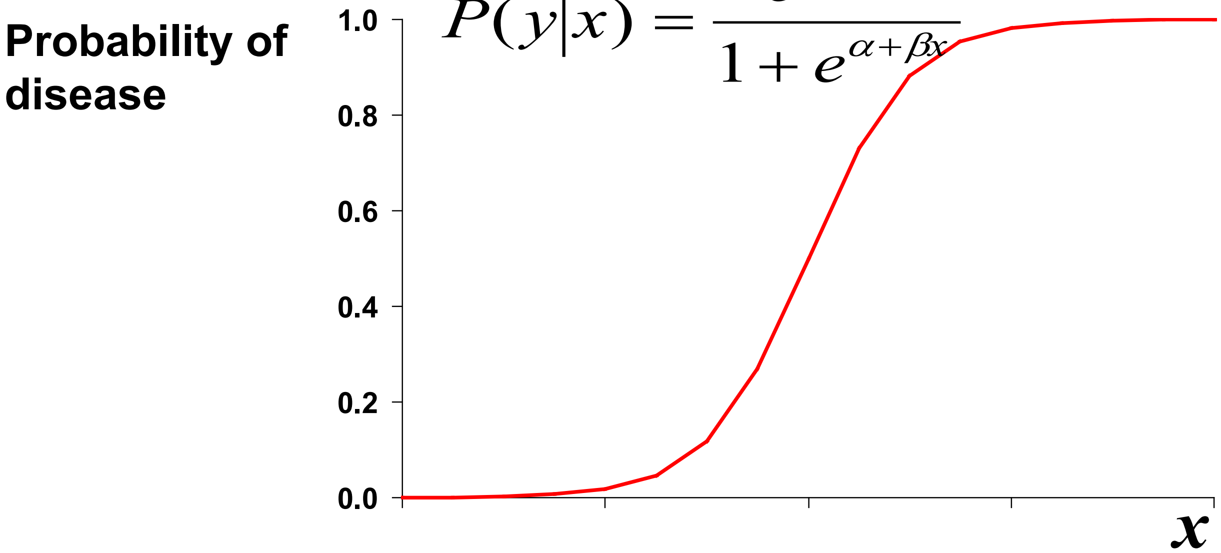
Logistic regression Data: An Example

Table : 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 function and Transformation

$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$



$$\underbrace{\ln\left[\frac{P(y|x)}{1-P(y|x)}\right]}_{\text{logit of } P(y|x)} = \alpha + \beta x$$

✓α = log odds of disease in unexposed

✓β = log odds ratio associated with being exposed

✓e^{β} = odds ratio

Linear versus Logistic Regression

Linear Regression

- Target is an interval attribute.
- Input attributes have any measurement level.
- Predicted values are the mean of the target attribute at the given values of the input attributes.

Logistic Regression

- Target is a categorical attribute.
- Input attributes have any measurement level.
- Predicted values are the probability of a particular level(s) of the target attribute at the given values of the input attributes.

Pros and Cons of Linear/Logistic Regression Models

Pros

+ Fast application

+simplicity, interpretability, scientific acceptance, and widespread availability

+Usually the first method to use for many problems

Cons

- many real-world phenomena do not correspond to the assumptions of a linear model; in these cases, it is difficult to produce useful results

-Cannot handle a large number of features or missing values

Conclusion: Use regression models only if the data is relatively clean and small.

Regression Modelling: Classification and Regression

Nonlinear Regression

&

Support Vector Machine

A quick tour

Generalized Linear Model

A flexible generalization of ordinary linear regression

(1) Allowing the linear model to be related to the response variable via a link function. What is the link between Y and $b_0 + b_1X$?

(a) Regular reg: identity

(b) Logistic reg: logit

(c) Poisson reg: log

(2) allowing the magnitude of the variance of each measurement to be a function of its predicted value. What is the distribution of Y given X?

(a) Regular reg: Normal (Gaussian)

(b) Logistic reg: Binomial

(c) Poisson reg: Poisson

Nonlinear Regression

- Linear regression is not appropriate if data exhibits non-linear dependencies
- But: can serve as building blocks for more complex schemes (i.e. model trees)
- Nonlinear regression is used to fit the non-linear dependencies.
- Some nonlinear models can be modeled by a polynomial function

Nonlinear Regression

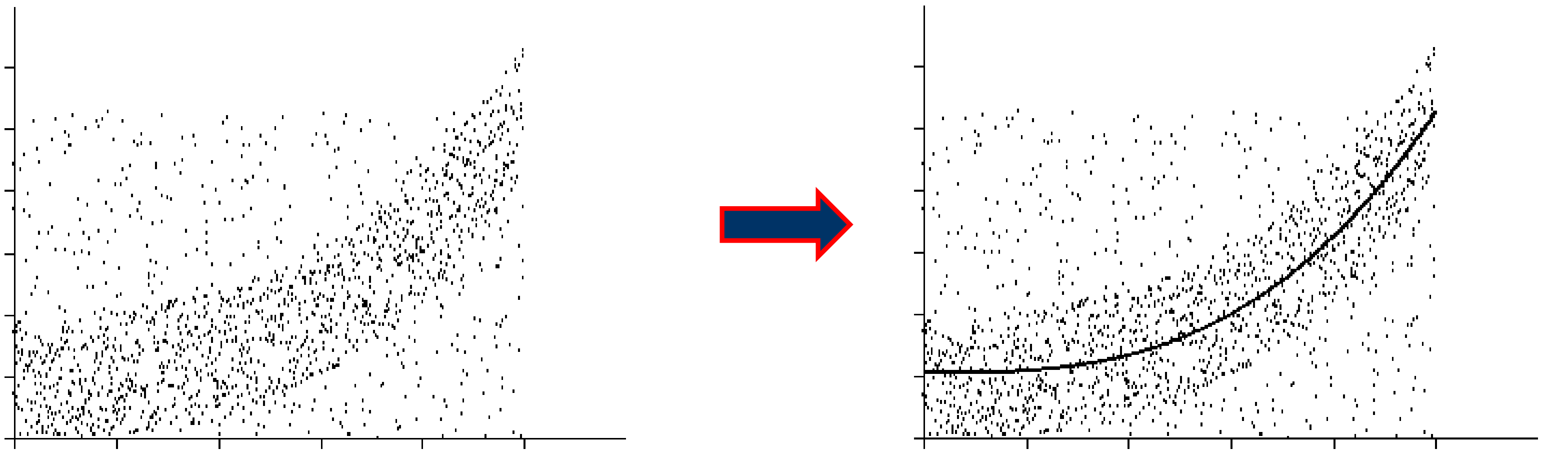
- A polynomial regression model can be transformed into linear regression model. For example,

  $$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

  convertible to linear with new variables: $x_2 = x^2$, $x_3 = x^3$

  $$y = w_0 + w_1 x + w_2 x_2 + w_3 x_3$$

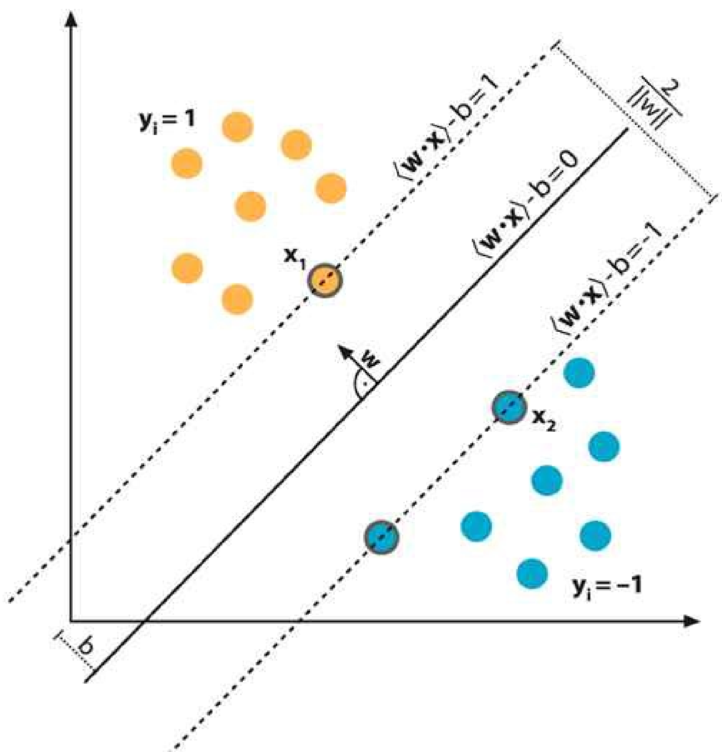
- Other functions, such as the power function, can also be transformed to a linear model
- Some models are intractable nonlinear (e.g., the sum of exponential terms)
  - possible to obtain least square estimates through extensive calculation on more complex formulae



Common Nonlinear function choices include Power, Logarithmic, Exponential, but any continuous function can be used.

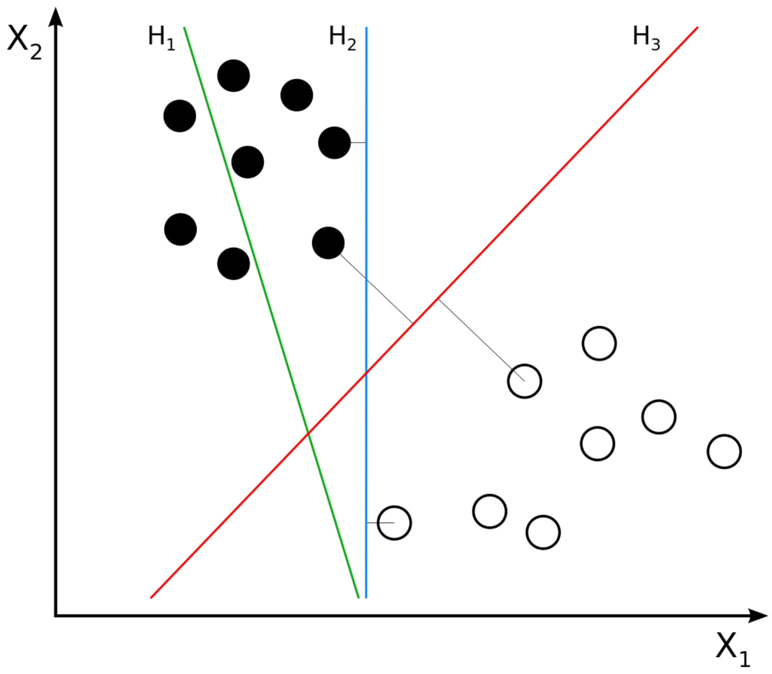
Support Vector Machines

- A Support Vector Machine (SVM) is a classifier that tries to **maximize the margin** between training data and the classification boundary (the plane defined by $X\beta = 0$)



Support Vector Machines

- The idea is that maximizing the margin **maximizes the chance that classification will be correct on new data**. We assume the new data of each class is near the training data of that type.



Summary

Final Remarks

- Two types of Predictive modelling
  - Classification: for categorical target attribute
  - Regression: for numerical target attribute
- Classification algorithms
  - Decision Tree, Neural Networks, Logistic Regression, Nearest-neighbour
  - Many others Naïve Bayes, Support Vector Machine, Genetic algorithms, etc
- Regression algorithms
  - Several regression functions

References

- Data Mining techniques and concepts by Han J et al, 2011.
- Discovering Data Mining, by Cabena, et al., 1997.
- Predictive Data Mining, by Weiss and Indurkhya, 1999.