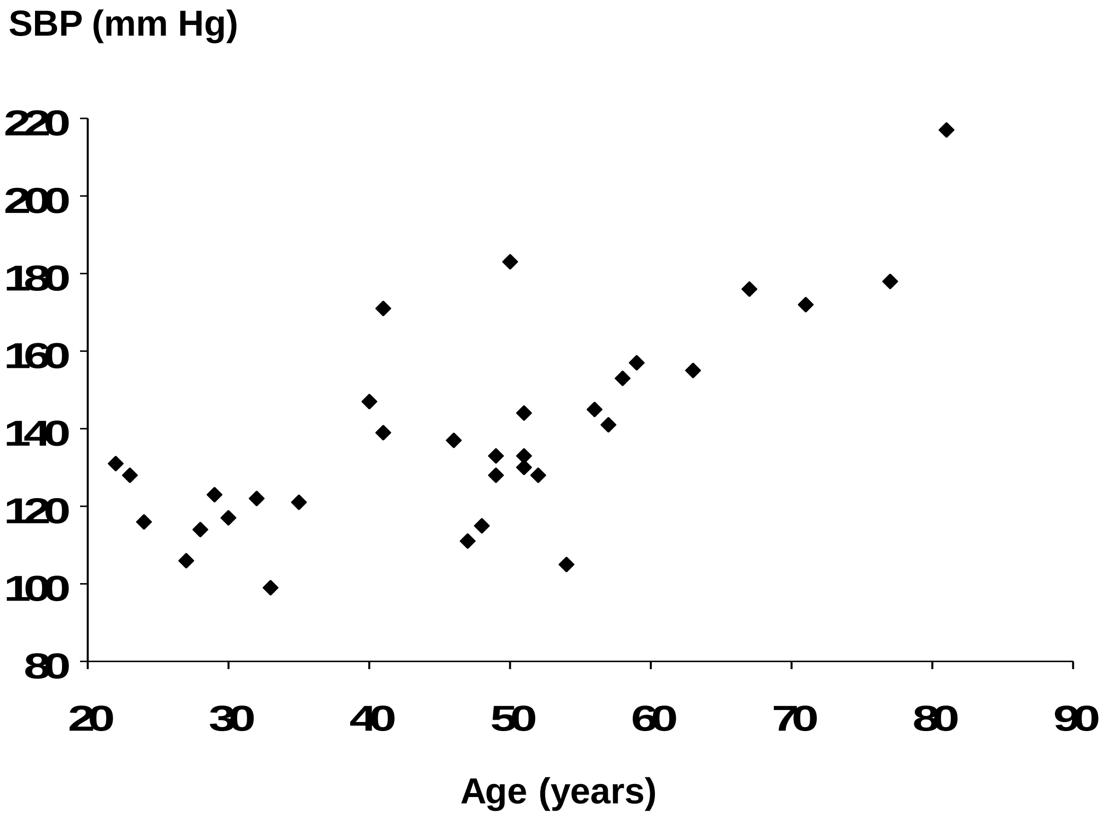
CMSC 471

ML: Regression

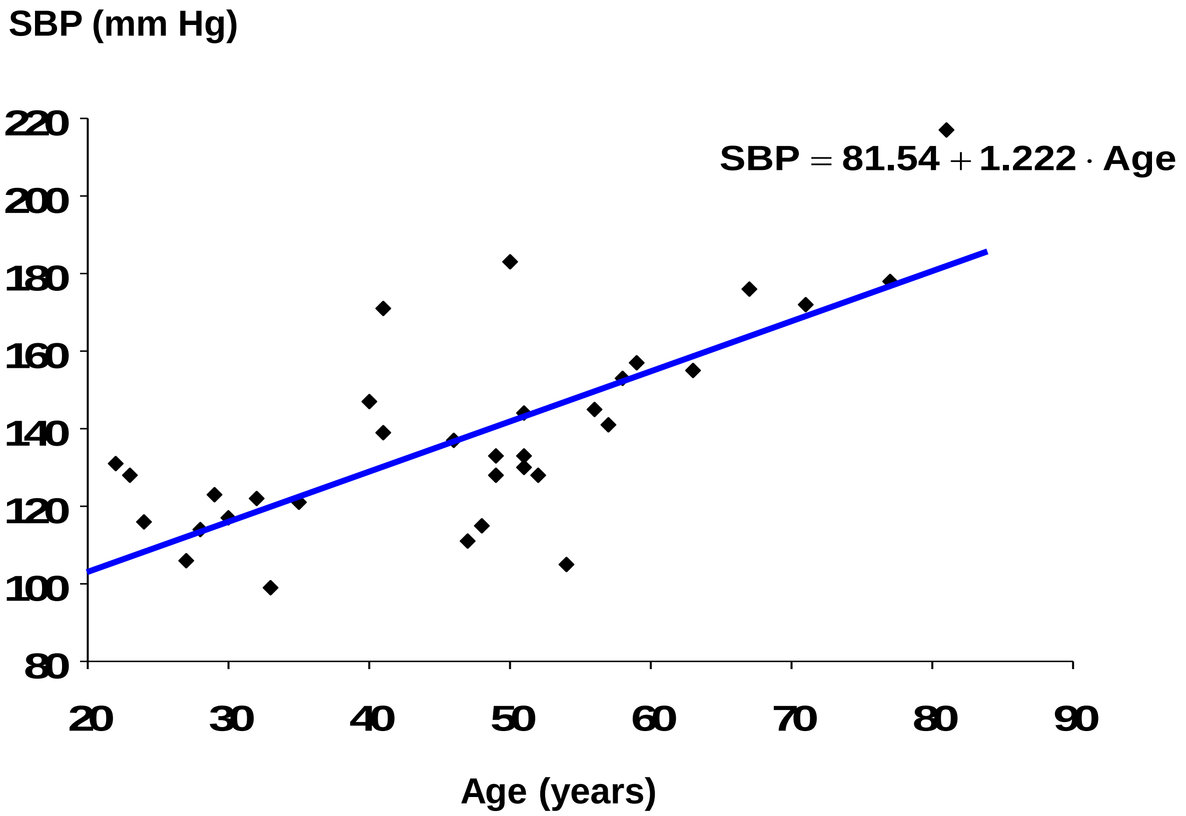
Simple linear regression

Table 1 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



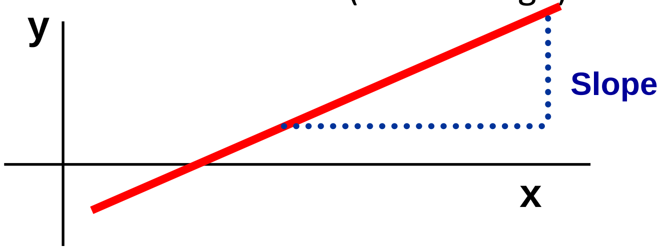
adapted from Colton T. Statistics in Medicine. Boston: Little Brown, 1974



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

Simple linear regression

- Relation between 2 continuous variables (SBP and age)



$$y = \alpha + \beta_1 x_1$$

- Regression coefficient β_1
  - Measures association between y and x
  - Amount by which y changes on average when x changes by one unit
  - Least squares method

Fitting equation to the data

- Finding the best curve = learning the best parameters
- Which is the best parameter?
  - One with minimum error
- Calculating error: [Mean squared error]

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\underbrace{Y_i}_{\text{actual}} - \underbrace{\hat{Y_i}}_{\text{predicted}}\right)^2.$$

Multiple linear regression

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

- Relation between a continuous variable and a set of i continuous variables

- Partial regression coefficients β_i
  - Amount by which y changes on average
    when x_i changes by one unit
    and all the other x_is remain constant
  - Measures association between x_i and y adjusted for all other x_i

- Example
  - SBP *versus* age, weight, height, etc

Multiple linear regression

$$\underline{\mathbf{y}} \quad = \quad \underline{\boldsymbol{\alpha} + \boldsymbol{\beta}_1\mathbf{x}_1 + \boldsymbol{\beta}_2\mathbf{x}_2 + \ldots + \boldsymbol{\beta}_i\mathbf{x}_i}$$

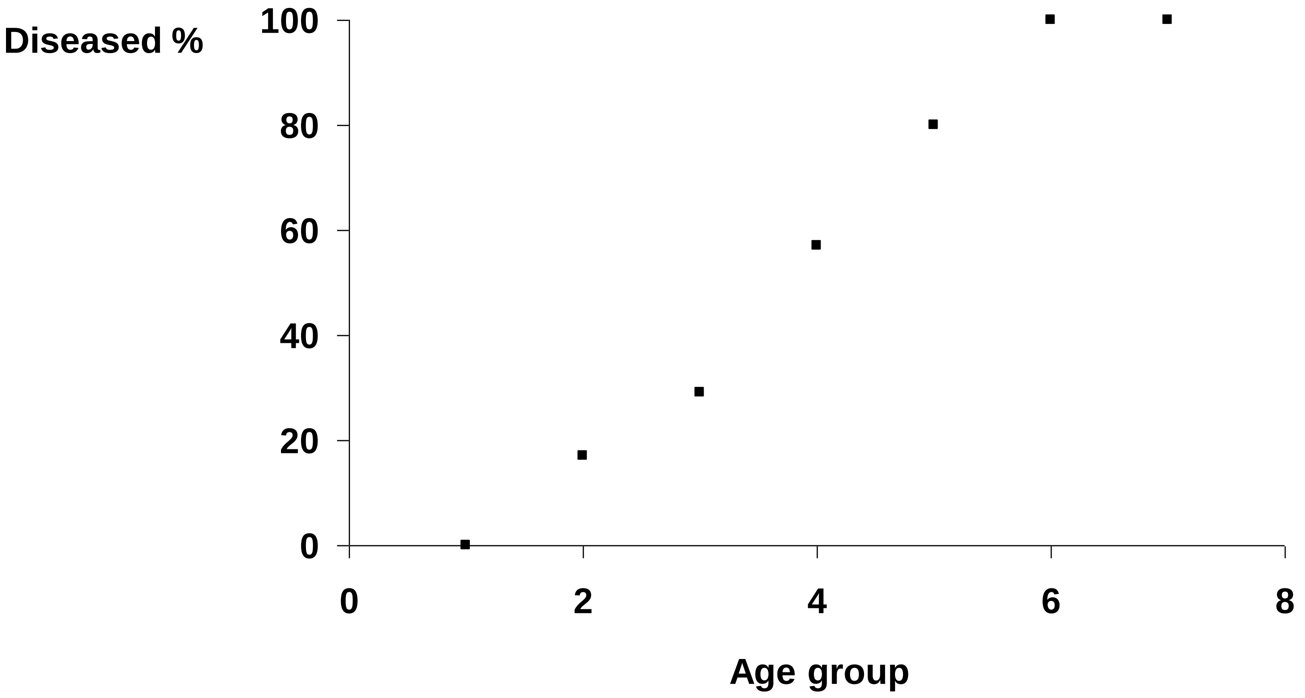
Predicted	Predictor variables
Response variable	Explanatory variables
Outcome variable	Covariables
Dependent	Independent variables

Logistic regression

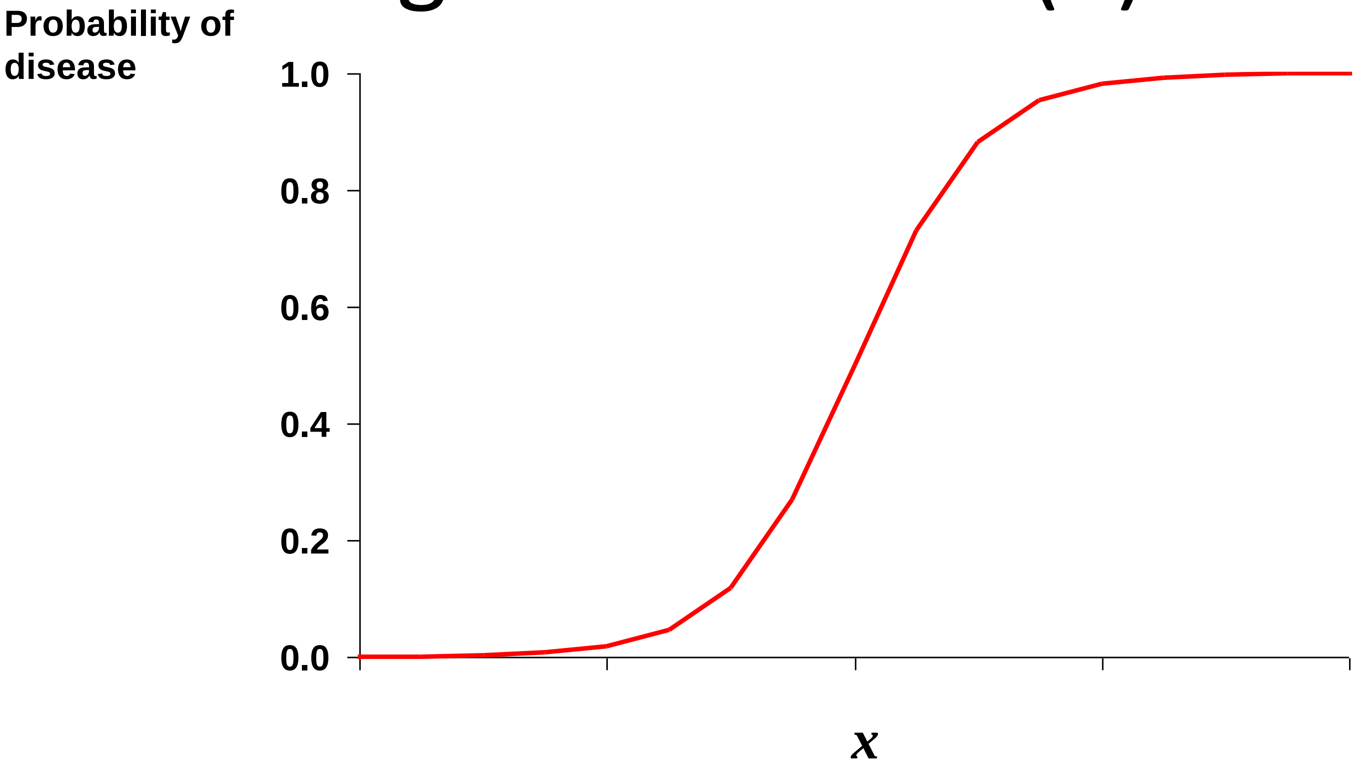
Table 2 Prevalence (%) of signs of CD according to age group

Age group	# in group	Diseased #	Diseased %
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

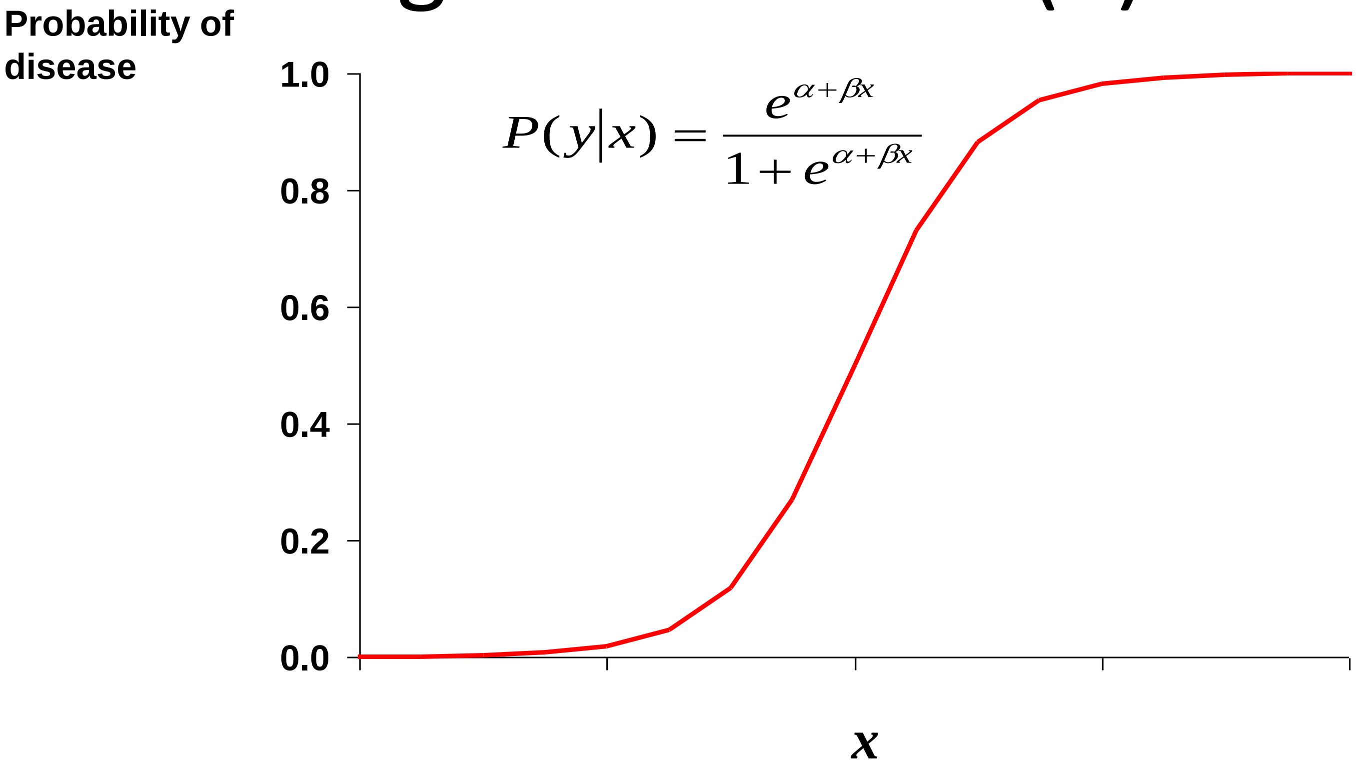
Dot-plot: Data from Table 2



Logistic function (1)



Logistic function (1)



Fitting equation to the data

- Can not use MSE
- Estimate Maximum likelihood
- Likelihood function
  - Estimates parameters α and β
  - Practically easier to work with log-likelihood

$$L(\mathrm{B}) = \ln\left[l(\mathrm{B})\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i)\ln\left[1 - \pi(x_i)\right] \right\}$$