Correlation and regression

Yongjua Laosiritaworn

Introductory on Field Epidemiology
6 July 2015, Thailand

Data

- Quantitative Exposure variable (*X)*
- Quantitative Disease variable (*Y)*
- Objective: To quantify the *linear* relationship between X and Y

Table 14.1 Synonyms for *explanatory variable* and *response variable*.

Explanatory Variable	→	Response Variable
X	→	*Y*
independent variable	→	dependent variable
factor	→	outcome
treatment	→	response
exposure	→	disease

Illustrative data (Doll, 1955)

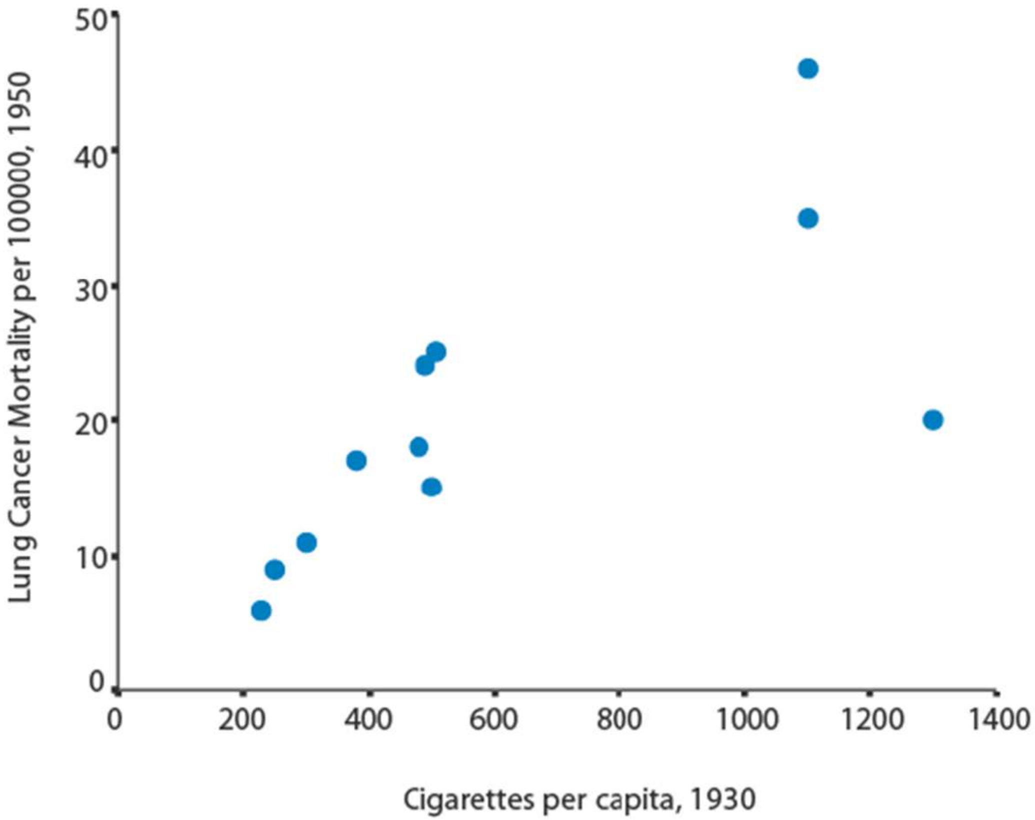
per capita cigarette consumption (X)

lung cancer mortality per 100,000 in 1950 (Y)

COUNTRY	CIG1930	LUNGCA
USA	1300	20
Great Britain	1100	46
Finland	1100	35
Switzerland	510	25
Canada	500	15
Holland	490	24
Australia	480	18
Denmark	380	17
Sweden	300	11
Norway	250	9
Iceland	230	6

$n = 11$

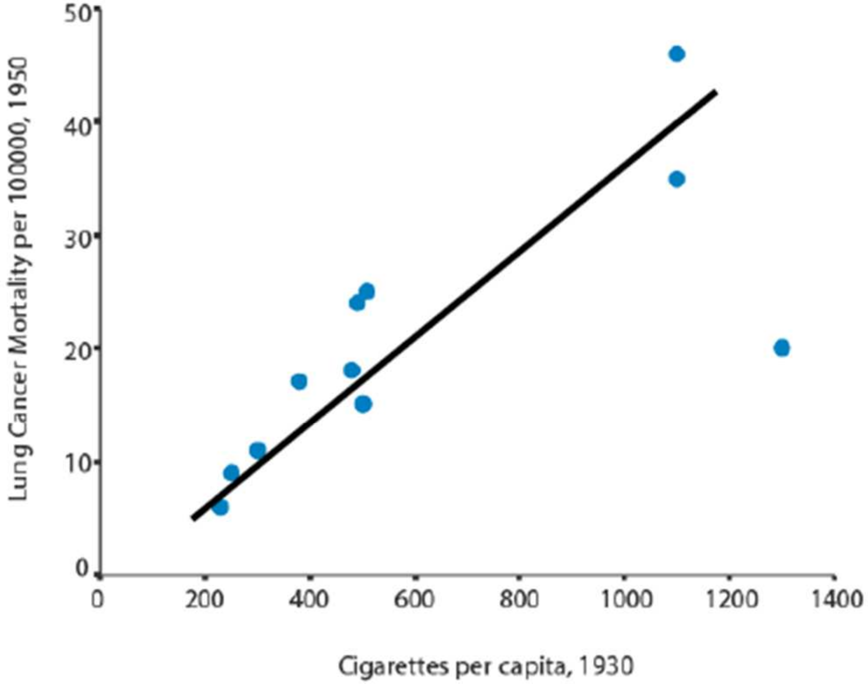
Scatter plot



Assess:
Form
Direction of association
Outliers
Strength of relation

n = 11

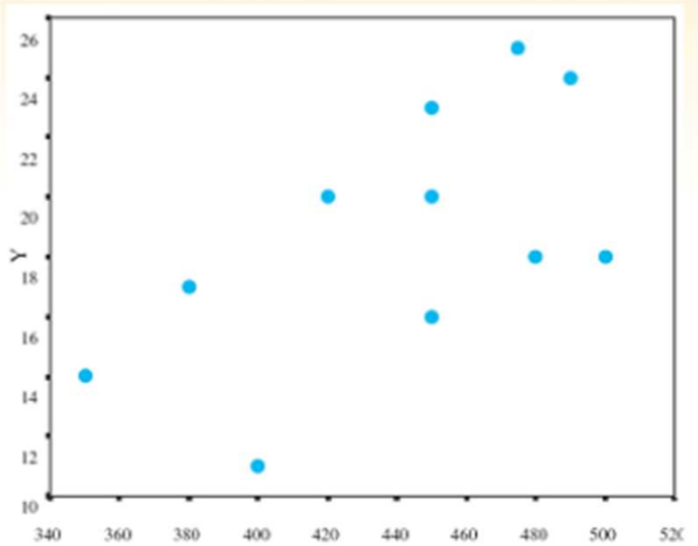
Doll, 1955



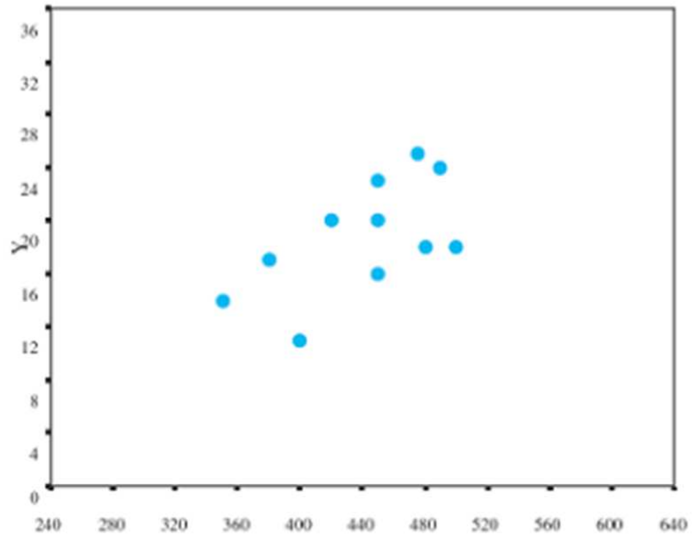
- Form: linear
- Direction: positive association
- Outlier: no clear outliers
- Strength: difficult to determine by eye

The **eye** is *not* a good judge of strength

Identical data sets on differently scaled axes



This relation appears to be weak



This relation appears strong

The different appearances in strength is an artifact of the axis scaling

Correlation coefficient, *r*

- $r \equiv$ *Pearson's product-moment correlation coefficient*
- Measures degree to which X and Y "go together"
- Always between −1 and 1
- $r \approx 0 \Rightarrow$ no correlation
- $r > 0 \Rightarrow$ positive correlation
- $r < 0 \Rightarrow$ negative correlation
- Closer r is to 1 or −1, the **stronger** the correlation

Karl Pearson

1857 - 1936

Interpretation of *r*

- **Direction of association:** positive, negative, ~0
- **Strength of association**
  - close to 1 or –1 ⇒ "strong"
  - close to 0 ⇒ "weak"
  - guidelines
    - if $|r| \geq .7$ ⇒ say "strong"
    - if $|r| \leq .3$ ⇒ say "weak"

r by hand

$$r = \frac{1}{n-1} \sum z_X z_Y$$

$$\text{where } z_X = \frac{x_i - \bar{x}}{s_X} \text{ and } z_Y = \frac{y_i - \bar{y}}{s_Y}.$$

- z quantify distance above or below mean in standard deviations units.
- When z scores track in same directions ⇒ *products* are positive
- When z scores track in opposite directions ⇒ *products* are negative

r by hand

Table 14.3 Calculation of correlation coefficient *r*, illustrative data.

$$\bar{x} = \frac{6640}{11} = 603.6364$$

$$\bar{y} = \frac{226}{11} = 20.54545$$

i	Country	X	$z_x = \frac{x_i - \bar{x}}{s_x}$	Y	$z_y = \frac{y_i - \bar{y}}{s_y}$	$z_X \cdot z_Y$
1	US	1300	1.840	20	−0.047	− 0.086
2	Great Britain	1100	1.312	46	2.171	2.847
3	Finland	1100	1.312	35	1.233	1.617
4	Switzerland	510	−0.247	25	0.380	−0.094
5	Canada	500	−0.274	15	−0.473	0.130
6	Holland	490	−0.300	24	0.295	−0.088
7	Australia	480	−0.327	18	−0.217	0.071
8	Denmark	380	−0.591	17	−0.302	0.179
9	Sweden	300	−0.802	11	−0.814	0.653
10	Norway	250	−0.934	9	−0.985	0.920
11	Iceland	230	− 0.987	6	−1.241	1.225
	Sums →	6640	0	226	0	7.373

$$r = \frac{1}{n-1} \sum z_x z_y = \frac{1}{11-1} \cdot 7.373 = 0.737$$

Coefficient of determination (r^2)

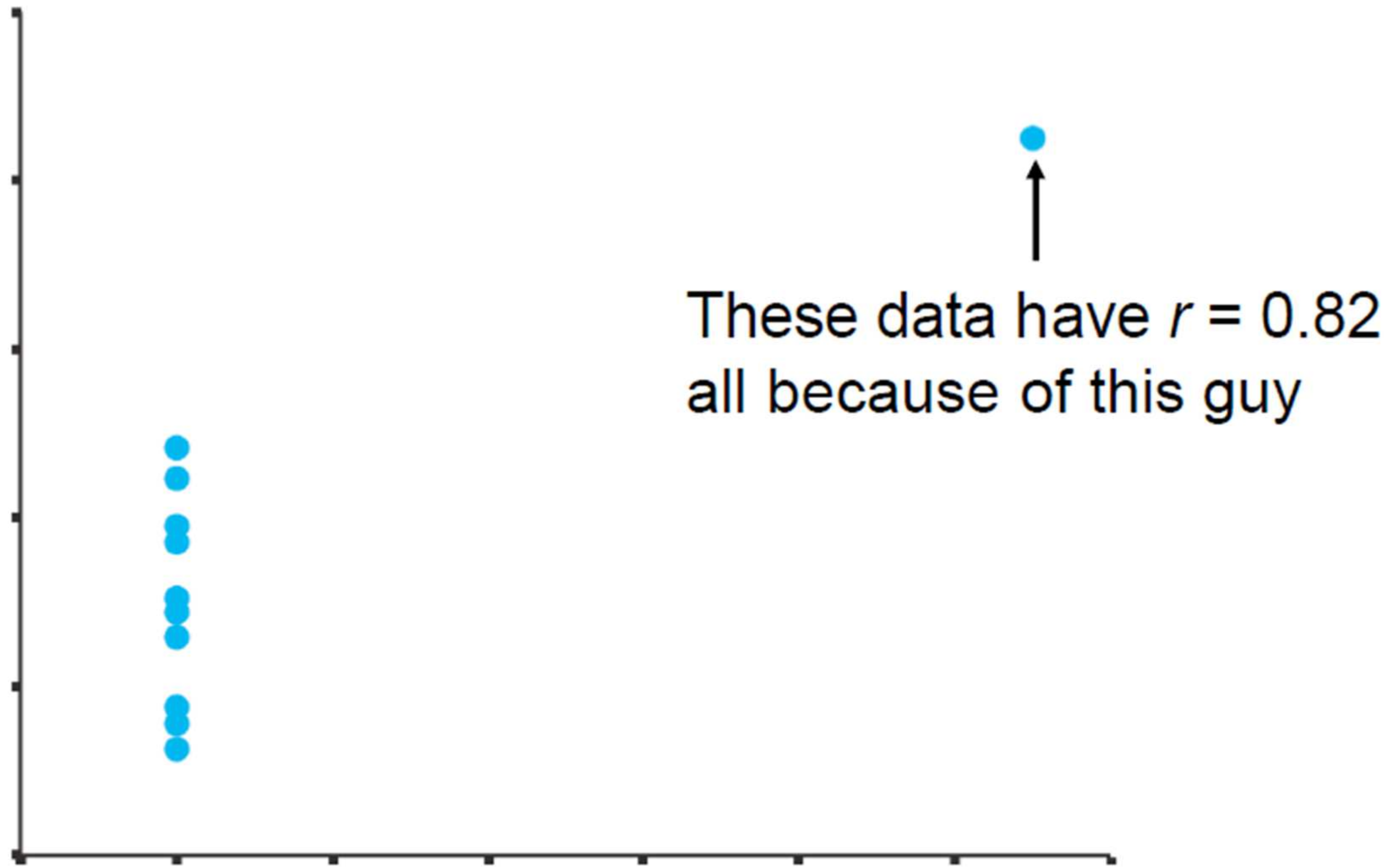
- Square the correlation coefficient $\Rightarrow$ r^2 = proportion of variance in Y mathematically explained by X
- Illustrative data: $r^2 = 0.737^2 = 0.54 \Rightarrow$ 54% of variance in lung cancer mortality is mathematically explained per capita smoking rates

Cautions

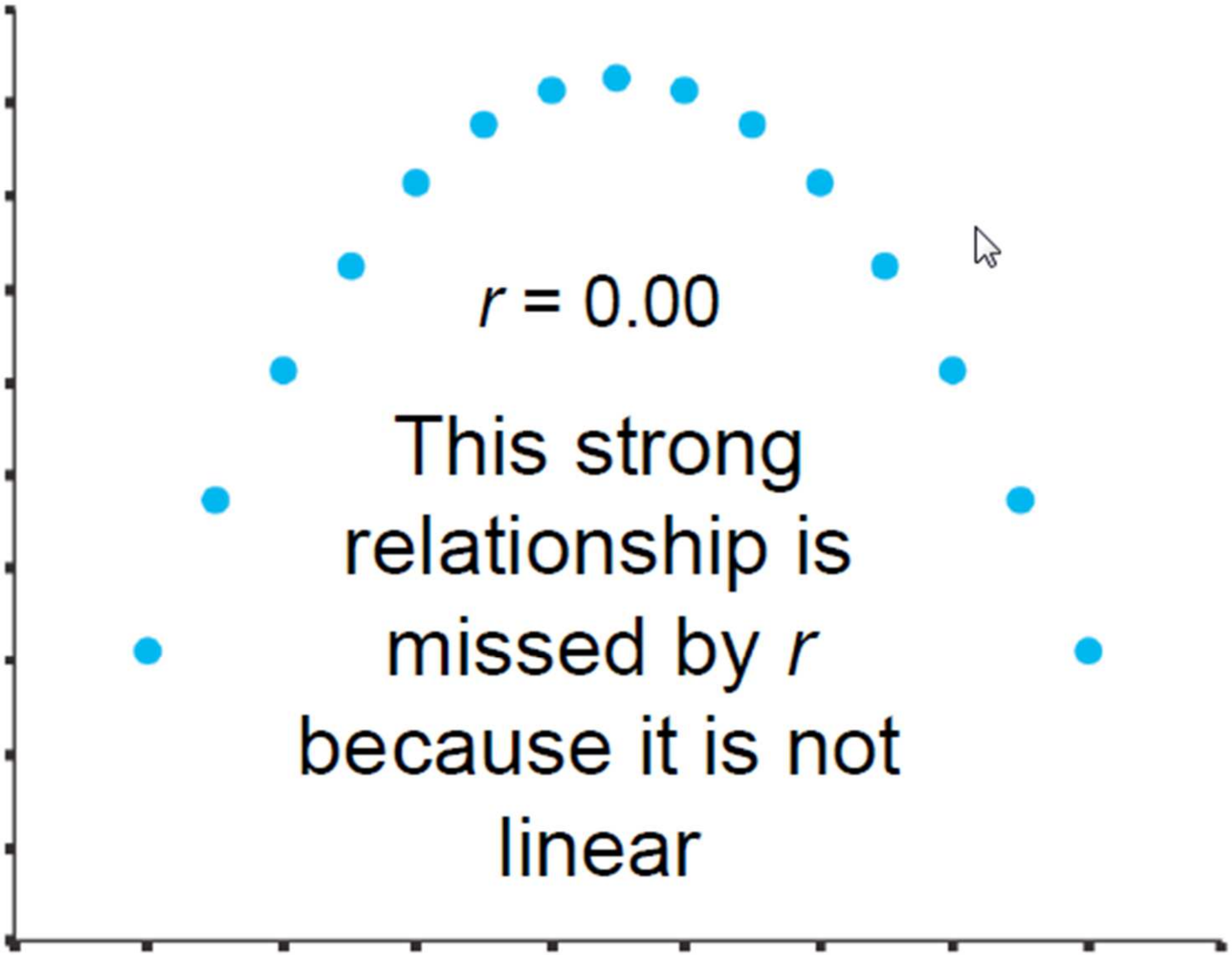
- Outliers
- Non-linear relationship
- Confounding (correlation is not causation)
- randomness

Outliers

Outliers can have profound influence on *r*

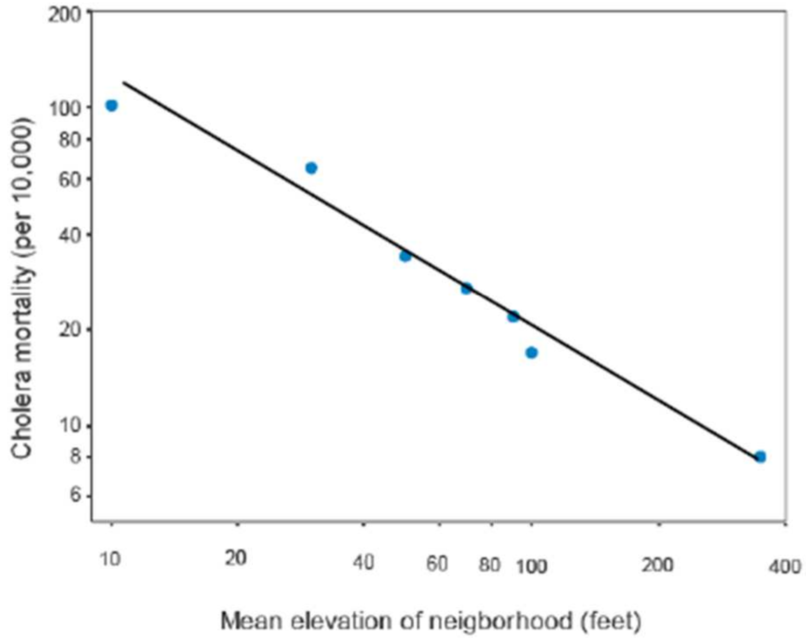


Non-linear relaitonship



Confounding

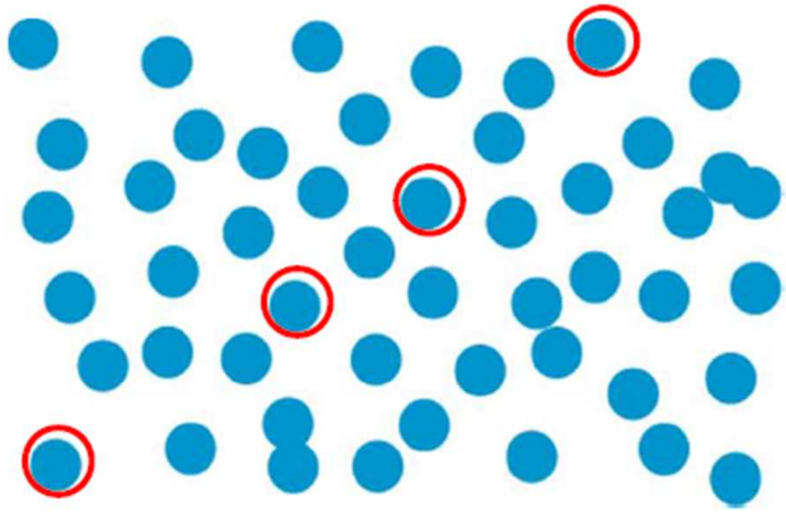
William Farr showed this strong negative correlation between cholera mortality and elevation above sea level in defense of miasma theory



However, he failed to account for the fact that people who lived at low elevations were more likely to drink from contaminated water sources (∴ confounding)

Randomness

Selection of specific data points would result in a false correlation



Need to do hypothesis testing!

Example of questions

- Estimate association between outcome and covariates
  - How cardiovascular disease (CVD) associate with smoking [and body mass index (BMI), age, etc.]
- Control of confounder(s)
  - How cardiovascular disease (CVD) associate with smoking after adjust for BMI, age, etc.
- Risk prediction
  - How to predict probability of getting CVD given information of BMI, age, etc.

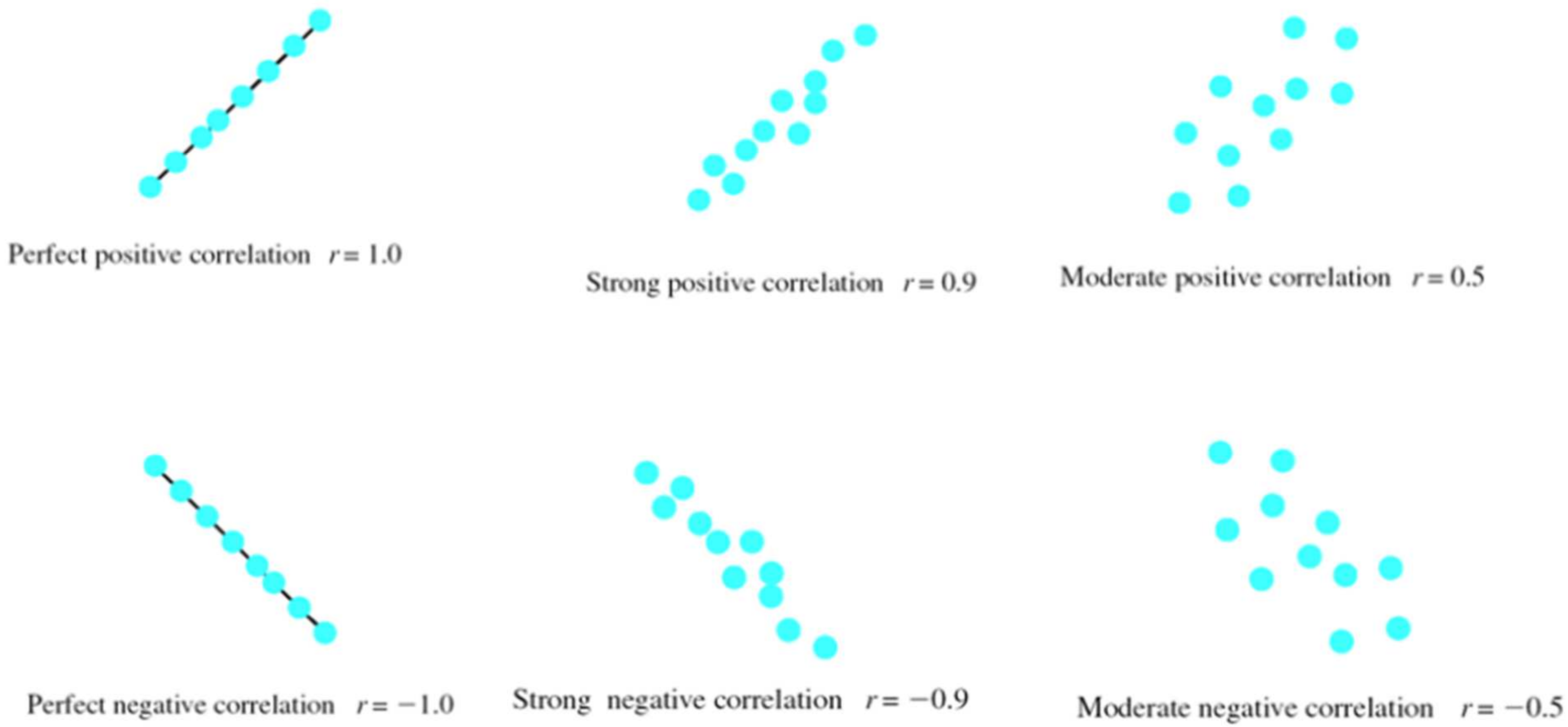
Result from bivariate analysis

Outcome is CVD, exposure is smoker

	CVD n=100	No CVD n=100	Odds ratio (95% CI)
Smoker	76	49	3.30 (1.80, 6.03)
Alcohol drinker	62	48	1.77 (1.01, 3.10)

Confounder ?

Correlation direction and strength



Sequence of analysis

- **Descriptive analysis**
  - Know your dataset
- **Bivariate analysis**
  - Identify associations
- **Stratified analysis**
  - Identify confounders and effect modifiers
- **Multivariable analysis**
  - Control for confounders and manage effect modifiers

Regression

- Regression is the study of dependence between an outcome variable (y) - the dependent variable, and one or several covariates (x) - independent variables

- Objective of using regression
  - Estimate association between outcome and covariates
  - Control of confounder(s)
  - Risk prediction

Result from bivariate analysis

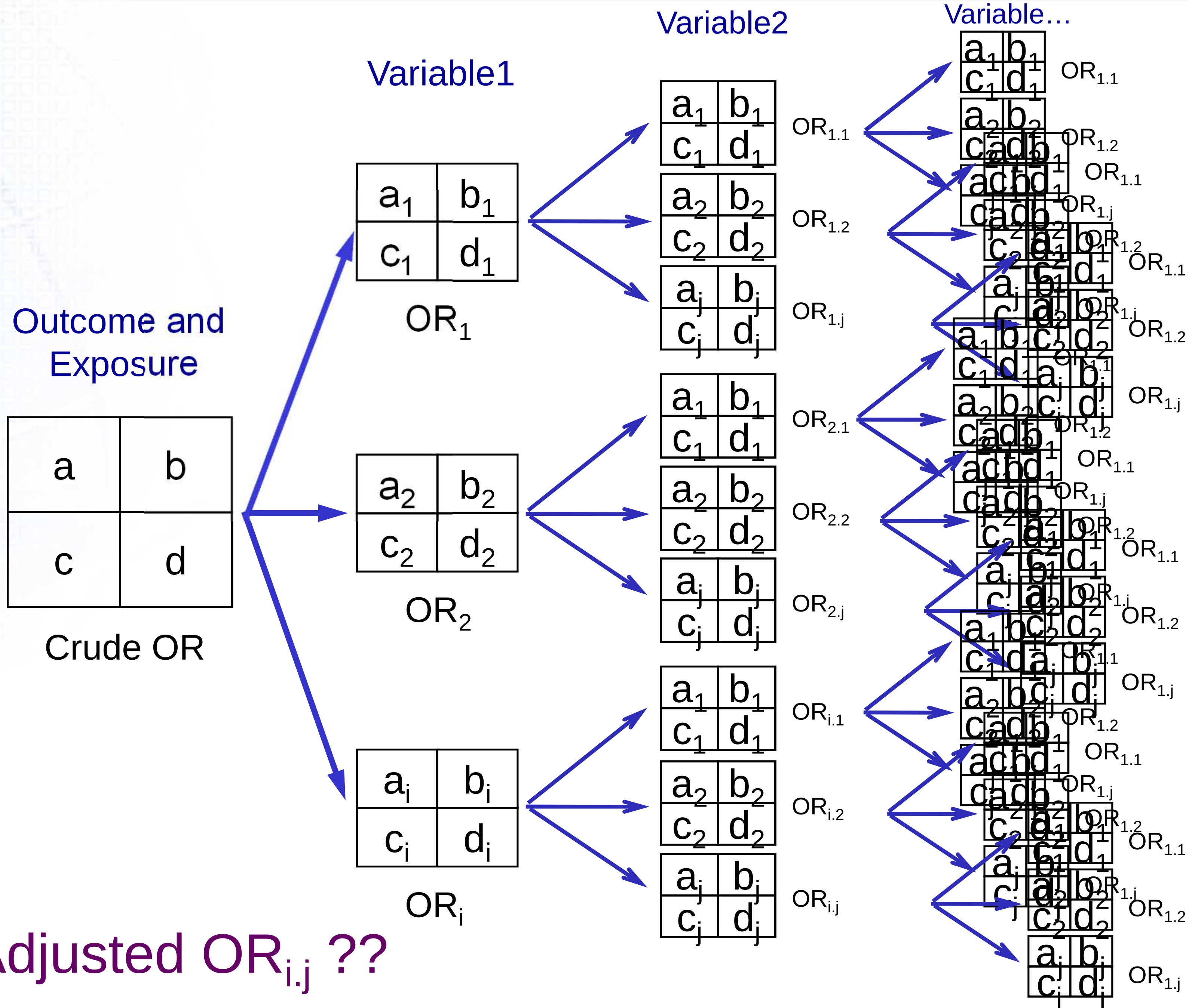
Outcome is CVD, exposure is smoker

	CVD n=100	No CVD n=100	p - value
Mean MBI (sd)	25.6(1.5)	21.6(1.6)	< 0.001
Mean age (sd)	63.0(7.2)	56.1(7.1)	< 0.001

Confounder ?

Stratify analysis

- General concept
- Linear equation
- Linear regression
  - Simple regress
  - Multiple regress
  - Model testing
  - EpiInfo & STATA
- Logistic regression
  - Introduction
  - Coefficients
  - Simple regress
  - Multiple regress
  - Model testing
  - Model building
  - EpiInfo & STATA



Adjusted $OR_{i.j}$??

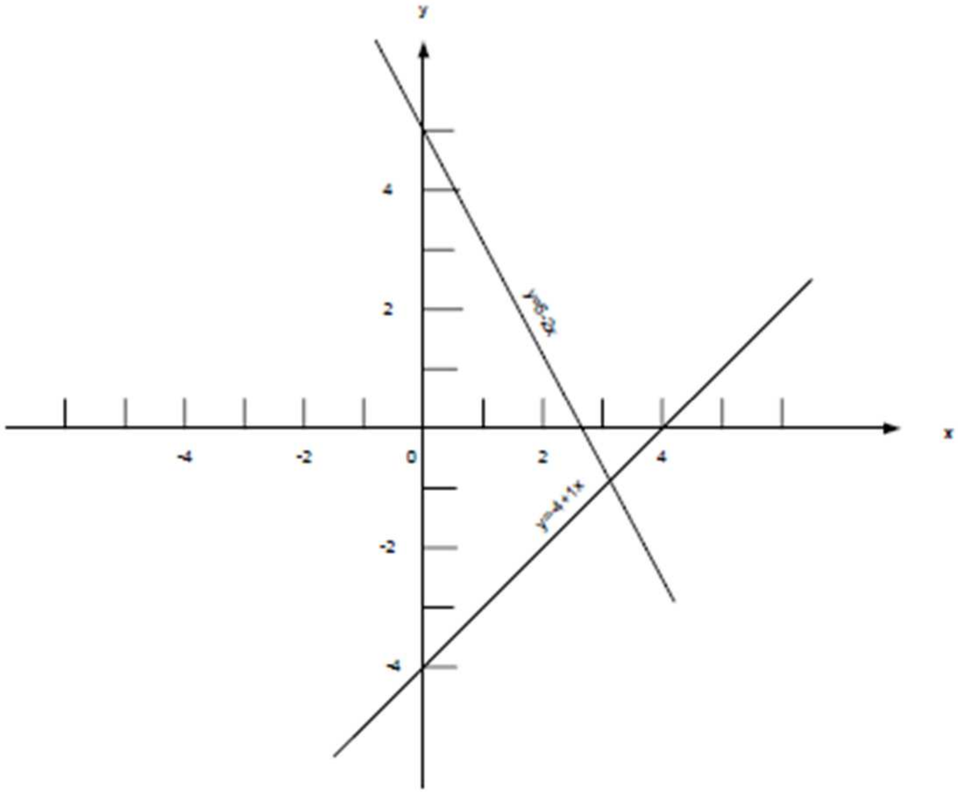
Regression

- It is a representation/modeling of the dependence between one or several variables

- Type of models
  - Linear regression
  - Logistic regression
  - Cox regression
  - Poisson regression
  - Loglinear regression

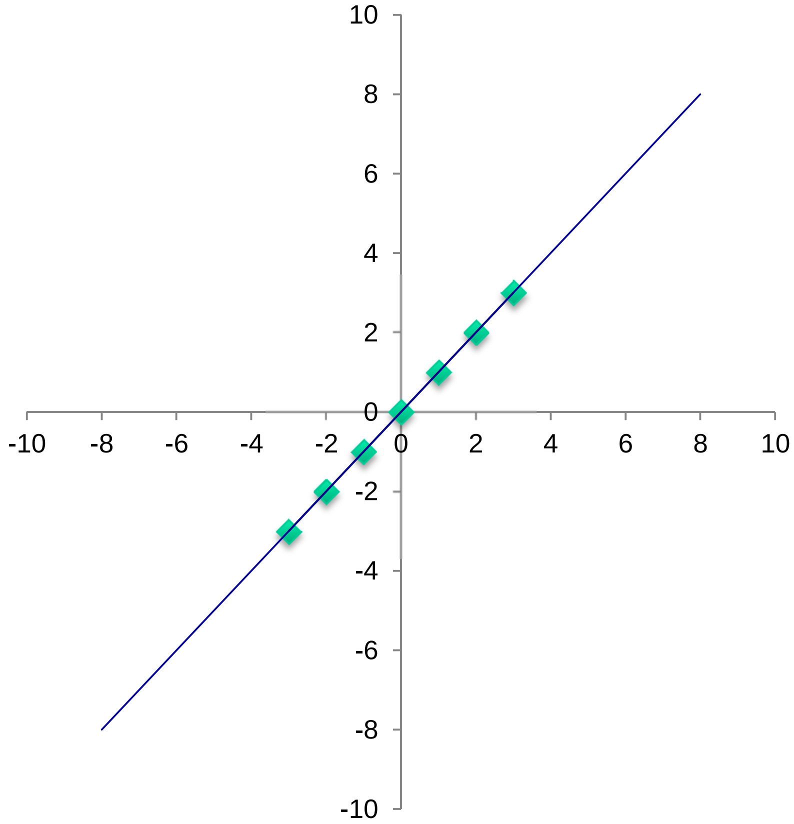
- Choices of the model depend on objectives, study design and outcome variables

Review of linear equation



x and y relationship example1

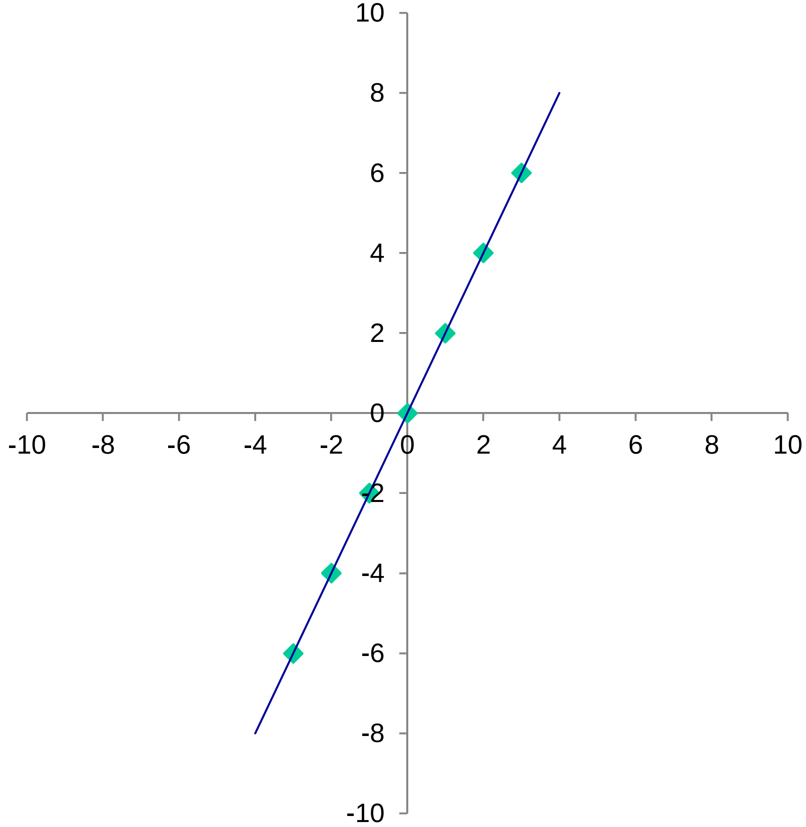
x	y
-3	-3
-2	-2
-1	-1
0	0
1	1
2	2
3	3



$$y = x$$

x and y relationship example2

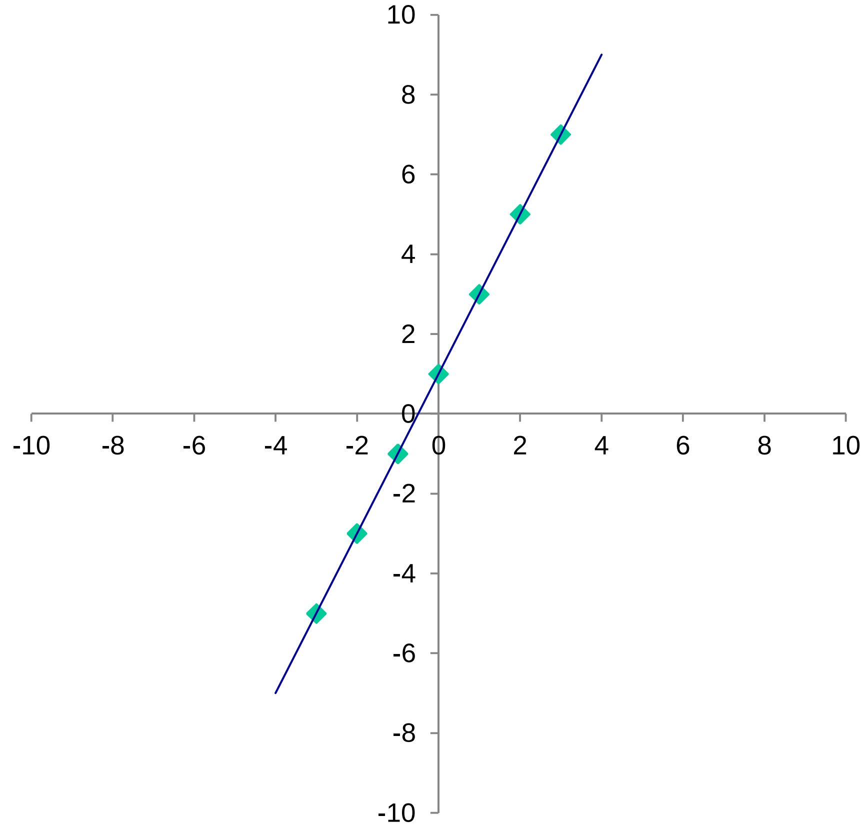
x	y
-3	-6
-2	-4
-1	-2
0	0
1	2
2	4
3	6



$$y = 2x$$

x and y relationship example3

x	y
-3	-5
-2	-3
-1	-1
0	1
1	3
2	5
3	7



$$y = 2x + 1$$

General form of linear equation

$$y = a + bx$$

- x is independent variable
- y is dependent variable
- a is a constant or y-intercept
  - The value of y when x = 0
- b is a slope
  - Amount by which y changes when x changes by one unit

Exercise: interpretation of slope

- $y = x$
  - x increase 1 → y …………
- $y = 2x$
  - x increase 1 → y …………
- $y = 2x + 1$
  - x increase 1 → y …………
- $y = -x + 1$
  - x increase 1 → y …………
- $y = -2x - 1$
  - x increase 1 → y …………

Linear regression

Linear regression

- Given an independent variable (x) and a dependent continuous variable (y), we computed *r* as a measure of linear relationship

- We want to be able to use our knowledge of this relationship to explain our data, or to predict new data.
  - This suggests linear regression (LR), a method for representing the dependent variable (outcome) as a linear function of the independent variables (covariates)
  - Outcome is continuous variable
  - Covariate can be either continuous or categorical variable

Linear regression

- Simple linear regression (SLR)
  - A single covariate
- Multiple linear regression (MLR)
  - Two or more covariate (with or without interactions)

Mathematical properties of a straight line

- A straight line can be described by an equation of the form

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

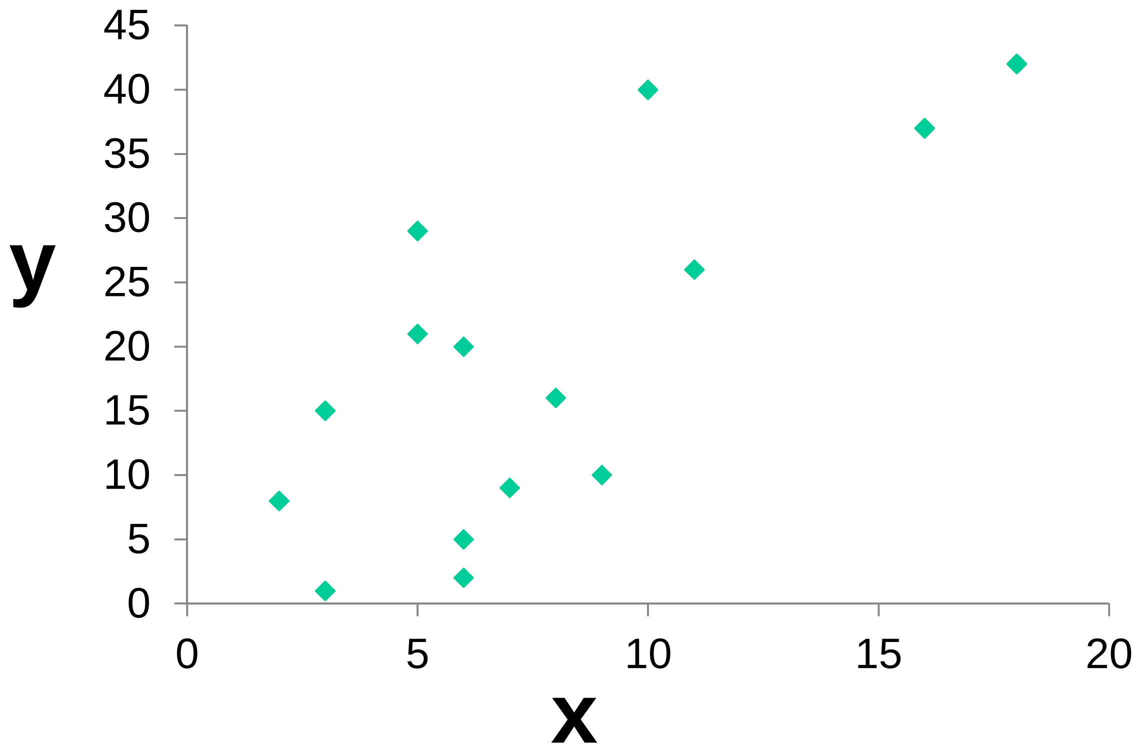
  - β_0 and β_1 are coefficients of the model
  - β_0 is called the y-intercept of the line: the value of y when x = 0
  - β_1 is called the slope: the amount of change in y for each 1-unit change in x

Example data with 15 epidemiologists

ID	x (Year of working)	y (Number of investigation)
1	3	15
2	6	2
3	3	1
4	8	16
5	9	10
6	6	5
7	16	37
8	10	40
9	2	8
10	5	21
11	5	29
12	6	20
13	7	9
14	11	26
15	18	42

Approach of SLR

- The following graph is a scatter plot of the example data with 15 epidemiologists
- We want to find a line that adequately represents the relationship between X (year of working) and Y (Number of outbreak investigation)



Algebraic approach

- The best fitting line is the one with the minimum sum of squared errors

- The way to find the best-fitting line to minimize the sum of squared errors is called *least-squares method*
  - Let $\hat{Y}_i$ denote the estimated outcome at X_i based on the fitted regression line, $\hat{Y}_i = \beta_0 + \beta_1 X_i$
  - β_0 and β_1 are the intercept and the slope of the fitted line
  - The error, or residual, is $e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$
  - The sum of the squares of all such errors is

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

Algebraic approach

- General concept
- Linear equation
- Linear regression
  - Simple regress
  - Multiple regress
  - Model testing
  - EpiInfo & STATA
- Logistic regression
  - Introduction
  - Coefficients
  - Simple regress
  - Multiple regress
  - Model testing
  - Model building
  - EpiInfo & STATA

➢ The least-squares estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{SP_{XY}}{SS_X}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Approach of SLR

ID	x	y	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
1	3	15	-4.67	-3.73	17.42	21.78
2	6	2	-1.67	-16.73	27.89	2.78
3	3	1	-4.67	-17.73	82.76	21.78
4	8	16	0.33	-2.73	-0.91	0.11
5	9	10	1.33	-8.73	-11.64	1.78
6	6	5	-1.67	-13.73	22.89	2.78
7	16	37	8.33	18.27	152.22	69.44
8	10	40	2.33	21.27	49.62	5.44
9	2	8	-5.67	-10.73	60.82	32.11
10	5	21	-2.67	2.27	-6.04	7.11
11	5	29	-2.67	10.27	-27.38	7.11
12	6	20	-1.67	1.27	-2.11	2.78
13	7	9	-0.67	-9.73	6.49	0.44
14	11	26	3.33	7.27	24.22	11.11
15	18	42	10.33	23.27	240.42	106.78
Total	115	281			636.67	293.33

$\bar{X} = 115/15 = 7.67$ $\bar{Y} = 281/15 = 18.73$

Algebraic approach

$$SP_{XY} = 636.67$$

$$SS_X = 293.33$$

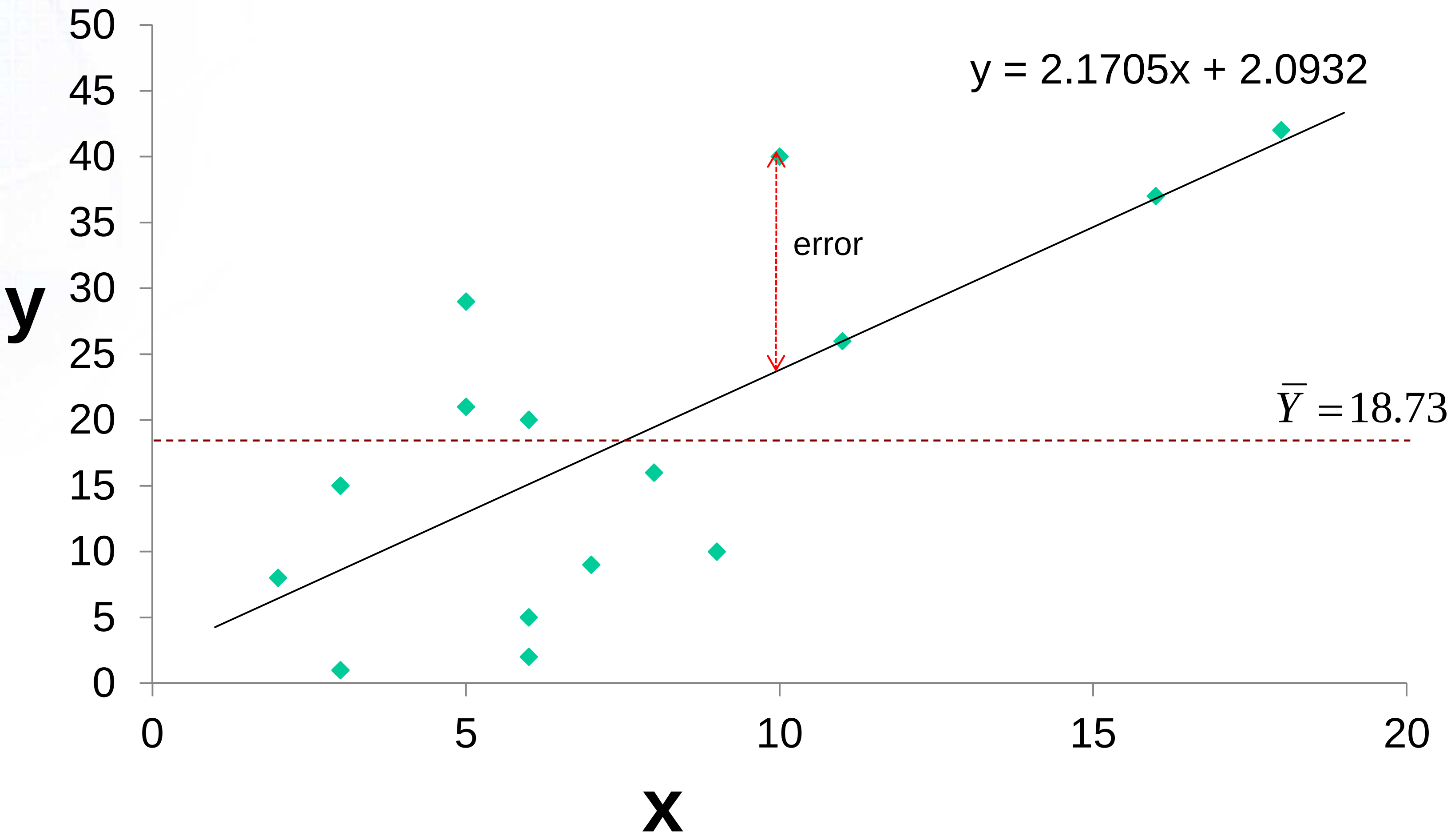
$$\hat{\beta}_1 = \frac{SP_{XY}}{SS_X} = \frac{636.67}{293.33} = 2.17$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 18.73 - (2.17)(7.67) = 2.09$$

$$Y = 2.17 + 2.09X$$

Approach of SLR

- Regression line and equation, graph showing error of data point between observed (10, 40) to expected (10, 23.80)



Interpretation of coefficients

$$Y = 2.17 + 2.09X$$

- β_0 : for epidemiologist with no working experience (X=0), average number of outbreak investigation is 2.17
  - Less meaningful for continuous variable of X, unless center it at some meaningful value

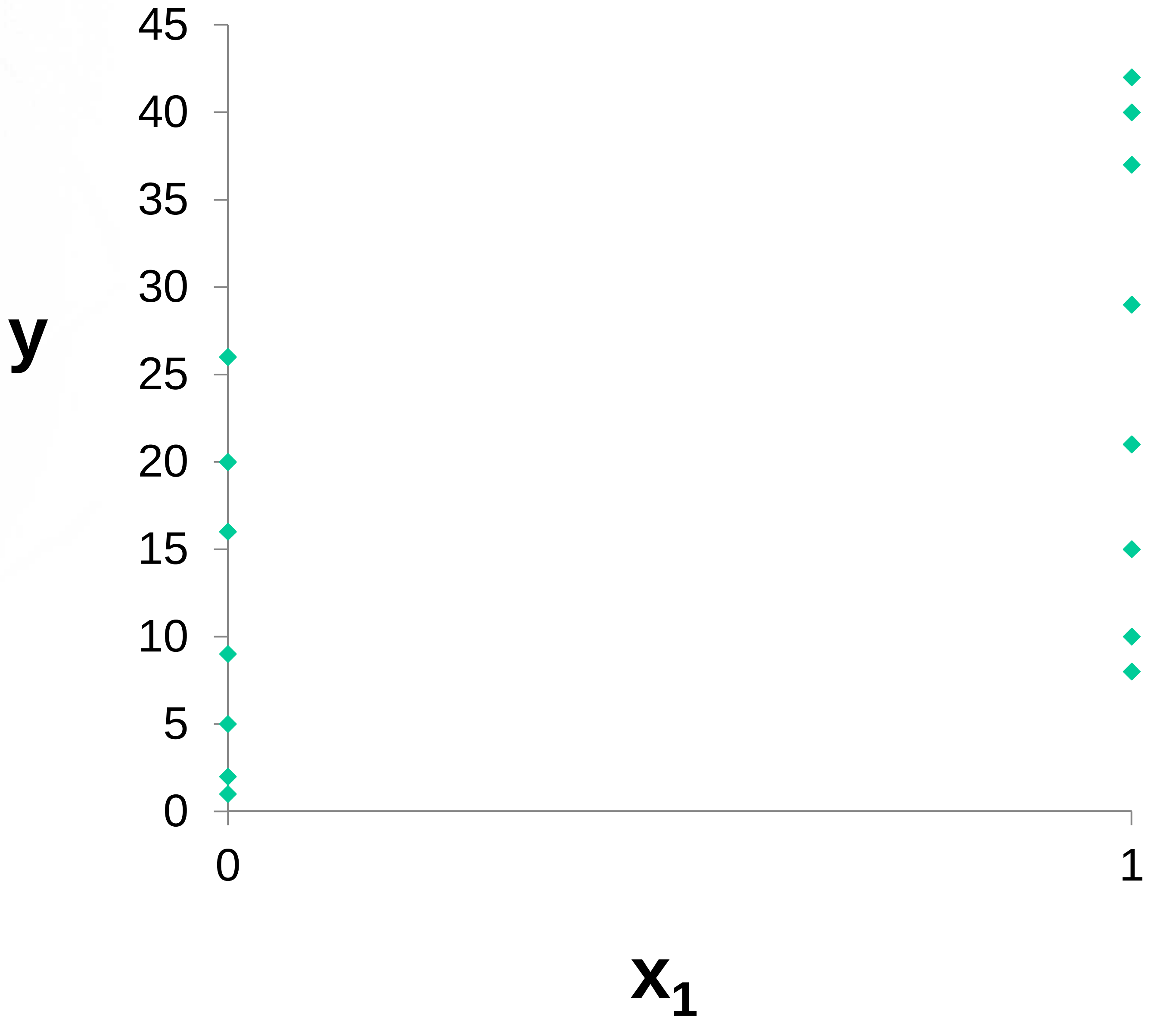
- β_1 : for additional one year of working experience, average number of outbreak investigation increase by 2.09

SLR for dichotomous variable

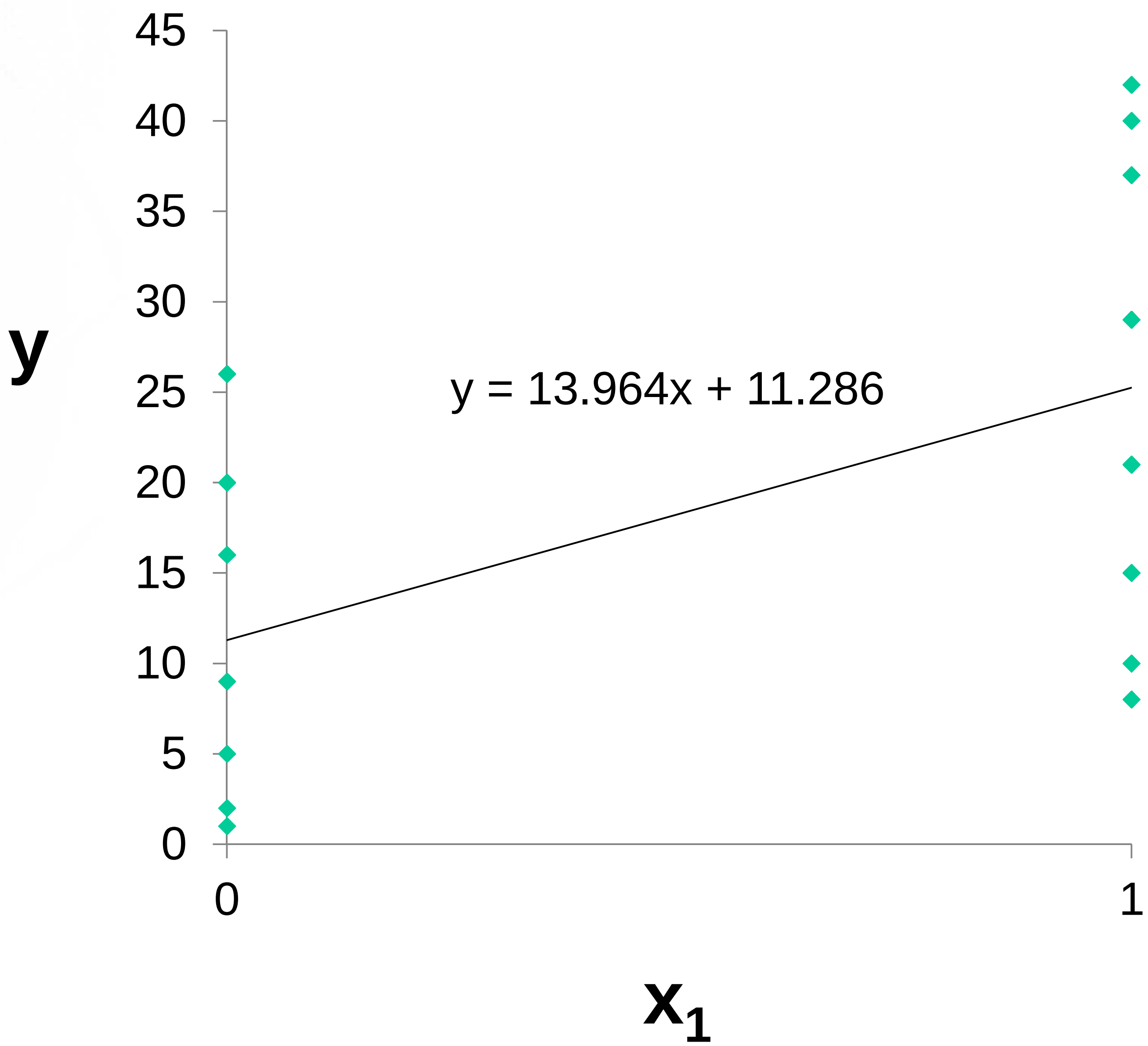
ID	x1 (training)	y (Number of investigation)
1	1	15
2	0	2
3	0	1
4	0	16
5	1	10
6	0	5
7	1	37
8	1	40
9	1	8
10	1	21
11	1	29
12	0	20
13	0	9
14	0	26
15	1	42

x1 : 0 = no training, 1 = ever attend training

SLR for dichotomous variable



SLR for dichotomous variable



Interpretation of coefficients

$$Y = 11.29 + 13.96X$$

- β_0 : for epidemiologist with no training(X1=0), average number of outbreak investigation is 11.29

- β_1 : for epidemiologist who ever attend training, average number of outbreak investigation increase by 13.96 (or average number of outbreak investigation is 25.25)

Multiple linear regression (MLR)

ID	X (Year of working)	X1 (training)	y (Number of investigation)
1	3	1	15
2	6	0	2
3	3	0	1
4	8	0	16
5	9	1	10
6	6	0	5
7	16	1	37
8	10	1	40
9	2	1	8
10	5	1	21
11	5	1	29
12	6	0	20
13	7	0	9
14	11	0	26
15	18	1	42

x1 : 0 = no training, 1 = ever attend training

Multiple linear regression (MLR)

- Also used *least-squares method* to estimate coefficients and draw best fitting line

- General form

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

  - β_0 is the value of y when all $x_i = 0$
  - β_i is amount of change in y for each 1-unit change in x_i adjusted for other covariates

Interpretation of coefficients

$$Y = -1.68 + 1.93X + 10.51X1$$

- β_0 : for epidemiologist with no working experience (X=0) and no training(X1=0), average number of outbreak investigation is -1.68
  - Less meaningful if there are continuous variables of X in the model, unless center it at some meaningful value
- β_1 : for additional one year of working experience, average number of outbreak investigation increase by 1.93 after adjusted for ever attend training
- β_2 : for epidemiologist who ever attend training, average number of outbreak investigation increase by 10.51 after adjusted for working experience

Coefficient of Determination (r^2 or R^2)

- A measure of the impact that X covariate(s) has in explaining the variation of Y
- $0 \leq r^2 \leq 1$
  - 1 : a perfect prediction
  - 0 : absolutely no association
- The higher the r^2 value, the more the variation in Y is able to be predicted by X. The higher the r^2, the smaller the errors of prediction
- Adjusted r^2 : a modification of r^2 that adjusts for the number of covariates in a model

Interpretation of r^2

$$Y = -1.68 + 1.93X + 10.51X1 \qquad r^2 = 0.69$$

- 69% of the variability of number of outbreak investigation is explained by number of year in working experience and ever attend training

Hypothesis testing of the model

- p-value or critical value approach
- Two parts of testing
  - Significant of overall model (do all covariates together can predict the outcome?)
    - Using ANOVA approach (overall F-test)
  - Significant of each β_i (adjusted for other covariates)
    - Using either partial F-test or partial t-test

ANOVA table

- General concept
- Linear equation
- Linear regression
  - Simple regress
  - Multiple regress
  - Model testing
  - EpiInfo & STATA
- Logistic regression
  - Introduction
  - Coefficients
  - Simple regress
  - Multiple regress
  - Model testing
  - Model building
  - EpiInfo & STATA

Source	Sum of square (SS)	df	Mean square (MS)	F statistics
Regression	SSR	k	MSR	MSR/MSE
Residual (Error)	SSE	n-k-1	MSE	
Total	SST	n-1		

$$SSR = \sum (\hat{Y}_i - \overline{Y})^2$$

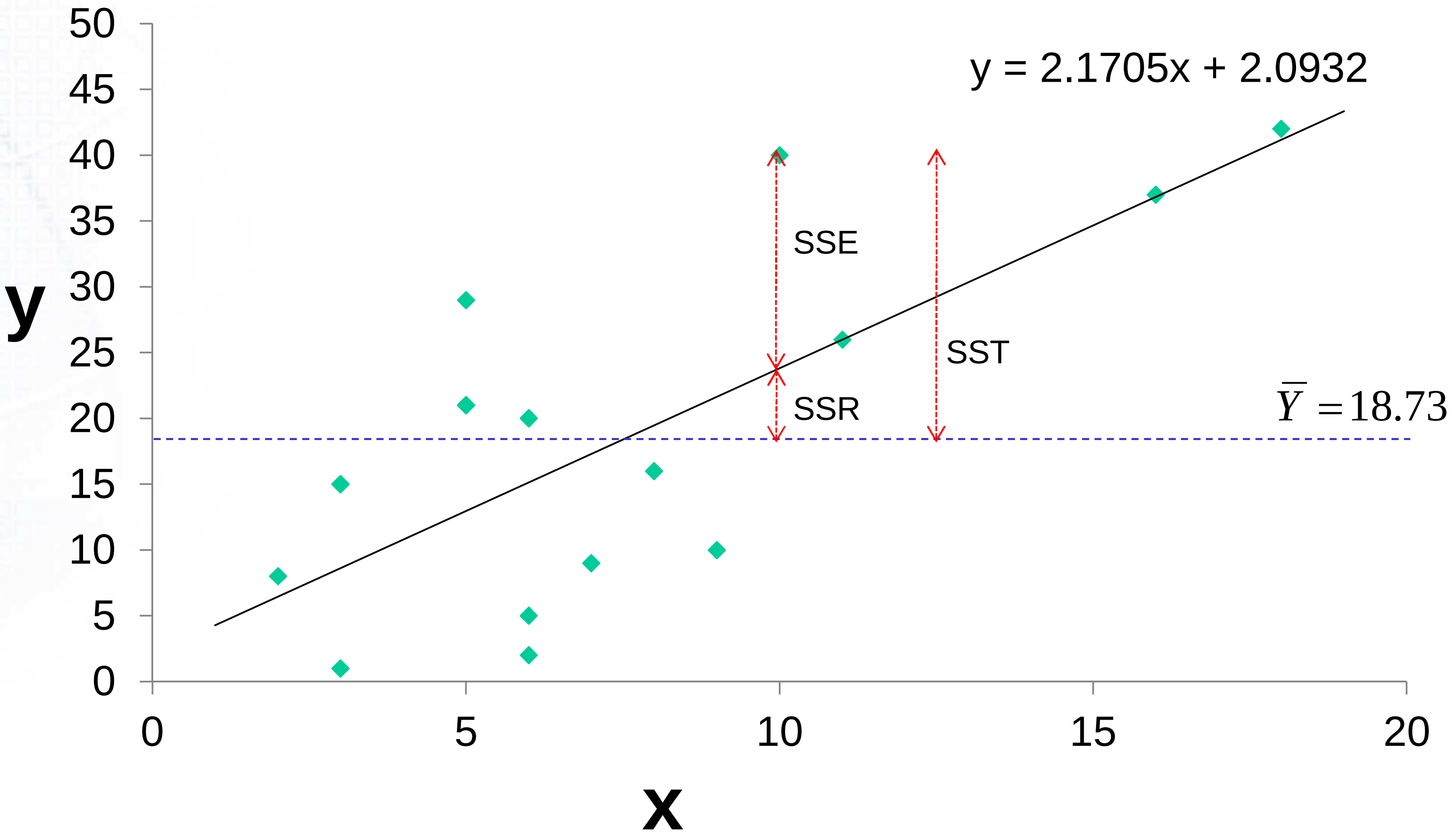
$$SSR = \sum (Y_i - \hat{Y}_i)^2$$

$$SST = \sum (Y_i - \overline{Y})^2 = SSR + SSE \longrightarrow SS_y$$

- k : number of covariate in the model
- n : number of observation in the model
- MSR = SSR/k
- MSE = SSE/(n-k-1)

$$r^2 = SSR / SST$$

Regression line and error



Note: sum of square (SS) is computed from all y_i, this figure just shows only 1 data point of y at (10, 40) as an example

Overall F test

- The ANOVA provides a test for

  $H_0 : \beta_1 = \beta_1 = \ldots = \beta_i = 0$

- Test statistics

$$F_{k,n-k-1} = \frac{MSR}{MSE}$$

- The decision rule is:

  Reject H_0 if $F > F_{(1-\alpha,\, k,\, n-k-1)}$ or p-value $< \alpha$

- Interpretation: if reject H_0

  There was a significant prediction of outcome by covariates taken together (at least one covariate is significant)

partial t test

- It provides a test for

  $H_0 : \beta_{i.\text{ given other covariates in the model}} = 0$

- Test statistics

$$t_{n-k-1} = \frac{\beta_i}{SE_{\beta_i}}$$

- The decision rule is:

  Reject H_0 if $t > t_{(1-\alpha/2,\, n-k-1)}$ or p-value $< \alpha$

- Interpretation: if reject H_0

  There was a significant prediction of outcome by x_i adjusted for other covariates in the model

partial F test

- Extra sum of squares
  - We refer to $SSR_{x2|x1}$ as the EXTRA SUM OF SQUARES due to including X2 in the model after X1 is already there
  - This calculation is done by subtraction, by comparing the SSR from two models, one referred to as the "FULL MODEL" and the other as the "REDUCED MODEL"

$$\underbrace{SSR_{X^*|X_1 \ldots X_p}}_{\text{Extra SS}} = \underbrace{SSR_{X_1 \ldots X_p, X}}_{\text{Full model SS}} - \underbrace{SSR_{X_1 \ldots X_p}}_{\text{Reduced model SS}}$$

- We use this extra SSR to calculate a partial F statistic for contribution of that variable in the order we designated

partial F test

- It provides a test for

  $H_0 : \beta_{i.\ \text{given other covariates in the model}} = 0$

- Test statistics

$$F_{k_1-k_2,n-k_1-1} = \frac{\dfrac{SSR_{Full} - SSR_{Reduced}}{k_1 - k_2}}{\dfrac{SSR_{Full}}{n - k_1 - 1}}$$

- The decision rule is:

  Reject H_0 if $F > F_{(1-\alpha,k_1-k_2,n-k_1-1)}$ or p-value < α

- Interpretation: if reject H_0

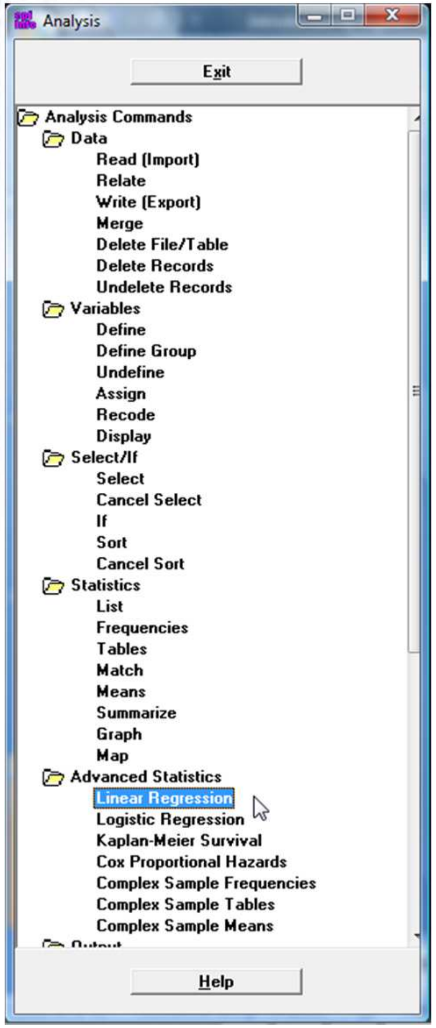
  There was a significant prediction of outcome by x_i adjusted for other covariates in the model

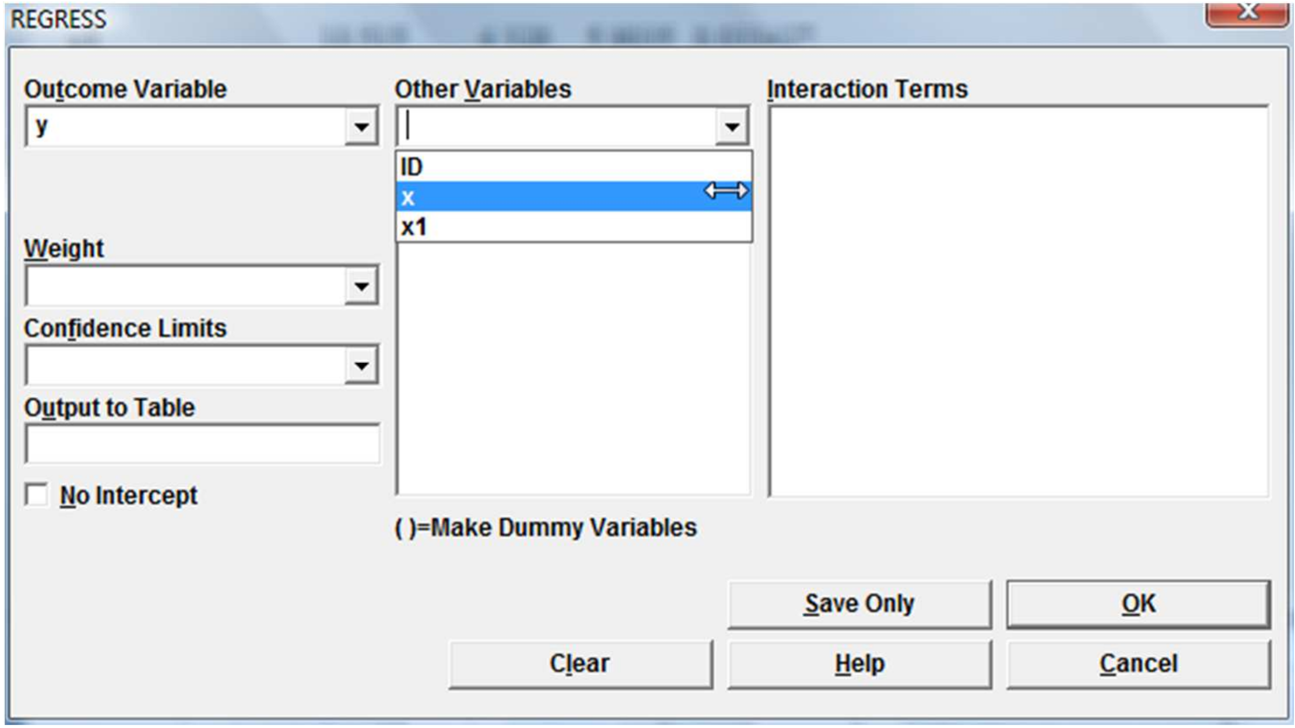
Example command in EpiInfo

- General concept
- Linear equation
- Linear regression
  - Simple regress
  - Multiple regress
  - Model testing
  - EpiInfo & STATA
- Logistic regression
  - Introduction
  - Coefficients
  - Simple regress
  - Multiple regress
  - Model testing
  - Model building
  - EpiInfo & STATA

SLR

REGRESS y = x





Example output in EpiInfo

➢ SLR

D:\???????\Course\Introduction to logistic (Monday)\OUT22.htm

Previous | Next | Last | History | Open | Bookmark | Print | Maximize

REGRESS y = x

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
x	2.170	0.561	14.9567	0.002238
CONSTANT	2.093	4.967	0.1776	0.680909

Partial F test

Correlation Coefficient: r^2= 0.53

Source	df	Sum of Squares	Mean Square	F-statistic
Regression	1	1381.856	1381.856	14.957
Residuals	13	1201.077	92.391	
Total	14	2582.933		

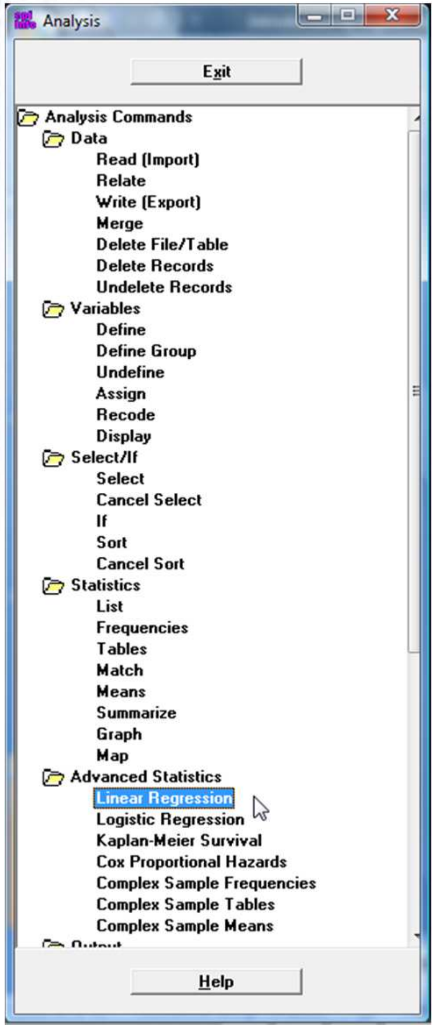
Overall F test

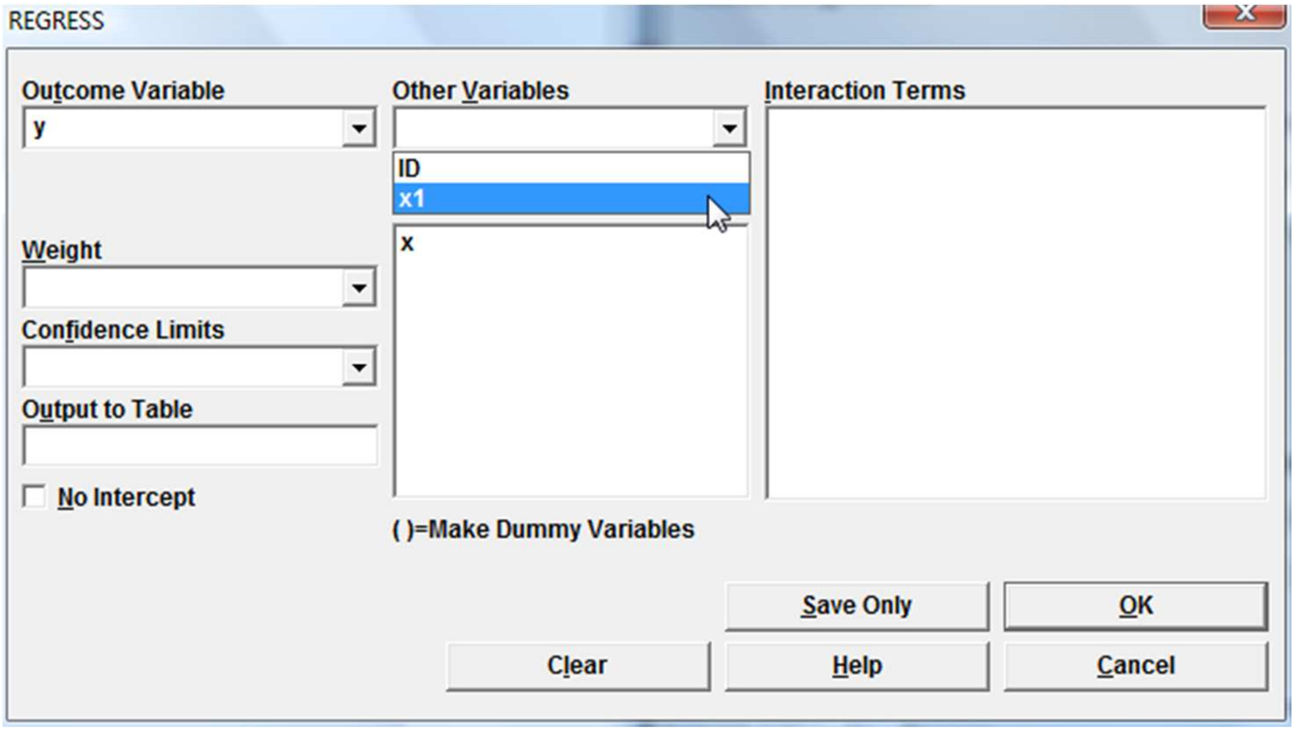
Example command in EpiInfo

- General concept
- Linear equation
- Linear regression
  - Simple regress
  - Multiple regress
  - Model testing
  - **EpiInfo & STATA**
- Logistic regression
  - Introduction
  - Coefficients
  - Simple regress
  - Multiple regress
  - Model testing
  - Model building
  - EpiInfo & STATA

➢ MLR

REGRESS y = x x1





Example output in EpiInfo

MLR

D:\???????\Course\Introduction to logistic (Monday)\OUT22.htm

Previous | Next | Last | History | Open | Bookmark | Print | Maximize

REGRESS y = x x1

Next Procedure

Linear Regression

Variable	Coefficient	Std Error	F-test	P-Value
x	1.931	0.488	15.6499	0.002249
x1	10.515	4.328	5.9035	0.033427
CONSTANT	-1.683	4.509	0.1393	0.716089

Partial F test

Correlation Coefficient: r^2= 0.69

Source	df	Sum of Squares	Mean Square	F-statistic
Regression	2	1777.898	888.949	13.251
Residuals	12	805.036	67.086	
Total	14	2582.933		

Overall F test

Example command & output in STATA

➢ SLR

reg y x

Stata Results

```
. reg  y x

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  1,    13) =   14.96
       Model |  1381.85606     1  1381.85606           Prob > F      =  0.0019
    Residual |  1201.07727    13  92.3905594           R-squared     =  0.5350
-------------+------------------------------           Adj R-squared =  0.4992
       Total |  2582.93333    14  184.495238           Root MSE      =   9.612

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   2.170455   .5612199     3.87   0.002     .9580126    3.382897
       _cons |   2.093182    4.96714     0.42   0.680    -8.637672    12.82404
------------------------------------------------------------------------------
```

Partial
t test

Overall
F test

Example command & output in STATA

➢ MLR

reg y x x1

Stata Results

```
. reg y  x x1

      Source |       SS       df       MS              Number of obs =      15
-------------+------------------------------           F(  2,    12) =   13.25
       Model |  1777.8978     2   888.9489           Prob > F      =  0.0009
    Residual | 805.035533    12  67.0862944           R-squared     =  0.6883
-------------+------------------------------           Adj R-squared =  0.6364
       Total | 2582.93333    14  184.495238           Root MSE      =  8.1906

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   1.931472   .4882394     3.96   0.002     .8676898    2.995254
          x1 |   10.51523    4.32778     2.43   0.032     1.085807    19.94465
       _cons |  -1.682741   4.508904    -0.37   0.716     -11.5068    8.141316
------------------------------------------------------------------------------
```

Partial
t test

Overall
F test

Logistic regression

Introduction

- Dichotomous outcome is very common situation in biology and epidemiology
  - Sick, not sick
  - Dead, alive
  - Cure, no cure

- What are problems if we use linear regression with dichotomous outcome?
  - For example
    - Outcome is cardiovascular disease (CVD): 0 = Absent, 1= Present
    - Exposure is body mass index (BMI)

Introduction

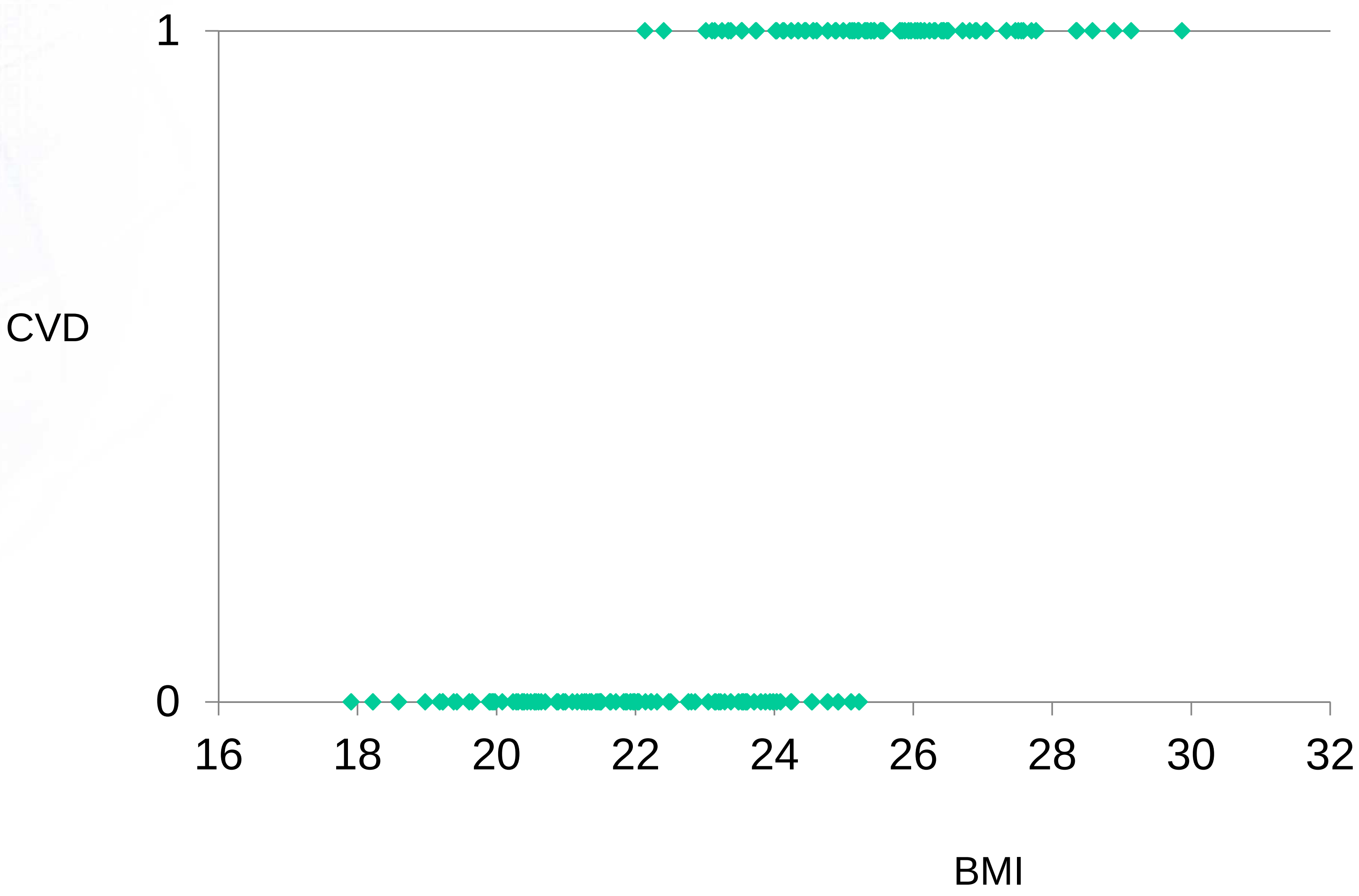
- A broad class of regression models, collectively known as the generalized linear model, has been developed to address multiple regression with a variety of dependent variable like dichotomies and counts
- Logistic regression is one generalized linear model for categorical outcome variables
- Logistic regression allows one to predict a dichotomous (or categorical) outcome from a set of variables that may be continuous, discrete (or categorical), dichotomous, or a mix.

BMI and CVD status of 200 subject

- General concept
- Linear equation
- Linear regression
  - Simple regress
  - Multiple regress
  - Model testing
  - EpiInfo & STATA
- Logistic regression
  - Introduction
  - Coefficients
  - Simple regress
  - Multiple regress
  - Model testing
  - Model building
  - EpiInfo & STATA

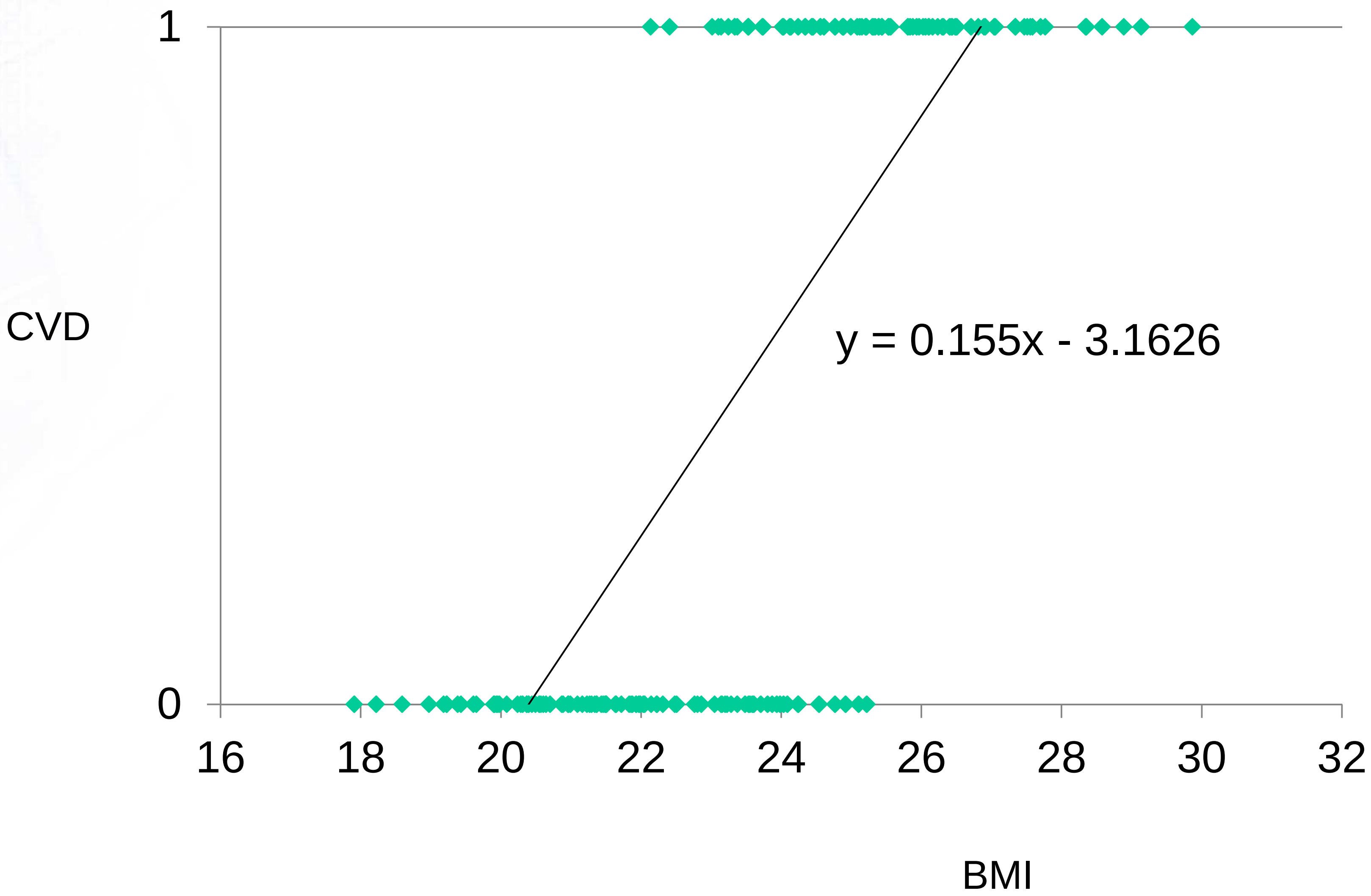
id	bmi	cvd	id	bmi	cvd	id	bmi	cvd	id	bmi	cvd	id	bmi	cvd	id	bmi	cvd	id	bmi	cvd	id	bmi	cvd
1	20.6	0	26	23.1	0	51	23.3	0	76	21.5	0	101	22.1	1	126	19.6	0	151	23.8	0	176	25.4	1
2	24.4	1	27	27.6	1	52	21.0	0	77	25.2	1	102	24.4	1	127	22.8	0	152	23.6	0	177	26.1	1
3	20.4	0	28	21.3	0	53	22.0	0	78	23.0	1	103	22.8	0	128	19.2	0	153	24.0	1	178	25.4	1
4	20.6	0	29	25.2	1	54	26.9	1	79	24.5	1	104	20.3	0	129	29.1	1	154	23.3	1	179	21.4	0
5	20.7	0	30	28.4	1	55	20.1	0	80	20.3	0	105	25.3	1	130	23.5	1	155	26.5	1	180	27.8	1
6	21.5	0	31	19.2	0	56	21.2	0	81	25.5	1	106	26.7	1	131	22.9	0	156	24.2	1	181	26.4	1
7	23.5	0	32	22.5	0	57	25.4	1	82	24.0	0	107	27.6	1	132	19.4	0	157	23.4	1	182	24.3	1
8	19.9	0	33	25.1	1	58	26.1	1	83	25.8	1	108	26.2	1	133	21.4	0	158	21.6	0	183	24.0	0
9	27.3	1	34	24.2	0	59	21.3	0	84	20.4	0	109	20.5	0	134	21.4	0	159	22.2	0	184	25.5	1
10	22.3	0	35	27.0	1	60	23.5	1	85	19.0	0	110	23.5	0	135	24.6	1	160	21.0	0	185	25.4	1
11	28.6	1	36	25.3	1	61	20.4	0	86	26.1	1	111	24.6	1	136	26.8	1	161	21.3	0	186	26.3	1
12	22.0	0	37	22.0	0	62	23.2	0	87	22.0	0	112	23.1	1	137	18.6	0	162	24.9	1	187	20.6	0
13	23.0	0	38	27.3	1	63	21.1	0	88	26.3	1	113	19.7	0	138	24.5	0	163	24.4	1	188	23.9	0
14	21.2	0	39	26.4	1	64	21.8	0	89	23.2	1	114	20.0	0	139	24.1	1	164	24.8	0	189	25.3	1
15	23.7	0	40	23.6	0	65	21.9	0	90	23.7	1	115	20.5	0	140	29.9	1	165	21.0	0	190	26.0	1
16	20.9	0	41	20.4	0	66	19.4	0	91	21.9	0	116	24.8	1	141	26.9	1	166	24.3	1	191	23.9	0
17	21.7	0	42	23.7	1	67	20.9	0	92	27.3	1	117	25.9	1	142	24.0	1	167	23.2	0	192	24.9	0
18	21.6	0	43	24.6	1	68	21.9	0	93	24.9	1	118	24.1	0	143	20.5	0	168	21.5	0	193	26.2	1
19	21.5	0	44	23.2	0	69	25.2	1	94	25.2	1	119	23.4	0	144	23.6	0	169	26.4	1	194	27.5	1
20	19.9	0	45	26.1	1	70	20.4	0	95	27.0	1	120	22.1	0	145	24.3	1	170	25.1	1	195	21.9	0
21	25.2	0	46	27.1	1	71	25.5	1	96	26.0	1	121	24.8	1	146	21.5	0	171	23.3	1	196	25.4	1
22	22.2	0	47	22.1	0	72	26.5	1	97	25.1	1	122	23.1	1	147	25.6	1	172	23.5	1	197	24.2	0
23	28.9	1	48	24.9	1	73	22.0	0	98	25.8	1	123	25.8	1	148	25.9	1	173	18.2	0	198	27.5	1
24	20.0	0	49	26.5	1	74	24.0	1	99	17.9	0	124	26.4	1	149	28.3	1	174	22.4	1	199	25.0	1
25	22.5	0	50	19.9	0	75	25.1	0	100	20.2	0	125	19.2	0	150	25.4	1	175	24.1	1	200	27.7	1

Example data and scatter plot



Example data and scatter plot

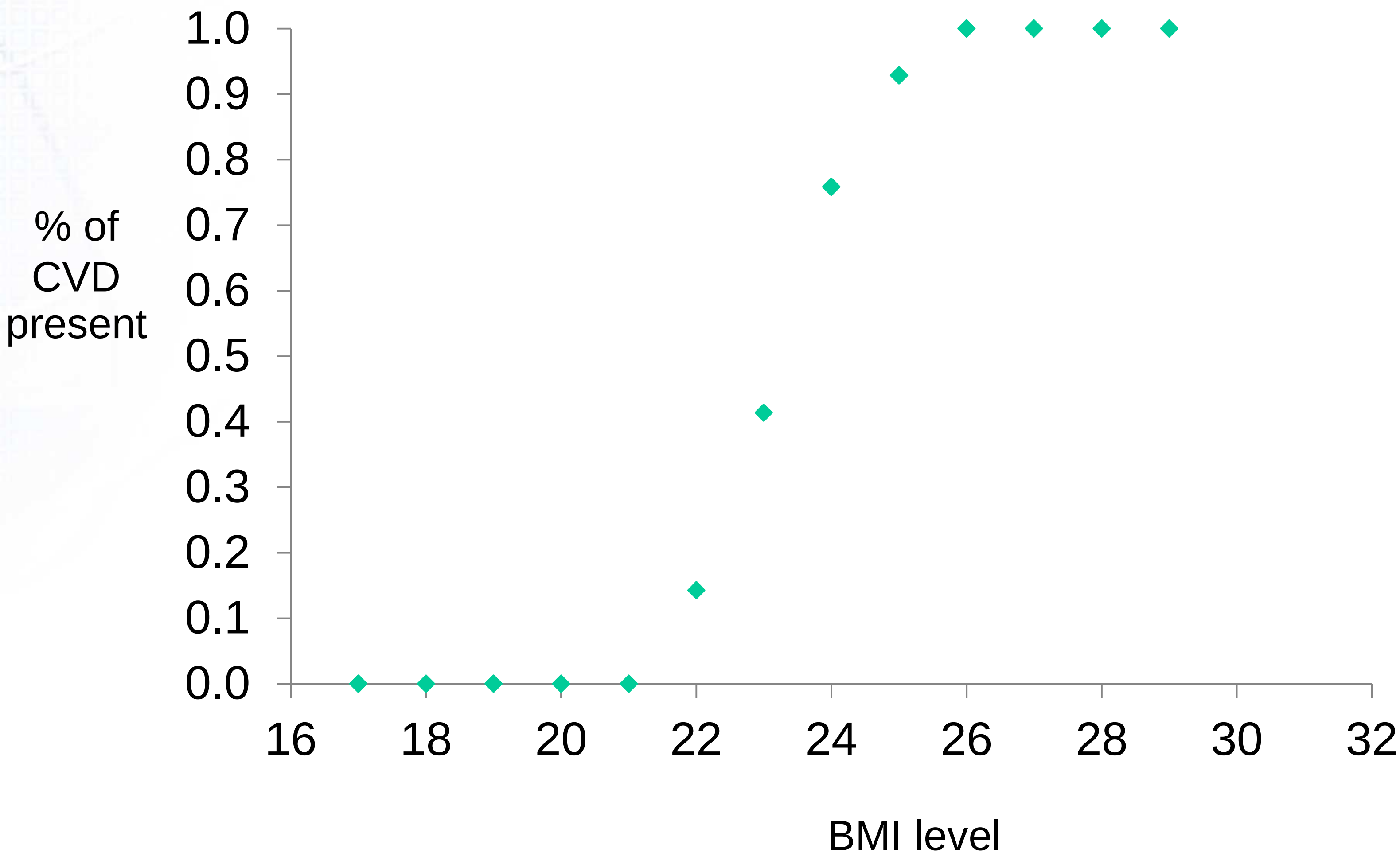
Can we model linear regression using least-square method ?



Frequency table of BMI level by CVD

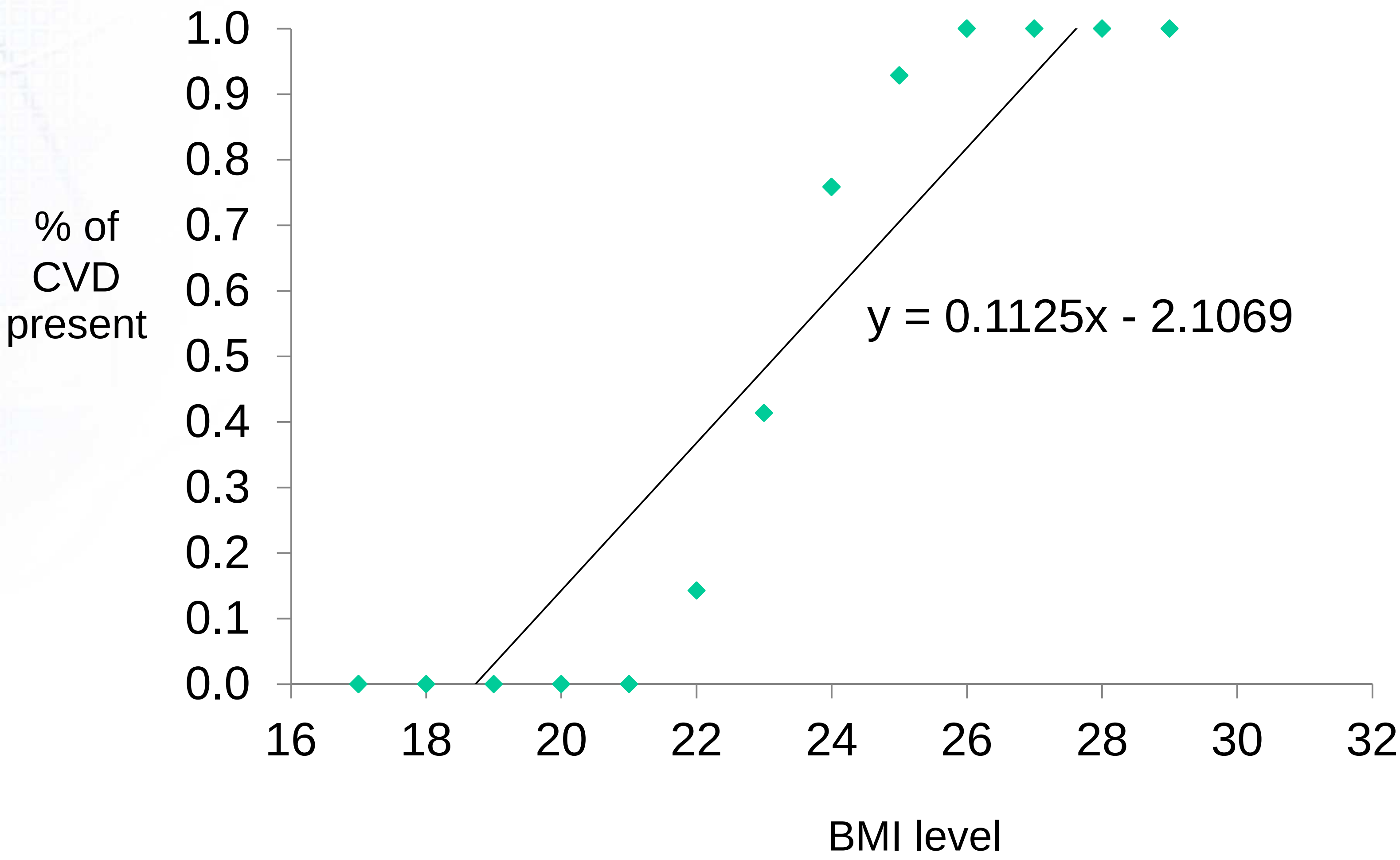
		CVD		
BMI	n	Absent	Present	Proportion
17 to 17+	1	1	0	0.0
18 to 18+	3	3	0	0.0
19 to 19+	12	12	0	0.0
20 to 20+	21	21	0	0.0
21 to 21+	25	25	0	0.0
22 to 22+	14	12	2	0.1
23 to 23+	29	17	12	0.4
24 to 24+	29	7	22	0.8
25 to 25+	28	2	26	0.9
26 to 26+	20	0	20	1.0
27 to 27+	12	0	12	1.0
28 to 28+	4	0	4	1.0
29 to 29+	2	0	2	1.0
TOTAL	200	100	100	0.5

Plot of percentage of subjects with CVD in each BMI level



Plot of percentage of subjects with CVD in each BMI level

Can we model linear regression using least-square method ?



Problem with linear regression

- Using dichotomous value of y (0 or 1)
  - Meaningless for predicted y
- Using probability of outcome as y (% of yes)
  - Predicted y can be less than 0 or greater than 1, but probability must lie between 0 and 1
  - Distribution of data seems to be s-shape (not straight line), resemble a plot of logistic distribution

Model a function of y

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

- Let
  - A denote the right side of equation
  - p or P(y|x) denote probability of getting outcome
  - q (or 1-p) denote probability of no outcome
- For left side of equation
  - p/q or p/(1-p) as y in the model

Model a function of y

$$\frac{p}{1-p} = A$$

$$p = A*(1-p) = A - Ap$$

$$p + Ap = A$$

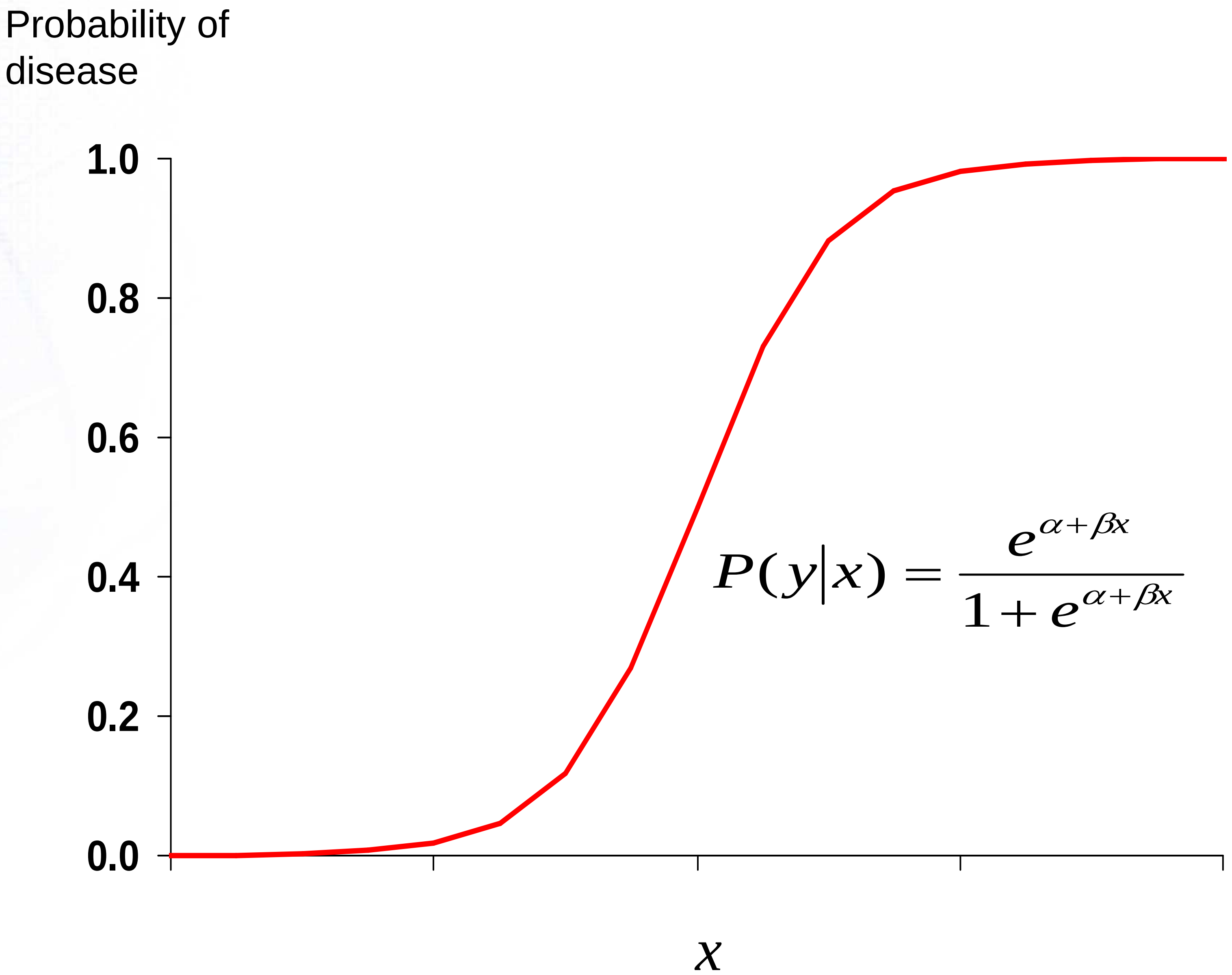
$$p(1+A) = A$$

$$p = \frac{A}{1+A}$$

$\frac{A}{1+A}$ belongs to the interval (0, 1) if A line between 0 and ∞

→ function of A → exponential of A → e^A

Logistic function



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

Model a function of y

$$\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

- The right-hand side of equation can take any value between -∞ and ∞
- The left-hand side of equation can take only value between 0 and ∞ (not linear model)
- Solution: take natural log (ln) on both sides
  - Both sides can take any value between -∞ and ∞ (as linear model)

Logit transformation

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

$$\underbrace{\ln\left[\frac{P(y|x)}{1-P(y|x)}\right]}_{\text{logit of } P(y/x)} = \beta_0 + \beta_i x_i$$

Advantages of Logit

- Properties of a linear regression model
- Logit between - ∞ and + ∞
- Probability (P) constrained between 0 and 1
- Directly related to odds of disease

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_i x_i \longrightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_i x_i}$$

e = The exponential function involves the constant with the value of 2.71828182845904 (roughly 2.72).

Fitting the equation of data

Maximum likelihood

- Yield values for the unknown parameters which maximize the probability of obtaining the observed set of data
- Maximum Likelihood Estimates (MLE) for β_0 and β_i

Likelihood function

- Express the probability of the observed data as a function of the unknown parameters
- The resulting estmators are those which agree most closely with the observed data
- Practically easier to work with log-likelihood: $L(\beta)$

  Let $\pi(x_i) = P(y \mid x)$

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

Logistic regression

- Simple Logistic regression
  - A single covariate
- Multiple logistic regression
  - Two or more covariate (with or without interactions)

Interpreting the estimated regression coefficients

- General concept
- Linear equation
- Linear regression
  - Simple regress
  - Multiple regress
  - Model testing
  - EpiInfo & STATA
- Logistic regression
  - Introduction
  - Coefficients
  - Simple regress
  - Multiple regress
  - Model testing
  - Model building
  - EpiInfo & STATA

- The simplest case is when the logistic regression model involves only one covariate (x), and that x takes only two values, 0(unexposed) and 1 (exposed)

- A logistic regression model for these data would correspond to

$$\ln\left[\frac{P(y|x_1)}{1-P(y|x_1)}\right] = \beta_0 + \beta_1 x_1$$

Interpreting the estimated regression coefficients

- For the exposed individuals (X1 = 1)

$$\ln\left[\frac{P(y|x_1=1)}{1-P(y|x_1=1)}\right]=\beta_0+\beta_1$$

- For the unexposed individuals (X1 = 0)

$$\ln\left[\frac{P(y|x_1=0)}{1-P(y|x_1=0)}\right]=\beta_0$$

Interpreting the estimated regression coefficients

- If we subtract the latter model equation (where X1 = 0) from the former (where X1 = 1)

$$\ln\left[\frac{P(y|x_1=1)}{1-P(y|x_1=1)}\right]-\ln\left[\frac{P(y|x_1=0)}{1-P(y|x_1=0)}\right]=(\beta_0+\beta_1)-\beta_0$$

$$\ln\left[\frac{P(y|x_1=1)}{1-P(y|x_1=1)}\div\frac{P(y|x_1=0)}{1-P(y|x_1=0)}\right]=\beta_1$$

	D+	D-
E+	a	b
E-	c	d

$$\ln\left[\frac{a/(a+b)}{b/(a+b)}\div\frac{c/(c+d)}{d/(c+d)}\right]=\beta_1$$

$$\frac{ad}{bc}=e^{\beta_1}$$

$$OR=e^{\beta_1}$$

- β = increase in natural log of odds ratio for a one unit increase in x

Interpreting the estimated regression coefficients

- General concept
- Linear equation
- Linear regression
  - Simple regress
  - Multiple regress
  - Model testing
  - EpiInfo & STATA
- Logistic regression
  - Introduction
  - Coefficients
  - Simple regress
  - Multiple regress
  - Model testing
  - Model building
  - EpiInfo & STATA

- β is amount of change in logit for each unit increase in X
- β is not interpreted directly, instead, e^{β} is interpreted
- e^{β} is equal to odds ratio; change in odds between a baseline group and a single unit increase in X
  - if $e^{\beta} = 1$ then there is no change in odds ratio
  - if $e^{\beta} < 1$ then odds ratio decrease
  - if $e^{\beta} > 1$ then odds ratio increase
  - $e^{\beta 0}$ is a baseline odds

Example data : CVD

Variable	Description	Value
id	Subject ID number	
gender	Gender	0=male,1=female
age	Age in year	
weight	Weigh in kilogram	
height	Height in centimeter	
bmi	Body mass index (kg/m^2)	
hypertension	Hypertension status	0=yes,1=no
dm	Diabetes status	0=yes,1=no
alcohol	Alcohol drinker	0=yes,1=no
smoke	Smoker	0=yes,1=no
cvd	CVD status	0=yes,1=no

Simple logistic regression

$$\text{logit}(y) = -38.15 + 1.61 * bmi$$

- Interpretation of β_1
  - log odds of CVD increase by 1.61 for a one unit increase in BMI
  - OR = $e^{1.61}$ = 5.0
  - The probability of getting CVD is 5.0 time as likely with an increase of BMI by one unit

Simple logistic regression

$$\text{logit}(y) = -0.75 + 1.19 * \mathit{Smoke}$$

- Interpretation of β_1
  - log odds of CVD increase by 1.19 for smoker
  - OR = $e^{1.19}$ = 3.3
  - Those who are smoker are 3.3 time as likely to get CVD compare to those who are non-smoker

Multiple logistic regression

$$\text{logit}(y) = -38.6 + 0.66 * smoke + 0.42 * alcohol + 1.6 * bmi$$

- Interpretation of β_1
  - log odds of CVD increase by 0.66 for smoker adjusted for alcohol and BMI
  - OR = $e^{0.66}$ = 1.9
  - Those who are smoker are 1.9 time as likely to get CVD compare to those who are non-smoker adjusted for alcohol and BMI

Pseudo R^2

- In linear regression we have the R^2 as a single agreed upon measure of goodness of fit of the model, the proportion of total variation in dependent variable accounted for by a set of covariates.
- No single agreed upon index of goodness of fit exists in logistic regression. Instead a number have been defined. These indices are sometimes referred to as Pseudo – R^2s.
  - R_L^2 :
  - Cox and Snell Index
  - Nagelkerke Index
- None of these indices have an interpretation as “proportion of variance accounted for” as in linear regression.
- It is important to indicate which index is being used in report.

Pseudo R²

- R_L^2 : a commonly used index
  - ranges between 0 and 1
  - can be calculated from the deviance(−2LL) from the null model (no predictors) and the model containing k predictors.
  - Interpreted as the proportion of the null deviance accounted for by the set of predictors

$$R_L^2 = \frac{D_{null} - D_k}{D_{null}}$$

Hypothesis testing of the model

- p-value or critical value approach
- Two parts of testing
  - Significant of overall model (do all covariates together can predict the outcome?)
    - Testing for overall fit of the model
  - Significant of each β_i (adjusted for other covariates)
    - Using Wald test

Overall fit of the model

- Measures of model fit and tests of significance for logistic regression are not identical to those in linear regression.

- In logistic regression, measures of deviance replace the sum of squares of linear regression as the building blocks of measures of fit and statistical tests. These measures can be thought of as analogous to sums of squares, though they do not arise from the same calculations.

- Each deviance measure in logistic regression is a measure of lack of fit of the data to a logistic model.

Overall fit of the model

- Two measures of deviance are needed
  - the null deviance, D_{null}, which is the analog of SS_Y (or SST) in linear regession.
    - D_{null} is a summary number of all the deviance that could potentially be accounted for. It can be thought of as a measure of lack of fit of data to a model containing an intercept but not predictors. It provides a baseline against which to compare prediction from other models that contain at least one predictor.
  - the model deviance from a model containing k predictors, D_k; it is the analog of SSE in linear regression.
    - It is a summary number of all the deviance that remains to be predicted after prediction from a set of k predictors, a measure of lack of fit of the model containing k predictors.
- If the model containing k predictors fits better than a model containing no predictors, then $D_k < D_{null}$. This is the same idea as in linear regression: if a set of predictors in linear regression provides prediction, then SSE < SST

Overall fit of the model

- The statistical tests built on deviances are referred to collectively as *likelihood ratio tests* because the deviance measures are derived from ratios of maximum likelihoods under different models.
- Standard notation for deviance : −2LL or −2 *log likelihood*
- Likelihood ratio test (LR)
  - $H_0 : \beta_1 = \beta_1 = \ldots = \beta_i = 0$
  - A likelihood ratio χ^2 test $= D_{null} - D_k$, which follows a χ^2 distribution with df = k
  - reject H_0 if observed is greater than 100(1- α) percentile of χ^2 distribution with k df

Testing of β

- If an acceptable model is found, the statistical significance of each of the coefficients is evaluated using the Wald test
  - Using chi-square test with 1 df

$$W_i = \frac{\beta_i^2}{S_{\beta_i}^2}$$

  - Using z test

$$W_i = \frac{\beta_i}{S_{\beta_i}}$$

Wald test

- It provides a test for

  $H_0 : \beta_{i.\ \text{given other covariates in the model}} = 0$

- Test statistics: chi-square or z test
- The decision rule is: Reject H_0 if
  - Wald's $\chi^2 > \chi^2_{(1-\alpha, df=k)}$, or Wald's $|z| > Z_{(1-\alpha/2)}$
  - p-value $< \alpha$
- Interpretation: if reject H_0

  There was a significant prediction of outcome by x_i adjusted for other covariates in the model

Model building

- Depends on study objectives
  - single hypothesis testing : estimate the effect of a given variable adjusted for all potential available confounders
  - exploratory study : identify a set of variables independently associated to the outcome adjusted for all potential available confounders
  - to predict : predict the outcome with the least varieble possible (parsimony principle)

Model building strategy

- Variable selection
- Approach
  - Hierarchical or sequential regression: investigator's judgement (preferred)
  - Best subset: computer's judgement (not recommend)
- Exclusion of least significant variables base on statitical test and no important variation of coefficient until obtaining a satisfactory model
- Add and test interaction terms
- Final model with interaction term if any
- Check model fit

Variable selection

- what is the hypothesis, what is (are) the variable(s) of interest
  - Define variables to keep in the model (forced)
- what are the confounding variables
  - Literature review
  - Confounding effect in the dataset (bivariate analysis)
- biologic pathway and plausibilty
- statistical (p-value)
  - Variable with p-value < 0.2
- Form of continuous variables
  - Original form as continuous
  - Categorized variable and coding

Hierarchical regression

- Assess whether a set of m predictors contribute significant prediction over and above a set of k predictors
  - likelihood ratio (LR) test
  - Deviance are computed for the k predictor model, D_k, and the (m + k) predictor model, D_{m+k}
  - The difference between these deviances is an LR test for the significance of contribution of the set of m predictors over and above the set of k predictors, with df = m
  - The LR tests are used to compare models that are nested : all the predictors in the smaller (reduced) model are included among the predictors in the larger (full) model.
- Backward elimination or forward selection approach

Likelihood ratio statistic

- Compares two nested models

$$\text{Log(odds)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \qquad \text{(model 1)}$$

$$\text{Log(odds)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad \text{(model 2)}$$

- LR statistic

-2LL model 2 *minus* -2LL model 1

LR statistic is a χ^2 with df = number of extra parameters in model

Example

- Full Model: smoke, alcohol, bmi
  - Deviance (-2LL): 100.03
- Reduced model: smoke, bmi
  - Deviance (-2LL): 100.39
- H_0: reduced model contribute significant prediction over full model
- Test statistics
  - LR test = 100.39 – 100.03 = 0.36
  - df = 3-2 = 1
- Critical value of $\chi^2_{(.95, df=1)}$ = 3.84
- Reject the H_0
- Conclusion: alcohol does no contribute significant prediction in the model

Interaction term

- If 2 covariates are x1 and x2
- Interaction is include in the model as x1*x2
- Assessing interaction use the same strategy of LR test
  - Full model : x1, x2, x1*x2
  - Reduced model : x1, x2
  - df = 3-2 = 1

Assumption

- No multicollinearity among covariates
- Linear relationships between continuous variables and a logit
- No outliers among errors

Check model fit

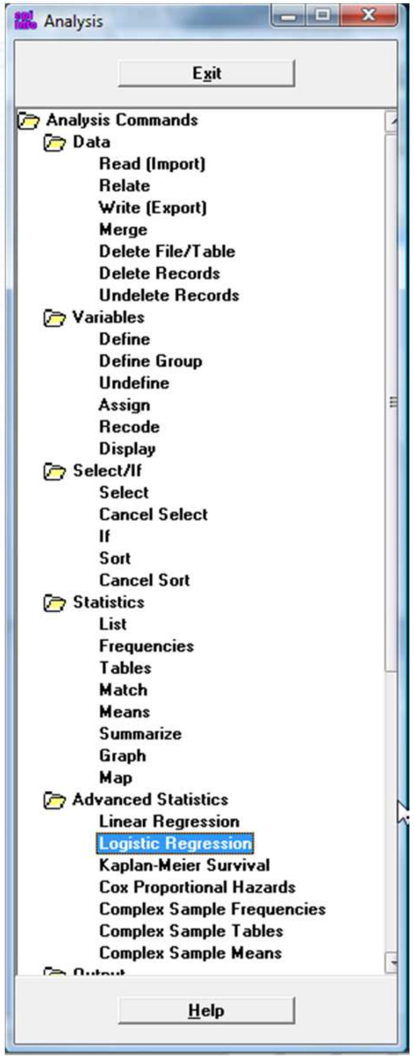
- Summary measure of Goodness-of-Fit
  - Pearson Chi-square statistics and deviance approach
  - Hosmer-Lemeshow test
  - Area under ROC curve
- Model diagnostics
  - Check whether the model violate the assumptions

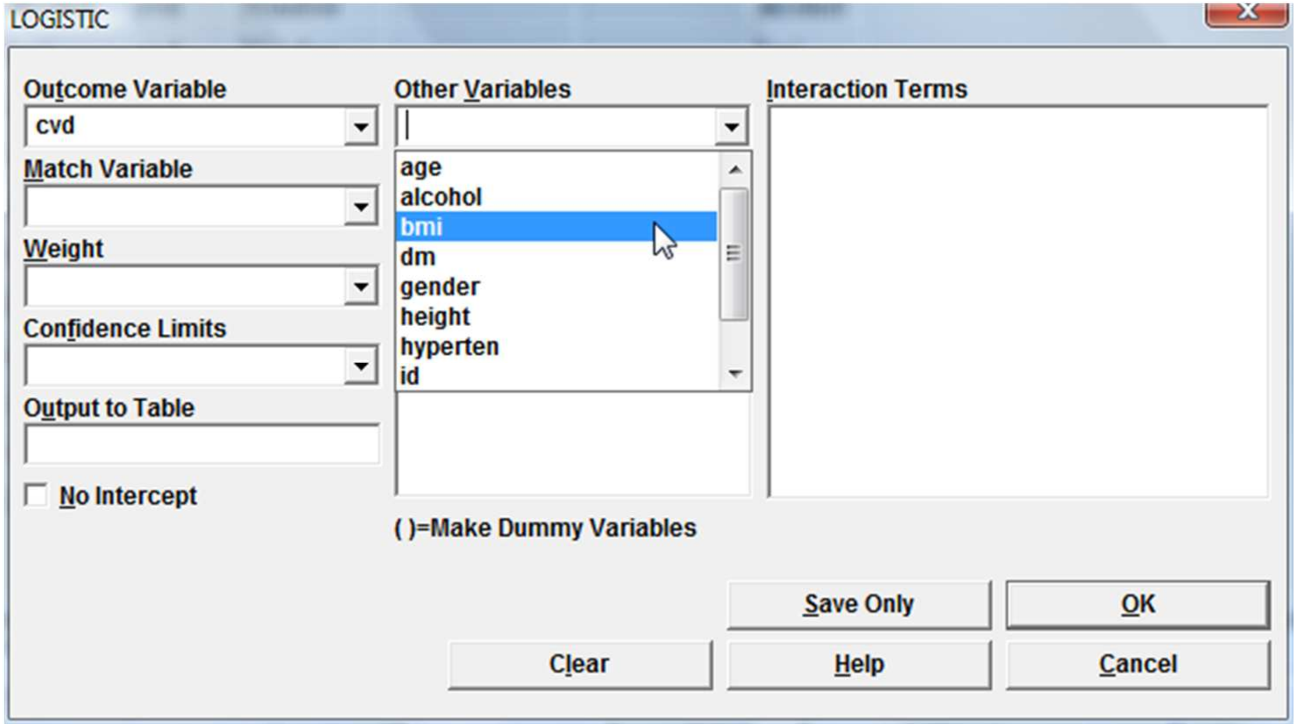
Example command in EpiInfo

- General concept
- Linear equation
- Linear regression
  - Simple regress
  - Multiple regress
  - Model testing
  - EpiInfo & STATA
- Logistic regression
  - Introduction
  - Coefficients
  - Simple regress
  - Multiple regress
  - Model testing
  - Model building
  - EpiInfo & STATA

➢ Simple logistic regression

LOGISTIC cvd = bmi





Example output in EpiInfo

➢ Simple logistic regression

D:\???????\Course\Introduction to logistic (Monday)\OUT24.htm

Previous | Next | Last | History | Open | Bookmark | Print | Maximize

LOGISTIC cvd = bmi

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
bmi	4.9967	3.1278	7.9823	1.6088	0.2390	6.7312	0.0000
CONSTANT	*	*	*	-38.1488	5.6958	-6.6977	0.0000

Wald test

Convergence:	Converged
Iterations:	8
Final -2*Log-Likelihood:	103.7093
Cases included:	200

Deviance (-2LL)

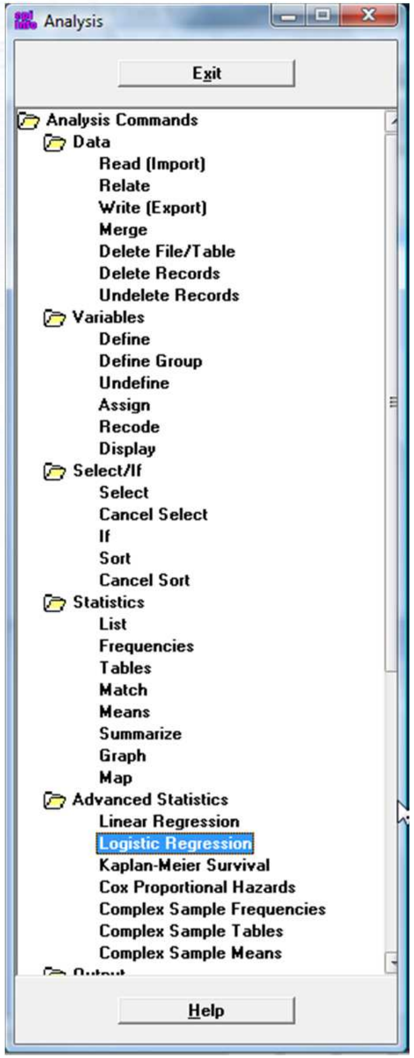
Test	Statistic	D.F.	P-Value
Score	119.7305	1	0.0000
Likelihood Ratio	173.5496	1	0.0000

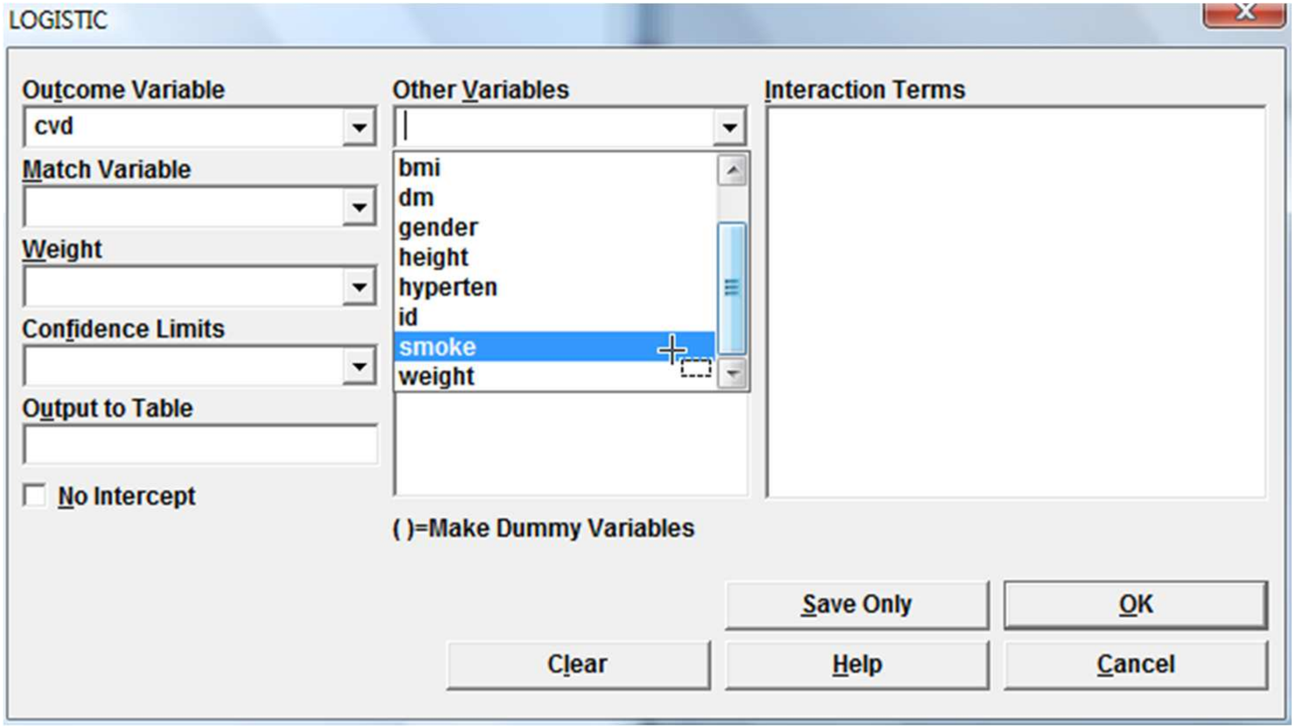
LR test (to null)

Example command in EpiInfo

➢ Simple logistic regression

LOGISTIC cvd = smoke





Example output in EpiInfo

➢ Simple logistic regression

D:\???????\Course\Introduction to logistic (Monday)\OUT24.htm

Previous | Next | Last | History | Open | Bookmark | Print | Maximize

LOGISTIC cvd = smoke

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
smoke	3.2959	1.8024	6.0271	1.1927	0.3080	3.8729	0.0001
CONSTANT	*	*	*	-0.7538	0.2475	-3.0452	0.0023

Wald test

Convergence:	Converged
Iterations:	4
Final -2*Log-Likelihood:	261.4390
Cases included:	200

Deviance (-2LL)

Test	Statistic	D.F.	P-Value
Score	15.5520	1	0.0001
Likelihood Ratio	15.8199	1	0.0001

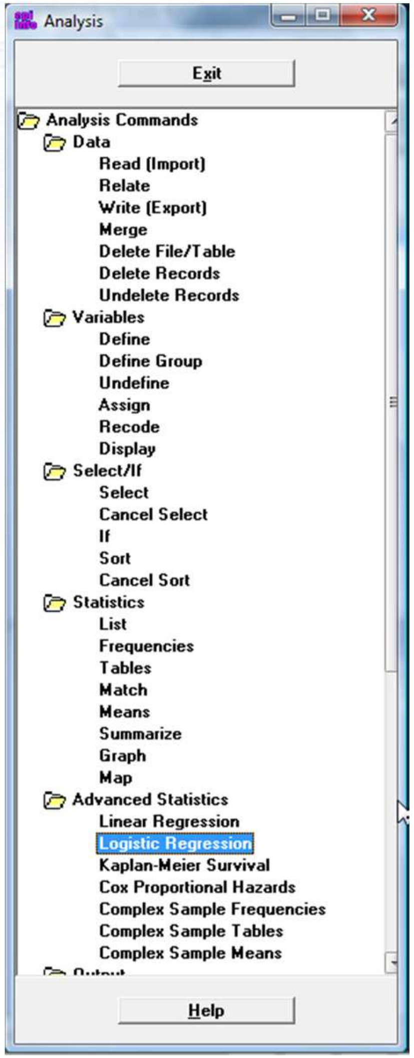
LR test (to null)

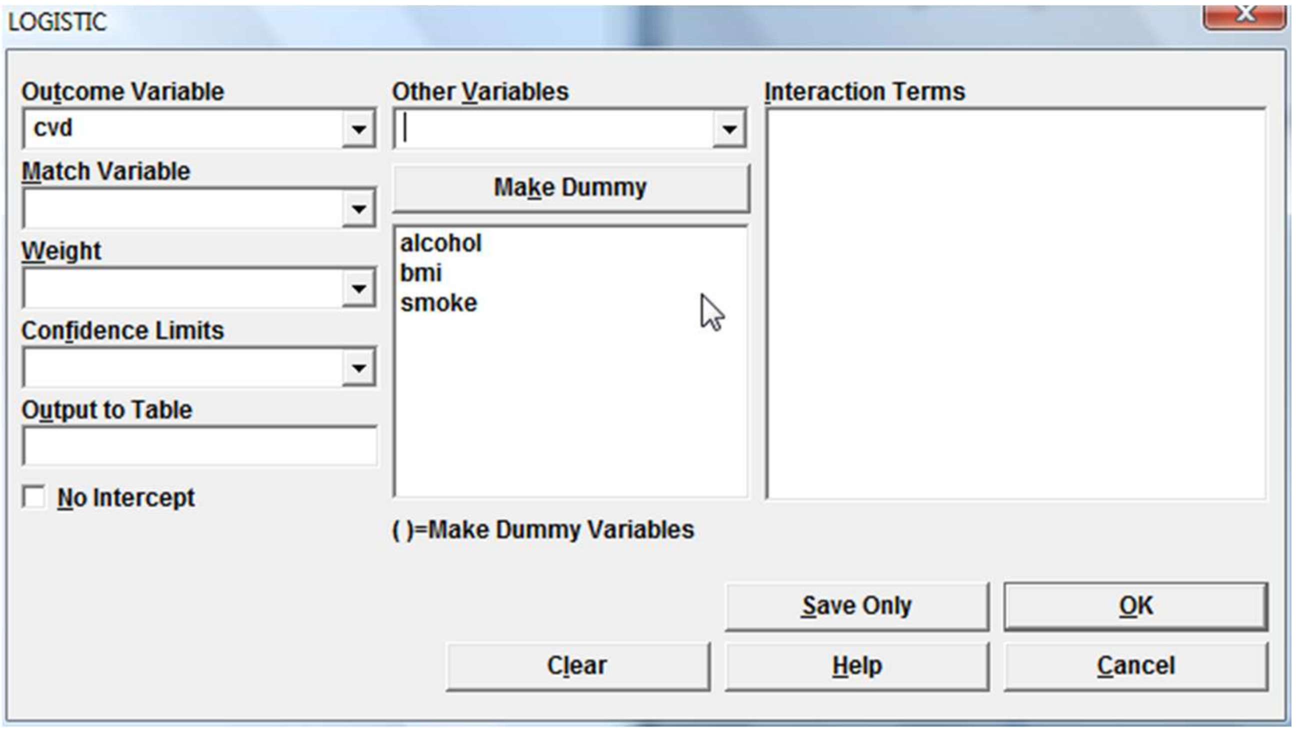
Example command in EpiInfo

- General concept
- Linear equation
- Linear regression
  - Simple regress
  - Multiple regress
  - Model testing
  - EpiInfo & STATA
- Logistic regression
  - Introduction
  - Coefficients
  - Simple regress
  - Multiple regress
  - Model testing
  - Model building
  - EpiInfo & STATA

➢ Multiple logistic regression

LOGISTIC cvd = alcohol bmi smoke





Example output in EpiInfo

➢ Multiple logistic regression

D:\???????\Course\Introduction to logistic (Monday)\OUT24.htm

Previous | Next | Last | History | Open | Bookmark | Print | Maximize

LOGISTIC cvd = alcohol bmi smoke

Next Procedure

Unconditional Logistic Regression

Term	Odds Ratio	95%	C.I.	Coefficient	S. E.	Z-Statistic	P-Value
alcohol	1.5280	0.3694	6.3201	0.4240	0.7244	0.5853	0.5584
bmi	4.9537	3.0524	8.0391	1.6001	0.2470	6.4772	0.0000
smoke	1.9337	0.4563	8.1948	0.6595	0.7368	0.8951	0.3708
CONSTANT	*	*	*	-38.6038	5.9242	-6.5163	0.0000

Wald test

Convergence:	Converged
Iterations:	8
Final -2*Log-Likelihood:	100.0327
Cases included:	200

Deviance (-2LL)

Test	Statistic	D.F.	P-Value
Score	122.1242	3	0.0000
Likelihood Ratio	177.2261	3	0.0000

LR test (to null)

Example command & output in STATA

➢ Simple logistic regression: compute coefficients

logit cvd bmi

Stata Results

```
. logit   cvd bmi

Iteration 0:   log likelihood = -138.62944
Iteration 1:   log likelihood = -69.416219
Iteration 2:   log likelihood = -56.009483
Iteration 3:   log likelihood = -52.374342
Iteration 4:   log likelihood = -51.869532
Iteration 5:   log likelihood = -51.854656
Iteration 6:   log likelihood = -51.854639

Logit estimates                                   Number of obs   =        200
                                                  LR chi2(1)      =     173.55
                                                  Prob > chi2     =     0.0000
Log likelihood = -51.854639                       Pseudo R2       =     0.6259

------------------------------------------------------------------------------
         cvd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   1.608781   .2390045     6.73   0.000     1.140341    2.077221
       _cons |  -38.14877   5.695823    -6.70   0.000    -49.31238   -26.98517
------------------------------------------------------------------------------
```

LR test
(to null)

LL

Wald test

Example command & output in STATA

➢ Simple logistic regression: compute OR

logit cvd bmi, or

Logistic cvd bmi

```
Stata Results

. logit   cvd bmi, or

Iteration 0:   log likelihood = -138.62944
Iteration 1:   log likelihood = -69.416219
Iteration 2:   log likelihood = -56.009483
Iteration 3:   log likelihood = -52.374342
Iteration 4:   log likelihood = -51.869532
Iteration 5:   log likelihood = -51.854656
Iteration 6:   log likelihood = -51.854639

Logit estimates                                   Number of obs   =        200
                                                  LR chi2(1)      =     173.55
                                                  Prob > chi2     =     0.0000
Log likelihood = -51.854639                       Pseudo R2       =     0.6259

------------------------------------------------------------------------------
         cvd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   4.996715   1.194237     6.73   0.000     3.127834    7.982255
------------------------------------------------------------------------------
```

LR test
(to null)

LL

Wald test

Example command & output in STATA

➢ Simple logistic regression: compute coefficients

logit cvd smoke

Stata Results

```
. logit cvd smoke

Iteration 0:   log likelihood = -138.62944
Iteration 1:   log likelihood = -130.72955
Iteration 2:   log likelihood =  -130.7195
Iteration 3:   log likelihood =  -130.7195

Logit estimates                                   Number of obs   =        200
                                                  LR chi2(1)      =      15.82
                                                  Prob > chi2     =     0.0001
Log likelihood =  -130.7195                       Pseudo R2       =     0.0571

------------------------------------------------------------------------------
         cvd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   1.192685    .307962     3.87   0.000     .5890905    1.796279
       _cons |  -.7537718   .2475369    -3.05   0.002    -1.238935   -.2686084
------------------------------------------------------------------------------
```

LR test (to null)

LL

Wald test

Example command & output in STATA

➢ Simple logistic regression: compute OR

logit cvd smoke, or

logistic cvd smoke

Stata Results

```
. logit cvd smoke, or

Iteration 0:   log likelihood = -138.62944
Iteration 1:   log likelihood = -130.72955
Iteration 2:   log likelihood =  -130.7195
Iteration 3:   log likelihood =  -130.7195

Logit estimates                                   Number of obs   =        200
                                                  LR chi2(1)      =      15.82
                                                  Prob > chi2     =     0.0001
Log likelihood =  -130.7195                       Pseudo R2       =     0.0571

------------------------------------------------------------------------------
         cvd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   3.295918   1.015017     3.87   0.000     1.802348     6.02718
------------------------------------------------------------------------------

.
```

LR test
(to null)

LL

Wald test

Example command & output in STATA

- General concept
- Linear equation
- Linear regression
  - Simple regress
  - Multiple regress
  - Model testing
  - EpiInfo & STATA
- Logistic regression
  - Introduction
  - Coefficients
  - Simple regress
  - Multiple regress
  - Model testing
  - Model building
  - EpiInfo & STATA

➢ Multiple logistic regression: compute coefficients

logit cvd smoke alcohol bmi

Stata Results

```
. logit cvd smoke alcohol bmi

Iteration 0:   log likelihood = -138.62944
Iteration 1:   log likelihood = -68.074226
Iteration 2:   log likelihood = -54.442985
Iteration 3:   log likelihood =   -50.6162
Iteration 4:   log likelihood = -50.036701
Iteration 5:   log likelihood = -50.016401
Iteration 6:   log likelihood = -50.016368

Logit estimates                                   Number of obs   =        200
                                                  LR chi2(3)      =     177.23
                                                  Prob > chi2     =     0.0000
Log likelihood = -50.016368                       Pseudo R2       =     0.6392

------------------------------------------------------------------------------
         cvd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   .6594565   .7367727     0.90   0.371    -.7845915    2.103504
     alcohol |   .4239548   .7243875     0.59   0.558    -.9958186    1.843728
         bmi |   1.600128   .2470412     6.48   0.000     1.115936     2.08432
       _cons |  -38.60376   5.924196    -6.52   0.000    -50.21497   -26.99255
------------------------------------------------------------------------------
```

LR test (to null)

LL

Wald test

Example command & output in STATA

➢ Multiple logistic regression: compute OR

logit cvd smoke alcohol bmi, or

logistic cvd smoke alcohol bmi

Stata Results

```
. logit cvd smoke alcohol bmi, or

Iteration 0:   log likelihood = -138.62944
Iteration 1:   log likelihood = -68.074226
Iteration 2:   log likelihood = -54.442985
Iteration 3:   log likelihood =   -50.6162
Iteration 4:   log likelihood = -50.036701
Iteration 5:   log likelihood = -50.016401
Iteration 6:   log likelihood = -50.016368

Logit estimates                                   Number of obs   =        200
                                                  LR chi2(3)      =     177.23
                                                  Prob > chi2     =     0.0000
Log likelihood = -50.016368                       Pseudo R2       =     0.6392

------------------------------------------------------------------------------
         cvd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   1.933741   1.424728     0.90   0.371     .4563061    8.194838
     alcohol |   1.527992   1.106859     0.59   0.558     .3694209    6.320056
         bmi |   4.953665   1.223759     6.48   0.000     3.052424     8.03912
------------------------------------------------------------------------------
```

LR test (to null)

LL

Wald test