SPH3101 Biostatistics for Public Health
Lecture 2
Simple linear regression

Kiesha Prem

Saw Swee Hock School of Public Health, National University of Singapore

TM-CM02

Biostatistics for Public Health

- quizzes
- 2 assignments

1. **Categorical data analysis**
   Data prep, descriptive & inferential statistics

2. **Hypothesis testing**
   Inferential statistics

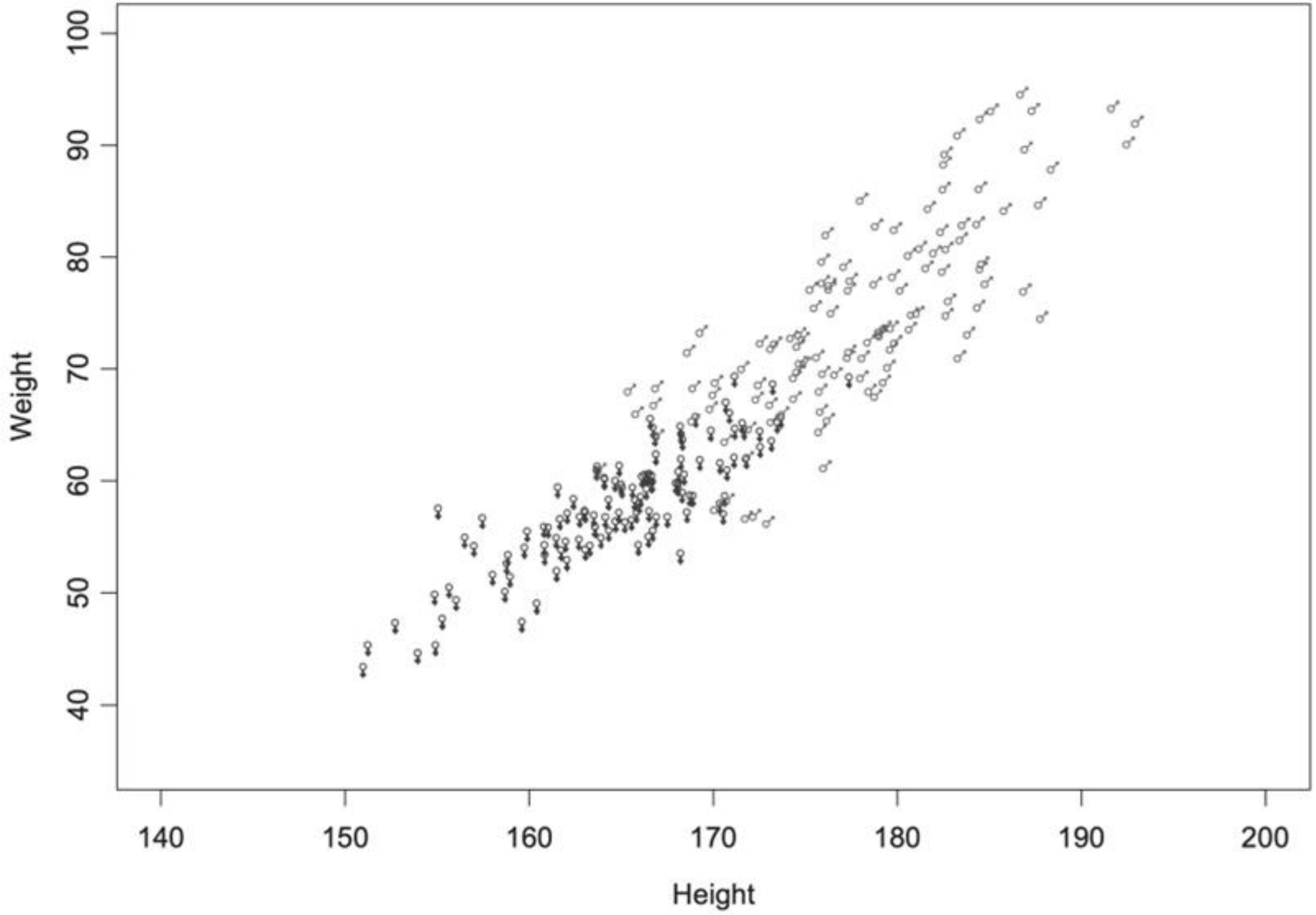
3. **Linear regression**
   Data prep, descriptive & inferential statistics, linear regression

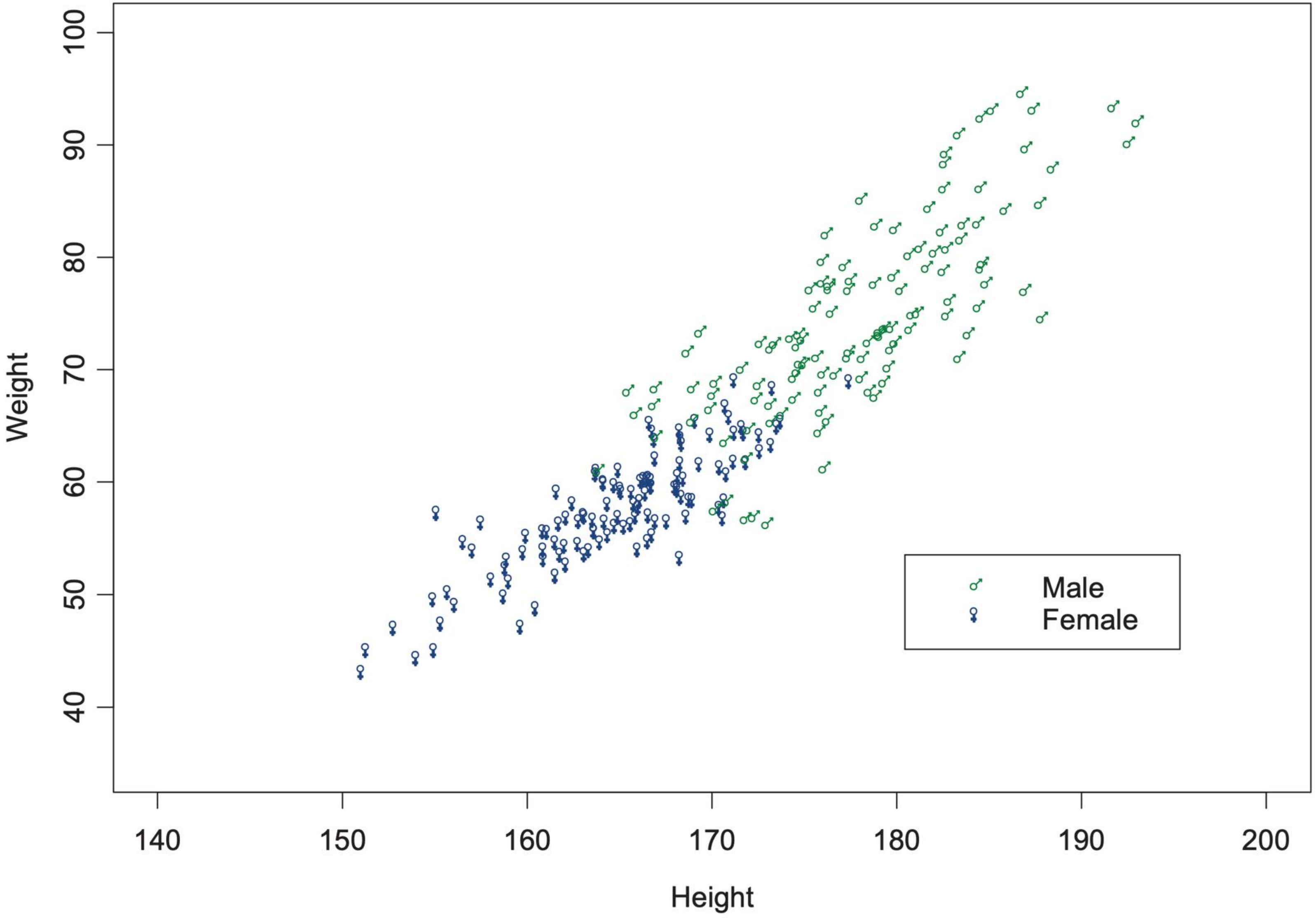
4. **Logistic regression**
   Data prep, descriptive & inferential statistics, logistics regression and others if we have time

NUS National University of Singapore
Saw Swee Hock School of Public Health

Learning objectives

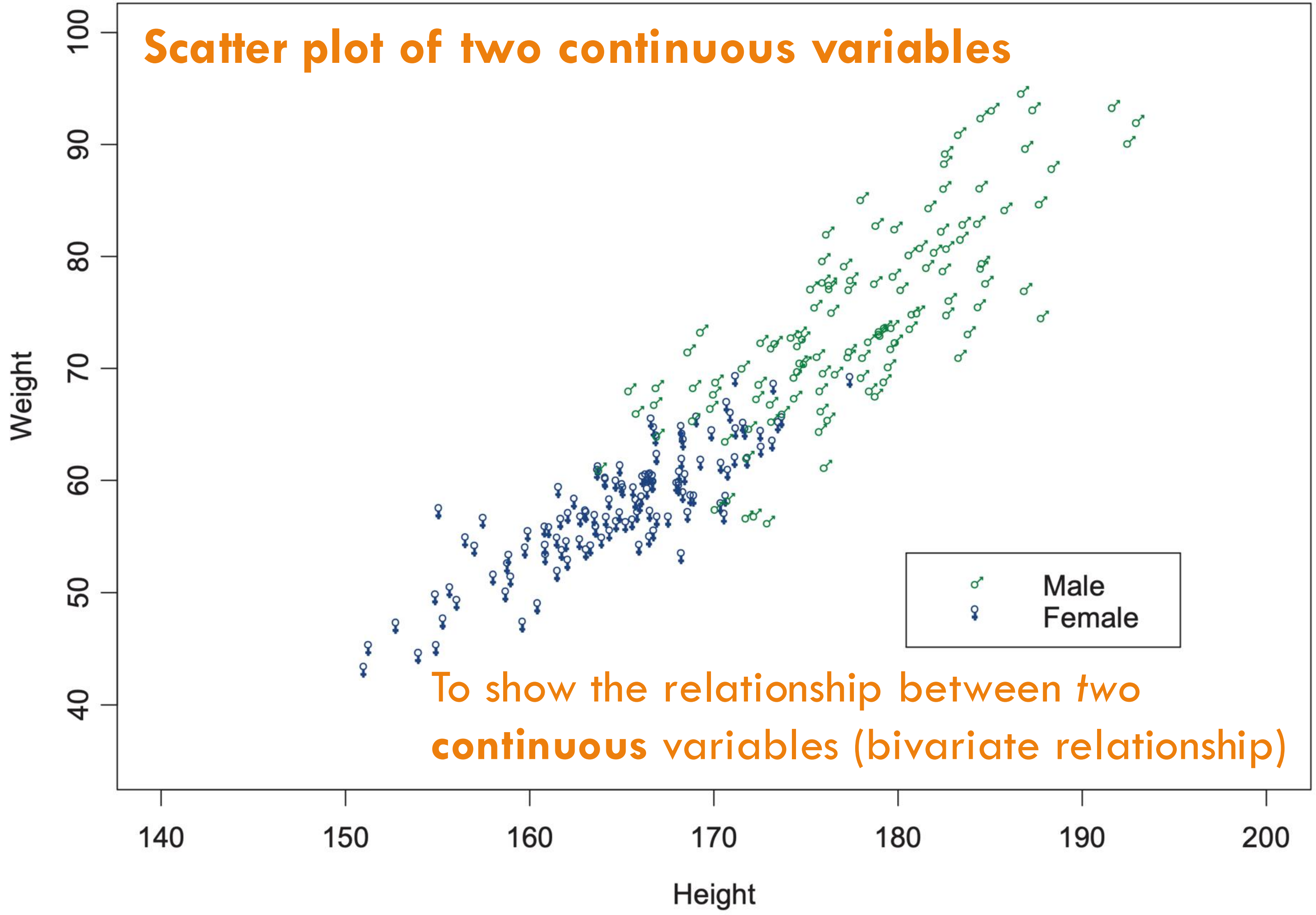
- Examine the **bivariate relationship** between continuous variables
  - Summarising the relationship using a **scatter plot**
- Quantify the **strength of the relationship** with the **correlation coefficient**
- Understand the basics of regression analysis and the **coefficients** β_0 and β_1
- Evaluate and verify the **assumptions** of regression analysis
- Make **inferences** about the **slope** and **correlation coefficient**
- Estimate **mean values** and **predict individual values** using regression analyses


Weight
100
90
80
70
60
50
40
140
150
160
170
180
190
200
Height

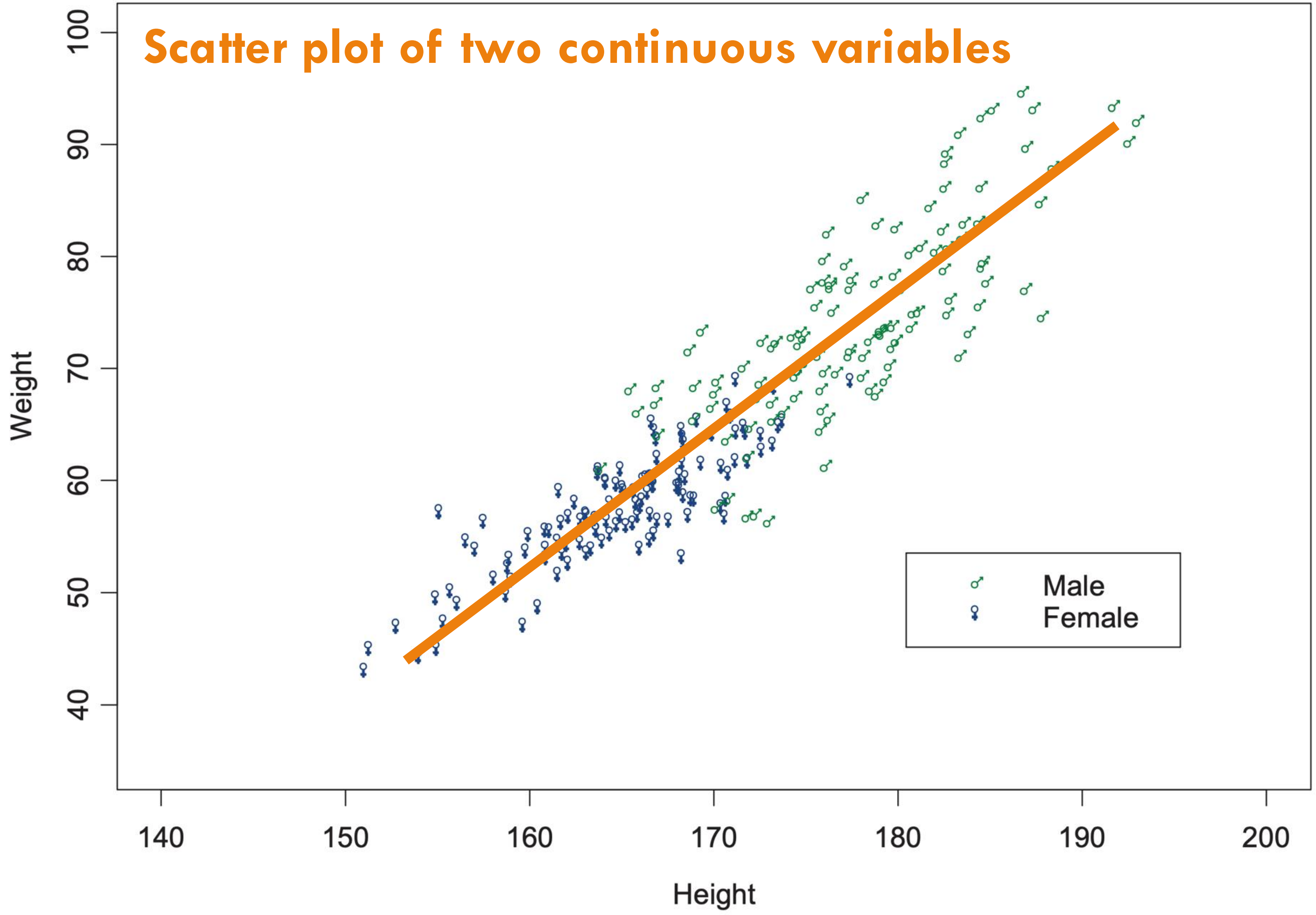


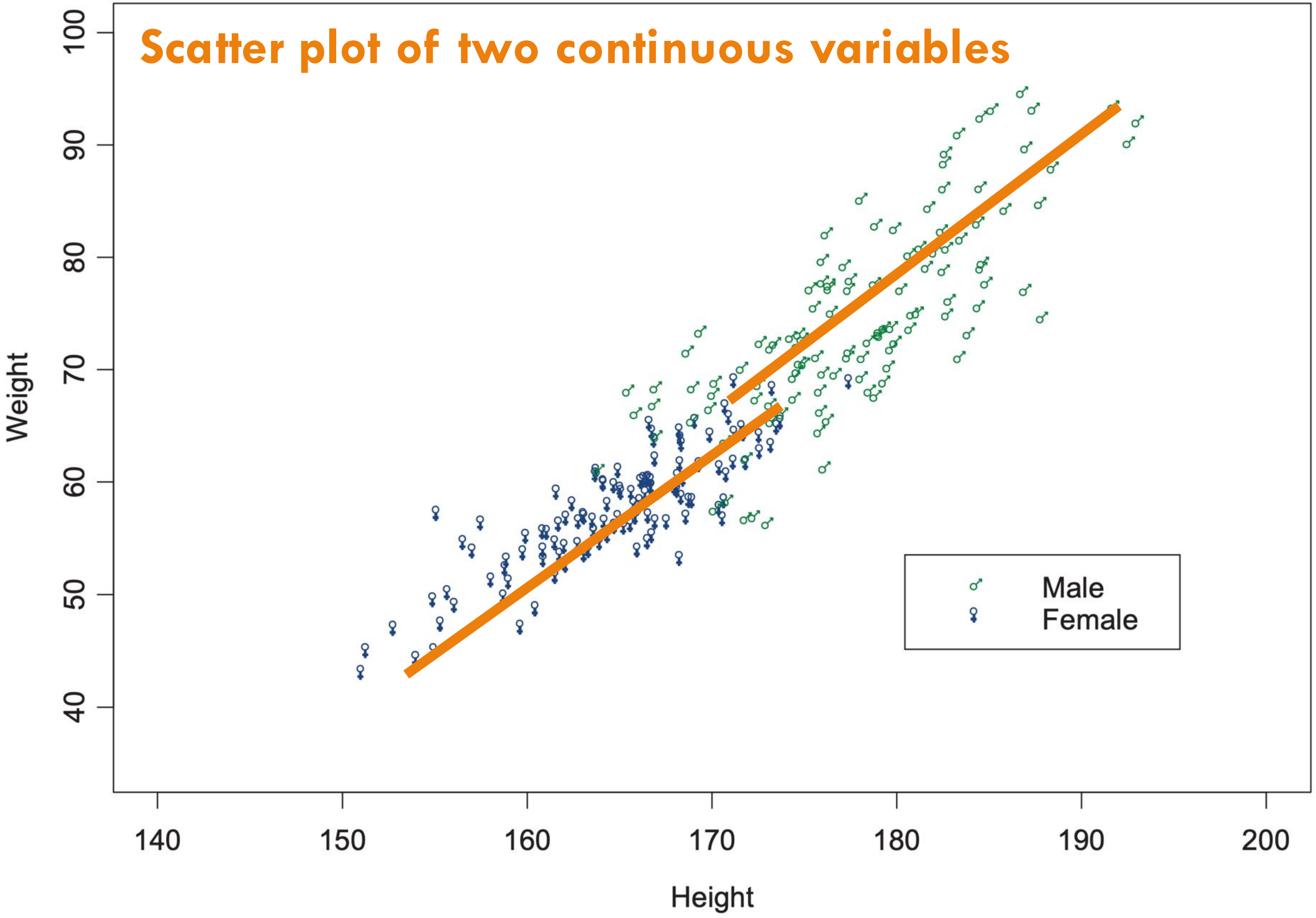
Weight
100
90
80
70
60
50
40
140
150
160
170
180
190
200
Height
Male
Female

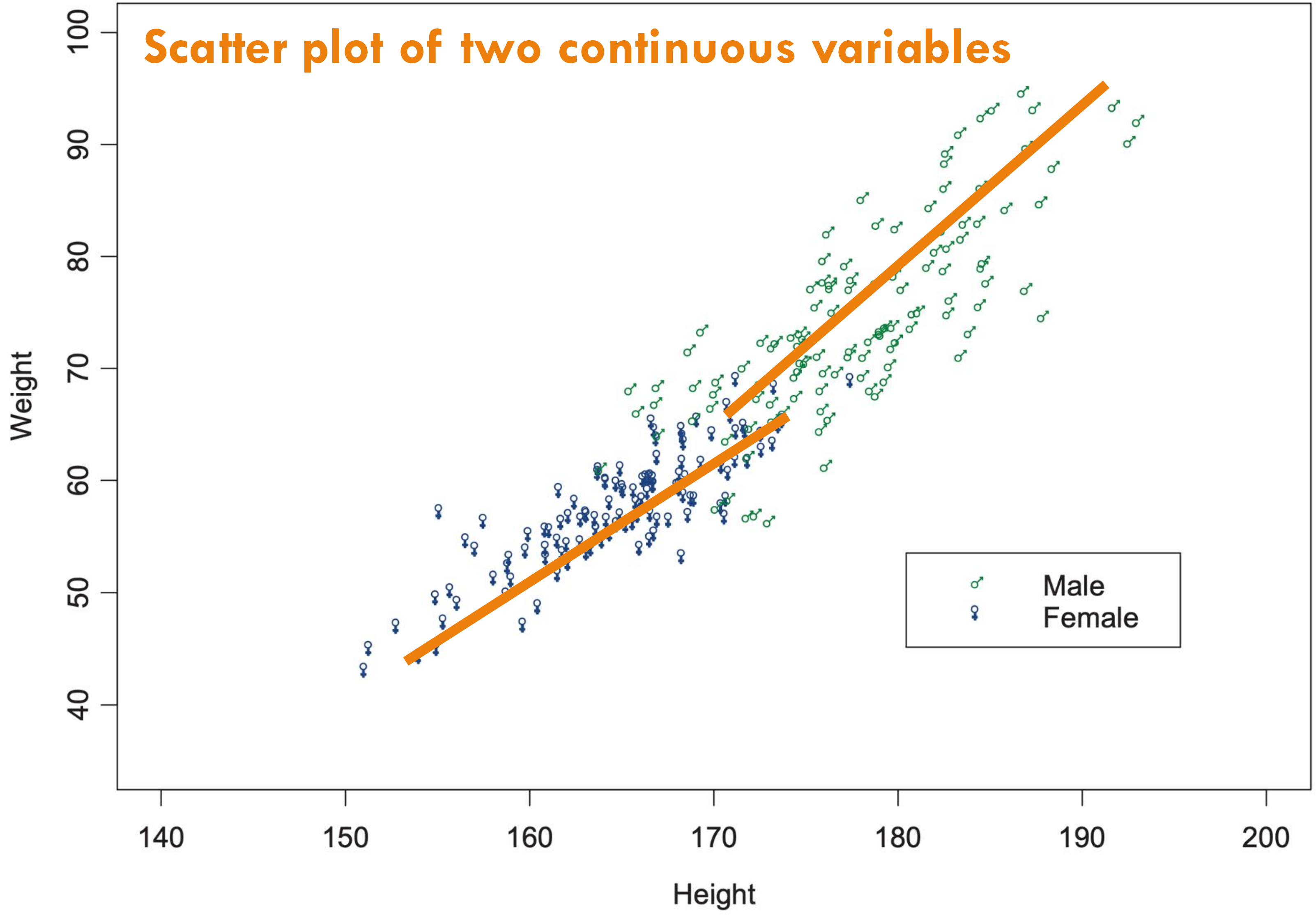
Scatter plot of two continuous variables



To show the relationship between *two* **continuous** variables (bivariate relationship)

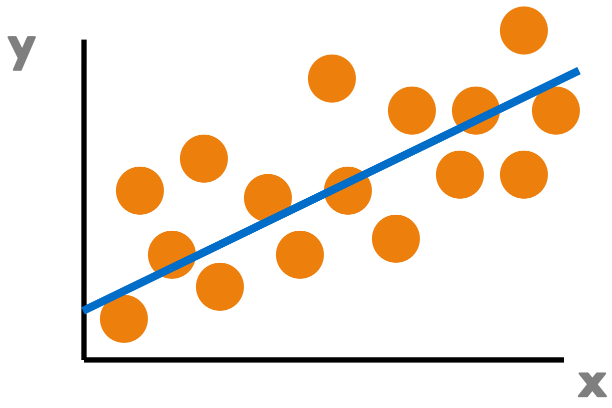

Scatter plot of two continuous variables
Weight
100
90
80
70
60
50
40
140
150
160
170
180
190
200
Height
Male
Female

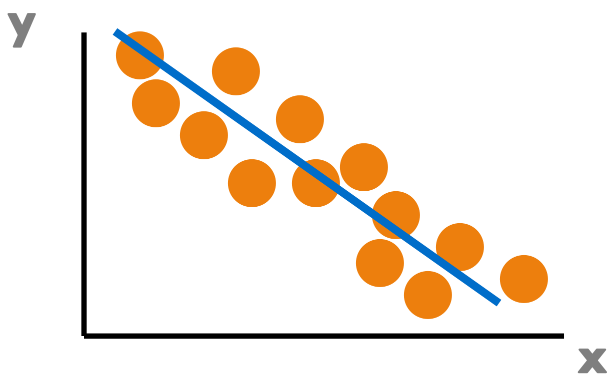

Scatter plot of two continuous variables
Weight
100
90
80
70
60
50
40
Height
140
150
160
170
180
190
200
Male
Female


Scatter plot of two continuous variables
Weight
100
90
80
70
60
50
40
140
150
160
170
180
190
200
Height
Male
Female

Correlation analysis

Linear relationships

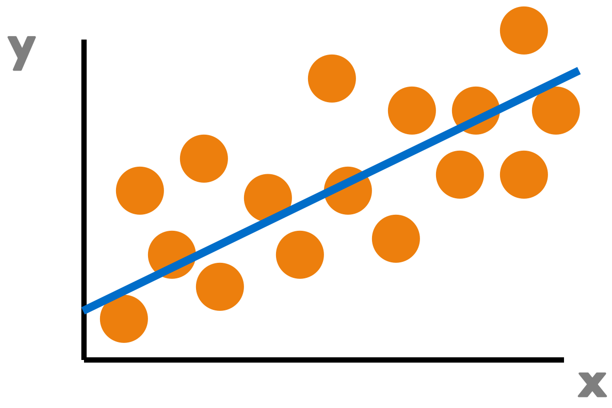


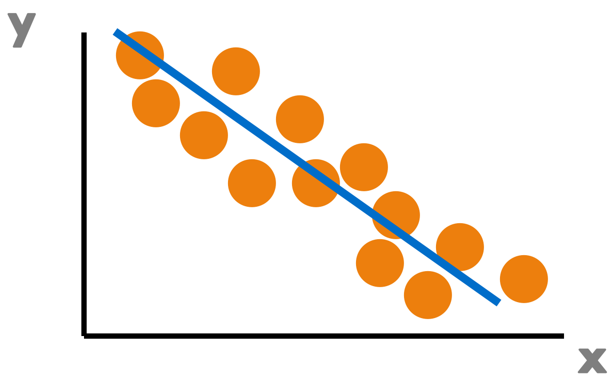


- Quantifies the strength of association between two continuous variables
- Range of correlation coefficient: **$-1 \leq r \leq 1$**
- *r* is a measure of scatter around an underlying **linear** trend
- Only concerned with the strength of the relationship

Types of relationship

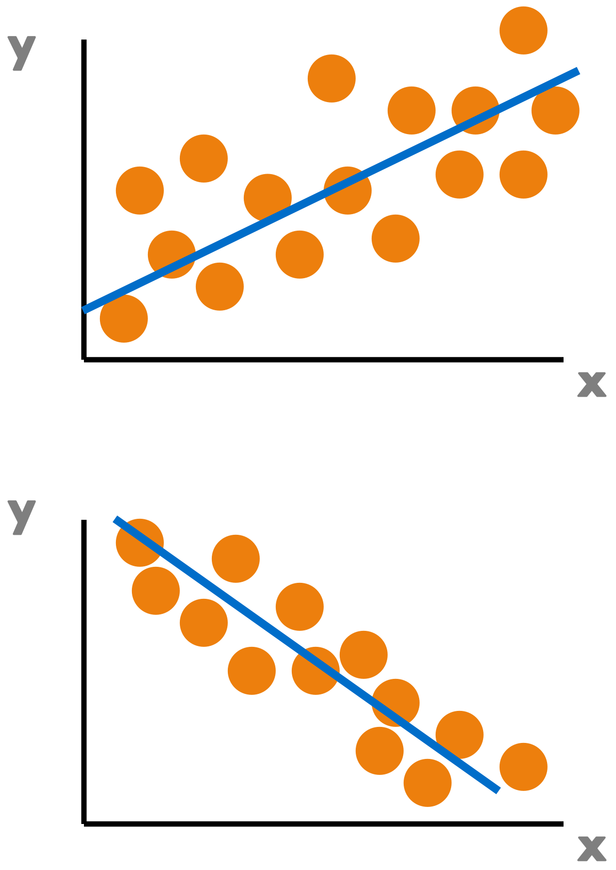
Linear relationships



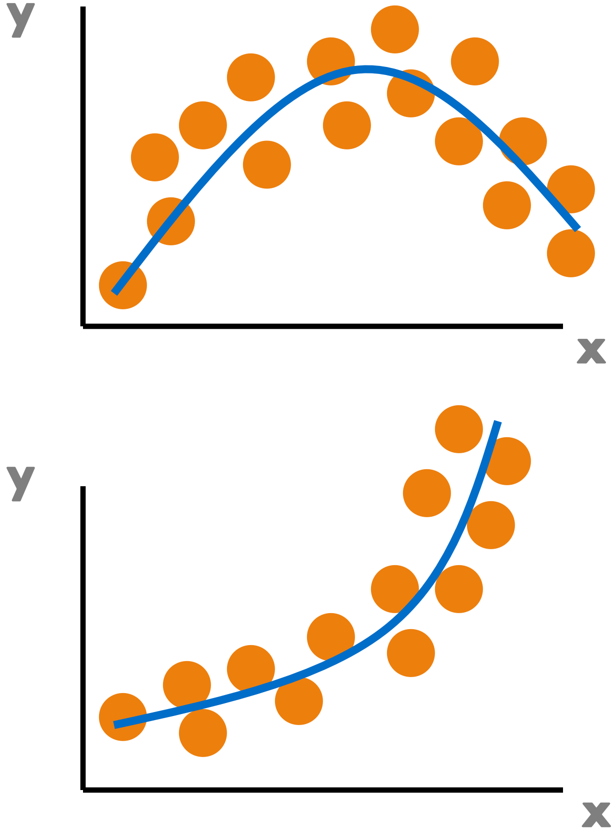


Types of relationship

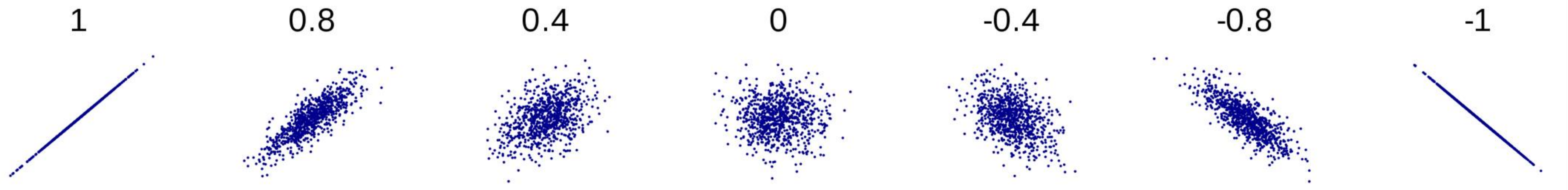
Linear relationships



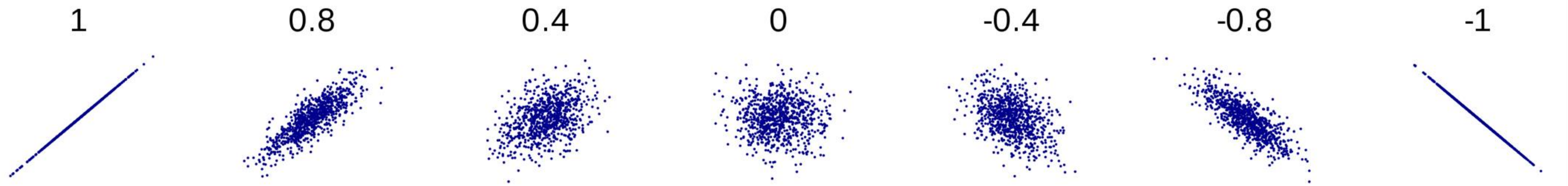
Nonlinear relationships



The correlation coefficient, *r*

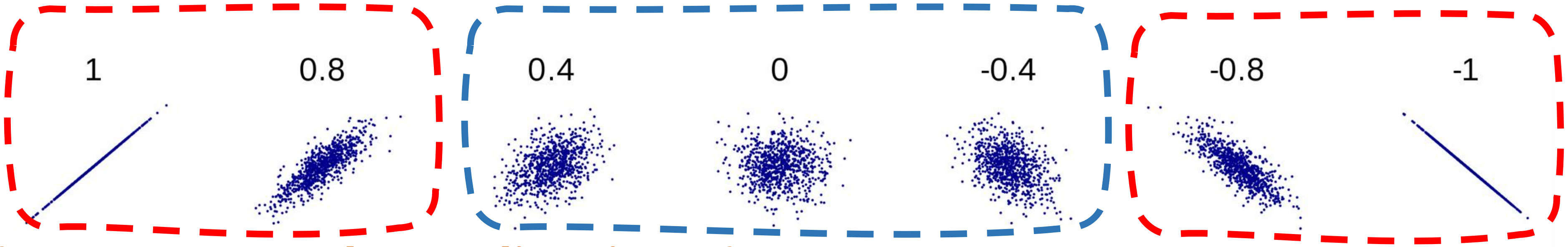


The correlation coefficient, *r*



- Reflects the **strength** and **direction** of a linear relationship

The correlation coefficient, *r*



- Reflects the **strength** and **direction** of a linear relationship

Positive relationship

Negative relationship

Strong relationship

Weak/no relationship

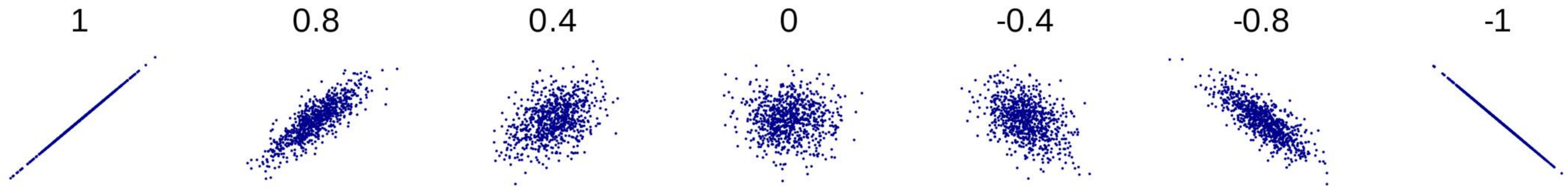
Strong relationship

The correlation coefficient, *r*

- Reflects the **strength** and **direction** of a linear relationship

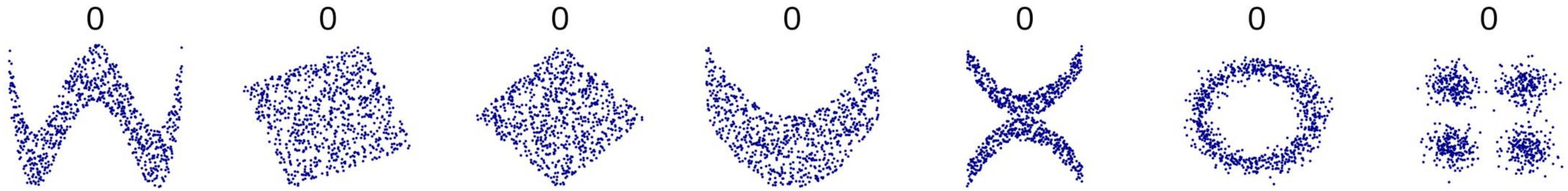
- Does not quantify the **slope** of the relationship

The correlation coefficient, *r*



- Reflects the **strength** and **direction** of a linear relationship

- Does not quantify the **slope** of the relationship



Nonlinear relationship

Two separate goals in regression

Prediction

- fitting a predictive model to an observed dataset, then using that model to make predictions about an outcome from a new set of explanatory variables;

Explanation

- fitting a model to explain the inter-relationships between a set of variables.

Regression analysis

Estimate the relationships between a dependent variable and independent variable(s)

- Explain the impact of changes in an independent variable on the dependent variable

Linear regression most common form of regression

Regression analysis

Estimate the relationships between a dependent variable and independent variable(s)

- Explain the impact of changes in an independent variable on the dependent variable

Linear regression most common form of regression

Dependent variable: the variable we wish to explain or predict, often called the *outcome* or *response* variable.

Independent variable: the variable used to explain or predict the dependent variable, often called *explanatory variable*, *covariates* or *predictors*.

Simple linear regression

Only **one** independent variable, X

A linear function describes the relationship between X and Y:

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

- Changes in Y are assumed to be related to changes in X

Simple linear regression

Dependent variable

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

Simple linear regression

Dependent variable

Independent variable

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

Simple linear regression

Dependent variable

Independent variable

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

Intercept

Slope (regression) coefficient

Simple linear regression

Dependent variable

Independent variable

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

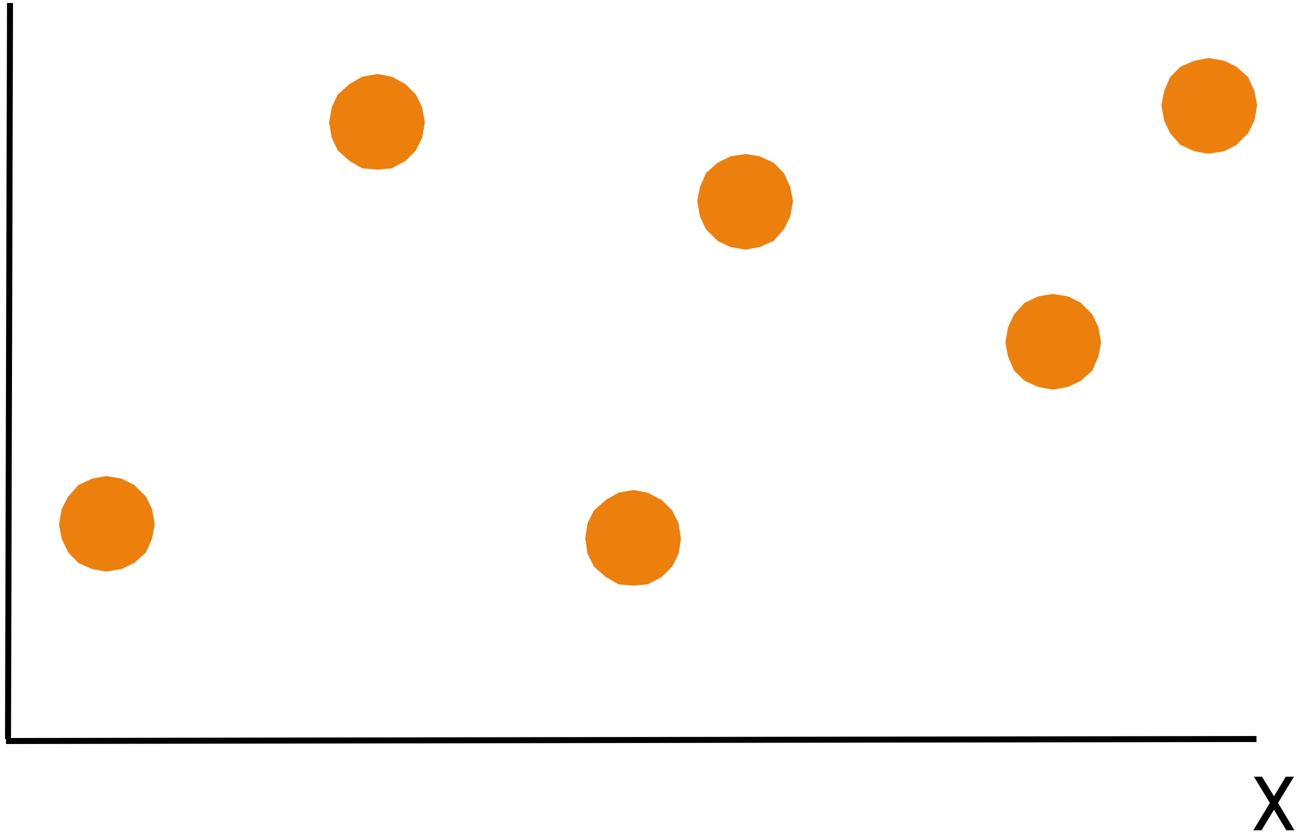
Random error

Intercept

Slope (regression) coefficient

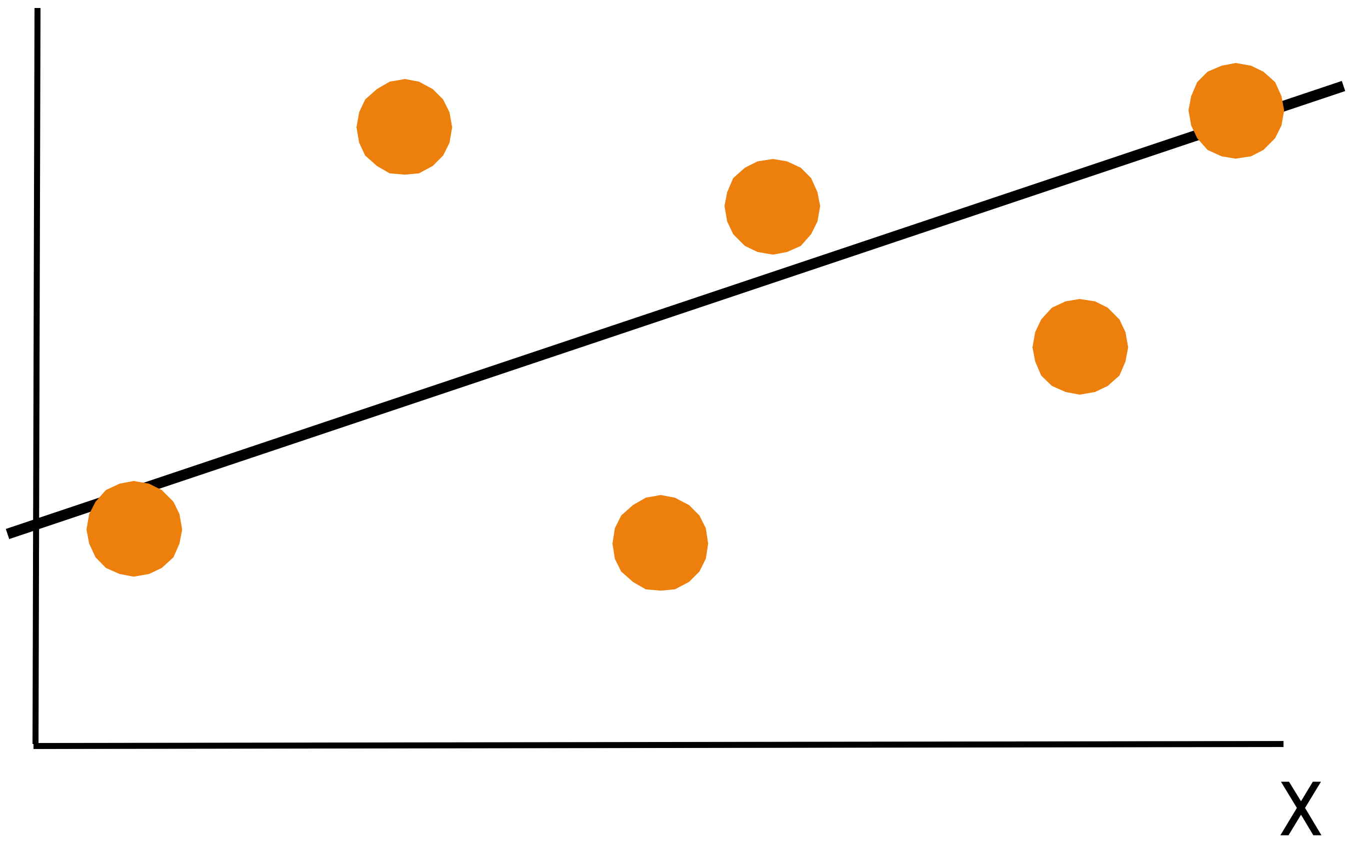
Simple linear regression

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



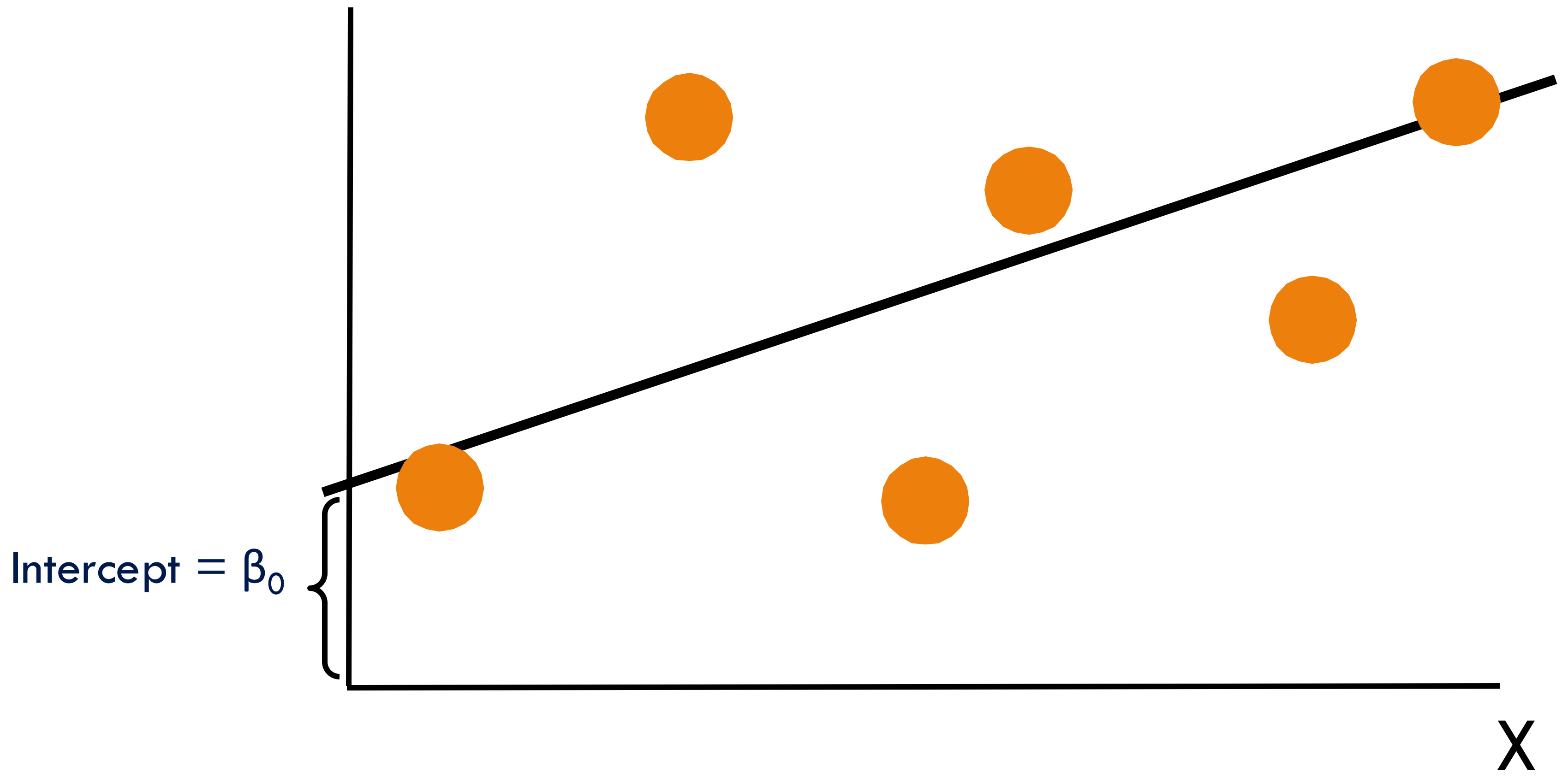
Simple linear regression

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



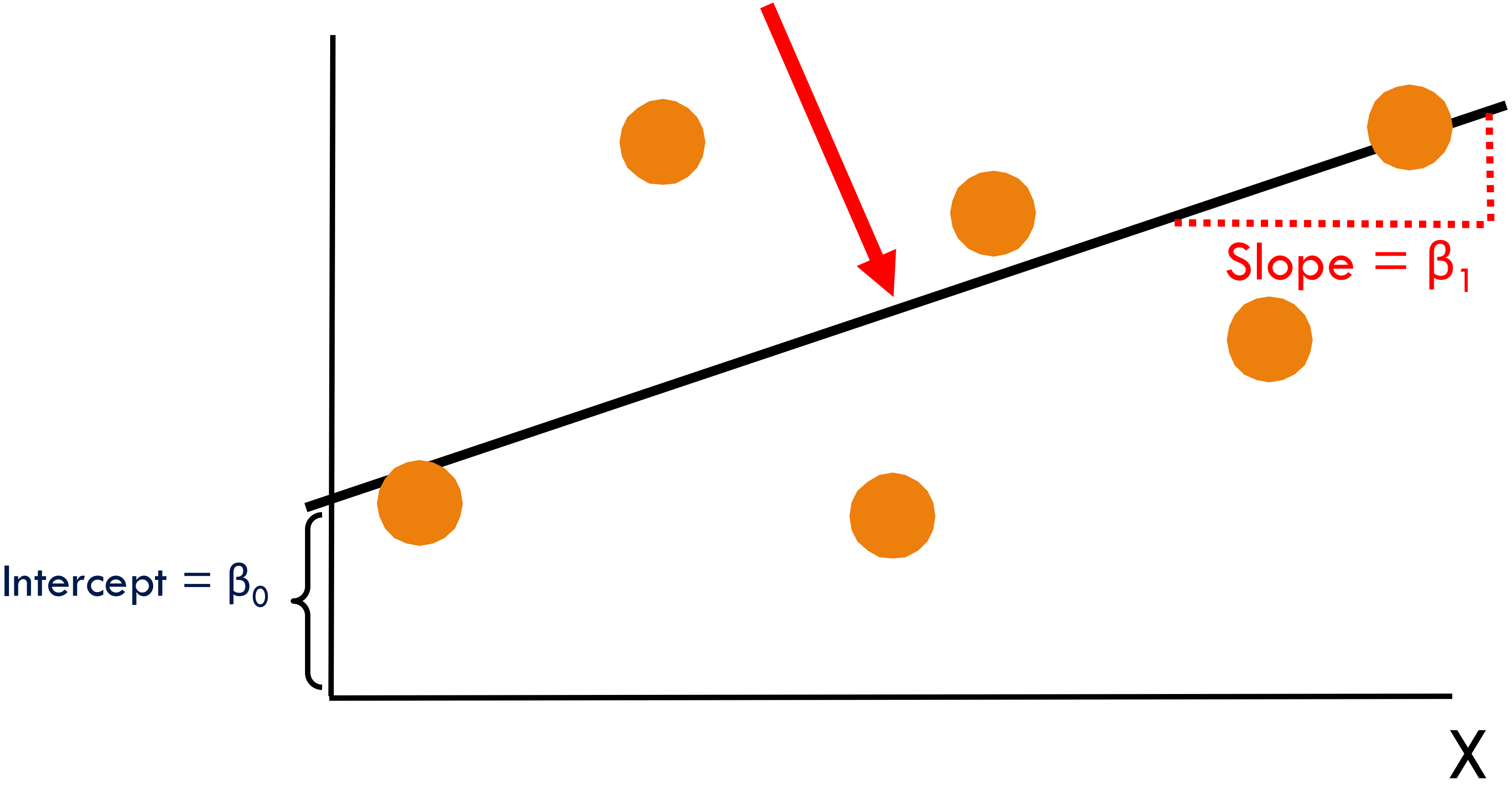
Simple linear regression

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



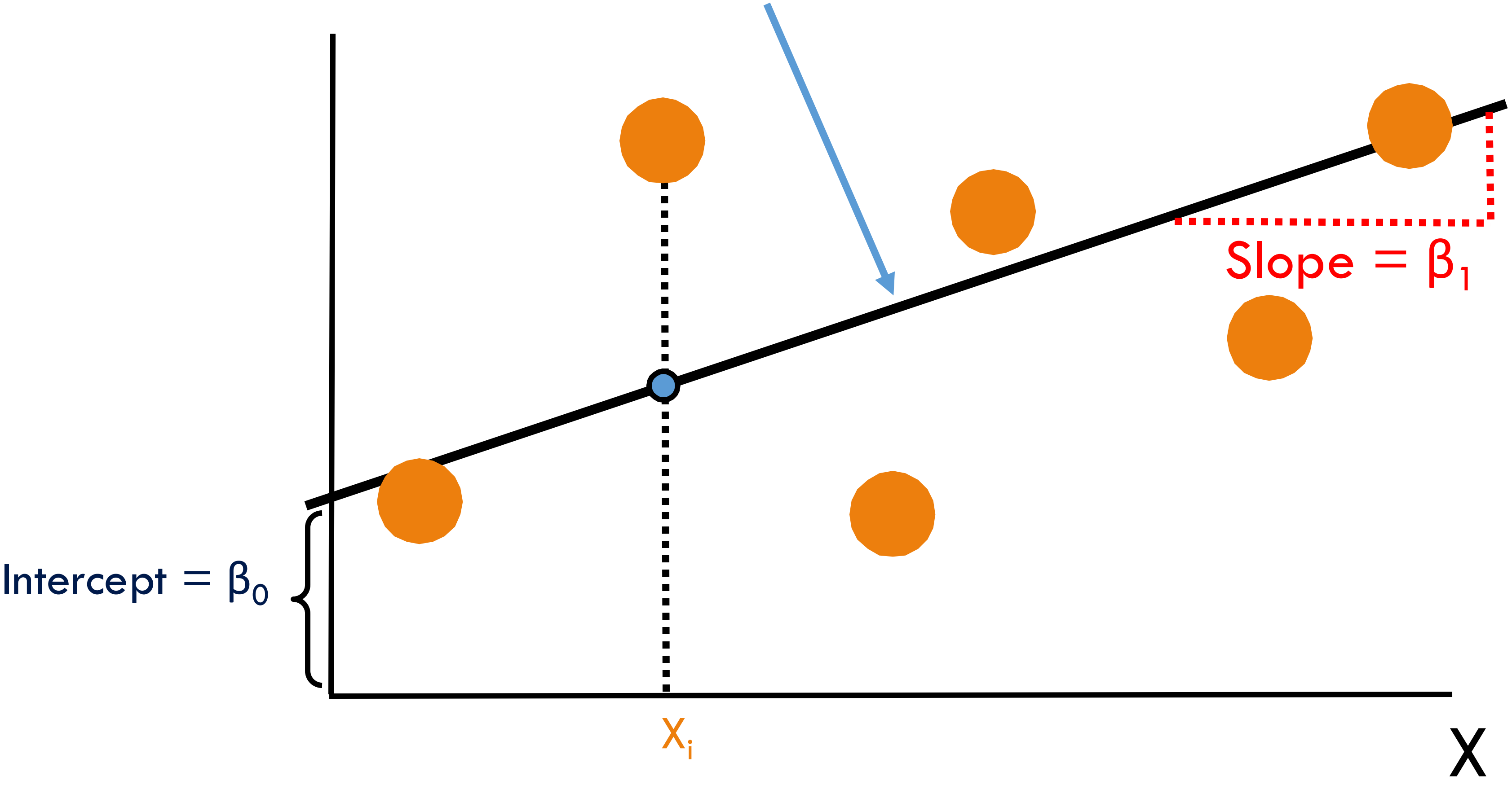
Simple linear regression

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



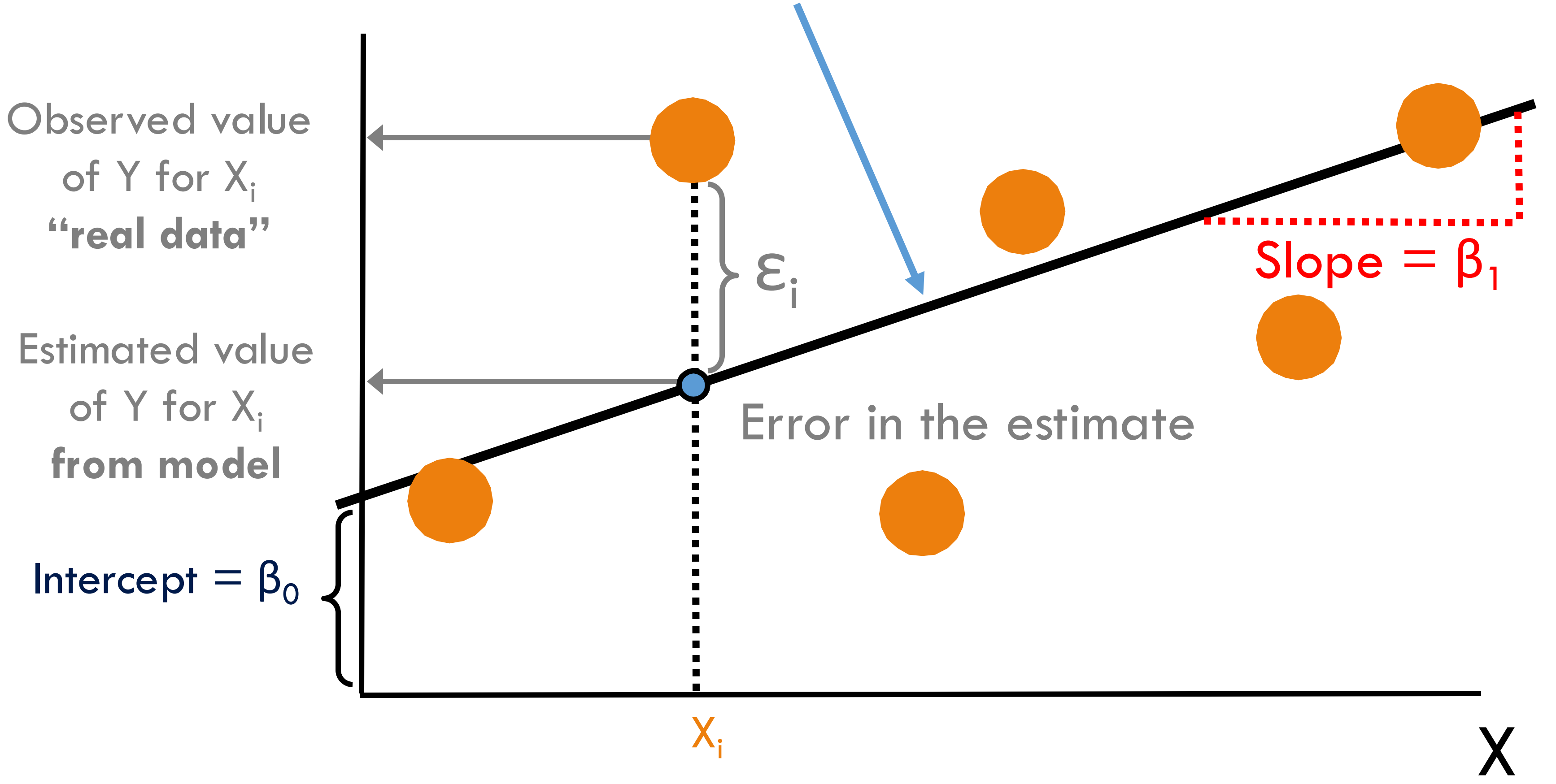
Simple linear regression

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



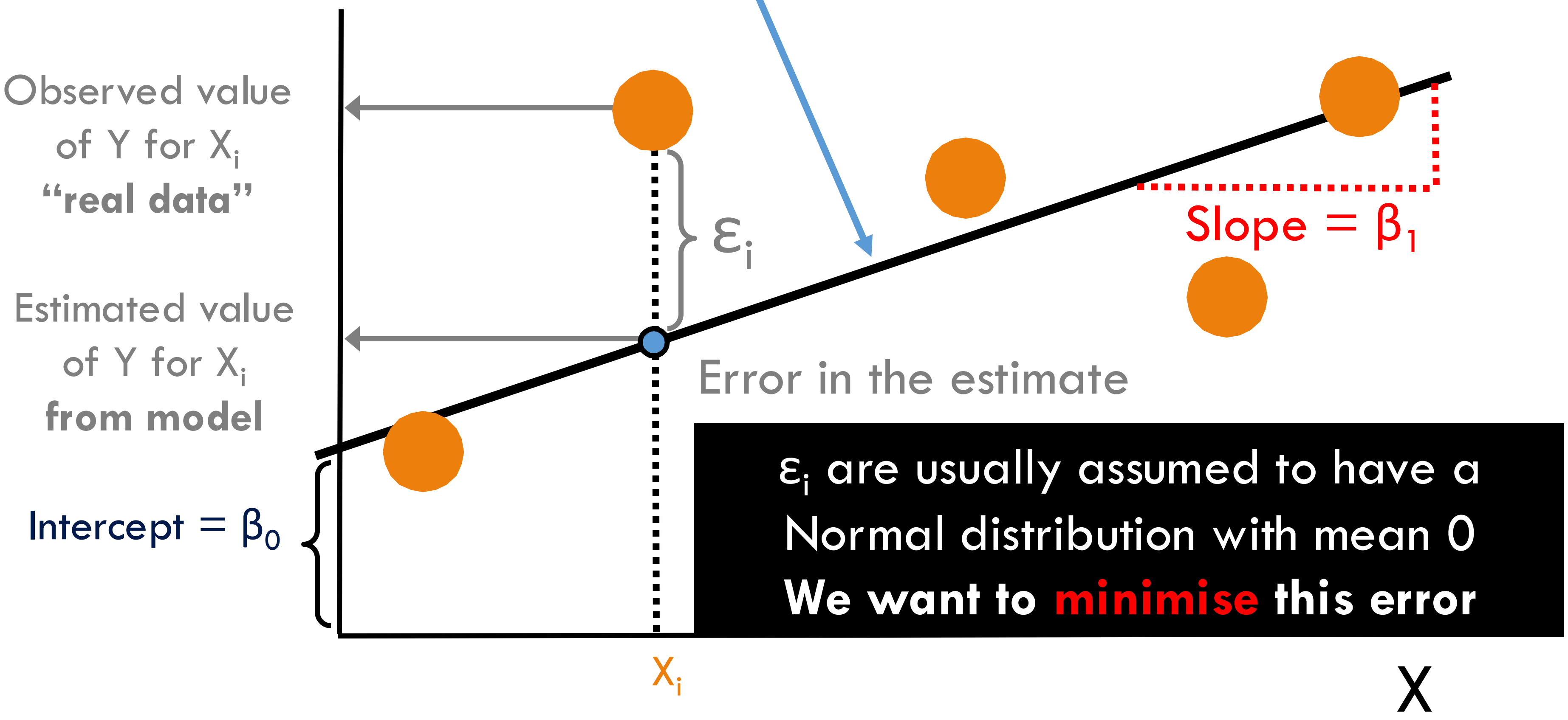
Simple linear regression

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



Simple linear regression

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

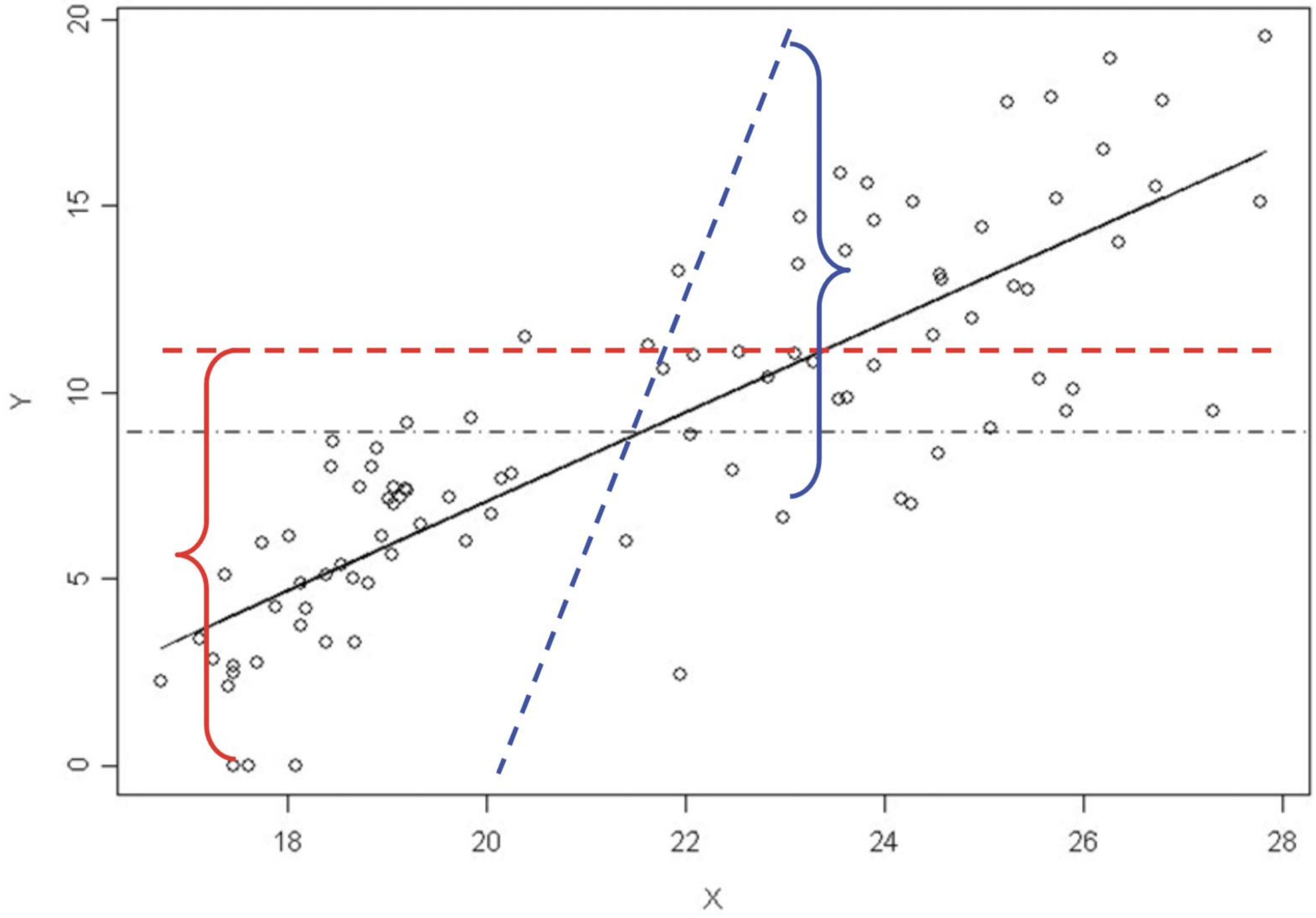


ε_i are usually assumed to have a Normal distribution with mean 0

We want to minimise this error

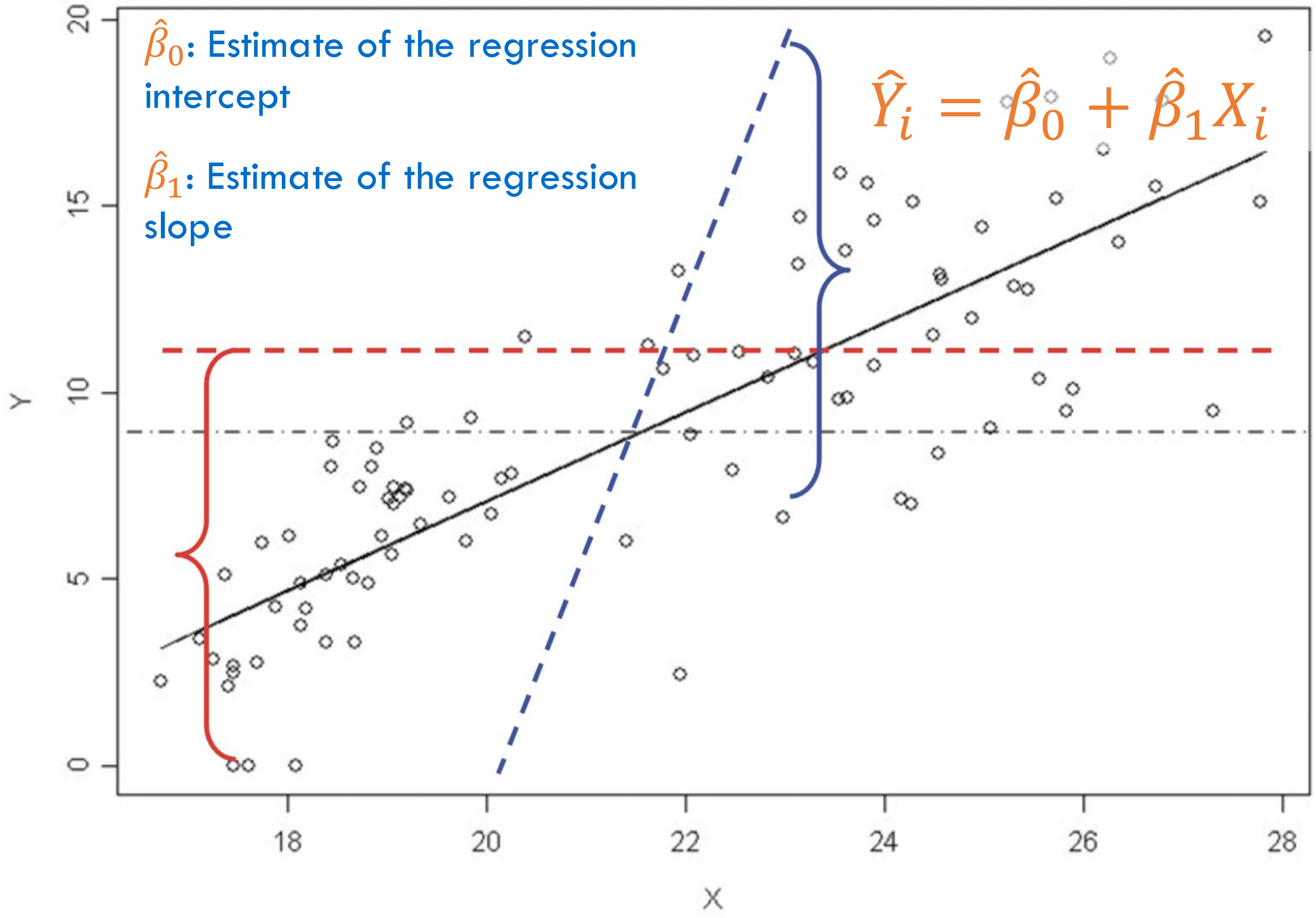
Minimising error

How do you know what the **best-fit line** is?



Minimising error

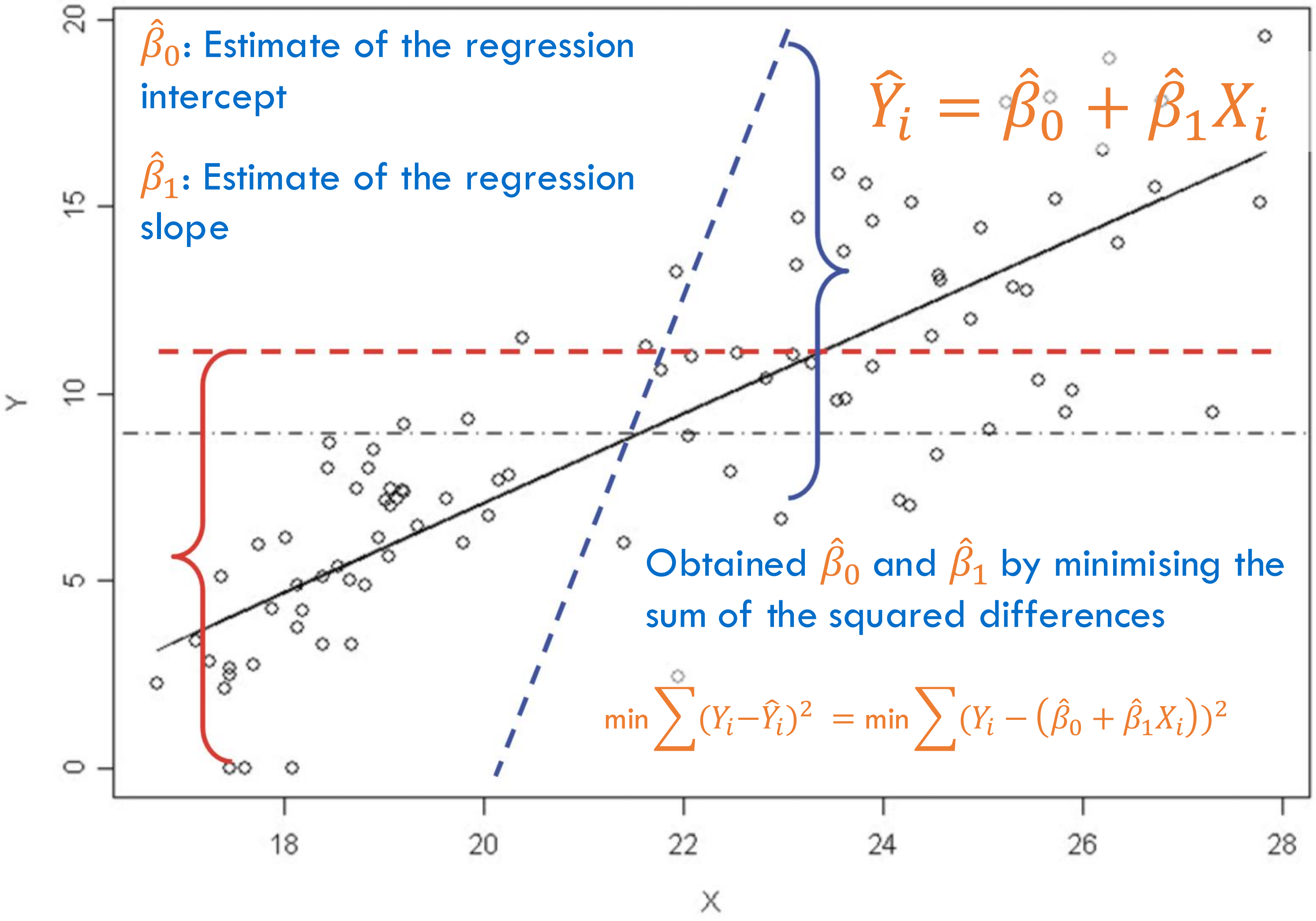
The equation provides an estimate of the population regression line.



Minimising error

Least Squares Method

The equation provides an estimate of the population regression line.



$\hat{\beta}_0$: Estimate of the regression intercept

$\hat{\beta}_1$: Estimate of the regression slope

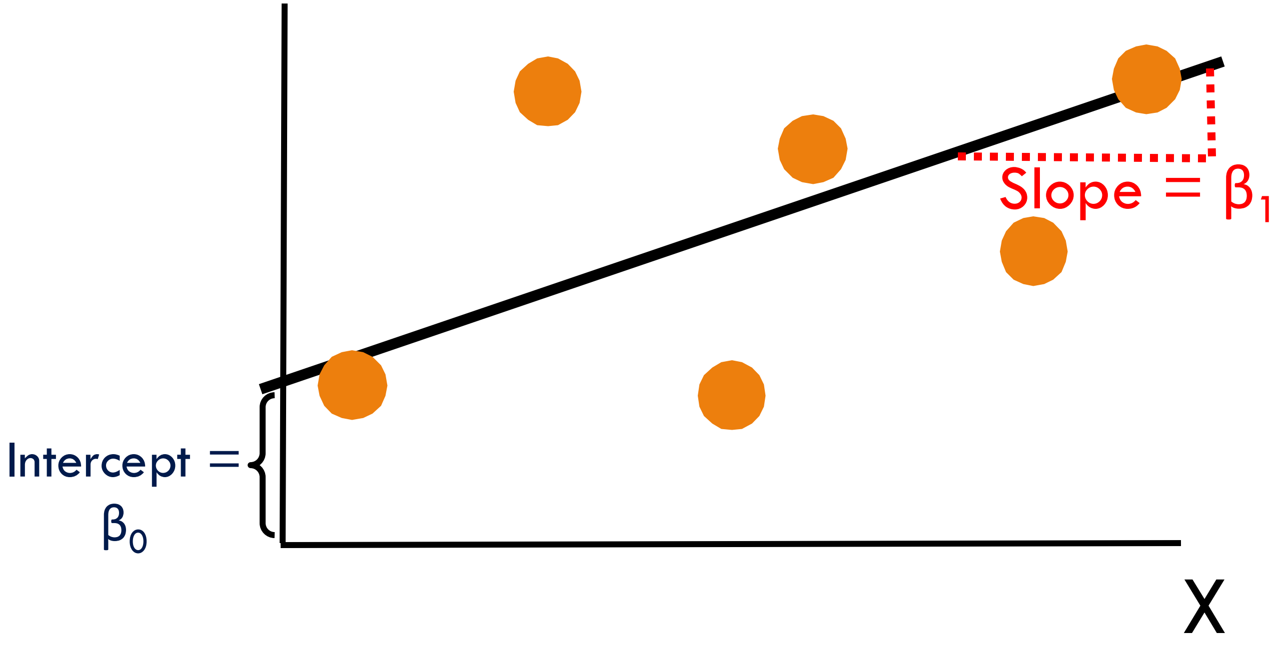
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Obtained $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimising the sum of the squared differences

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

Interpreting regression coefficients

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$



$\hat{\beta}_0$ least-square estimate of the regression intercept

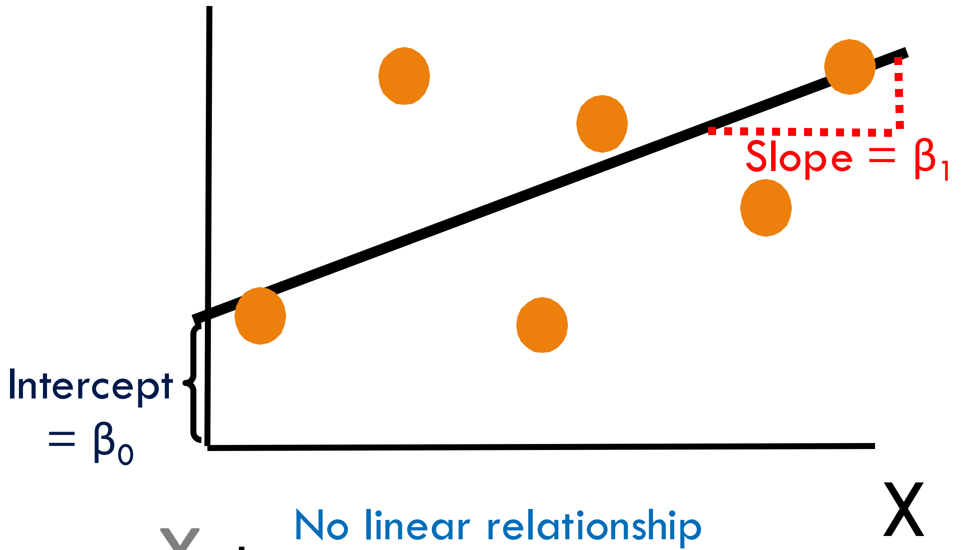
- estimated mean value of Y when x = 0

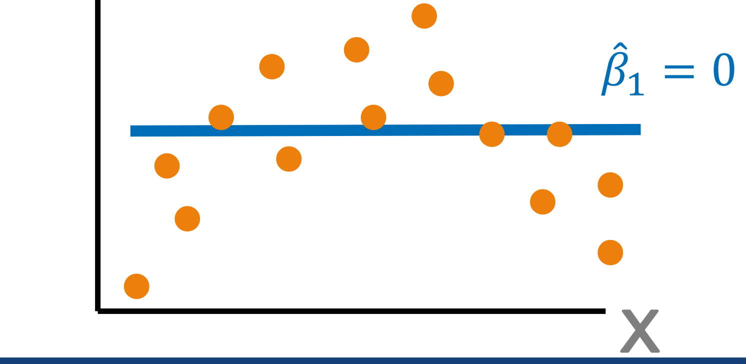
$\hat{\beta}_1$ least-square estimate of the regression slope

- estimated change in y when x changes by one unit

Inference about the slope

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$





Hypothesis tests

or confidence intervals

- To test the significance or "contribution" of an independent variable (*X*) to the dependent variable (*Y*)

- Is there a linear relationship between X and Y?

We are testing for zero slope:

If $\hat{\beta}_1 = 0$: then the X does not influence the value of Y

Inference about the slope

Hypothesis tests

or confidence intervals

- To test the significance or "contribution" of an independent variable (*X*) to the dependent variable (*Y*)

- Is there a linear relationship between X and Y?

We are testing for zero slope:

If $\hat{\beta}_1 = 0$: then the X does not influence the value of Y

Inference about the slope

T-test

Null hypothesis

H$_0$: $\hat{\beta}_1 = 0$

no linear relationship

Alternative hypothesis

H$_1$: $\hat{\beta}_1 \neq 0$

linear relationship may exist

Hypothesis tests

or confidence intervals

- To test the significance or "contribution" of an independent variable (*X*) to the dependent variable (*Y*)

- Is there a linear relationship between X and Y?

We are testing for zero slope:

If $\hat{\beta}_1 = 0$: then the X does not influence the value of Y

Inference about the slope

T-test

Null hypothesis

H_0: $\hat{\beta}_1 = 0$

no linear relationship

Alternative hypothesis

H_1: $\hat{\beta}_1 \neq 0$

linear relationship may exist

Test

$$t = \frac{(\hat{\beta}_1 - 0)}{SE\,(\hat{\beta}_1)}$$

with t distribution of d.f. = n - 2

Hypothesis tests

or confidence intervals

- To test the significance or "contribution" of an independent variable (*X*) to the dependent variable (*Y*)

- Is there a linear relationship between X and Y?

We are testing for zero slope:

If $\hat{\beta}_1 = 0$: then the X does not influence the value of Y

Simple linear regression

An example: factors associated with cardiovascular risk

A community-based cross-sectional on CVD

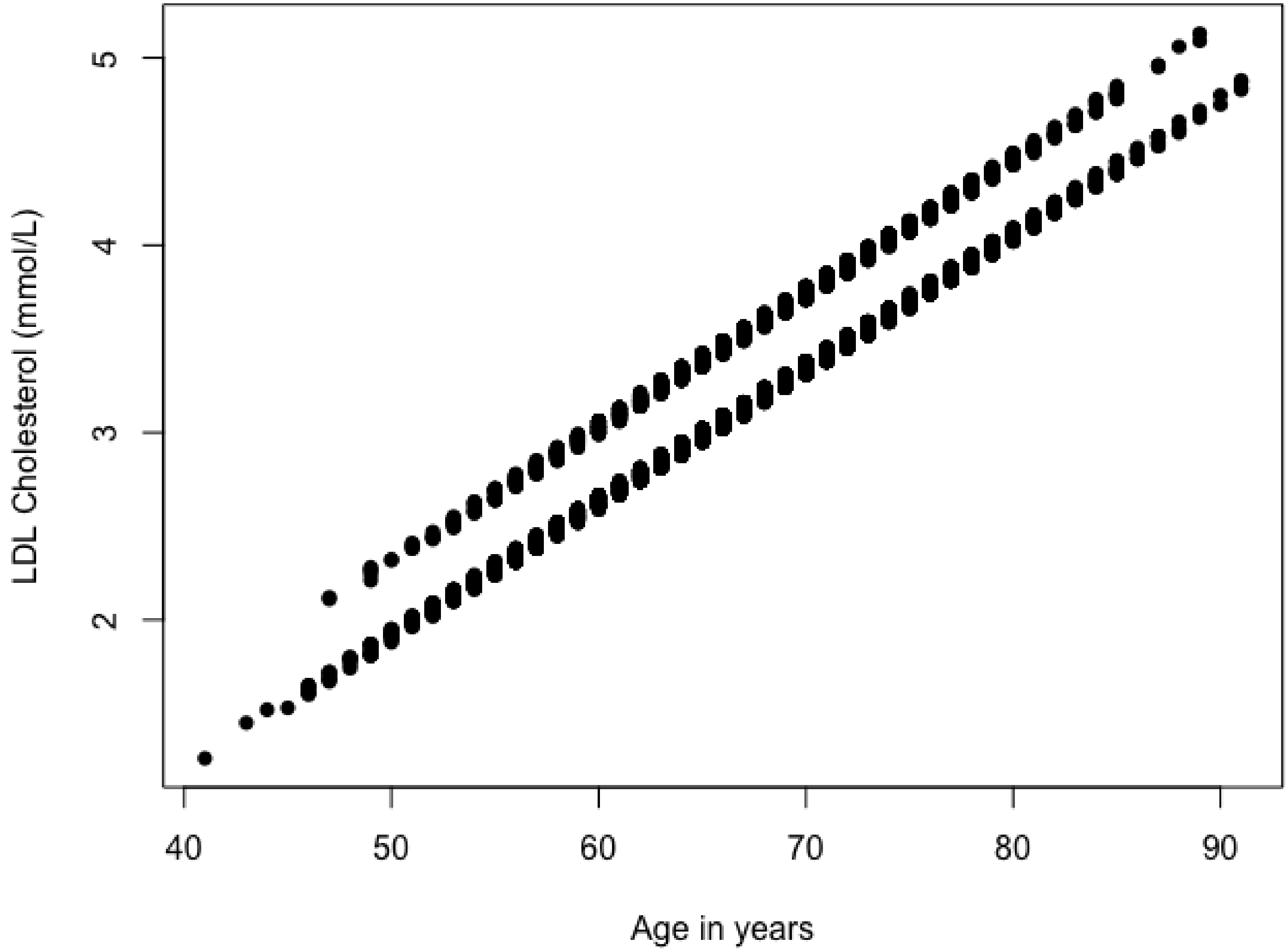
A community-based cross-sectional survey of 10000 respondents aged ≥40 years in Singapore was conducted to examine the factors associated with cardiovascular risk

Disproportionate stratified sampling of ethnic groups was undertaken (only individuals of Chinese, Malay or Indian ethnicity were included)

Data collected

1. Age - in completed years (in integers)
2. Gender (Male/Female)
3. BMI – measured height and weight (calculated to one decimal place)
4. Ethnicity (Chinese/Malay/Indian)
5. Smoking status – self-reported by participants (Daily smoker/Occasional smoker/Ex-smoker/Never smoker)
6. LDL cholesterol – measured from fasting blood samples obtained from participants (available up to to two decimal places)
7. Presence of cardiovascular disease– self-reported by participants as having being diagnosed by a physician (Yes/No)

LDL Cholesterol ~ Age



LDL Cholesterol ~ Age

```
> m_ldl_age = lm(chp$ldl ~ chp$age)
> summary(m_ldl_age)

Call:
lm(formula = chp$ldl ~ chp$age)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.13569 -0.10771 -0.08370 -0.05917  0.34480 

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)    
(Intercept) -1.5666710  0.0167344  -93.62 <0.0000000000000002 ***
chp$age      0.0714981  0.0002448  292.05 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1714 on 9998 degrees of freedom
Multiple R-squared:  0.8951,	Adjusted R-squared:  0.8951 
F-statistic: 8.529e+04 on 1 and 9998 DF,  p-value: < 0.00000000000000022
```

LDL Cholesterol ~ Age

```
> m_ldl_age = lm(chp$ldl ~ chp$age)
> summary(m_ldl_age)

Call:
lm(formula = chp$ldl ~ chp$age)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.13569 -0.10771 -0.08370 -0.05917  0.34480 

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)    
(Intercept) -1.5666710  0.0167344  -93.62 <0.0000000000000002 ***
chp$age      0.0714981  0.0002448  292.05 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1714 on 9998 degrees of freedom
Multiple R-squared:  0.8951,	Adjusted R-squared:  0.8951 
F-statistic: 8.529e+04 on 1 and 9998 DF,  p-value: < 0.00000000000000022
```

estimate the relationship between the independent and dependent variables

p-value used in testing the null hypothesis

This section summarises the overall model fit

R-Squared: proportion of variance in dependent variable which can be predicted from the independent variables

LDL Cholesterol ~ Age

```
> m_ldl_age = lm(chp$ldl ~ chp$age)
> summary(m_ldl_age)

Call:
lm(formula = chp$ldl ~ chp$age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.13569 -0.10771 -0.08370 -0.05917  0.34480

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)
(Intercept) -1.5666710  0.0167344  -93.62 <0.0000000000000002 ***
chp$age      0.0714981  0.0002448  292.05 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1714 on 9998 degrees of freedom
Multiple R-squared:  0.8951,	Adjusted R-squared:  0.8951
F-statistic: 8.529e+04 on 1 and 9998 DF,  p-value: < 0.00000000000000022
```

There is a 0.0714 mmol/L increase in mean LDL cholesterol levels for every 1-year increase in age ($p < 0.001$).

estimate the relationship between the independent and dependent

p-value used in testing the null hypothesis

This section summarises the overall model fit

R-Squared: proportion of variance in dependent variable which can be predicted from the independent variables

LDL Cholesterol ~ Age

NUS National University of Singapore | Saw Swee Hock School of Public Health

```
> m_ldl_age = lm(chp$ldl ~ chp$age)
> summary(m_ldl_age)

Call:
lm(formula = chp$ldl ~ chp$age)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.13569 -0.10771 -0.08370 -0.05917  0.34480 

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)    
(Intercept) -1.5666710  0.0167344  -93.62 <0.0000000000000002 ***
chp$age      0.0714981  0.0002448  292.05 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1714 on 9998 degrees of freedom
Multiple R-squared:  0.8951,	Adjusted R-squared:  0.8951 
F-statistic: 8.529e+04 on 1 and 9998 DF,  p-value: < 0.00000000000000022
```

There is a 0.0714 mmol/L increase in mean LDL cholesterol levels for every 1-year increase in age ($p < 0.001$).

estimate the relationship between the independent and dependent ...s

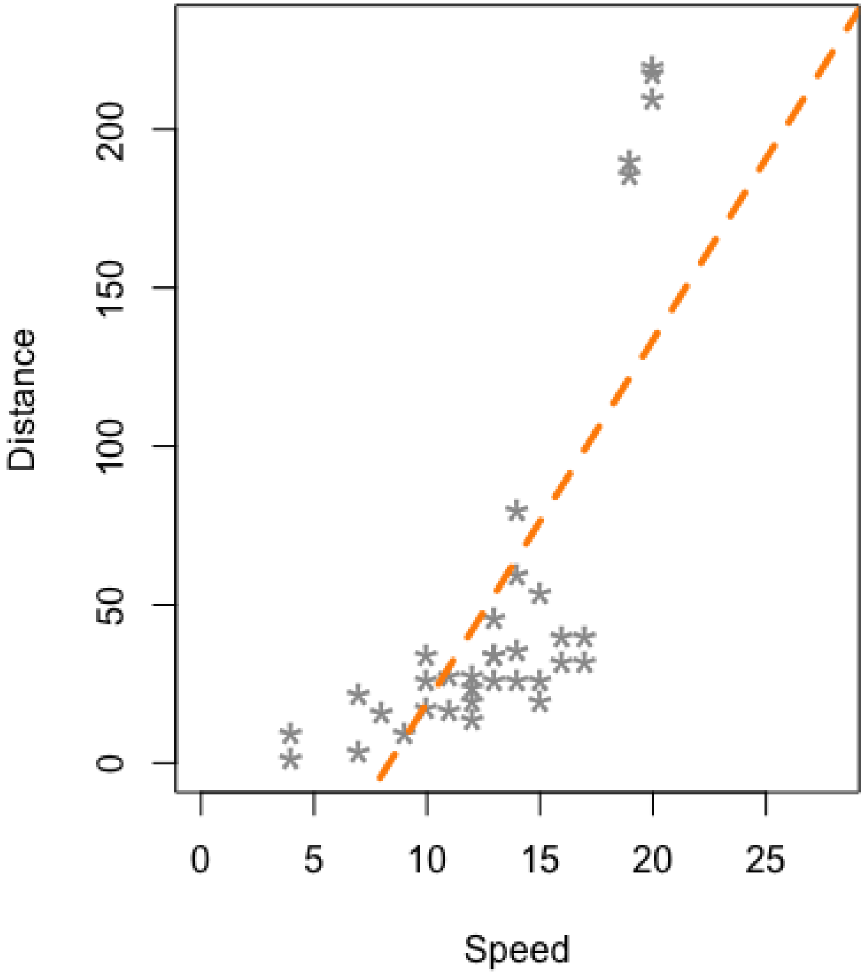
p-value used in testing the null hypothesis

This section summarises the overall model fit

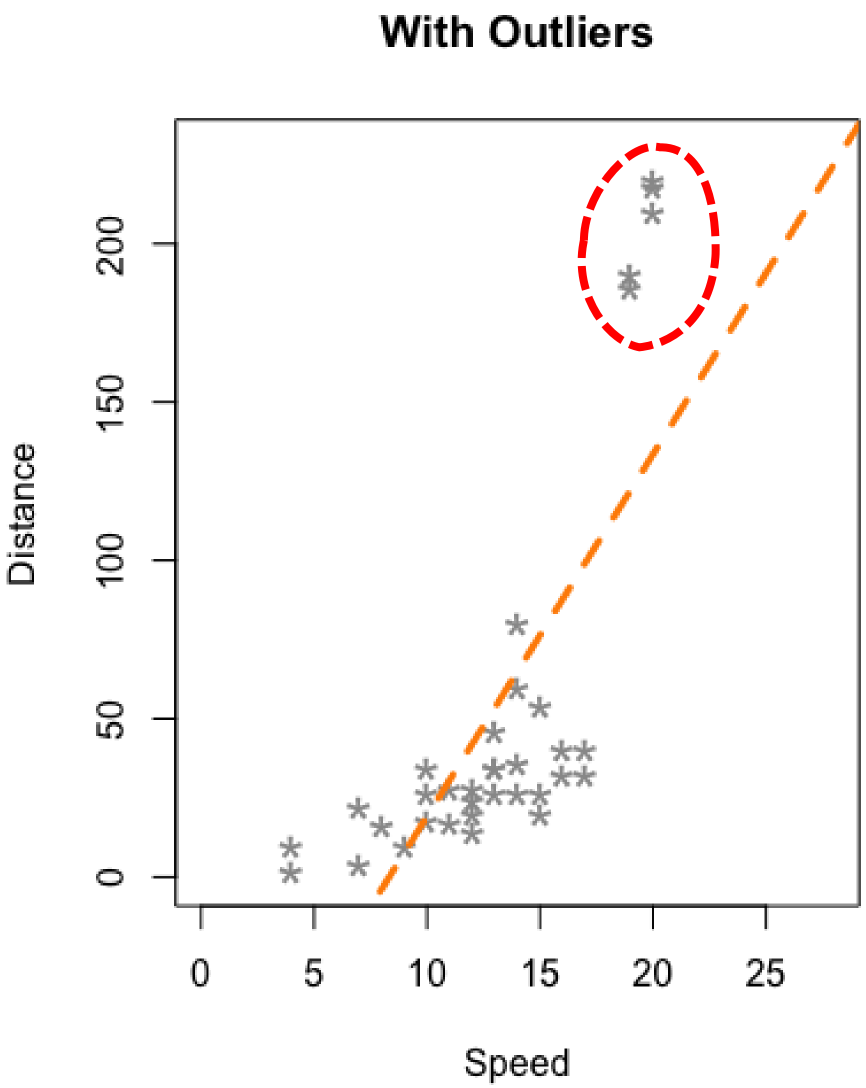
R-Squared: proportion of variance in dependent variable which can be predicted from the independent variables

R-Squared

The proportion of the variability in the y that is accounted for by the linear relationship with the x variable.

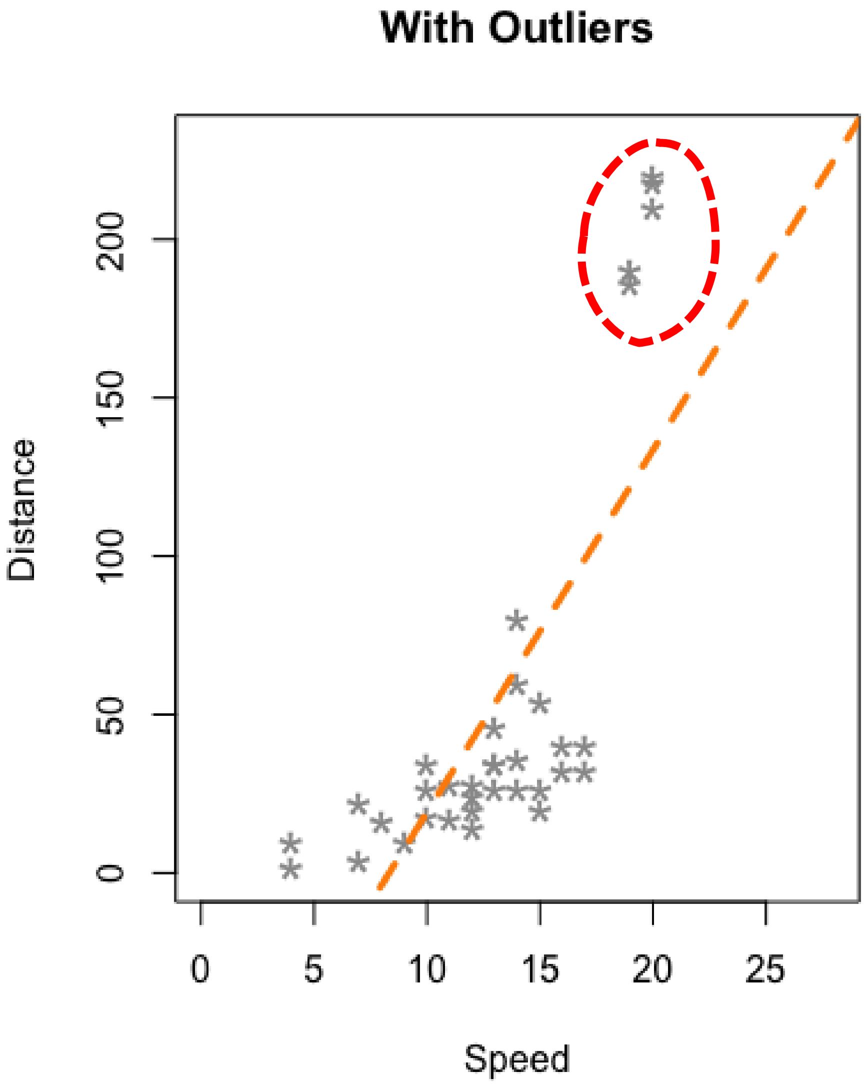

Distance
Speed
0
50
100
150
200
0
5
10
15
20
25

Outliers



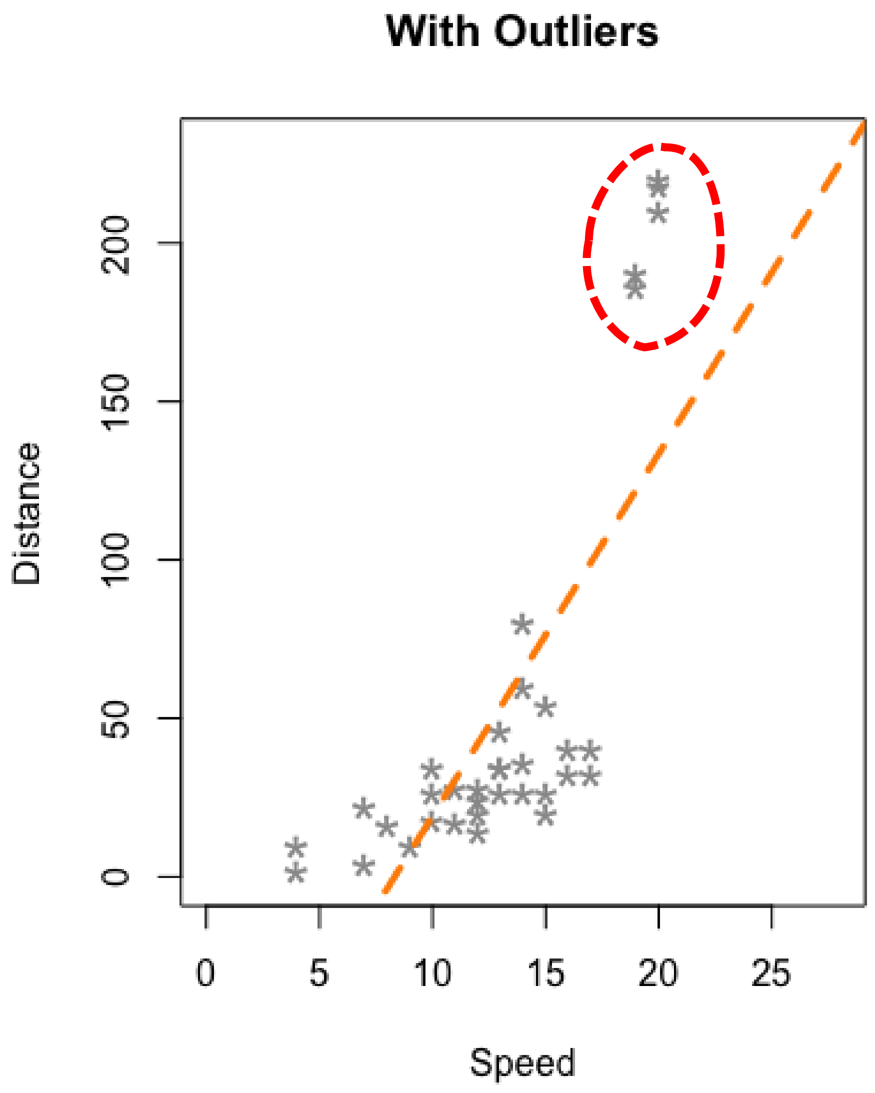
Turning Discovery into Healthier Communities

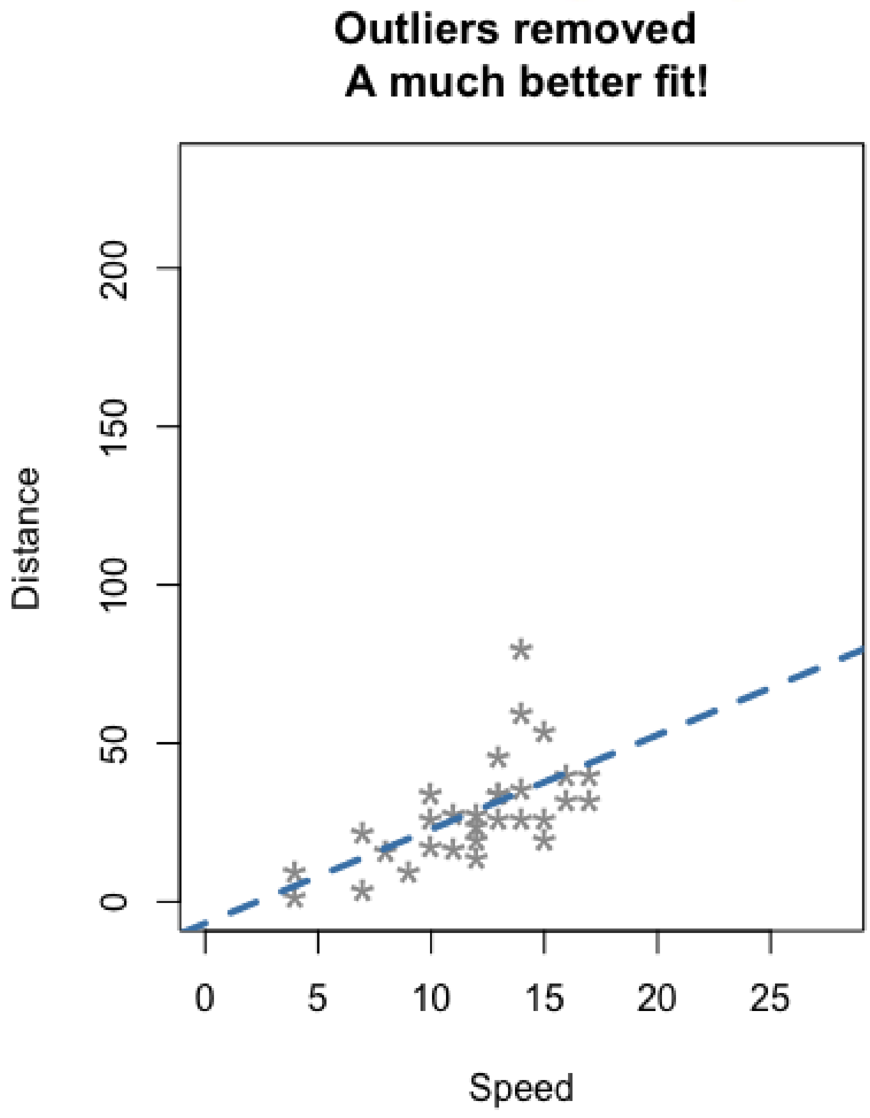
Outliers



Outliers

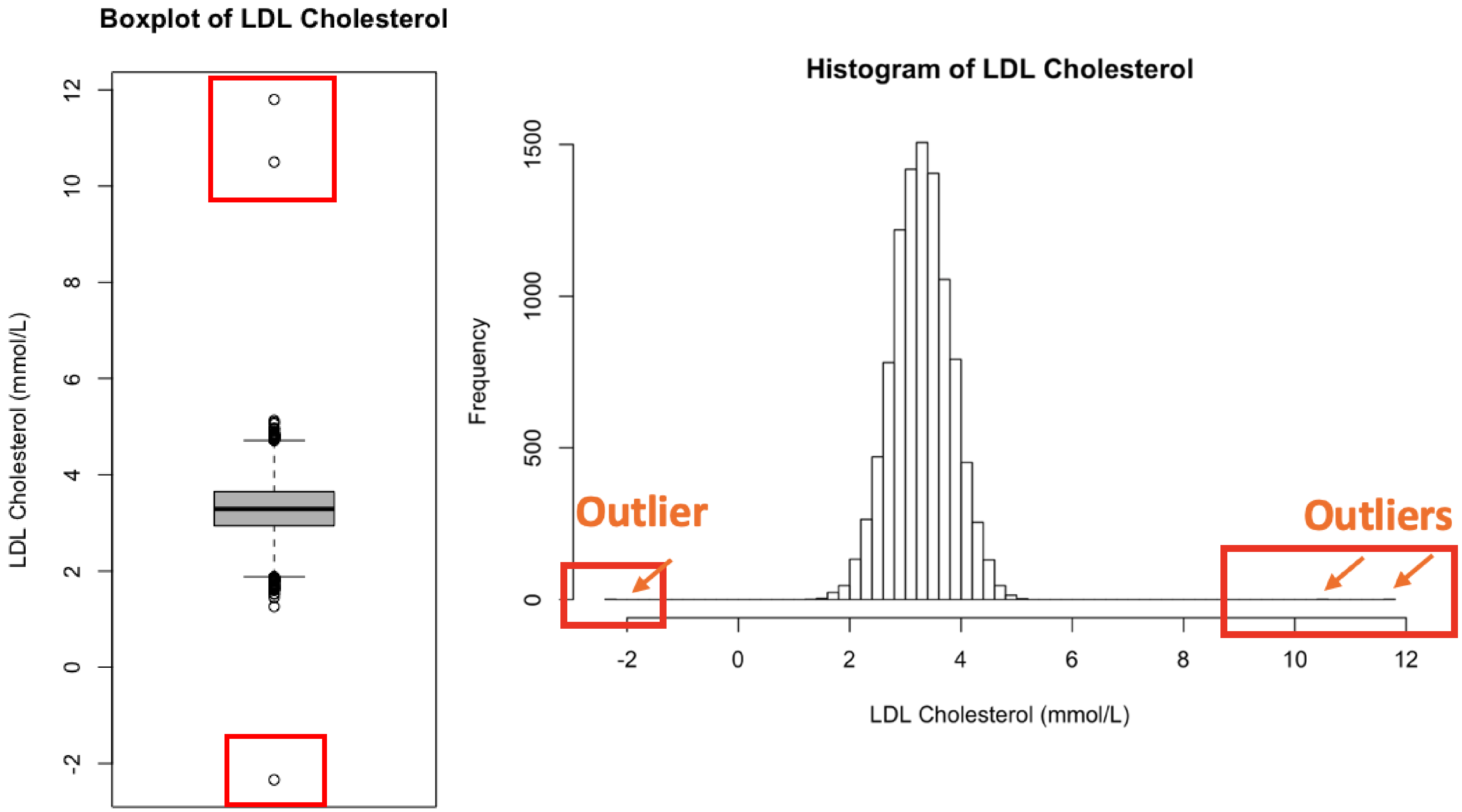
Saw Swee Hock School of Public Health





Outliers

Plot the **box plot** Plot the **histogram**



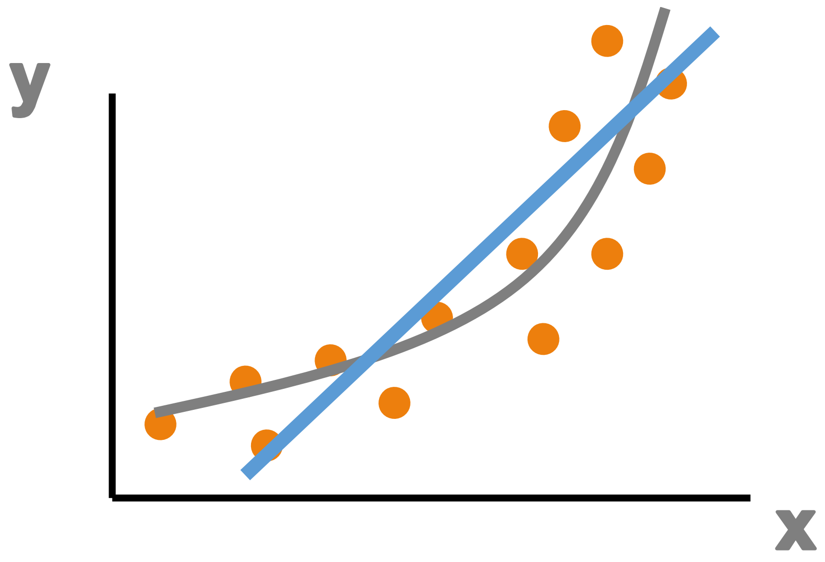
Linear regression diagnostics

Important to check!

Linearity

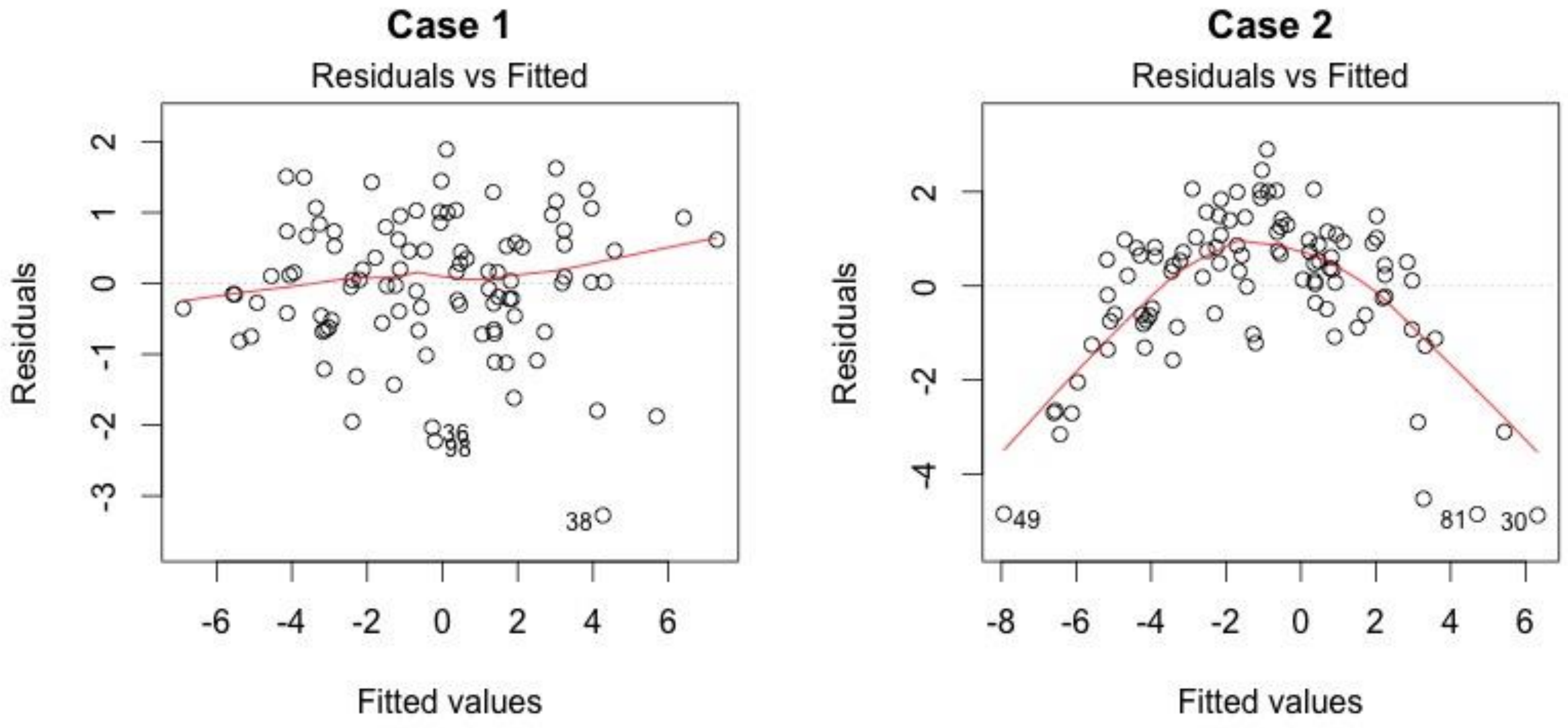
- A straight line may be an inadequate model
- Contamination from outliers from different populations (more on Friday)

- Resulting estimates misleading, biased
- Possible transformations or polynomial variables



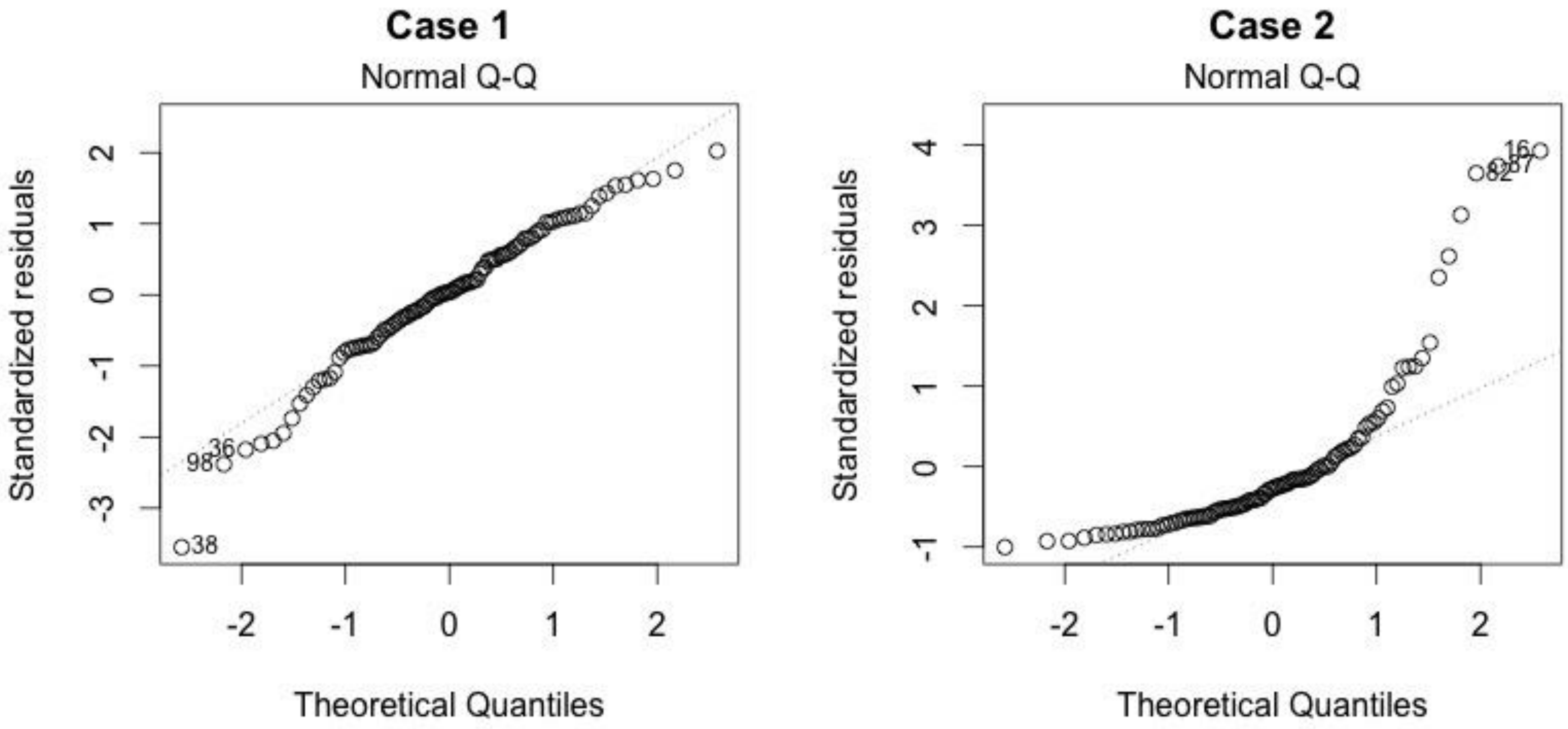
Residuals

Linear vs nonlinear relationship



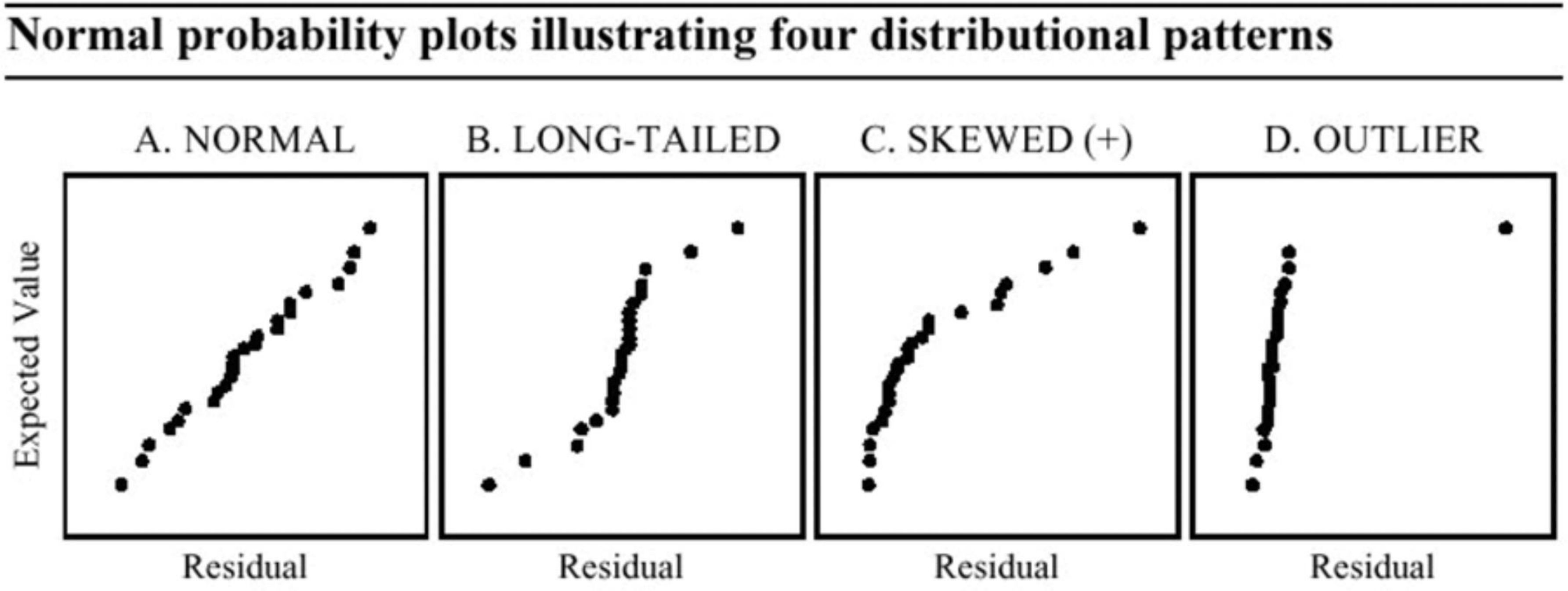
Residuals

Normality



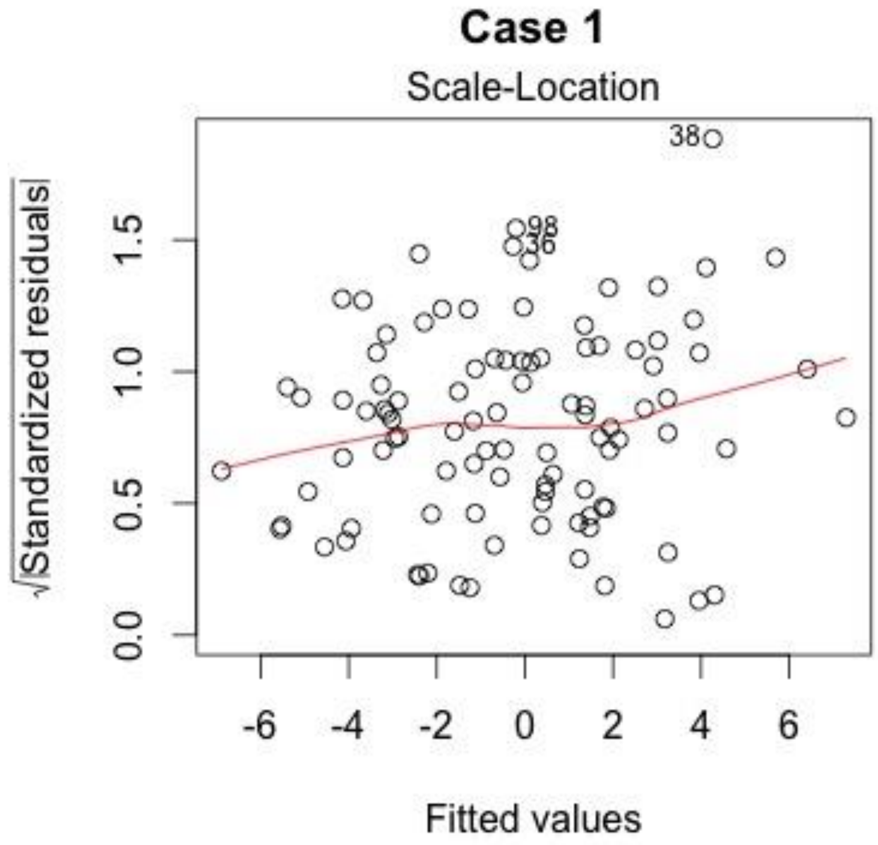
Residuals

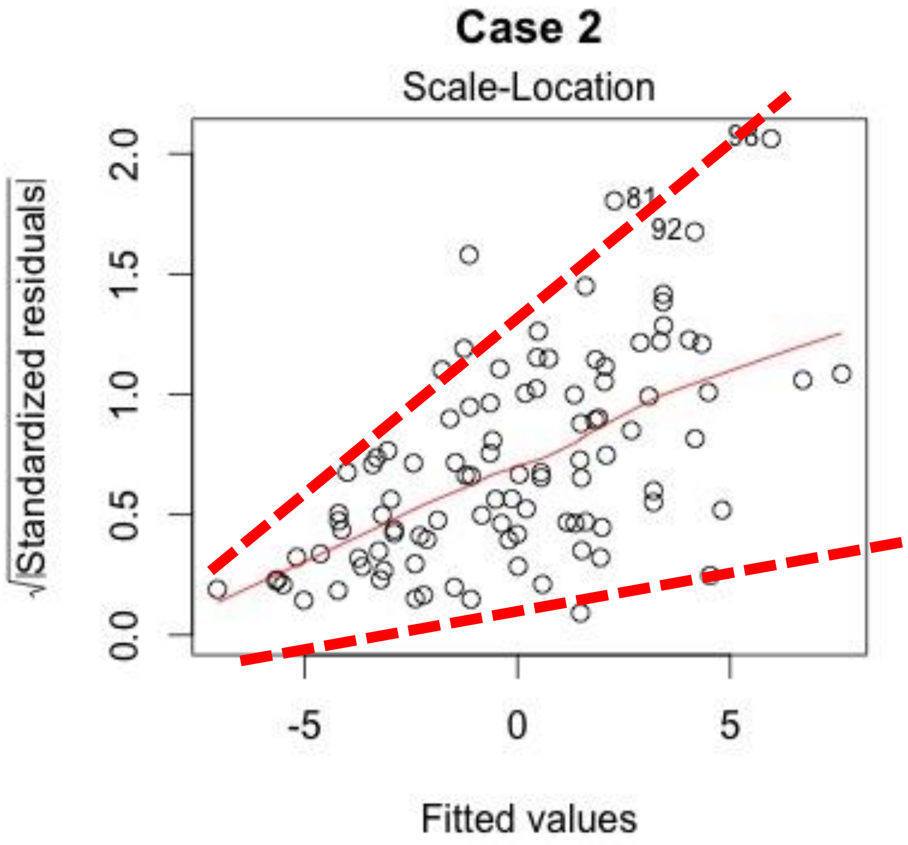
Normality



Residuals

Homoscedasticity (equal variance)





Do the assumptions matter?

Lack of normality of the residuals

- unlikely to be serious

Lack of constant variance of the residuals

- unlikely to be serious

Both will have some influence on the final p-value

Do the assumptions matter?

Lack of linearity

- more serious, and would suggest a transformation of y before fitting the regression equation on x

Lack of independence of the residuals

- may be serious if the data involve repeated measures of individuals

The presence of outliers is potentially very serious

Significant tests

Steps for conducting significance tests:

1. **State the null hypothesis** (H_0)
2. **State the alternate hypothesis** (H_α)
3. **Calculate test statistic** (parameter of interest divided by standard error)
4. **Look up and interpret p-value:**
   - Remember that statistical significance is not equivalent to medical or biological significance!
   - Interpret a p-value in terms of the level of evidence (α) against the null hypothesis.

NUS National University of Singapore | Saw Swee Hock School of Public Health

TM-CM02

Biostatistics for Public Health

- quizzes
- 2 assignments

1

Categorical data analysis

Data prep, descriptive & inferential statistics

2

Hypothesis testing

Inferential statistics

3

Linear regression

Data prep, descriptive & inferential statistics, linear regression

4

Logistic regression

Data prep, descriptive & inferential statistics, logistics regression and others if we have time

Thank you