MORU Tropical Health Network

Basic R programing

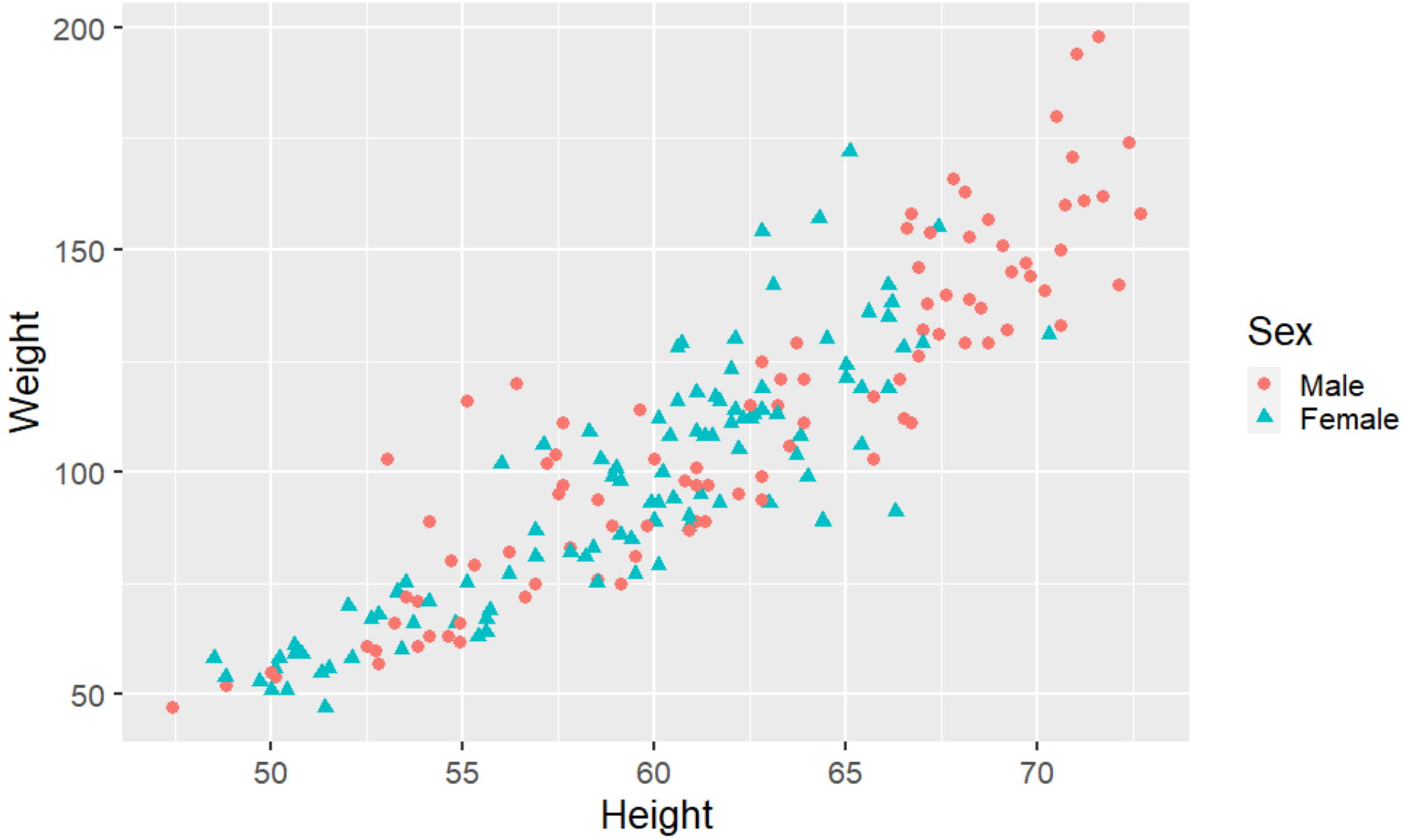
11-12 Jan 2021

Lecture 6 (13:00-14:30): Continuous data analysis (Linear modelling)

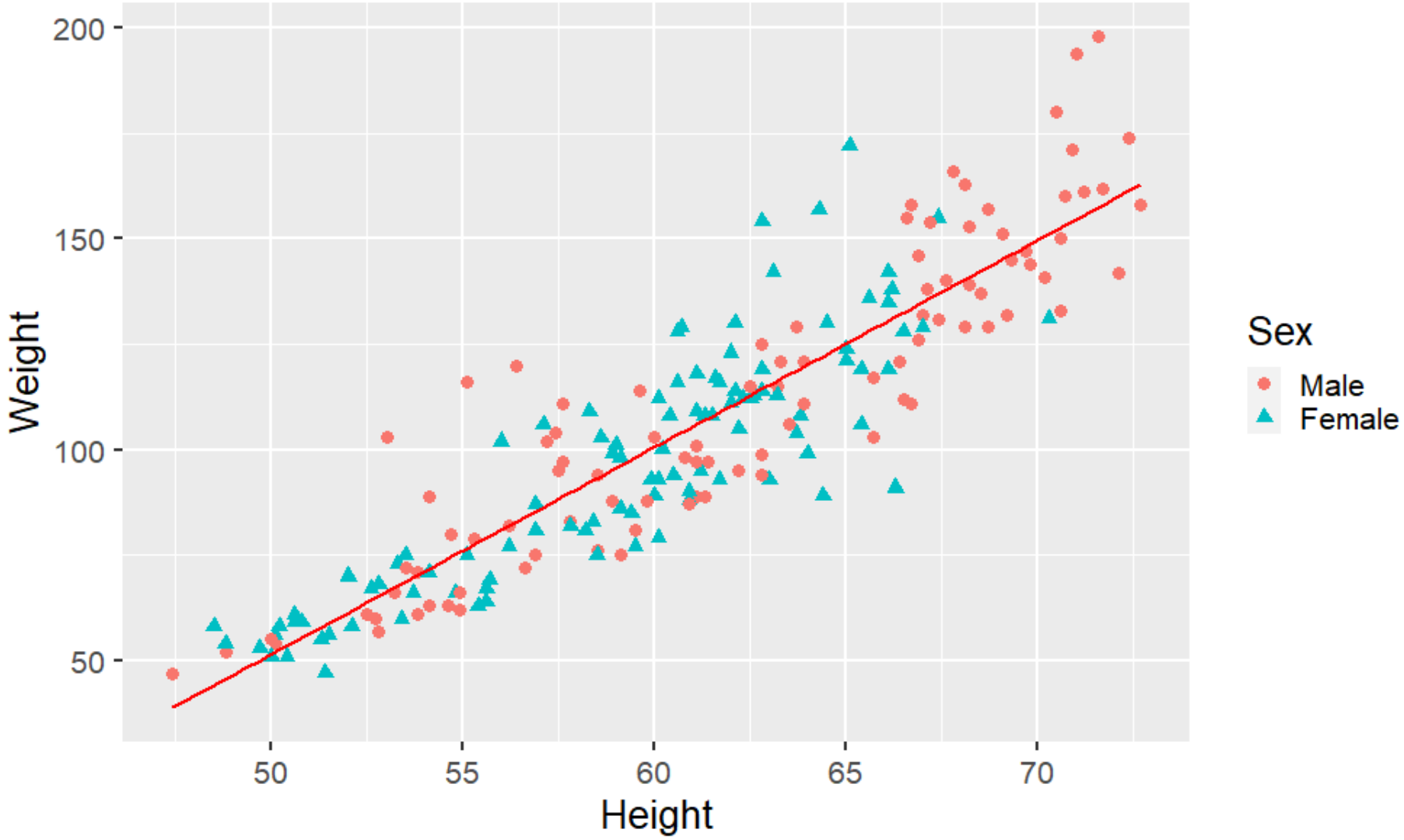
Dr. Palang Chotsiri

palang@tropmedres.ac

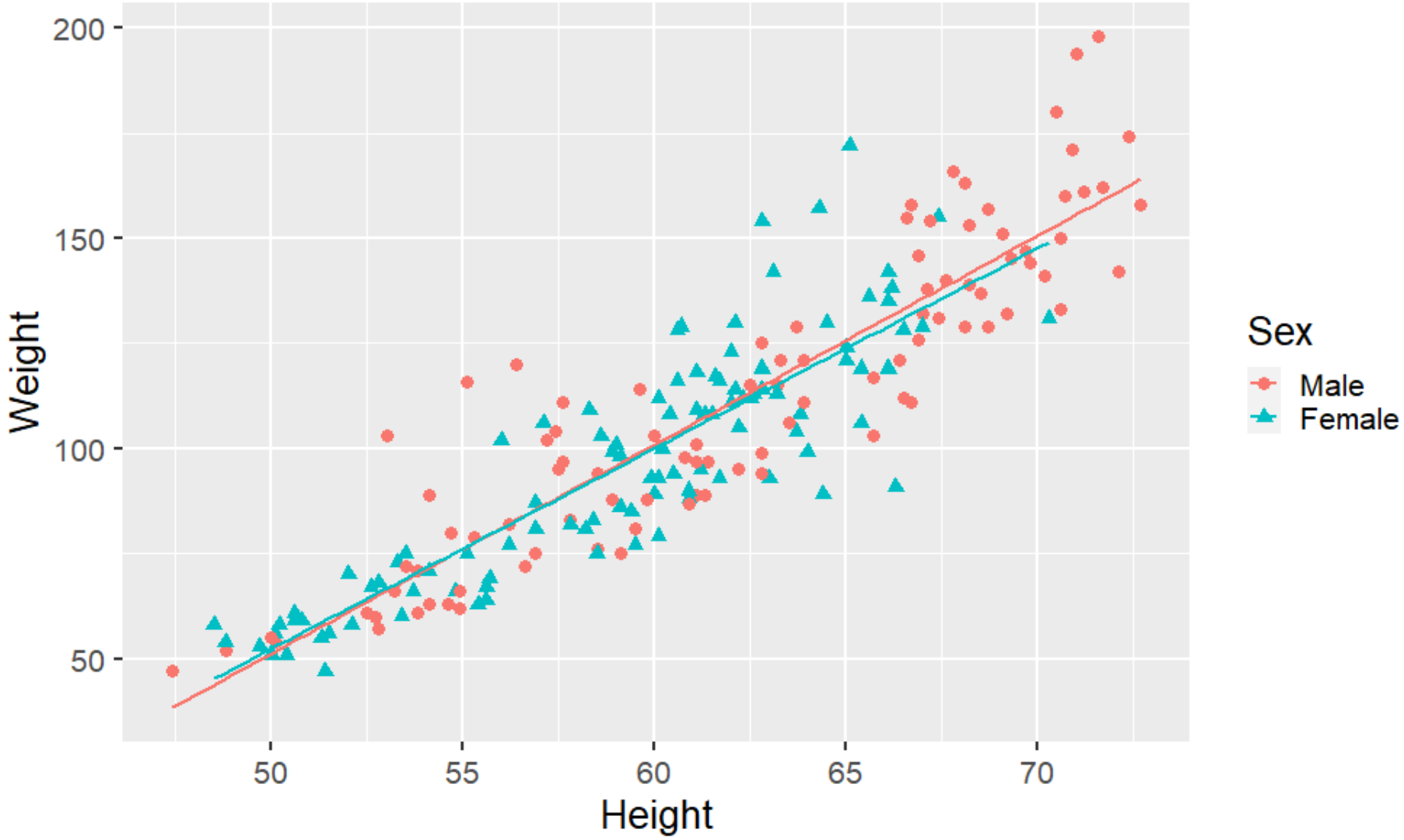
Height and Weight



Height and Weight



Height and Weight



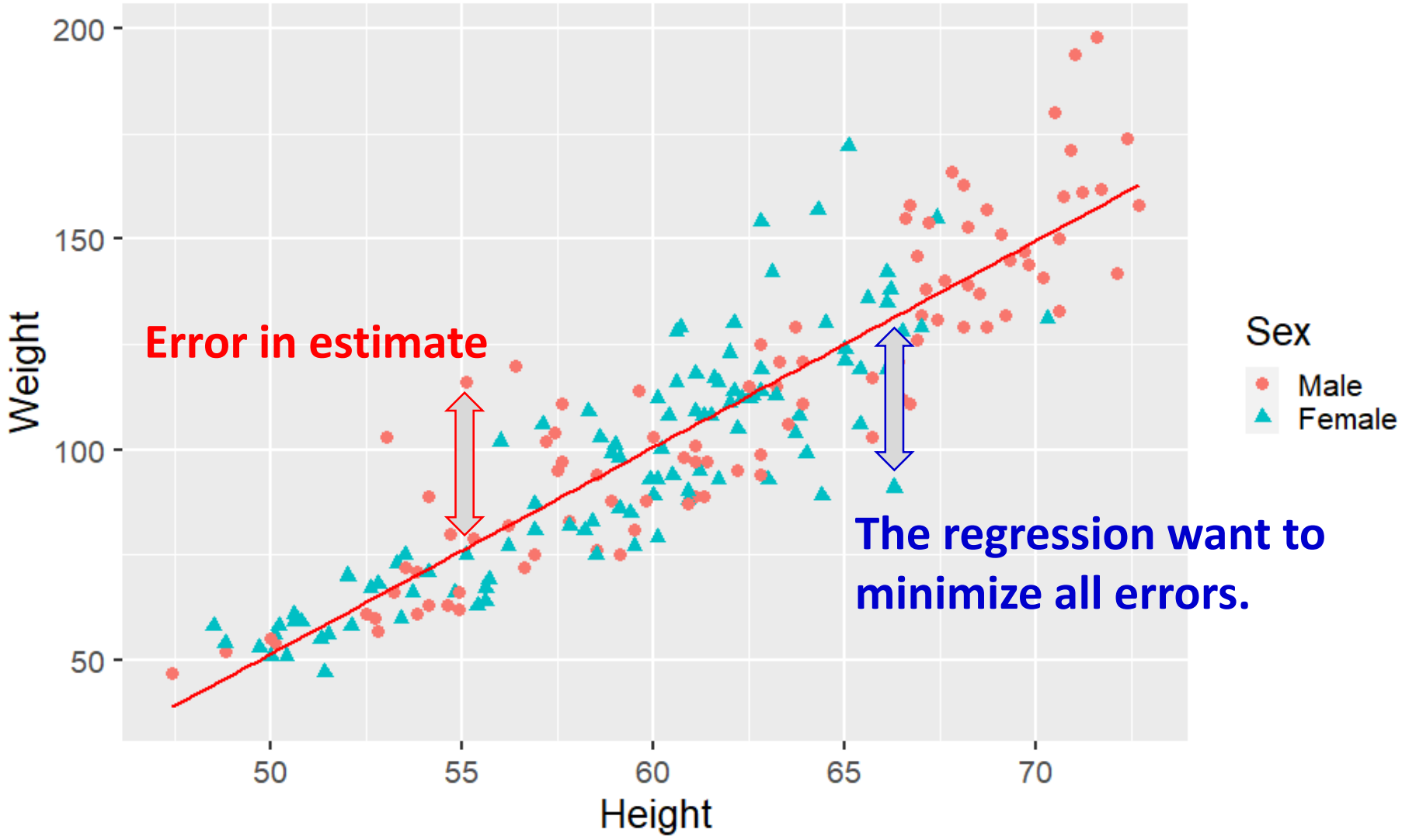
Simple linear regression

- Apparently linear relation, can we quantify this relation?
- Statistical modelling describing the relationship between height and weight with a straight line equation:

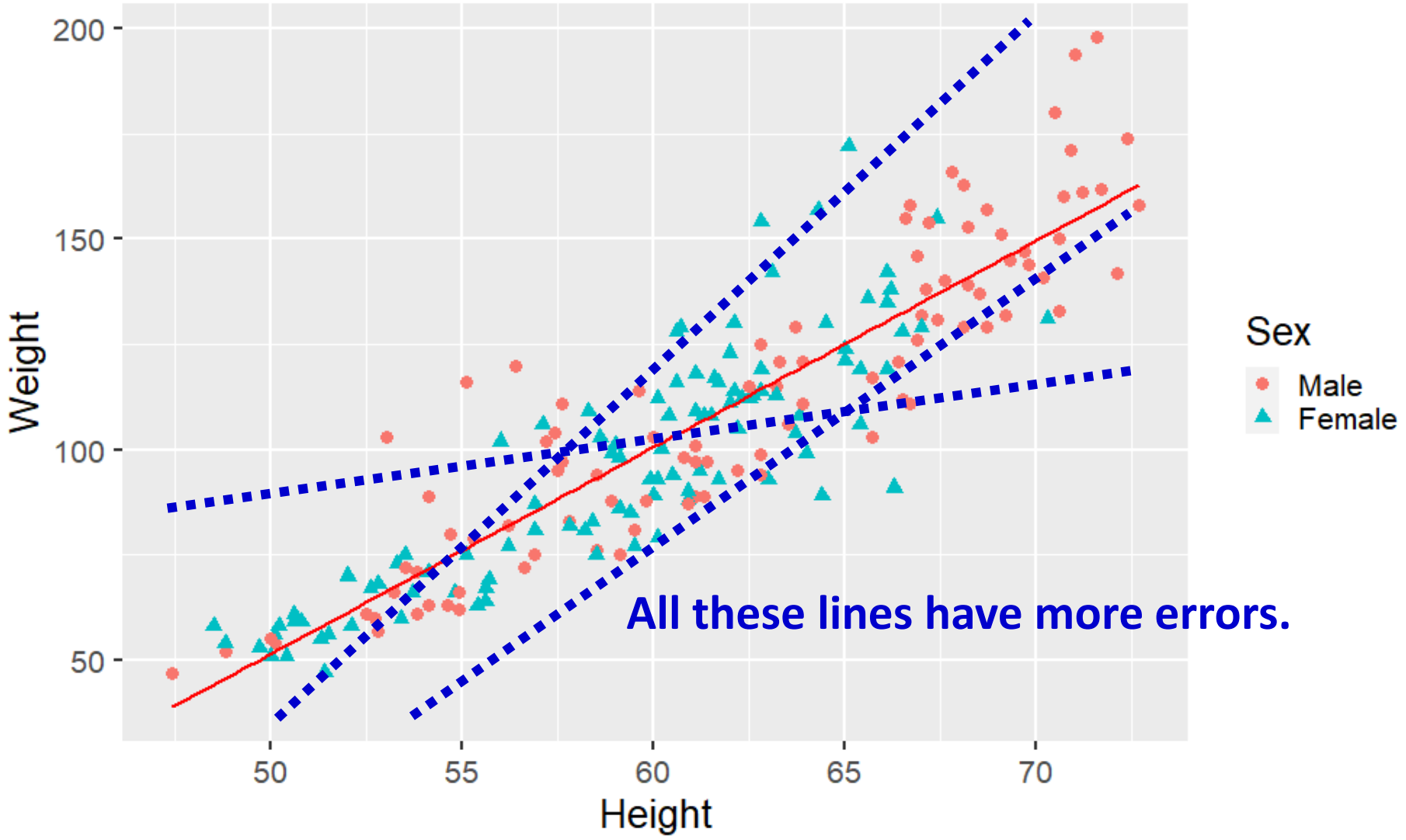
$$y = \alpha + \beta x + \epsilon$$

- y is dependent on x, and therefore refer to y as the dependent variable or the response; x is the explanatory variable.
- ϵ is the error, assumed to be 0 on average.

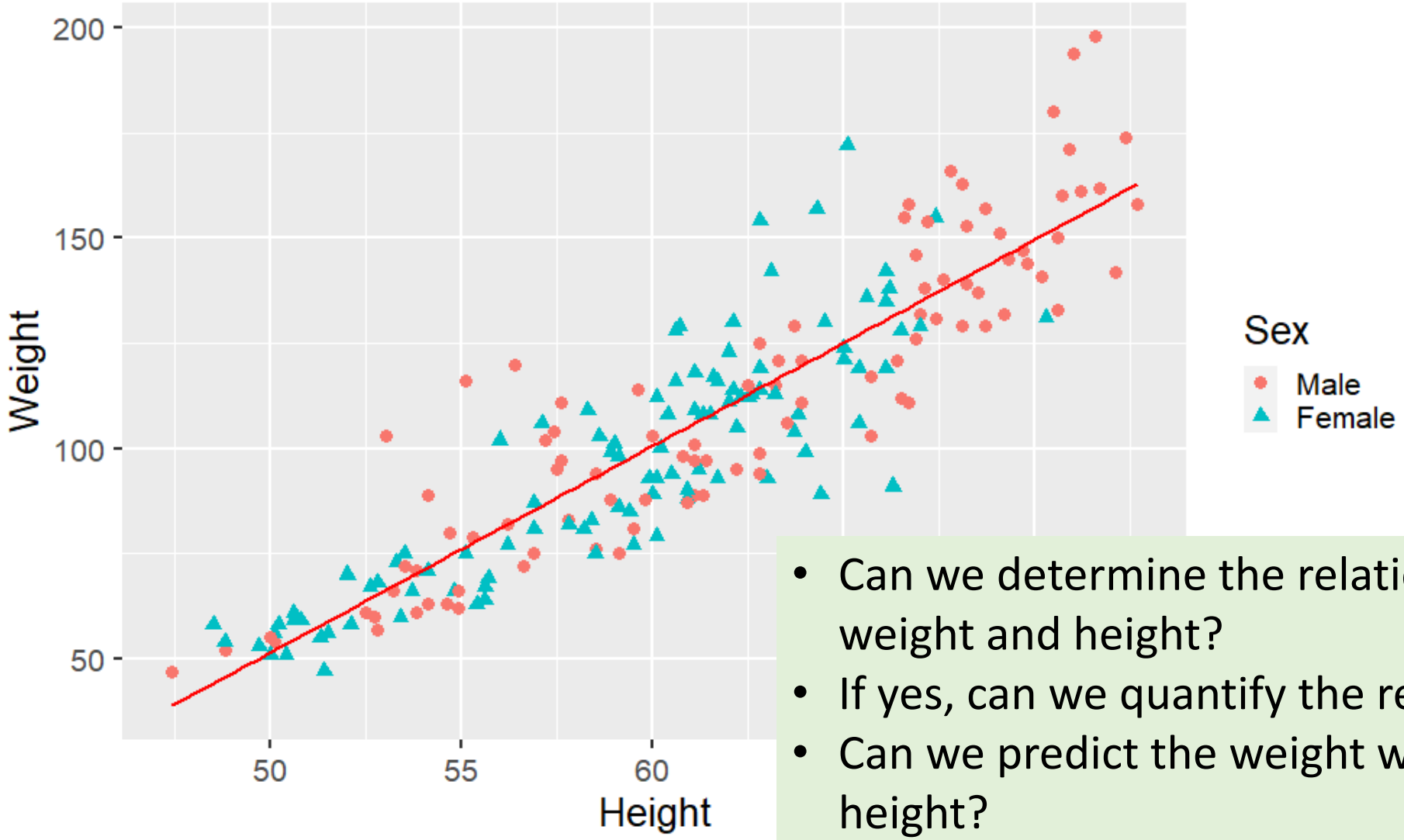
Mathematic of linear regression



Mathematic of linear regression



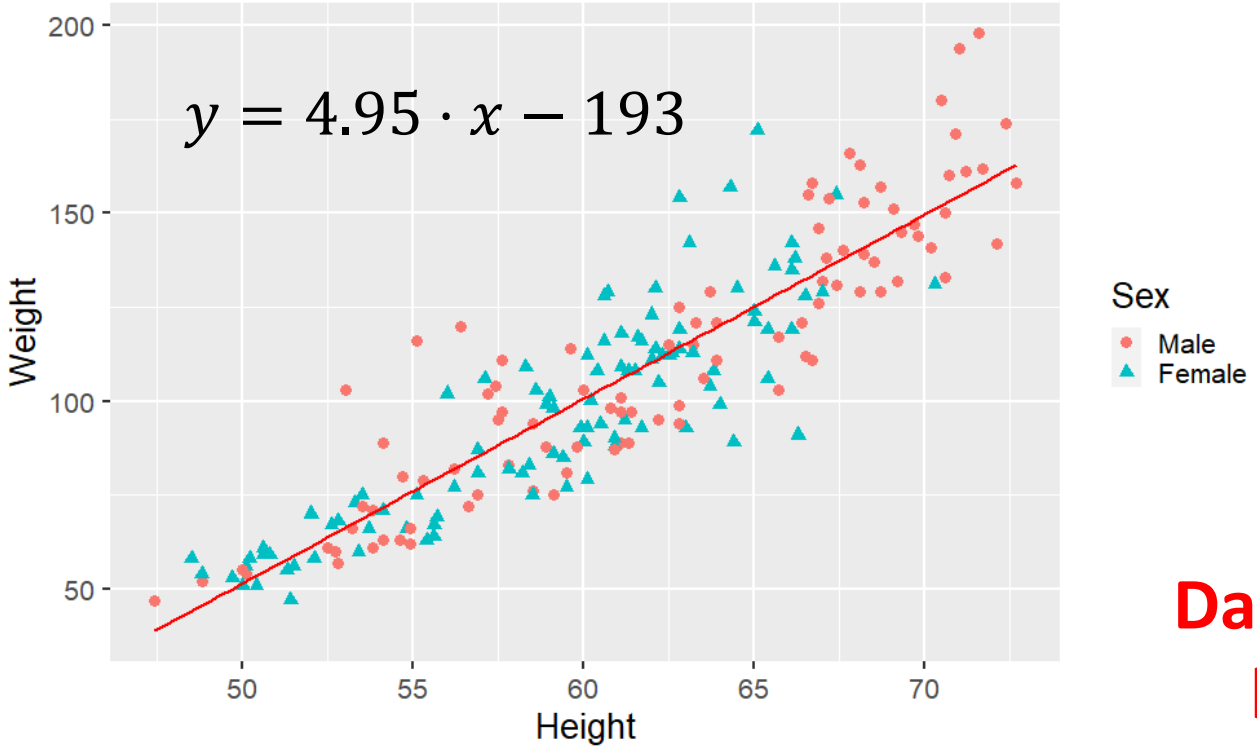
Simple linear regression



- Can we determine the relationship of weight and height?
- If yes, can we quantify the relationship?
- Can we predict the weight when given height?

Extrapolation of Data

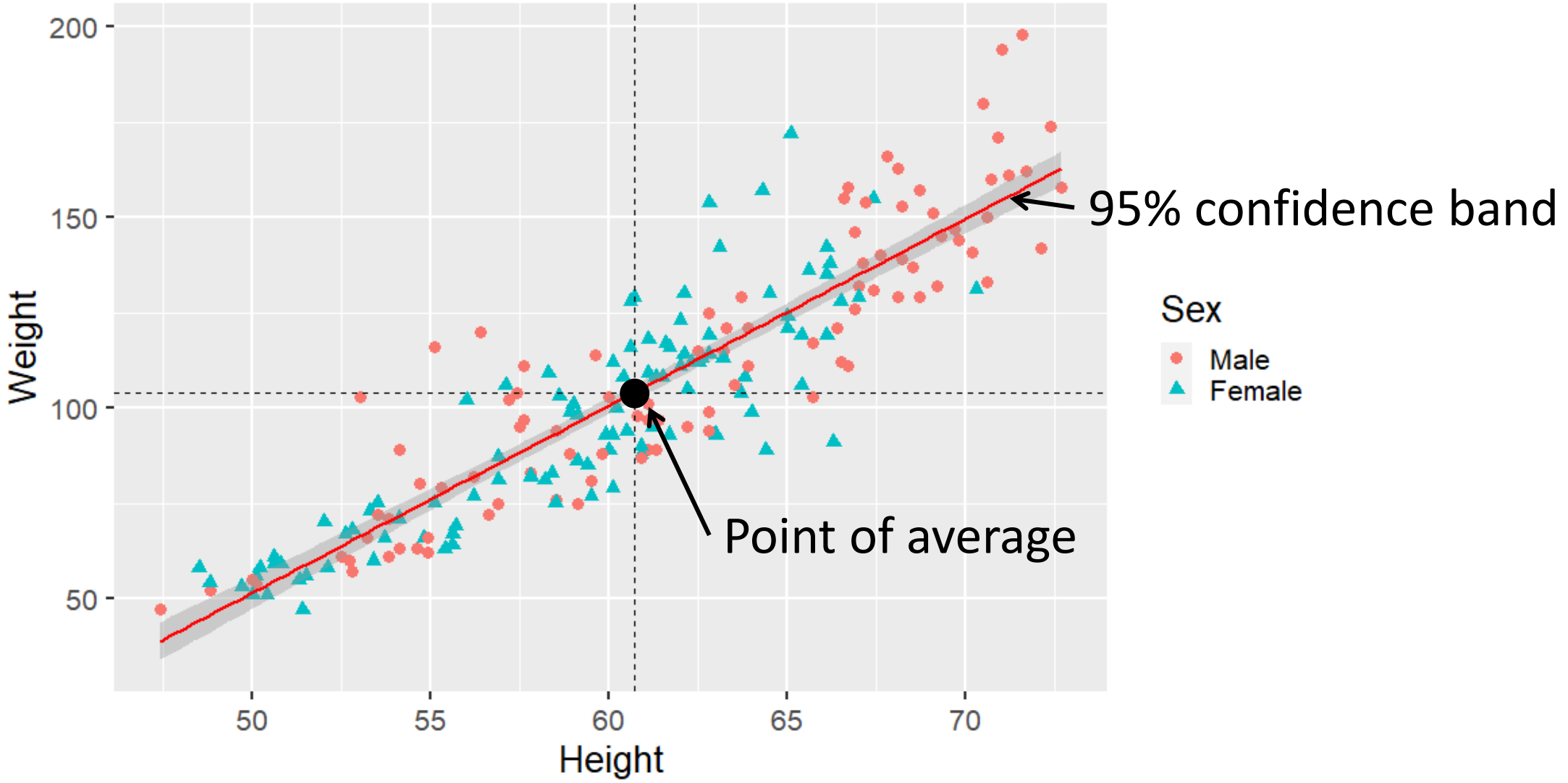
- Often convenient to extrapolate result to data outside range of regression, and just as often erroneous.



What's the children weight when height = 30?

Dangerous to extrapolate the results beyond the range of regression

Confidence band



Statistical inference in linear regression

- Test the significance (or 'contribution') of an independent variable (x) to the dependent variable (y) via hypothesis tests (or confidence intervals).

$$y = \alpha + \beta x + \epsilon$$

- Consider null hypothesis of:

$$H_0: \ \beta = 0$$

- Tests linear relationship using t-tests.
- Often performed by default by software in regression.

Multiple regression and linear modelling

- More than one explanatory variable, example age, gender, ethnic groups and height
- Interested to find how these variables affect weight.
- Mathematically complicated, but conceptually identical to finding the coefficients which minimises the errors (easy with a computer)

$$Weight = \alpha + \beta_1 Height + \beta_2 Age + \beta_3 I(Male) + \beta_4 I(Smoke) + \epsilon$$

- Notice the difference for categorical variables like gender and smoke. $I(\dots)$ represents an indicator variable, taking the value 1 when the condition in the bracket is satisfied, and zero otherwise.

Linear Modelling

- Statistical approach to explain a response, or some function of the response variable, as a linear combination of the other explanatory variables.

Regression	Response
Multiple linear regression	Numerical response
Logistic regression	Binary categorical response
Multinomial logistic regression	Categorical variable with multiple outcomes
Poisson (log-linear) regression	Count/Rate response
Cox proportional hazard regression	Survival response

Model selection

- In linear modelling, the main focus usually is in identifying the explanatory variables that contribute significantly in explaining the response variable.

$$Weight = \alpha + \beta_1 Height + \beta_2 Age + \beta_3 I(Male) + \beta_4 I(Race) + \epsilon$$

- There will be variables that are not useful/informative in explaining how Weight changes.
- Pointless to include these variables in the model, and statistically wasteful as well since they use up precious information to estimate the βs.

Model selection

- There are multiple approaches for selecting the optimal or near-optimal model.

- Eg.
  - Forward selection
  - Backward selection
  - Stepwise selection
  - Etc.

Coefficient of determination

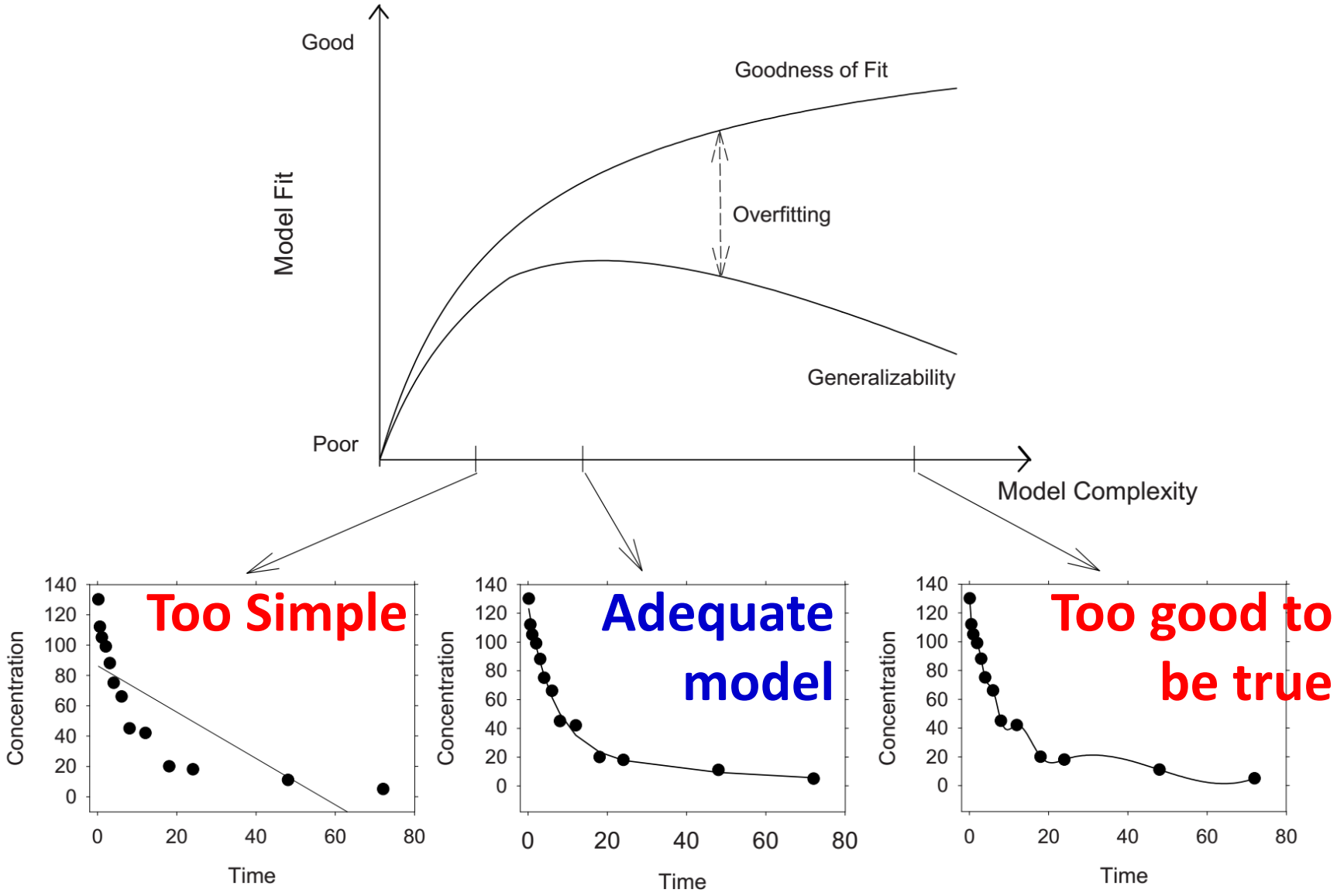
- R^2 is percentage of total response variation explained by explanatory variable

$$R^2 = 100 \times \left(\frac{SSE_{regression}}{SSE_{total}} \right)$$

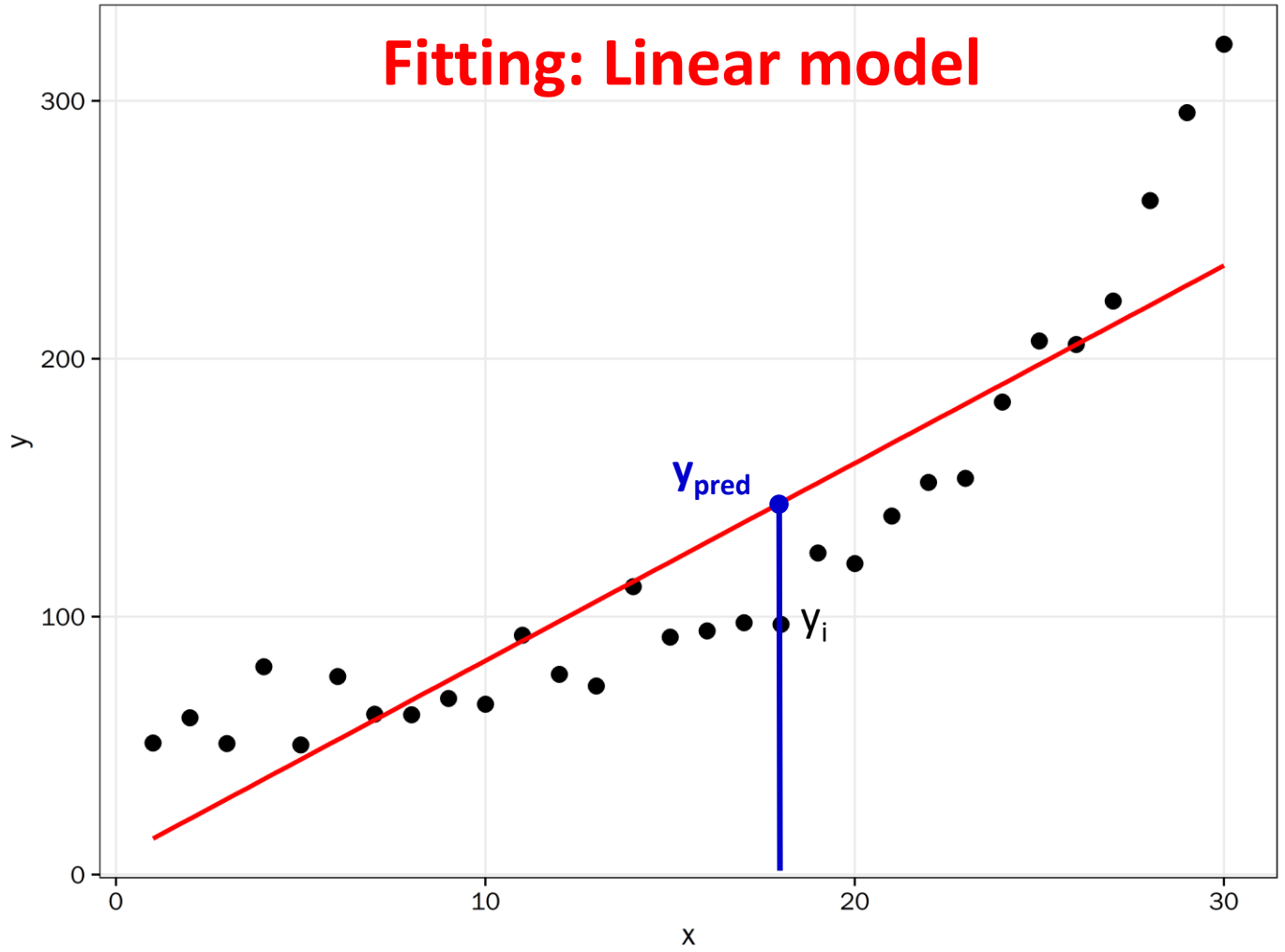
- Low R^2 indicates that not much of variation in data can be explained by regression model
- Commonly reported at the end of the regression analysis to indicate how well the mod
- For example:
- Height explains 80% of the variation in Weight el is doing to explain the response.

Linear regression diagnostic: linearity

- Possible violations:
  - –Straight line may be inadequate model
  - –Contamination from outliers from different populations
- Resulting estimates misleading, biased
- Degree of biased-ness depends on degree of violation of assumption
- Possible transformations or polynomial variables

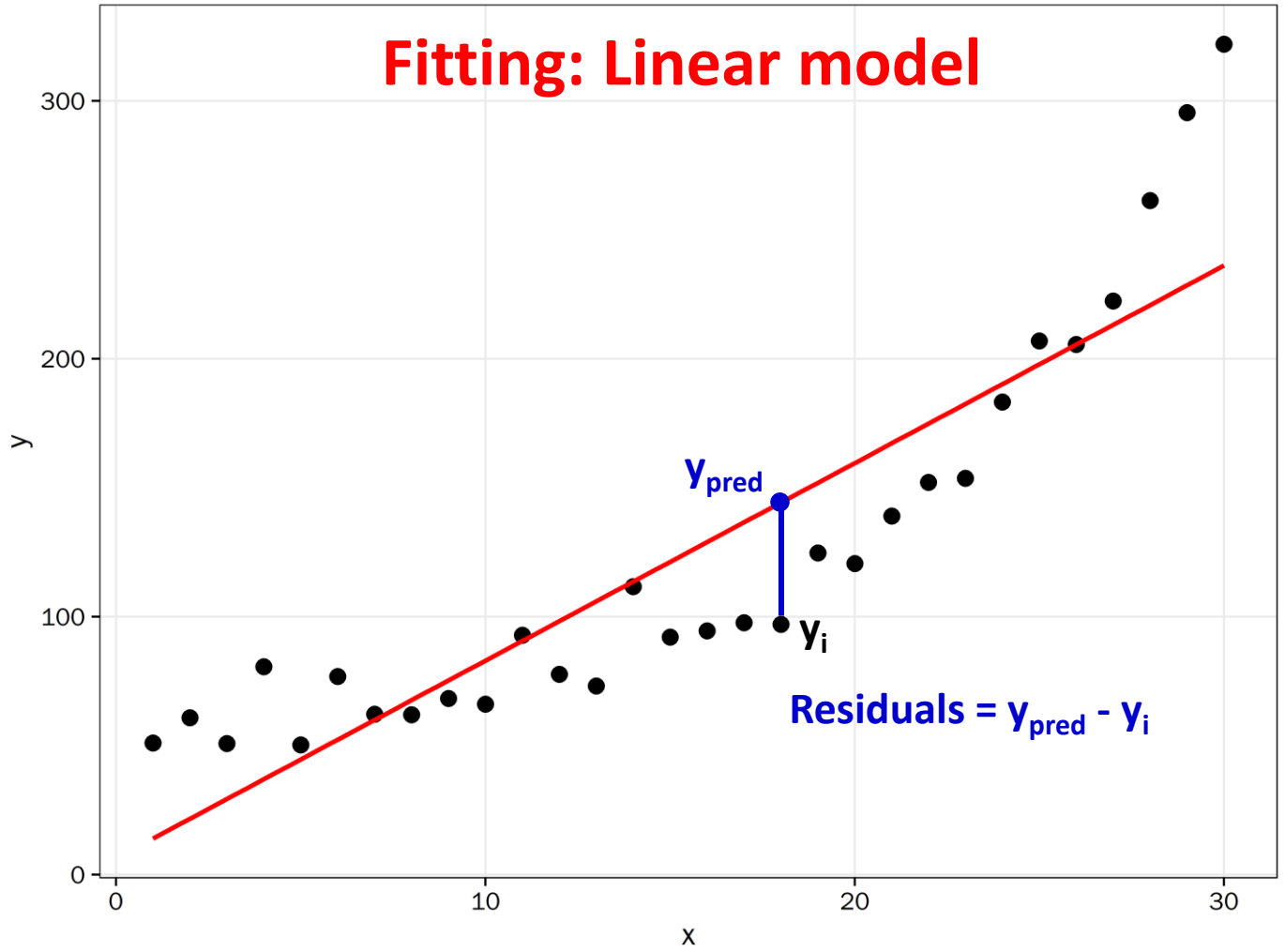

Good
Goodness of Fit
Model Fit
Overfitting
Generalizability
Poor
Model Complexity
Too Simple
Adequate model
Too good to be true
Concentration
Time
0
20
40
60
80
100
120
140

Example



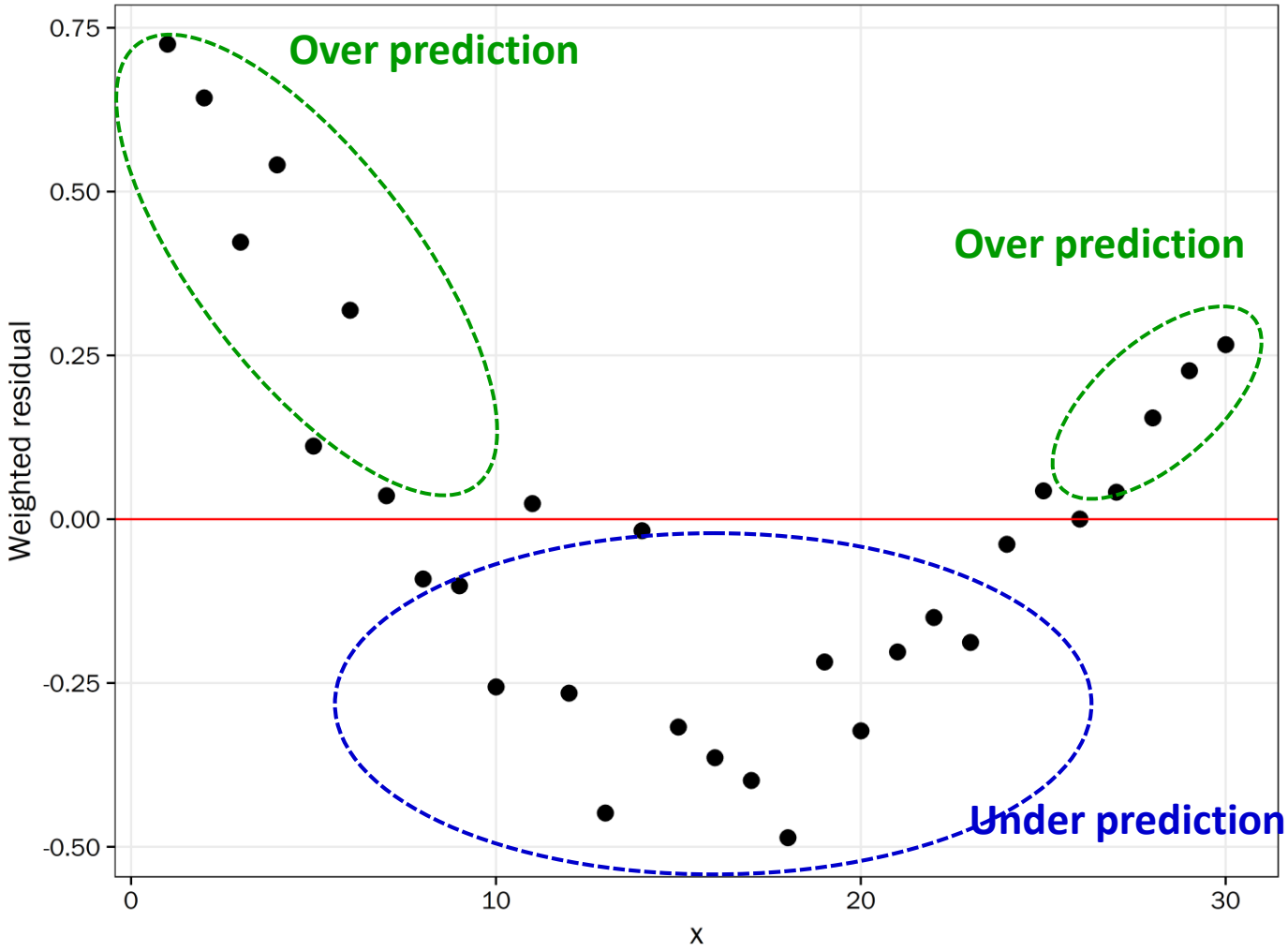
Observed: X~Y

Example



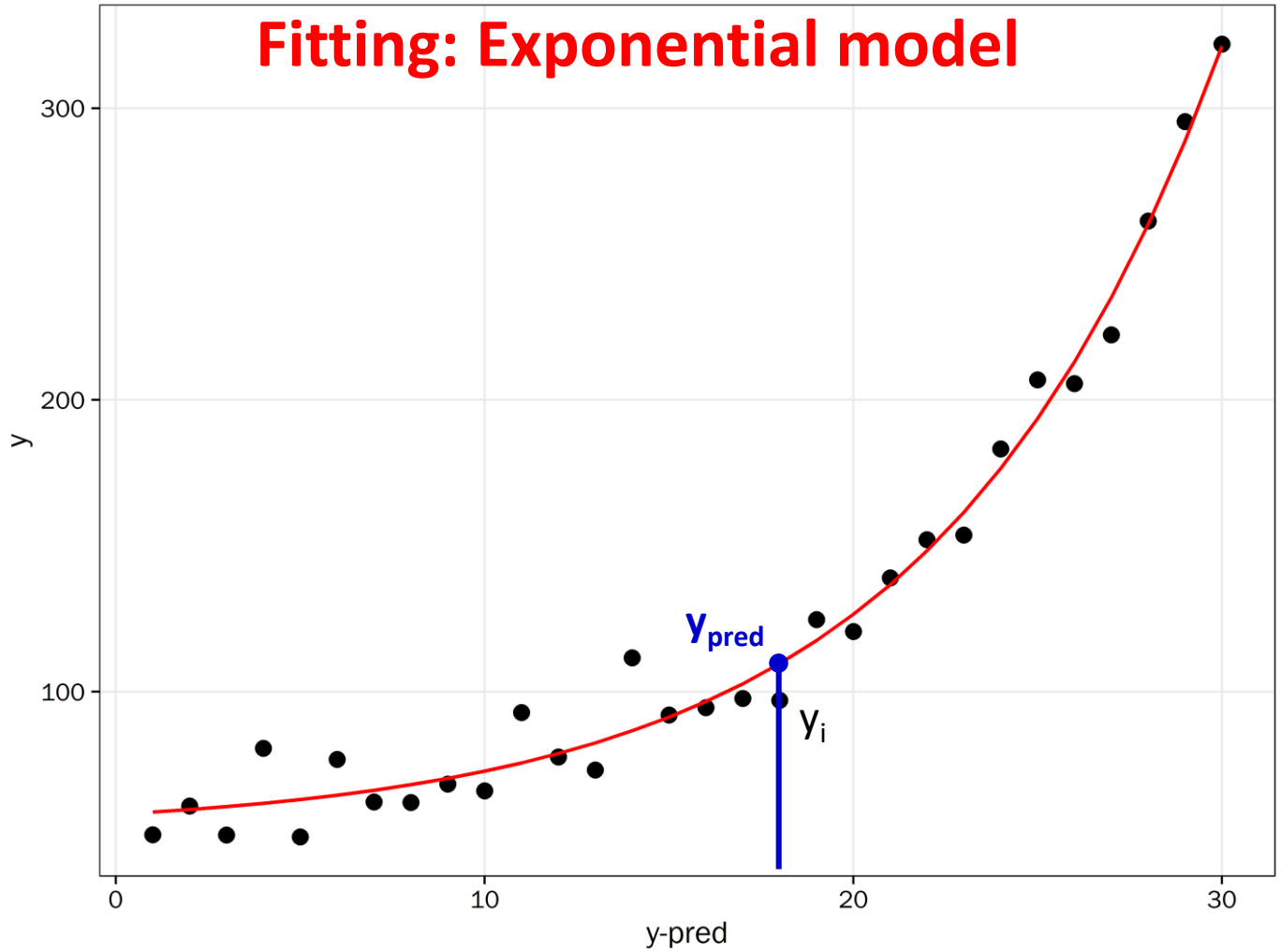
Observed: X~Y

Example



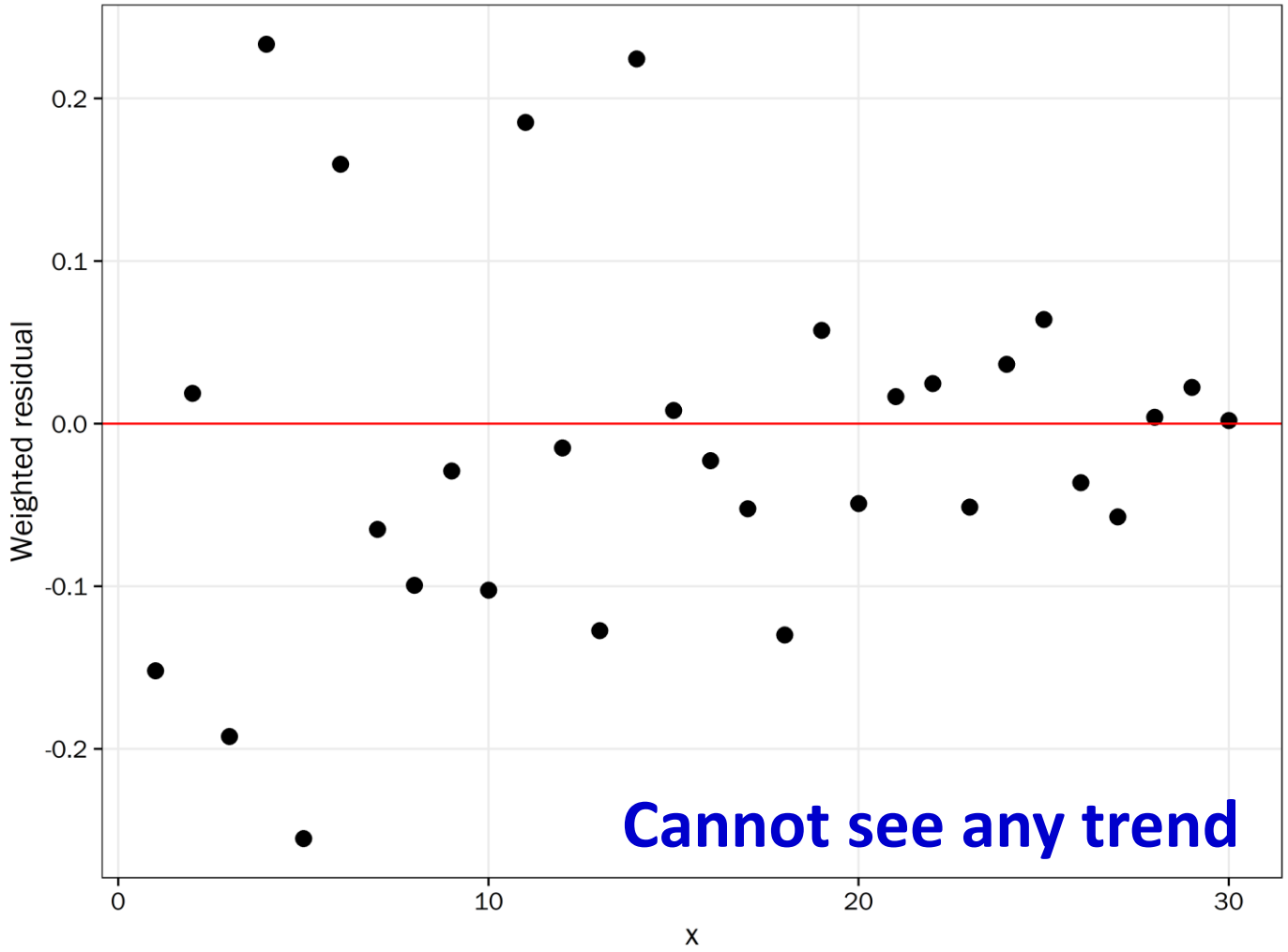
X ~ Predicted Errors

Example : A better model fit



Observed: X~Y

Example: A better model fit



X ~ Predicted Errors

MORU Tropical Health Network

Basic R programing

11-12 Jan 2021

Hand On Day 2 (14:45-16:00)

Dr. Palang Chotsiri
palang@tropmedres.ac

Hand On Day 2

- Load the data “covid_analytic_clinical_data.csv” (read data information at “covid_analytic_clinical_data.doc”)
- Check for the 10 first lines
- Get a summary of each variable

Hand On Day 2

- Solution: ***HandOn_day2_solution.R***

Day 2 Wrap-up

What did you learn today?