# Linear Regression

Wan Nor Arifin

Biostatistics and Research Methodology Unit, Universiti Sains Malaysia.

email: wnarifin@usm.my

# Introduction

1. A statistical method to model relationship between:
    - outcome: numerical variable.
    - predictors/independent variables: numerical, categorical variables.
2. A type of Generalized Linear Models (GLMs), which also includes other outcome types, e.g. categorical and count.

# Introduction

3. Basically, the linear relationship is structured as follows,

$$numerical\ outcome = numerical\ predictors + categorical\ predictors$$

# Simple linear regression (SLR)

1. Model *linear* (straight line) relationship between:
   - outcome: numerical variable.
   - a predictor: numerical variable (only).

   *Note: What if the predictor is a categorical variable? Remember, we already handled that with one-way ANOVA.*

# Simple linear regression (SLR)

2. Formula,

$$numerical\ outcome = intercept + coefficient \times numerical\ predictor$$

in short,

$$\hat{y} = \beta_0 + \beta_1 x_1$$

where $\hat{y}$ is the predicted value of the outcome y.

# Multiple linear regression (MLR)

1. Model *linear* relationship between:
   - outcome: numerical variable.
   - predictors: numerical, categorical variables.

   *Note: MLR is a term that refers to linear regression with two or more numerical variables. Whenever we have both numerical and categorical variables, the proper term for the regression model is General Linear Model. However, we will use the term MLR in this workshop.*

# Multiple linear regression (MLR)

2. Formula,

$$\begin{aligned} numerical\ outcome = {} & intercept + coefficients \times numerical\ predictors \\ & + coefficients \times categorical\ predictors \end{aligned}$$

in a shorter form,

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

where we have $k$ predictors.

# Multiple linear regression (MLR)

Whenever the predictor is a categorical variable with more than two levels, we use dummy variable(s). This can be easily specified in R using `factor()` if the variable is not yet properly specified as such. There is no problem with binary categorical variable.

For a categorical variable with more than two levels, the number of dummy variables (i.e. once turned into several binary variables) equals number of levels minus one. For example, whenever we have four levels, we will obtain three dummy (binary) variables.

# Analysis steps

1. Library
2. Load data
3. Data exploration
    - descriptive
    - plots (if relevant)
4. Univariable
5. Multivariable
    - all selected
    - stepwise
    - confounder

# Analysis steps

6. Multicollinearity, MC
7. Interaction
8. Model fit: residuals
9. Interpretation
10. Equation
11. Prediction