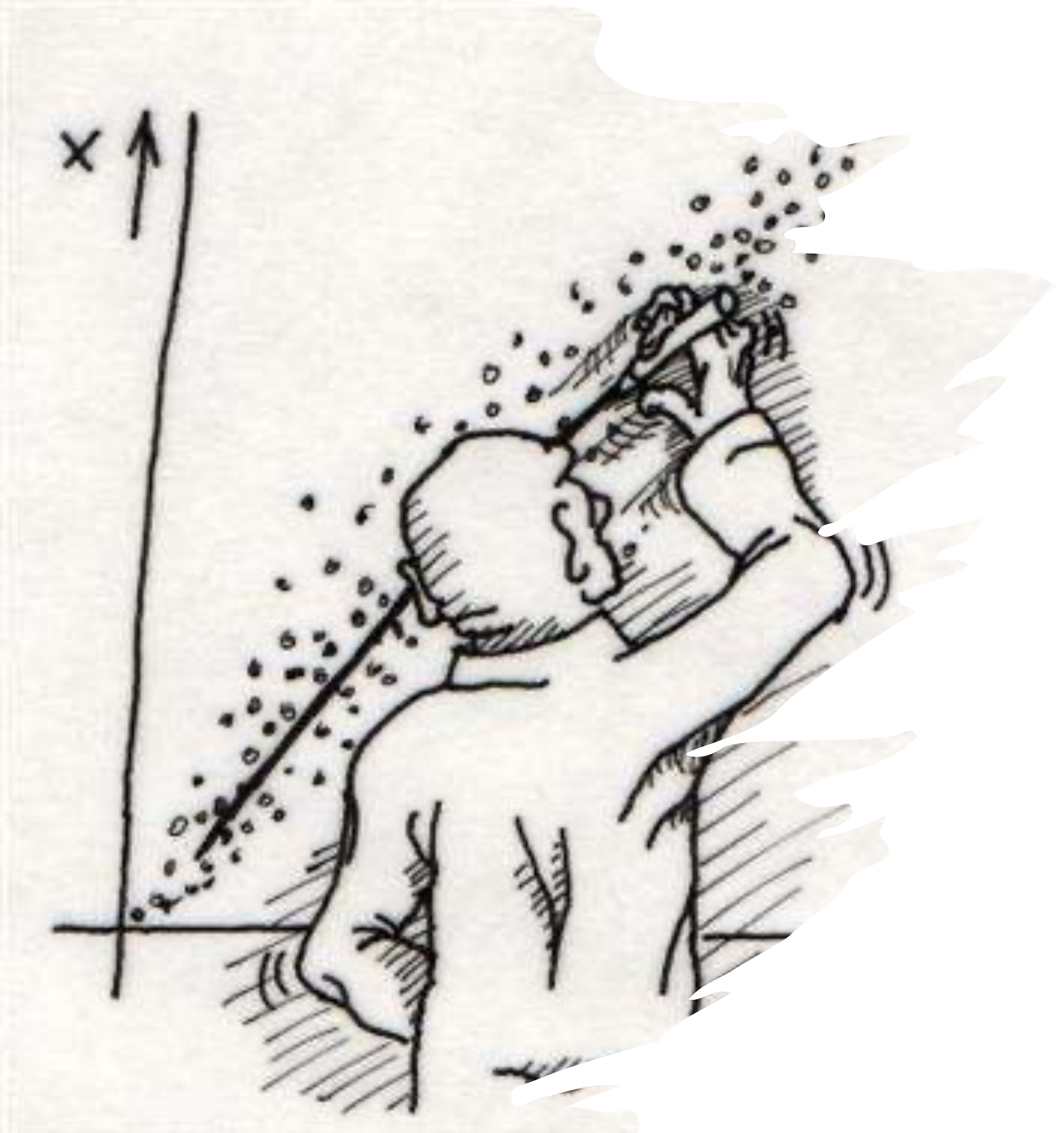
THE UNIVERSITY OF MELBOURNE

https://www.menti.com/al94qs1px1j6

Linear Models

SARITHA KODIKARA



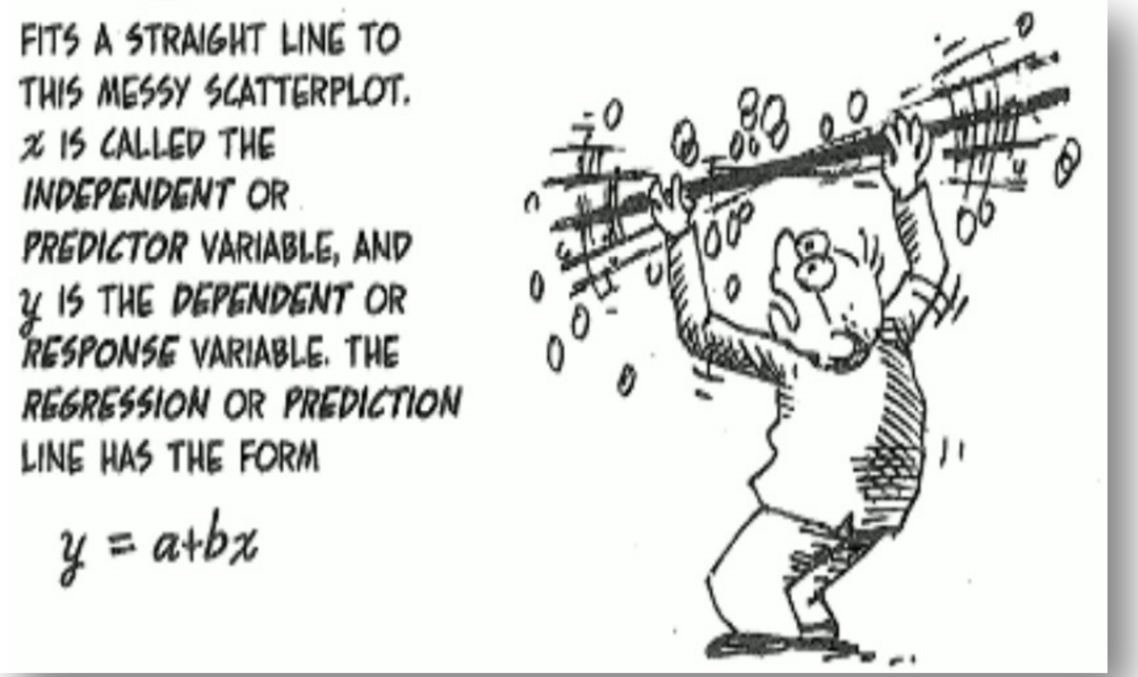
Objectives of this workshop

- ❑ To be able to understand the fundamental principles of linear models.
- ❑ To evaluate the assumptions of linear models and know what to do if the assumptions are violated.
- ❑ To be able to understand the difference between linear models and linear mixed models.
- ❑ To be able to apply linear models and linear mixed models in R.

Linear Models

❑ Linear Models are used to :

- Predict the value of a dependent variable based on the value of at least one independent variable
- Explain the impact of changes in an independent variable on the dependent variable



❑ Dependent variable/ response variable/ outcome variable (Y) :

the variable we wish to predict or explain

❑ Independent variables/ regressor variables/ predictor variables/ explanatory variable (X):

the variable used to predict or explain the dependent variable

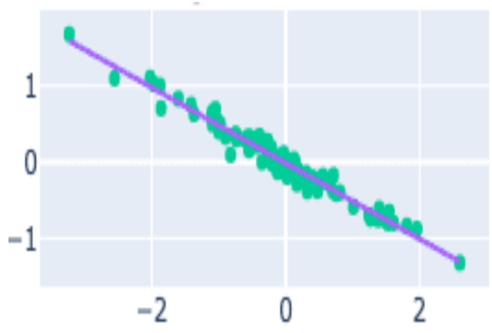
Simple Linear Regression Model

❑ One of the most well-known examples of a linear model is the simple linear regression

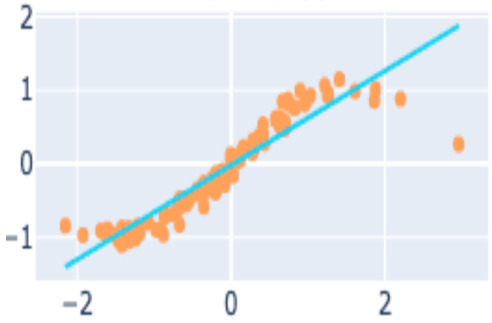
❑ Only one independent variable, X

❑ Relationship between X and Y is described by a linear function

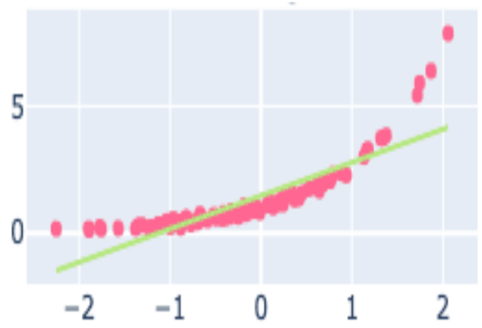
Linear



Linear



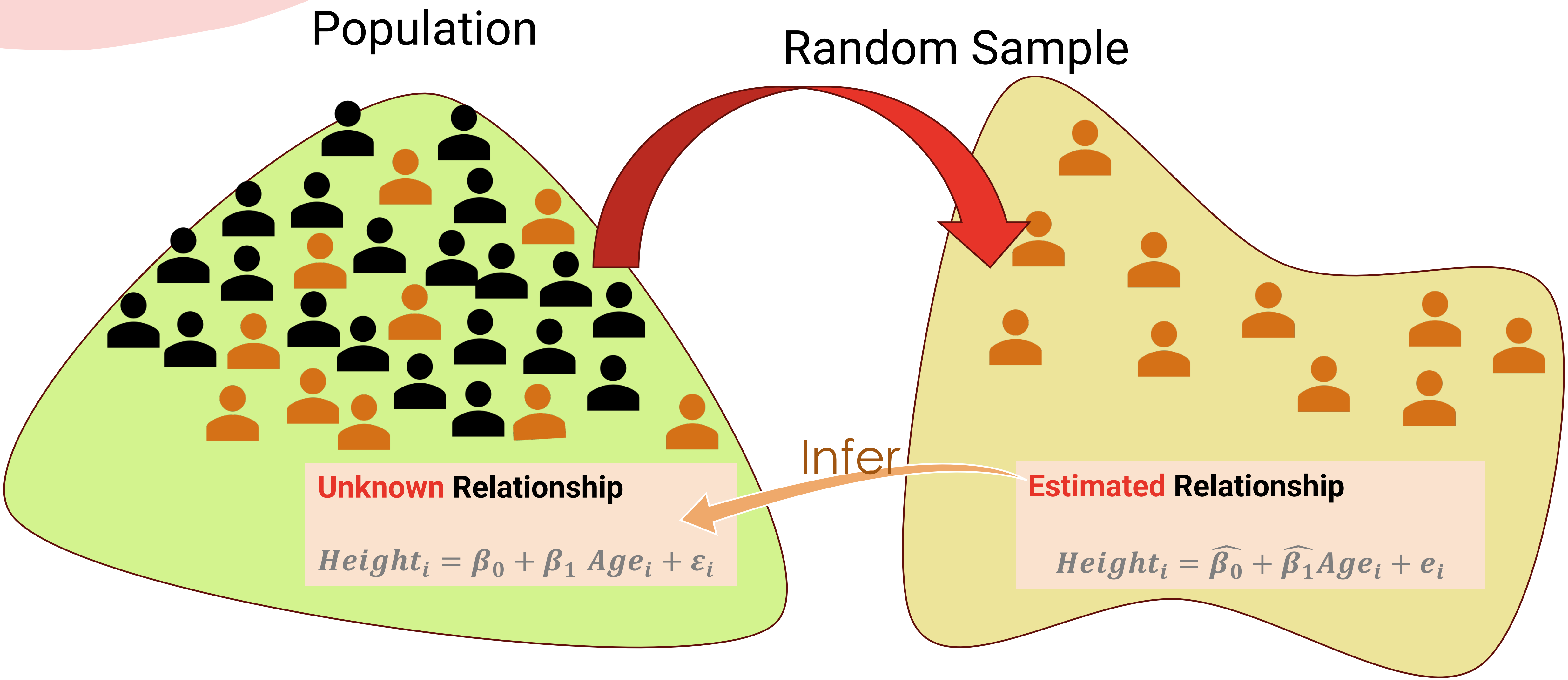
Non-Linear



Non-Linear

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

Population & Sample Regression Models



see previous MIG workshop 'Expt design'

Simple Linear Regression Model

Dependent Variable

Population Y intercept

Population Slope Coefficient

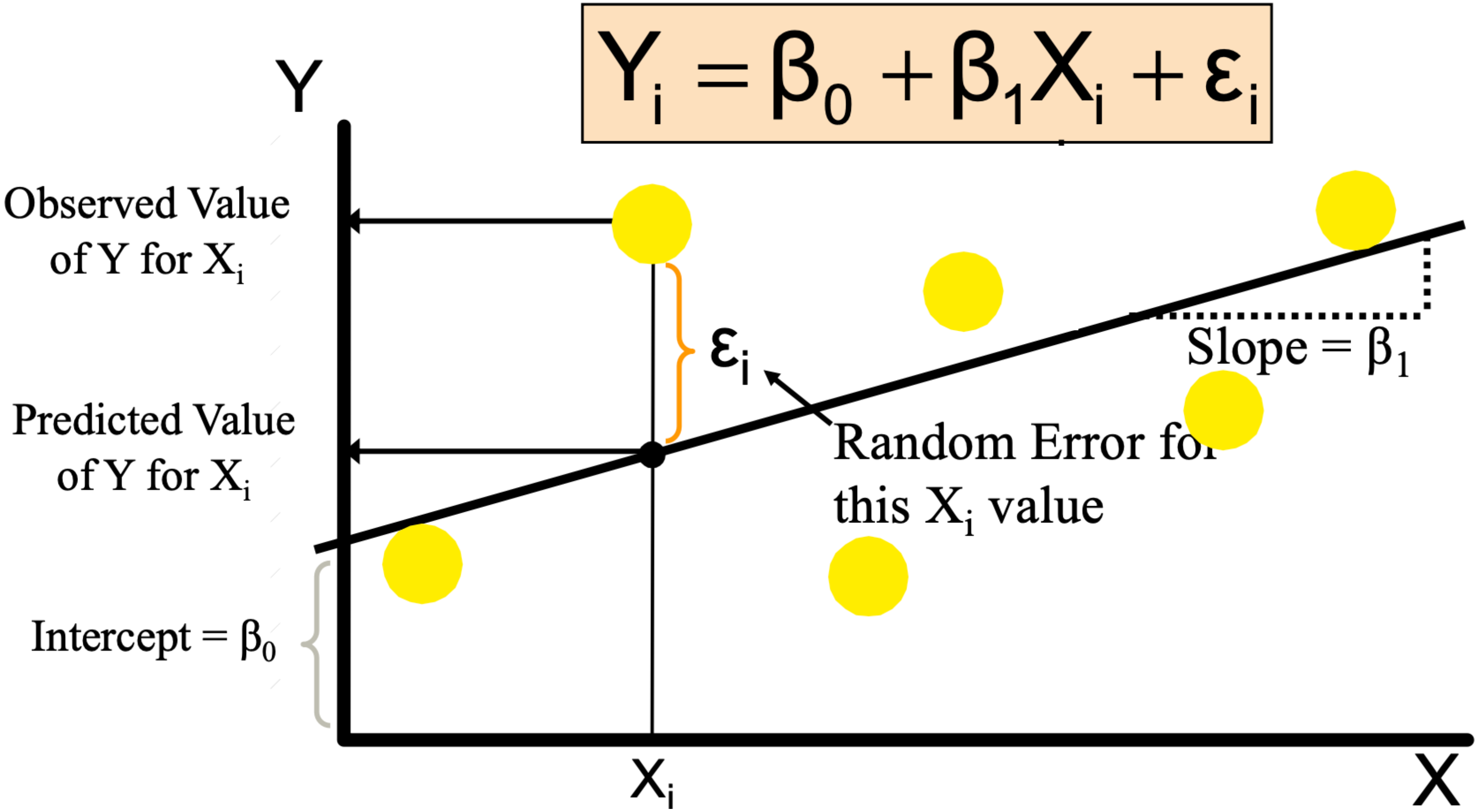
Independent Variable

Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

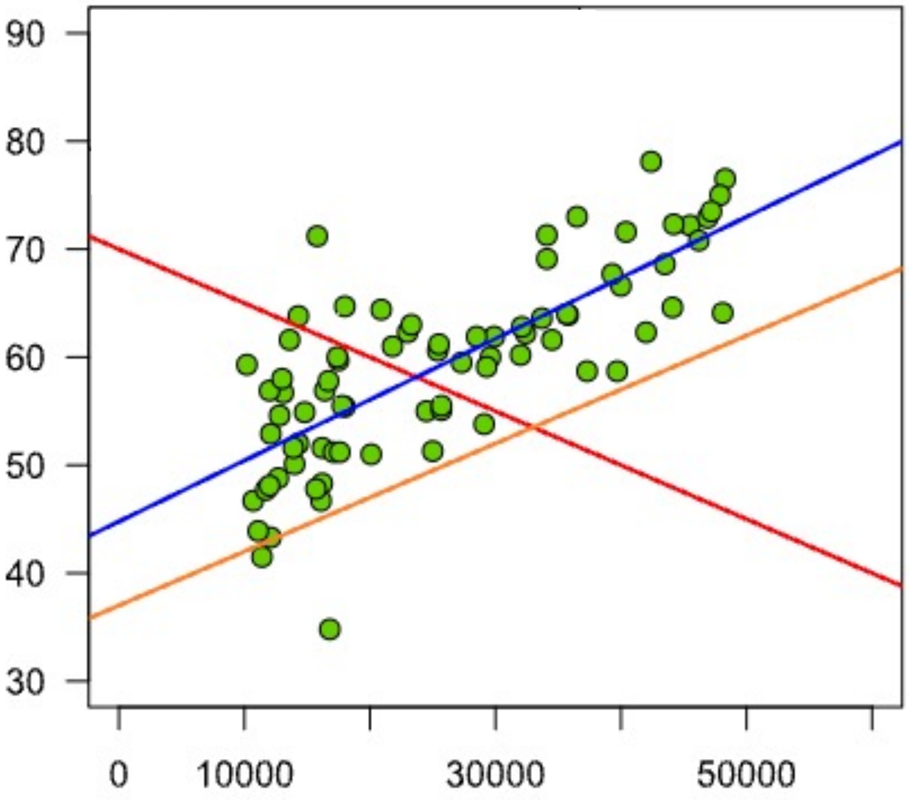
Simple Linear Regression Model

$$\widehat{Y}_i = \beta_0 + \beta_1 X_i$$



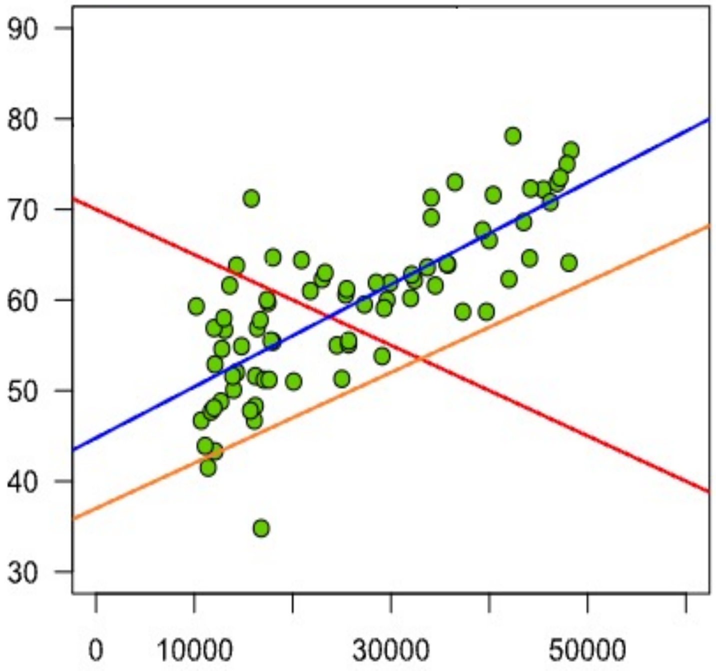
Simple Linear Regression Model

❑ How would you draw a line through the points? How do you determine which line 'fits best'?

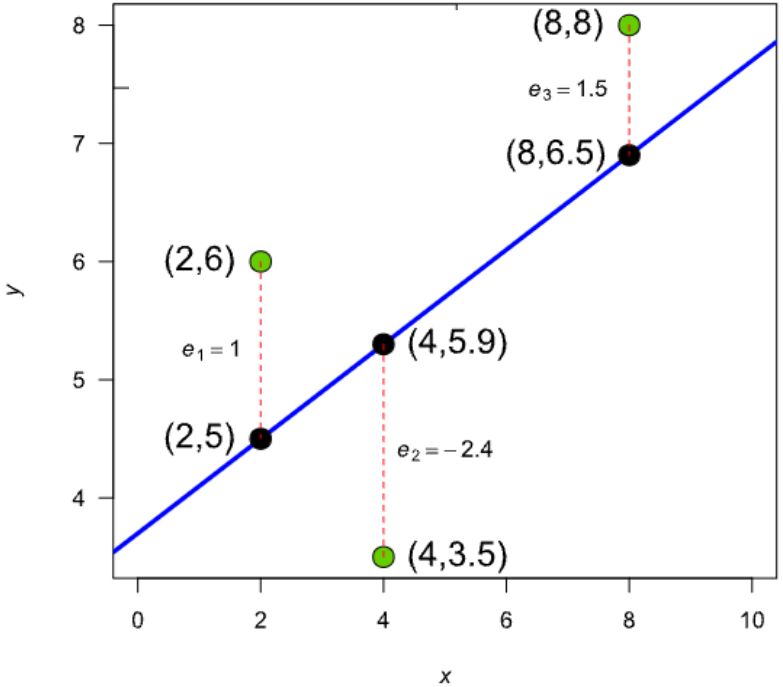


Simple Linear Regression Model

- ❑ How would you draw a line through the points? How do you determine which line 'fits best'?



The smaller the sum of squared differences the better the fit of the line to the data.



fitting a model to the data such that the **sum of squared residuals is minimized**

Assumptions of the linear model

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

- ❑ Linear relationship between response and predictor

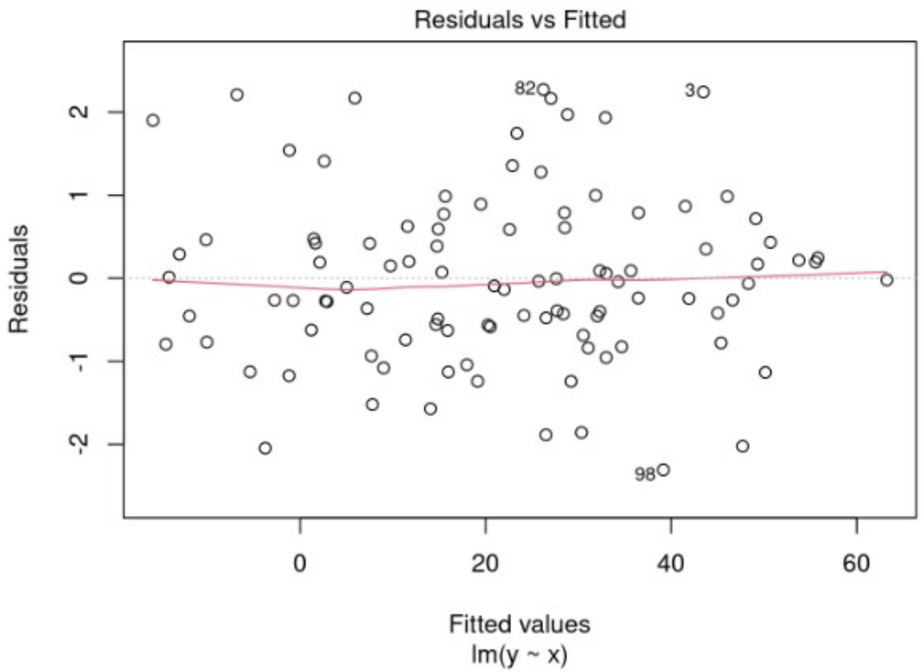
- ❑ Errors follow a normal distribution with mean 0 and constant variance (Homoscedasticity) → $\varepsilon_i \sim N(0, \sigma^2)$

- ❑ Errors are independent from each other

We can verify assumptions using 4 diagnostic plot

Diagnostic plot 1 - Residuals vs Fitted

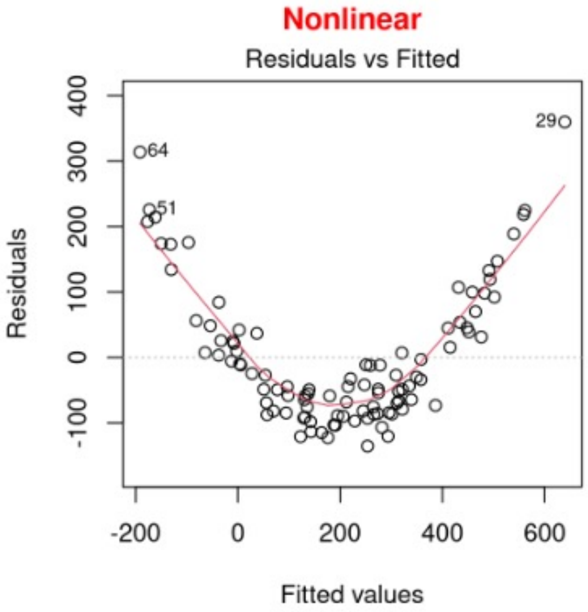
❑ **What we hope to see**: Random scatter, no pattern

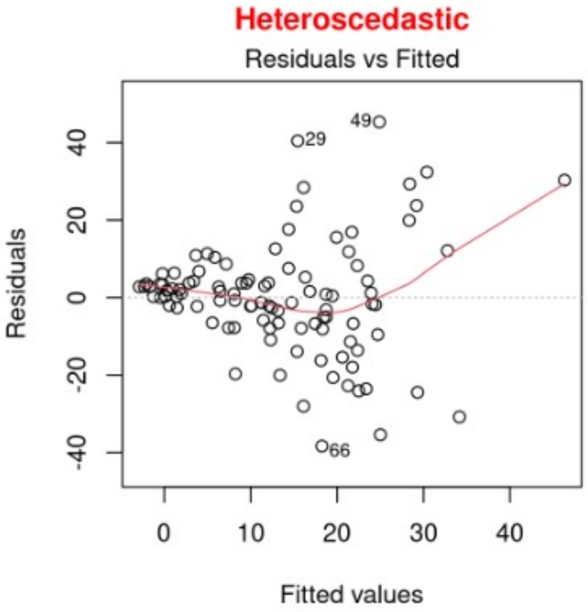


❑ Why: Shows whether residuals are independent and identically distributed

Diagnostic plot 1 - Residuals vs Fitted

❑ What should make you suspicious:



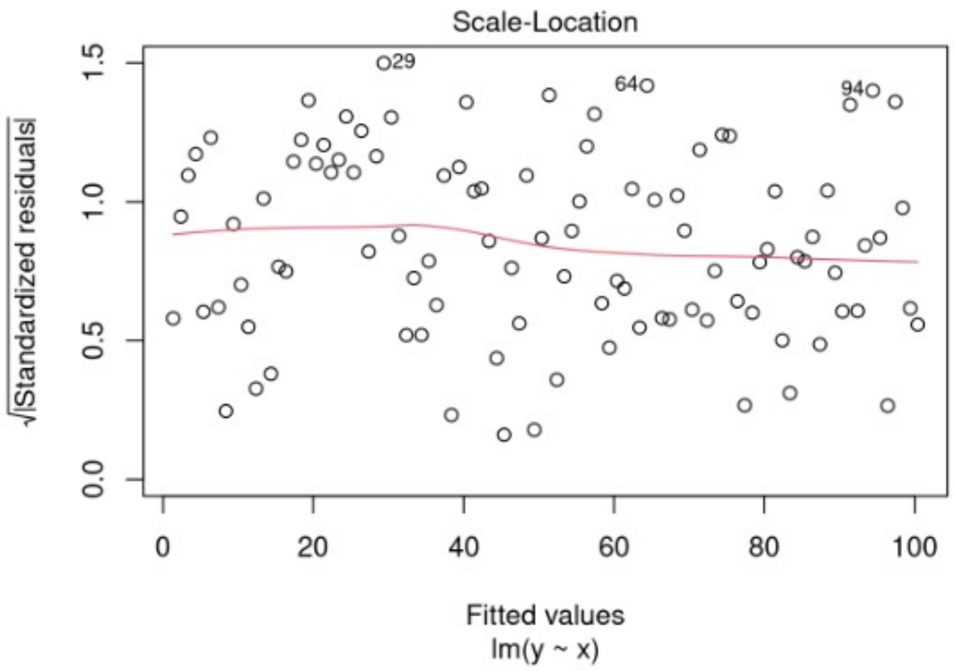


❑ What can you do:

Use a generalized linear model (GLM)
Transforming the response and/or predictor variables

Diagnostic plot 2 - Scale-Location

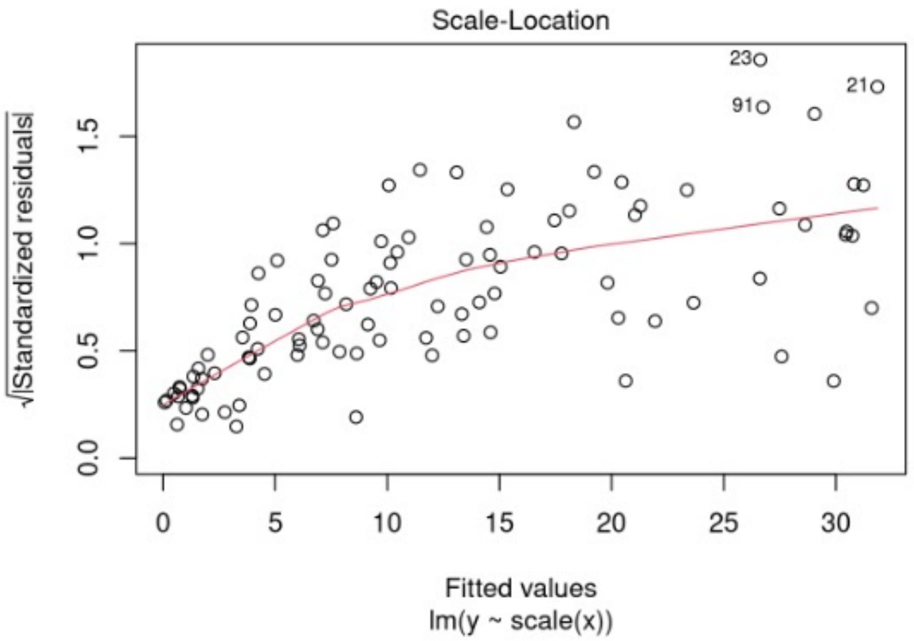
❑ **What we hope to see**: Random scatter, no pattern



❑ Why: Violations of assumptions are sometimes easier to detect than in the first plot

Diagnostic plot 2 - Scale-Location

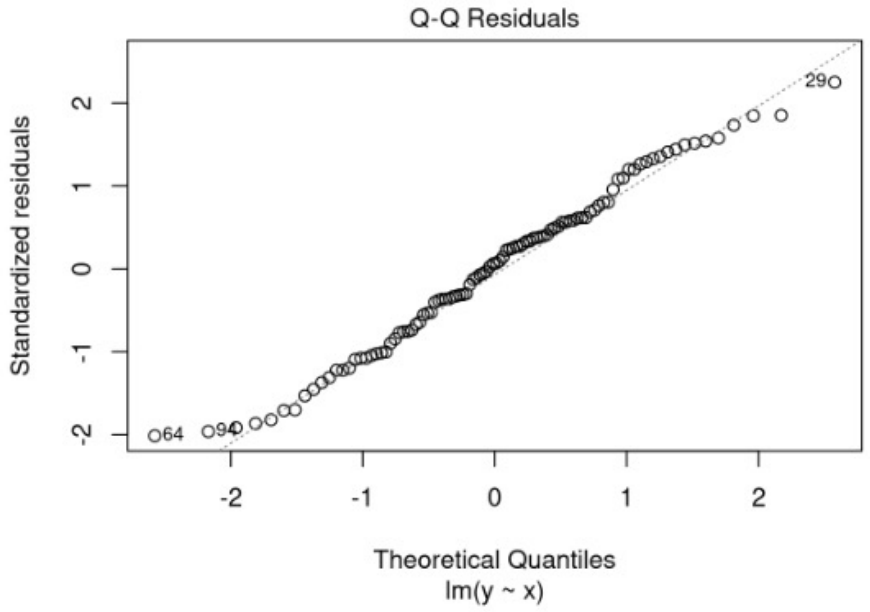
❑ What should make you suspicious:



Strong pattern in the residuals

Diagnostic plot 3 - Normal Quantile-Quantile

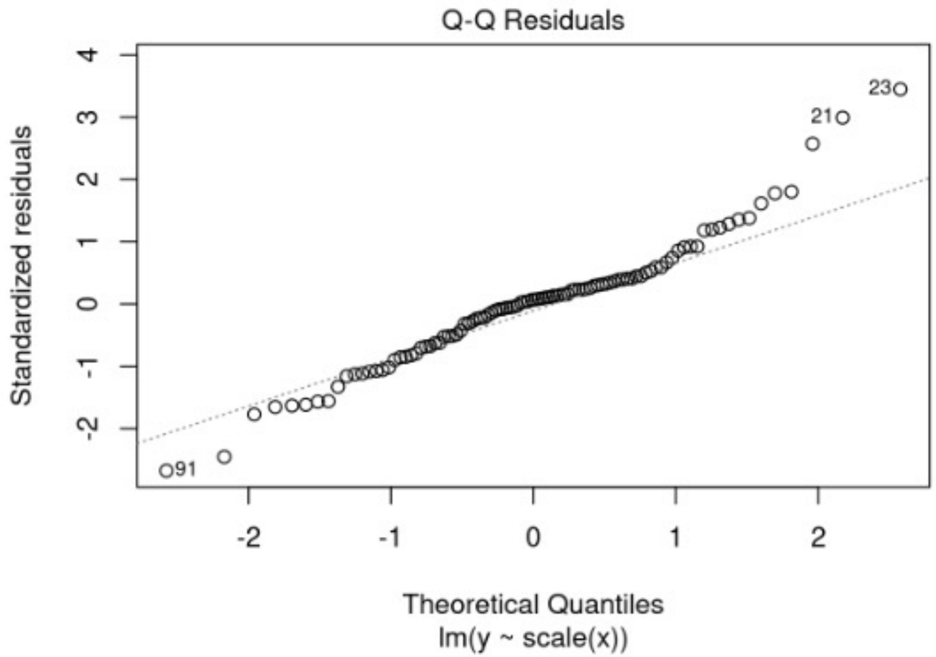
❑ **What we hope to see**: Points clearly on the diagonal line



❑ Why: Compares the distribution (quantiles) of the residuals with a standard normal distribution

Diagnostic plot 3 - Normal Quantile-Quantile

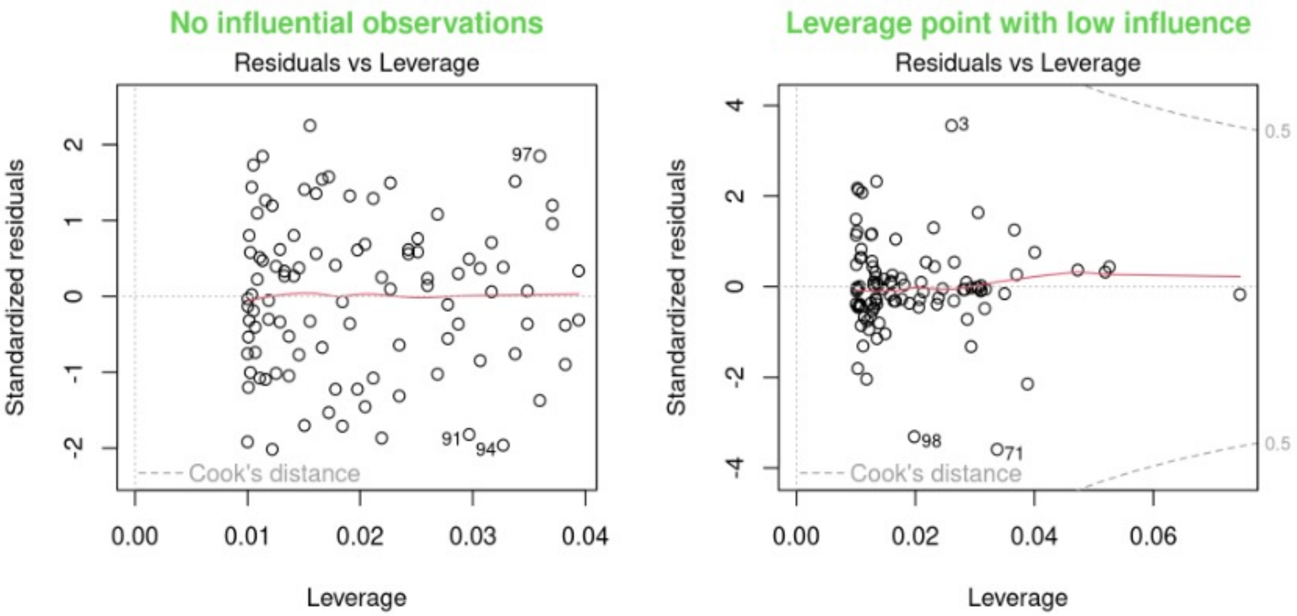
❑ What should make you suspicious:



Residuals do not follow a normal distribution

Diagnostic plot 4 - Residuals vs. Leverage

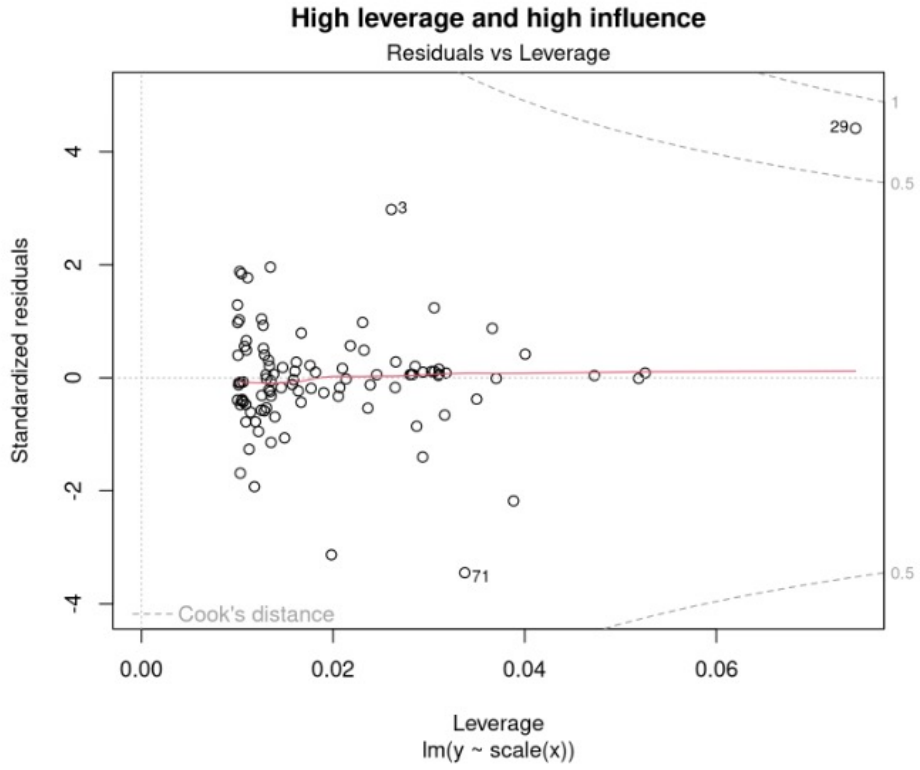
❑ **What we hope to see**: No leverage points with high influence



❑ Why: The model should not depend strongly on single observations

Diagnostic plot 4 - Residuals vs. Leverage

❑ What should make you suspicious:



Diagnostic checking in Practice

- For example, if you have fitted 1000 Linear models for each gene expression, do you need to check 4000 plots?

  Technically YES, but practically **NO**

- What to do?
  - Check most important assumptions (i.e. normality)
  - Instead of looking at 1000 qq plots

    use a statistical test to check normality (e.g. Shapiro-Wilk test)

Interpreting $\widehat{\beta_0}$ and $\widehat{\beta_1}$

- $\widehat{\beta_0}$ (Intercept): estimated mean value of Y when the value of X is zero

- $\widehat{\beta_1}$(Slope): estimated change in the mean value of Y as a result of a one-unit increase in X

Hypothesis testing for β_0 and β_1

```
Call:
lm(formula = Y ~ X)


Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.107e+03  1.928e+02   10.93 1.28e-14 ***
X           6.085e-01  5.969e-02   10.19 1.35e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $H_0 : \beta_0 = 0 \;\; Vs. \;\; H_a : \beta_0 \neq 0$

  If p value for the intercept is < 0.05, we REJECT H_0

- $H_1 : \beta_1 = 0 \;\; Vs. \;\; H_a : \beta_1 \neq 0$

  If p value for the X variable is < 0.05, we REJECT H_0

see previous MIG workshop 'Expt design'

Linear Mixed Model

- Linear mixed models are an extension of simple linear models to allow both fixed and random effects

Fixed effect : Any effects or variables that we are specifically interested in studying

Random effects: Any effects or variables that we are not specifically interested in, but we need to account for in our model to avoid bias. (Blocking variables)

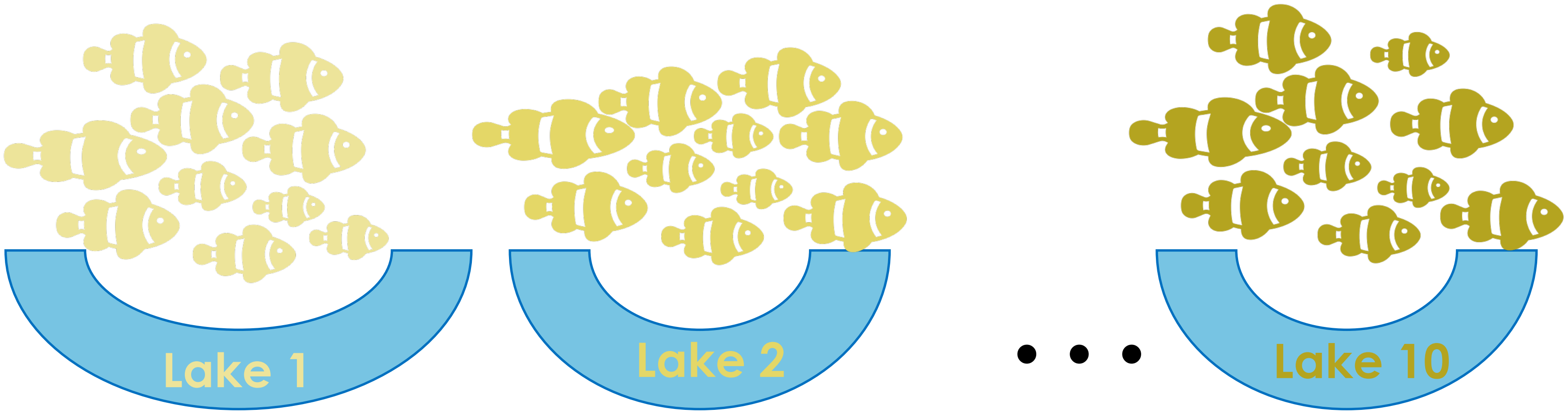
see previous MIG workshop 'Expt design'

Linear Mixed Model- Example

- ❑ **Research question**

Does fish trophic position increase with fish size?

- ❑ To answer this question, researchers measures 10 fish, sampled across 6 different lakes

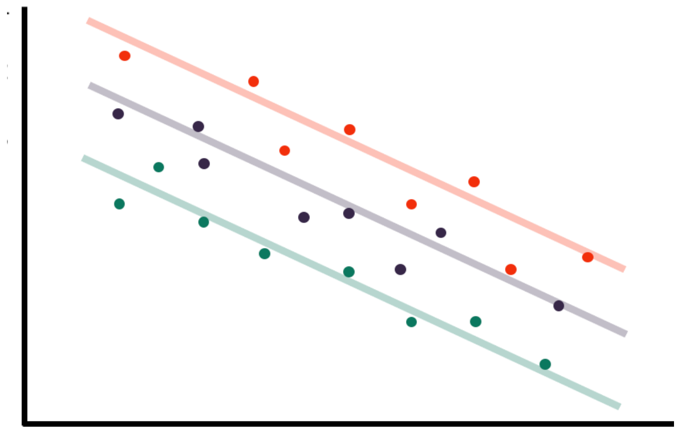


Fixed effect ??
Random effect ??

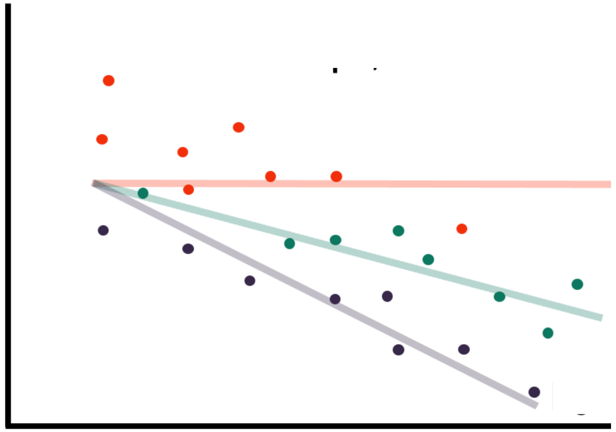
Linear Mixed Model

❑ Different structures for random effects in the model

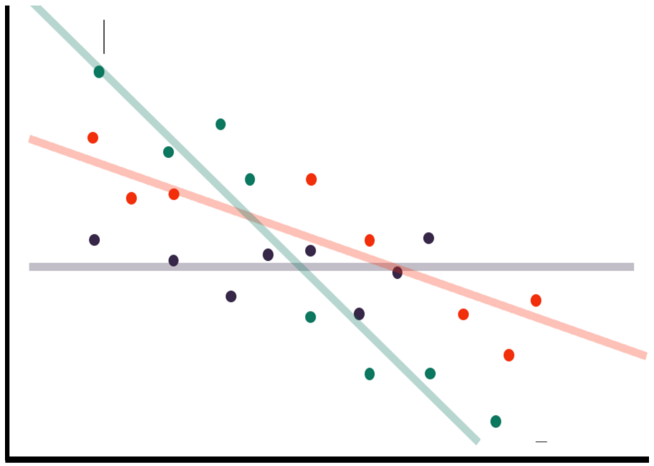
❑ random effect at the intercept



❑ random effect at the slope



❑ random effect at the intercept and slope



❑ The Akaike Information Criterion corrected (AICc) can be used for model selection.

Other Types of Models

- Multiple linear regression :-
  Only difference to simple linear regression: several independent variables are included in the model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

- Generalized models :-
  Can used when the normality assumption is violated

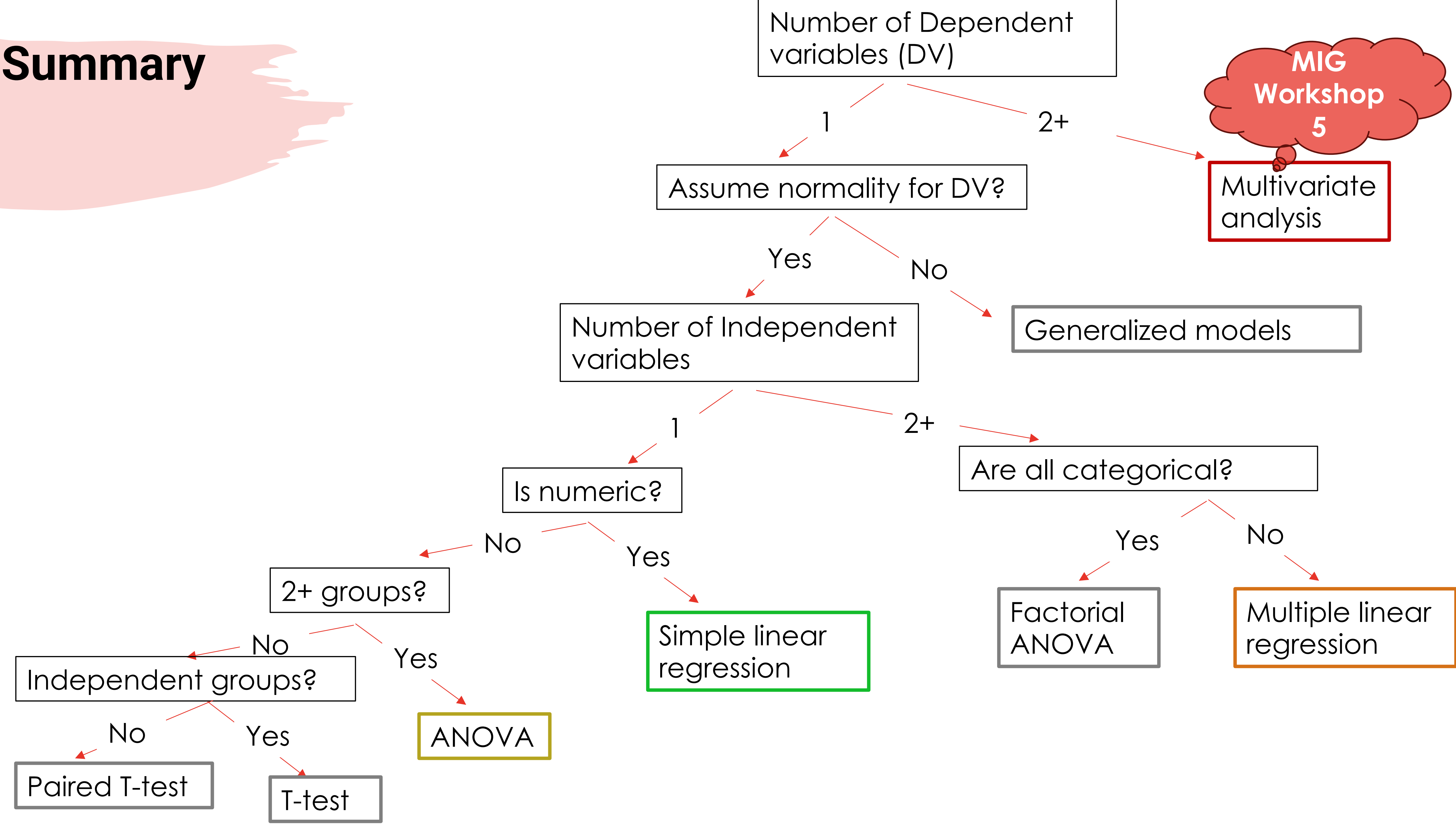
- ANOVA :-
  When you want to compare the means of three or more groups

$$H_0: \mu_{group1} = \mu_{group2} = \mu_{group3}$$

Summary

Statistical Model	When to Use
Simple Linear Regression	one continuous dependent variable and one independent variable
Multiple Linear Regression	one continuous dependent variable and two or more independent variables
ANOVA (Analysis of Variance)	Compare the means of three or more groups Continuous dependent variable, and categorical independent variable
Linear Mixed Models	both fixed and random effects. It is used when data is collected in groups or clusters.
Generalized Linear Models	dependent variable that does not have a normal distribution
Generalized Linear Mixed Models	both fixed and random effects, and the dependent variable does not have a normal distribution

Summary



Number of Dependent variables (DV)
1
2+
MIG Workshop 5
Multivariate analysis
Assume normality for DV?
Yes
No
Generalized models
Number of Independent variables
1
2+
Are all categorical?
Yes
No
Factorial ANOVA
Multiple linear regression
Is numeric?
No
Yes
Simple linear regression
2+ groups?
No
Yes
ANOVA
Independent groups?
No
Yes
Paired T-test
T-test