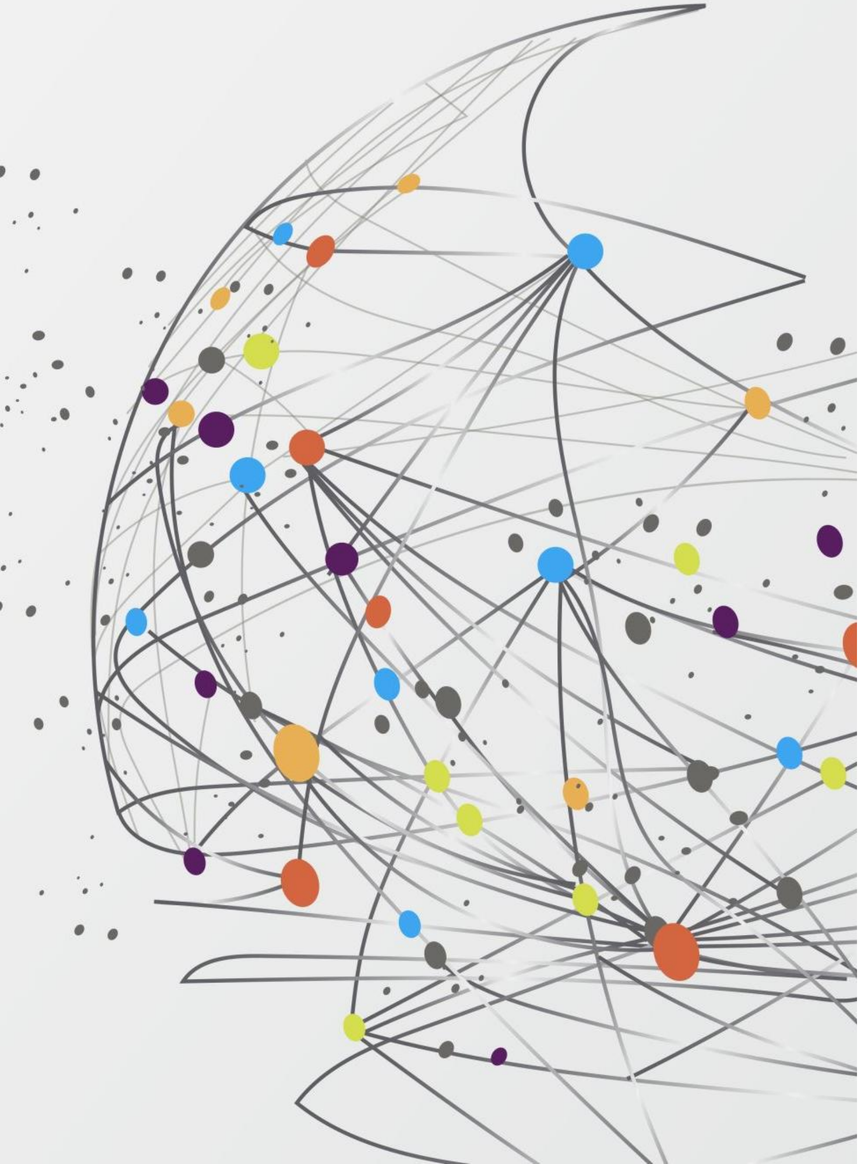
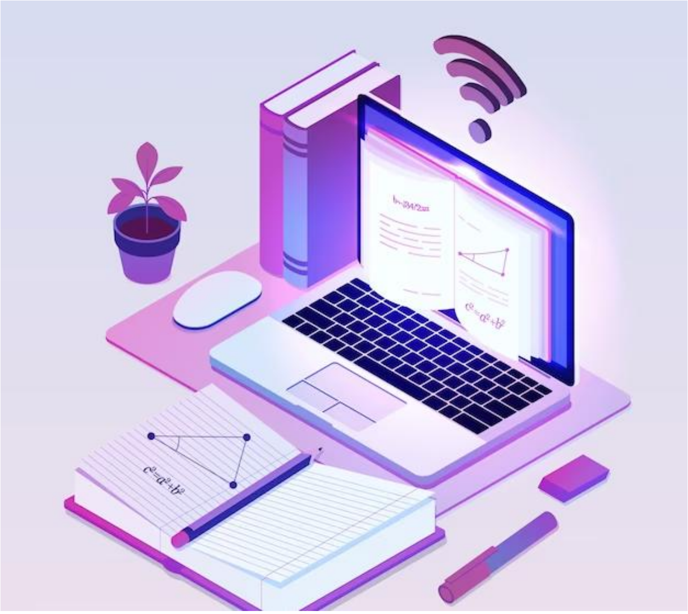


Intro to Linear Models

BioData Training School 2025
December 10-13, 2025, Tirana, Albania

Overview



- One and two-way ANOVA
- Multiple linear regression
- Fitting the model
- Checking the assumptions
- Assessing model performance

The problem

- We want to investigate the relationship between cholesterol levels and several predictors, including systolic blood pressure (SBP), diastolic blood pressure (DBP), age, BMI, race, and gender. The goal is to determine how these factors influence cholesterol levels and assess the strength of their associations.

What is the purpose of fitting a model?

- To explain the relationship between the response and the predictors.
- To predict the response based on the predictors. Often, a good model will do both.

One-way ANOVA as a linear model

One-Way ANOVA tests whether the means of **k groups** are equal.

It can be expressed as a **linear regression model** with categorical predictors.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_{k-1} X_{(k-1)i} + \varepsilon_i$$

ANOVA F-test = test that all $\beta_1 = \cdots = \beta_{k-1} = 0$

Y_i :response variable
X_{gi} :dummy variables for group membership
β_0 :mean of reference group
β_g :difference between group $g$and reference

$$\varepsilon_i \sim N(0, \sigma^2)$$

Two-way ANOVA as a linear model

Two-Way ANOVA evaluates effects of Factor A, Factor B, and their interaction.

Also fits naturally in a linear model framework.

$$Y_{ij} = \beta_0 + \alpha_a + \beta_b + \left(\alpha\beta\right)_{ab} + \varepsilon_{ij}$$

- Main effects: Do levels of A or B change the mean outcome?
- Interaction: Does the effect of A depend on B?
- F-tests compare each set of coefficients to 0.

Multiple Linear regression

- When there are two or more independent variables used in the regression analysis, the model is not simply linear but a multiple regression model.

Dependent Variable (Response Variable)

Independent Variables (Predictors)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_p X_p + \varepsilon$$

Y intercept

Slope Coefficient

Error Term

- Y is always continuous
- The covariates X can be:
  - Continuous variable (age, weight, etc.)
  - Dummy variables coding a categorical covariate.

Significance and interpretation of coefficients

- The coefficients can be interpreted after testing their significance.

- If the independent variable X_i increases by one unit and all other predictors are constant, the dependent variable Y increases by β_i.

Checking the assumptions

1. Linear relationship between the dependent and the independent variables.
2. Multicollinearity, no strong correlation between independent variables.
3. Residual values are normally distributed
4. Homoscedasticity assumes that the variance of the residual errors is similar across the values of each independent variable.

Check model performance

- **Coefficient of determination R-Square**

R-squared is the proportion of the variance in the response variable that can be explained by the predictor variables.

Variance of the predicted values

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2}$$

Variance of the observed values

$$R_{adj}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$$

- **Root mean squared error**

$$\mathrm{RMSE}(\mathrm{model}, \mathrm{data}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

y_i are the actual values of the response
y_i are the predicted values using the fitted model

Linear models step by step in R

Go to Practical_LinearModels R notebook