STA130 Winter 2022, Tutorial 6

Ente Kang

University of Toronto

2022-03-07

Linear Regression

- Also known as the *line of best fit*
- Key assumption:
    - There is a linear relationship between the predictor and target variables
    - We will see that this is often *violated*
- This week, we will study OLS with 1 predictor

The model

Assuming a linear relationship between y_i and x_i, there is a **population** regression line

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- β_0: Intercept
- β_1: slope
- ϵ_i: error (noise)

We want to approximate this using the data that we have, which will be our **fitted** regression line

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are found using optimization, by minimizing

$$\sum_i (y_i - \hat{y}_i)^2$$

Interpretation

- Assessing the model fit
  - R^2 is the metric that we will use
  - It is between 0 and 1
  - Tells us the percentage of variation in our y_i's that are explained by the regression line
- How do we interpret the model?
  - Recall: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - β_1: Given an unit increase in x_i, it is associated with a β_1 increase/decrease in y_i
- **Important**
  - This is **not** a causal relationship, only an **association**
  - **Do not say**, an unit increase in x_i **causes** a β_1 increase/decrease in y_i
  - To interpret it causally, we can use **econometrics**, which is out of the scope of this course. I am happy to discuss this with anyone who is interested

Thank you and see you next week