Tips for Lab Sessions

- Get preparation: get familiar / try to tackle the problems **before** coming to lab.
- Do **ALL** the problems!
- Format you answer: use **Markdown** to organize your answer / conclusion.
- Watch the time: the DDL is at **XX:20 PM**, not XX:30 PM.
- Learn to Google for **usage of basic functions**.
- Do NOT mail your answer: submit your work **ONLY** via **NTU-Learn**.

Lab 4.
Linear Regression

The Simplest Machine Learning Model

Linear Regression

- Goal: how to make **quantitative prediction** on Y given a good variable X (i.e. $\mathrm{corr}(X, Y)$ is high)?

- Rationale: **find coefficients** (a, b) such that

$$Y \approx a + b\,X$$

holds on training data.

- How: **fit** a linear regression model; **evaluate** the model performance on the testing data.

Linear Regression Workflow (Lab Mark Checkpoints)

1. Split your Dataset: **randomly split** the dataset into training v.s. test dataset.

2. Fit & Evaluate Linear Model: **fit** a linear regression model with **different variable** X on the **training** set; **evaluate** the model on the **test** set.

3. Model Selection: **find the 'best'** model according to some performance metric.

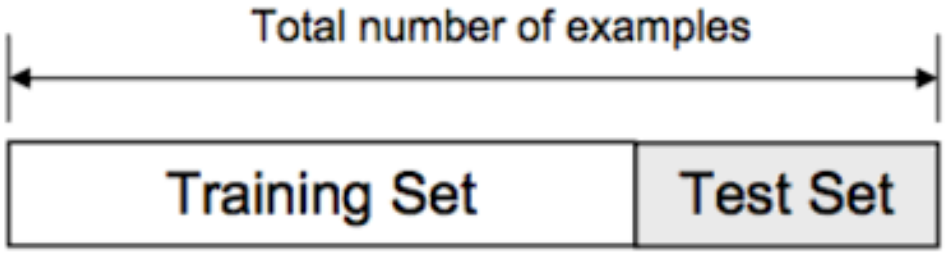
4. Refine model: fit regression model again on **outliers-free** data.

Mark Checkpoint 1: Split the dataset

- **Randomly** split the train / test dataset.

- Print the **shape** of train / test dataset

```
# Import the required function from sklearn
from sklearn.model_selection import train_test_split
```



Mark Checkpoint 2: Fit Linear Models

- **Fit** a linear model with **training data**.

```
# Import LinearRegression model from Scikit-Learn
from sklearn.linear_model import LinearRegression
```

- Google for basic attributes of LinearRegression().

- **Plot** the regression line with **test data.** Google for how to add plots on a plot.

Mark Checkpoint 3: Model Selection

- **Compute** goodness of fit: MSE, R^2 on **training & test** data.
- **Explain** in your word: Do you think this MSE is too large to be accurate?
- **Compare** multiple univariate models w.r.t **both** MSE and R^2.

Mark Checkpoint 4: Refine your Model

- **Identify** the row indices & **count** the number of outliers.
- **Visualize** the outliers via boxplot.

- **Remove** the outliers using drop() function.

- **Repeat** the previous split + fit + evaluate process with the clean data.

- **Compare** this model with previous ones.