**Machine Learning 2018**

Dependent variables

Independent variables

In machine learning, we tend to use independent variable to predict a dependent variable.

**PART 1**

**Data Preprocessing**

We need to do this in every model. Split it to test and train data.

**Feature Scaling**

Variables are not on the same scale.

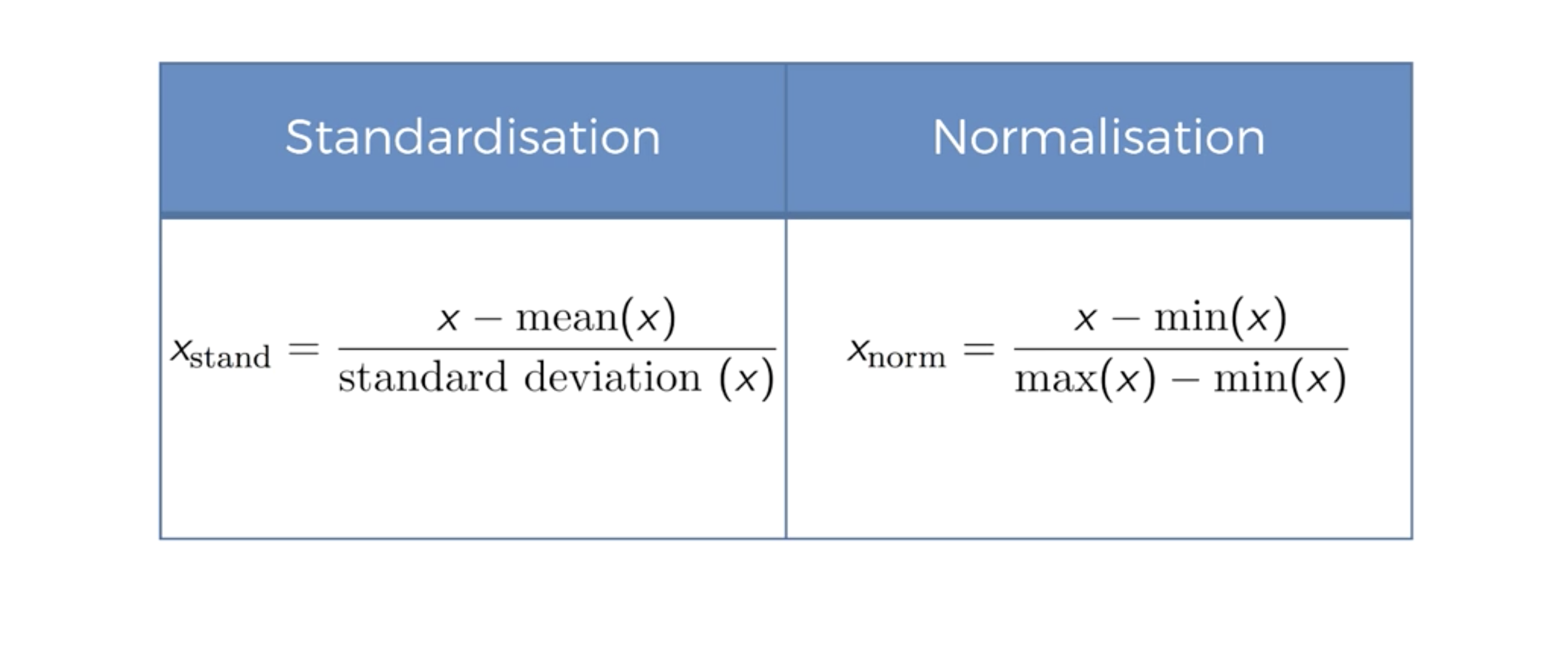
For example age from 27 to 44, or salary from 44 – 90k.

It is because distance metrics is Euclidean distance.

Age – x coordinate, Salary – y coordinate.

Euclidean distance calculated based on that.

For example, the distance will be dominated by the salary.



Fit\_transform()

Fit it to training set, then transform it.

Do we need to scale dummy variables?

* It depends on the context.
* Won’t break the model if you don’t scale.

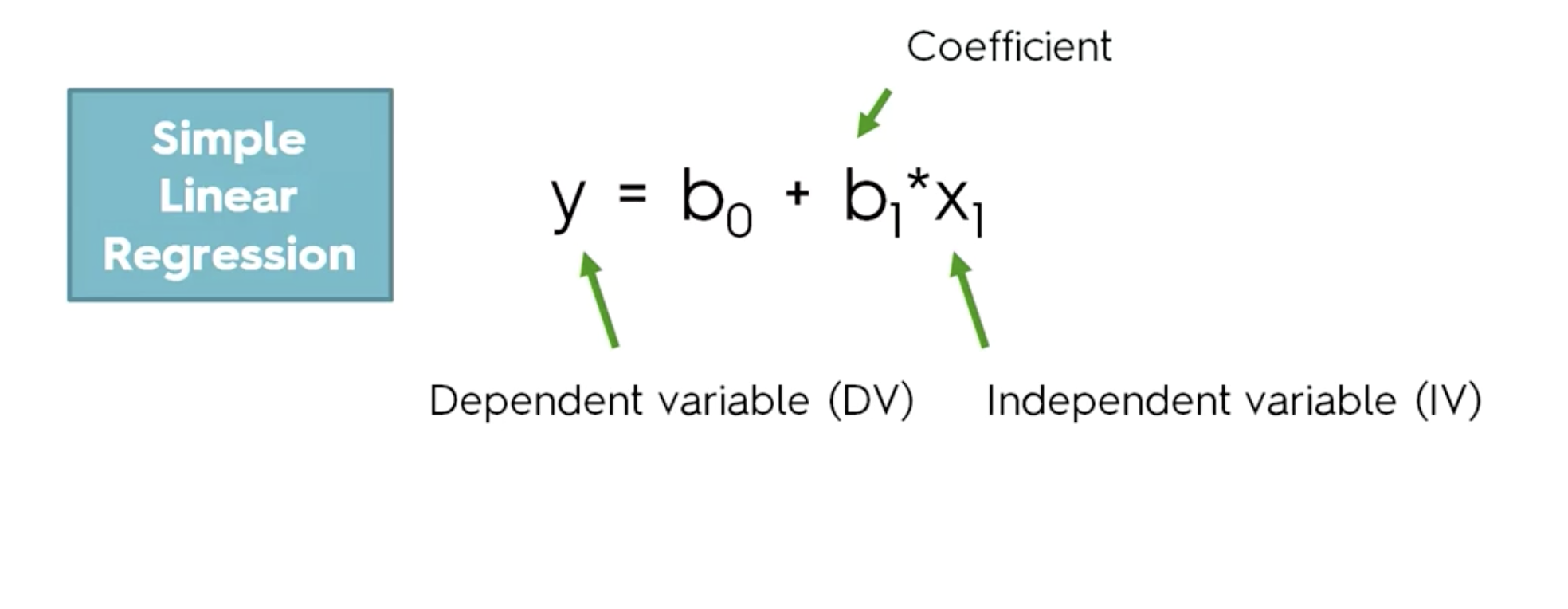
Now, all the variables are in the same range. Betweenn -1 and + 1 ish.

Algorithms may converge much faster.

**Part 2 – Regression**

**Part 2A – Simple Linear Regression**

Eg. Finding years of experience and salary.



Y = salary

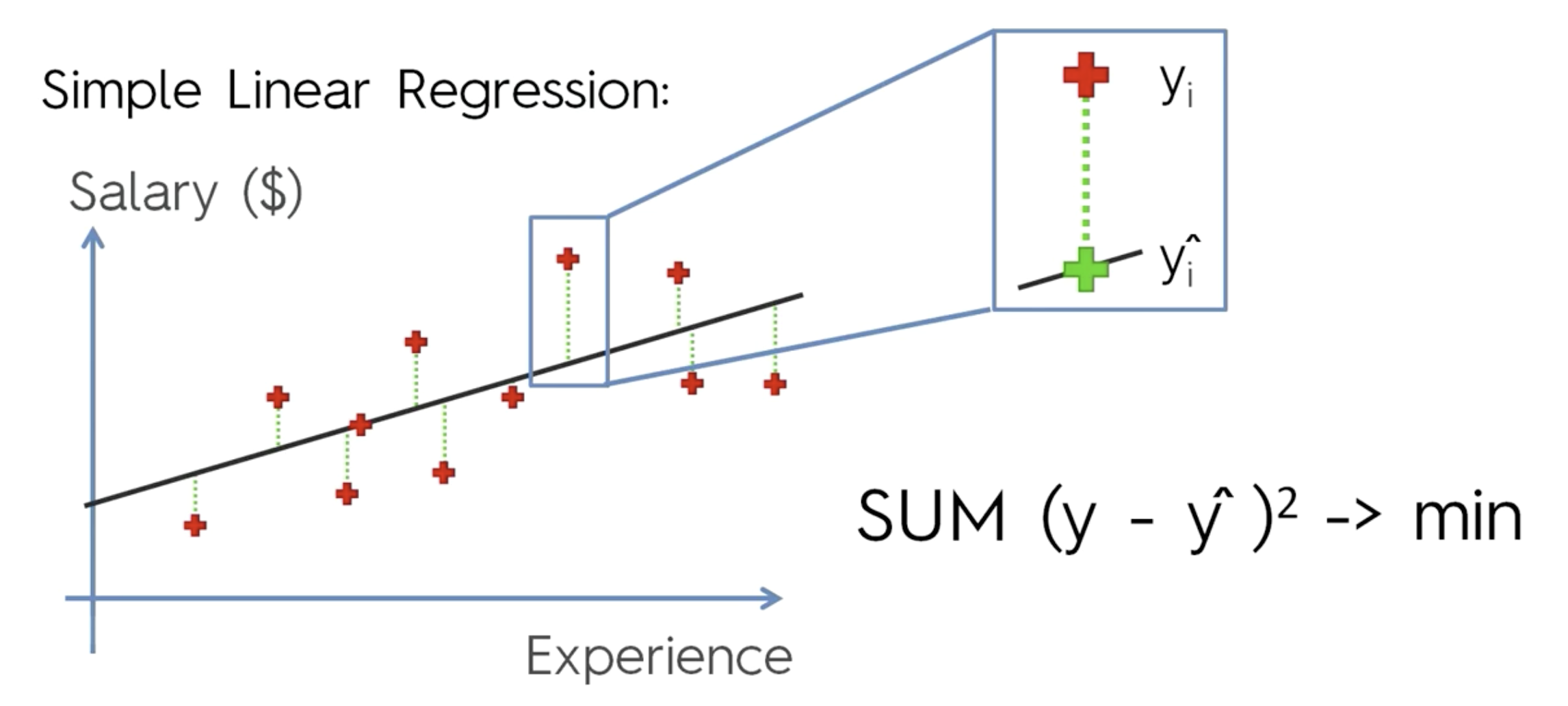
X1 = experience

Increase in x, how does it change y?

B0 is when the graph is at 0, for example no experience.

Yi and Yi hat.

Ordinary least squares



The model finds the least error.

The green is the difference between predicted and actual.

First thing we need to do is import the model from scikit learn.

Import the sklearn.linear\_model import Linear Regression



In R:

regressor = lm(formula = Salary ~ YearsExperience,

data = training\_set)

You can type summary(regressor) to find out the statistical data.

The number of stars indicate the statistical significance.

0 – no statistical significance

3 – high statistical significance

P value – lower value, higher significance, good value is 5%,

over 5% then its not really significant.

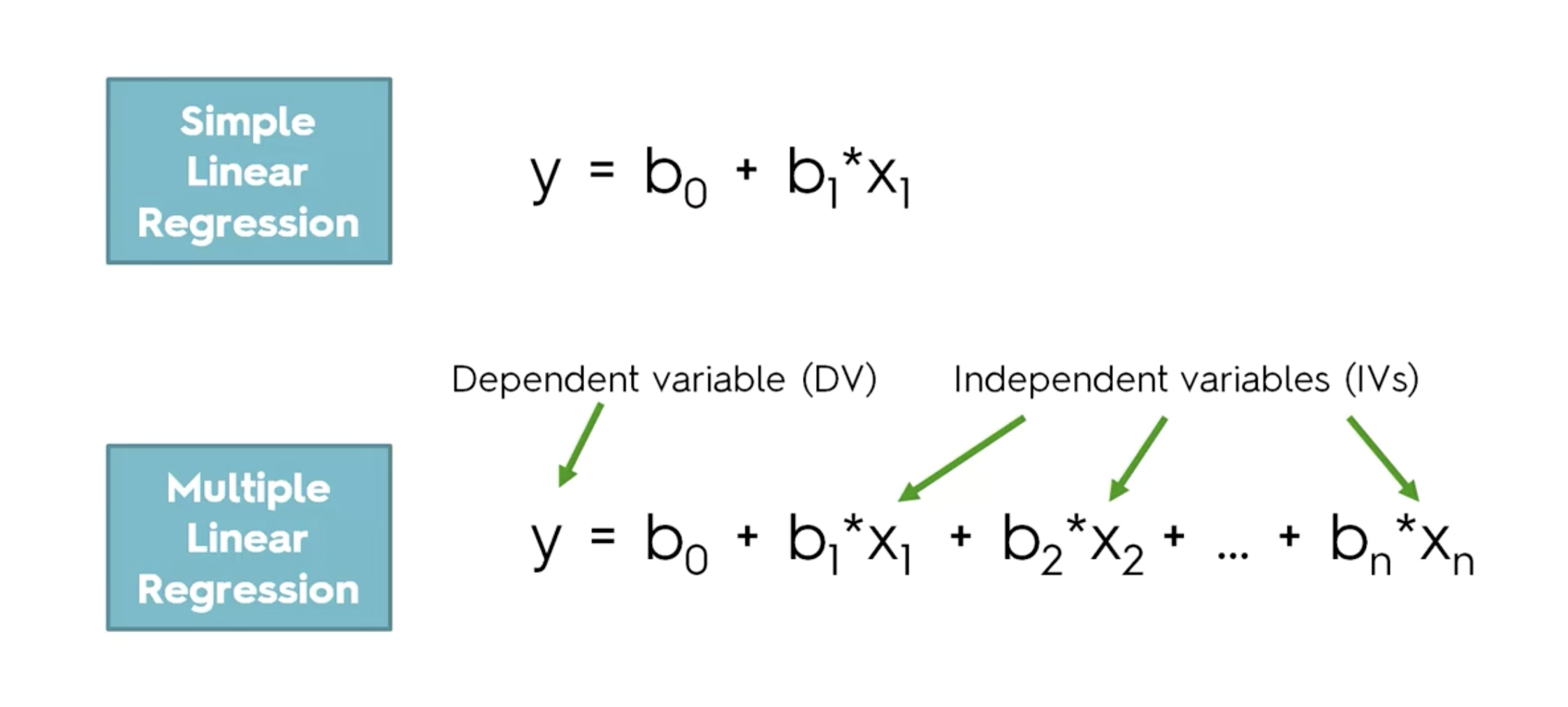
**Part 2B**

Multiple Linear Regression

Profit based on multiple areas.

Assumptions of Linear regression:

1. Linearity
2. Homoscedasticity
3. Multivariate normality
4. Independence of errors
5. Lack of multicollinearity



Duplicating a variable.

Multiple linearity. Dummy variable trap.

Cannot have constant and have 2 variables.

Like New York and California – 0 and 1 respectively.

Always omit one dummy variable.

So say b1 = b1 – b2

Its always a 1 or a 0, and they are the same.

Building a model.

1. Garbage in – garbage out : if there are a lot of variables there may be many garbage stuffs.
2. At the end of the day, you will need to explain these variables. Like why each variable affects the dependent variable. It might not be practical to explain 1000s of them.

We need knowledge of p-value.

5 methods of building models.

1. All in
2. Backward elimination
3. Forward selection
4. Bidirectional elimination
5. Score comparison

2,3,4 are stepwise regression. Some people may also say that 4 is bidirectional elimination.

**1. All in cases**

Just throw in all your variables.

When to use?

* When you have prior knowledge, domain knowledge, or someone gave you these variables.
* Or you simply have to.
* Preparing for backward elimination.

**2. Backward elimination**

Step 1: select a significance level

Step 2: fit the full model with all possible predictors

Step 3: consider the predictor with the highest p-value (if p> SL, don’t go to step 4, finish the iteration)

Step 4: remove predictor

Step 5: fit model without variable.

Repeat step 3.

This sounds like leave one out validation in machine learning class taken in comp30027.