Introduction to Regression Analysis

Regression Analysis is a supervised machine learning algorithm using which we try to find *linear correlations between an outcome variable Y and a set of regressors X*, where Y is also called target variable or dependent variable while the set of regressors is also called features or independent variables.

We represent X as (x1, x2, ...xp)' where x1, x2.. represent some features.

For example, You want to build a regression model to predict the **hourly wages of a worker** using variables like **education level, gender, experience, skillset etc**.

Here, the hourly wage is the outcome variable (Y) that needs to be predicted and education level, gender, experience, skillset etc. are a set of regressors or a set of features (X) using which we want to predict the hourly wages.

Using regression analysis we can answer 2 questions,

1. **The Prediction question:** How can we use the set of regressor X to predict Y well?
2. **The Inference question:** How does the predicted value(Y) change if we only change 1 component or 1 feature of X, keeping all other components constant.

So in the above hourly wage example:

The Prediction question will be: how will we use the set of regressors X, namely **education level, gender, experience, skill set,** etc, to predict hourly wages (Y). The inference question will become, for example: how does gender affect the wage of a worker?

For the inference question we divide the set of target regressors X into 2 parts:

X = (D,W)

Where **D is the target regressor;** in the above example, **gender** is the target regressor as we want to check its effect by keeping all other job-relevant features constant.

W represents controls or confounders: which are the remaining job-relevant characteristics, namely: **education level, experience, skillset etc.**