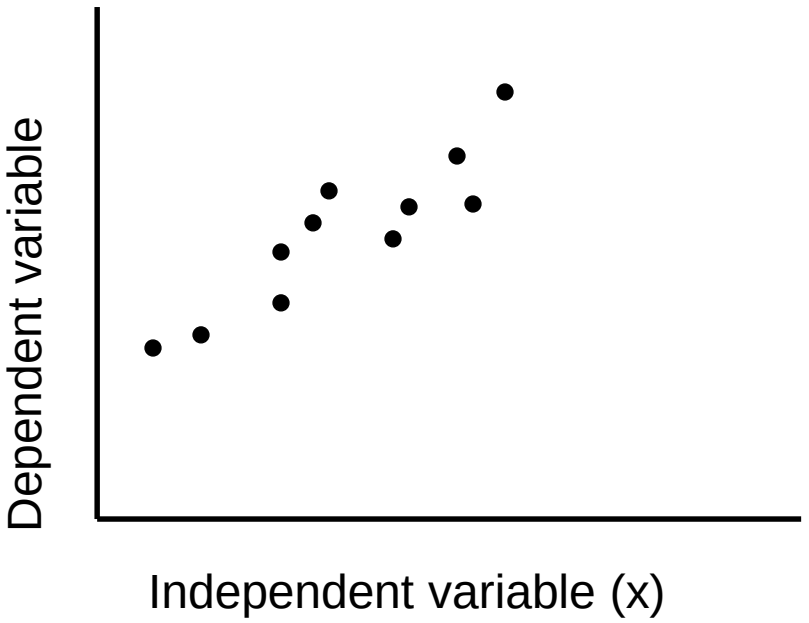
Machine Learning

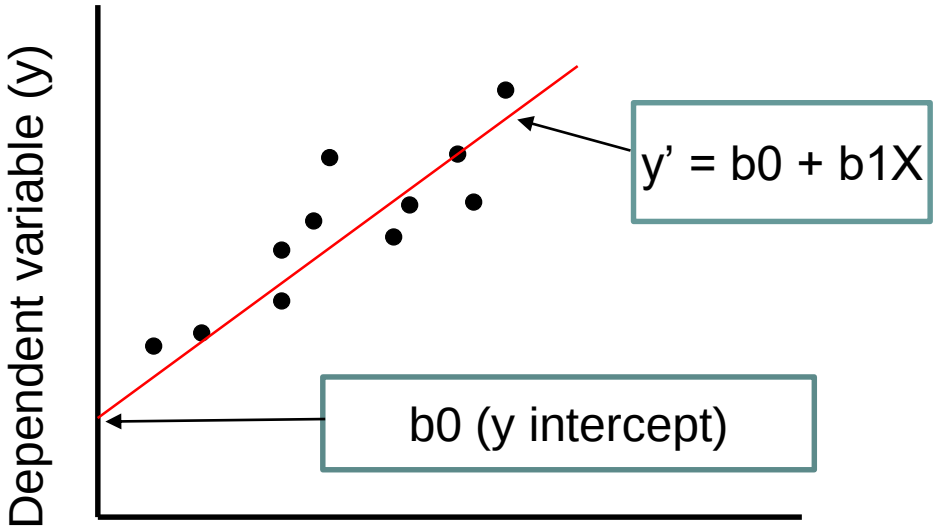
Regression

Regression

- Regression is the attempt to explain the variation in a dependent variable (or response variable) using the variation in independent (explanatory) variables

- If the independent variable(s) sufficiently explain the variation in the dependent variable, the model can be used for prediction.





Simple Linear Regression

- The term "linear" in the name “linear regression” refers to the fact that the method models data with linear combination of the explanatory (independent) variables.

- In simple linear regression there is only one explanatory variable

- Simple linear regression can be expressed in any of the following ways:

  - response = intercept + constant ∗ explanatory
  - y = c + m*x (more commonly y = m*x + c)
  - y = a + b * x
  - $y = \beta_0 + \beta_1 * x_1$

- In its most basic form fits a straight line. The model is designed to fit a line that minimizes the squared differences (also called errors or residuals.)

Correlation

- Before generating a regression model, we need to understand the degree of relationship between Y and X

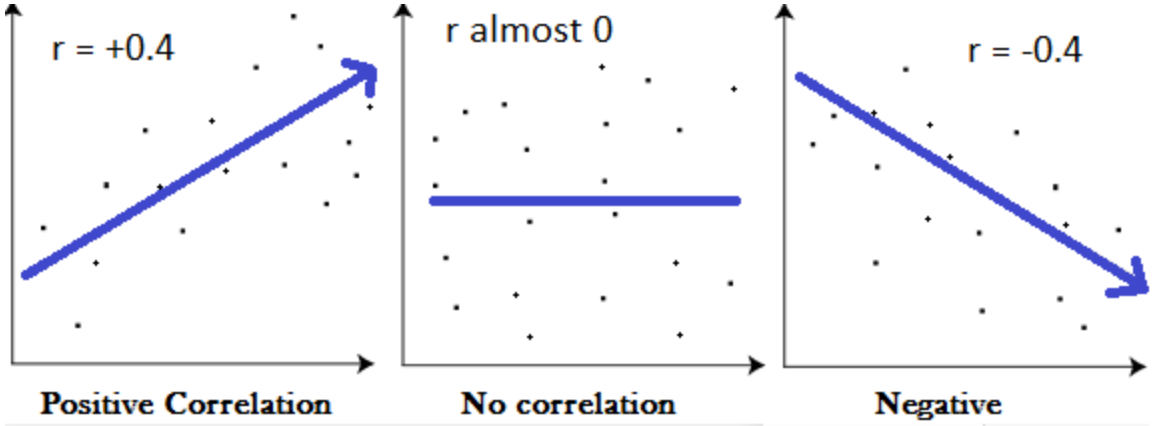
- Correlation between two variables indicates how closely their relationship follows a straight line. Pearson's correlation coefficient is commonly used to measure strength of linear relationship. It ranges between -1 and +1.

- Correlation of extreme possible values of -1 and +1 indicate a perfectly linear relationship between X and Y whereas a correlation of 0 indicates absence of linear relationship. Practically, we may not observe such a perfect relationships in business data.

Coefficient of Correlation

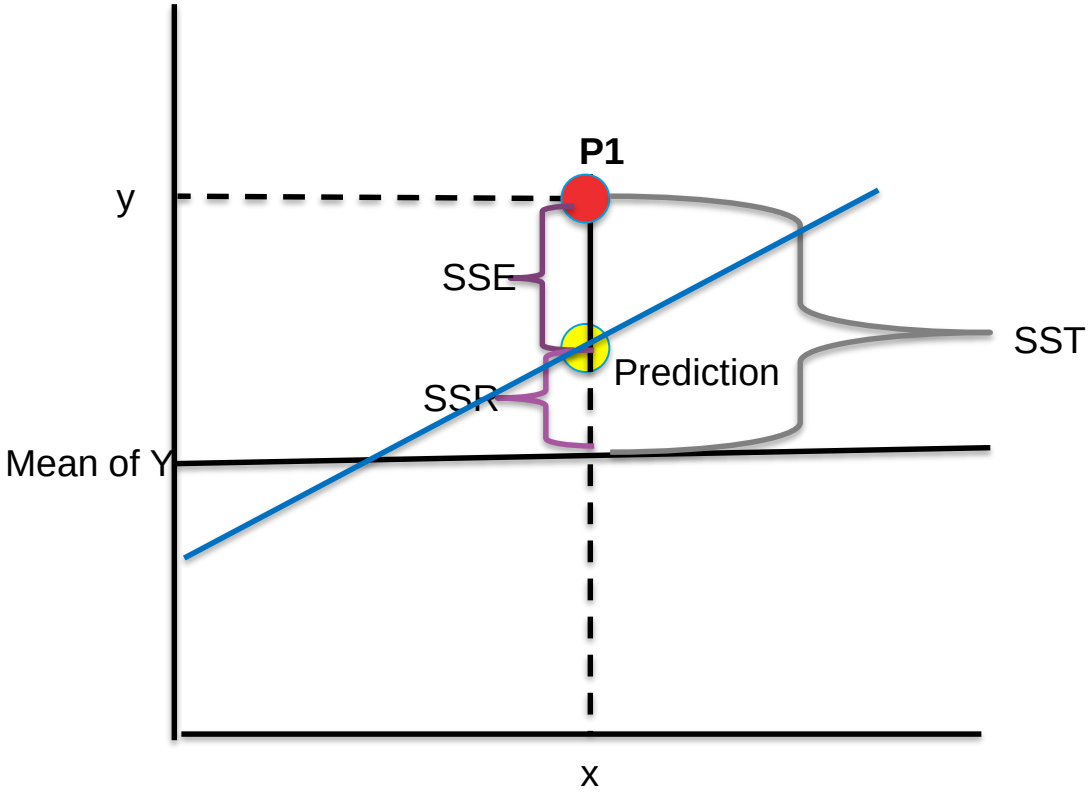
$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

- Cov(x,y) is covariance of x and y
- Sx is standard deviation of x
- Sy is standard deviation of y



Regression

- If there is meaningful correlation between x and y, we need to fit a line to build a model. But there are infinite number of lines that can be fit. Which one should we consider as the model
- The line with the lowest total sum of squared prediction errors is considered as best fit line
- This value is called **the Sum of Squares of Error, or SSE**.



Regression

- The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.

- The proportion of total variation (SST) that is explained by the regression (SSR) is known as the Coefficient of Determination, and is often referred to as R^2 .

- $R^2 = \frac{\mathbf{SSR}}{\mathbf{SST}}$

- The value of R^2 can range between 0 and 1, and the higher its value the more accurate the simple linear regression model is. It is often referred to as a percentage.

Multiple Linear Regression

- More than one independent variable can be used to explain variance in the dependent variable, as long as they are not linearly related.

- A multiple regression takes the form:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

where k is the number of variables

Feature Selection

- Multiple methods exist for feature selection. A common method is “Backward Selection” or “Backward Elimination”

- In this method, model starts with all X variables in the model. At each step, the X that is the least significant (highest p-value) is removed. Continue the process until all variables are significant. The user decides the significance level (generally 0.05).

- Other methods include “Forward Selection”, “Stepwise Selection”

- All these methods face criticism regarding reliability of p-value, especially with multicollinearity

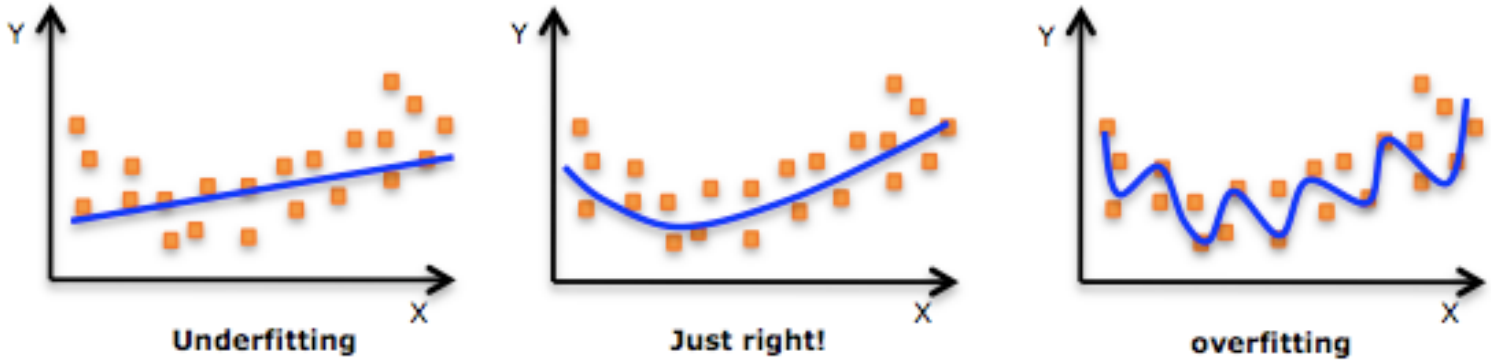
Linear Regression

- Advantages –
  - Very intuitive and easy to understand method
  - Simple to implement

- Disadvantages -
  - Linear regression models relationships between dependent and independent variables that are linear
  - Outliers can have significant effect of regression model
  - Assumes no multicollinearity - independence between attributes

Overfitting

- Overfitting:
  - Learn the “data” and not the underlying function
  - Performs well on the data used during the training and poorly with new data.

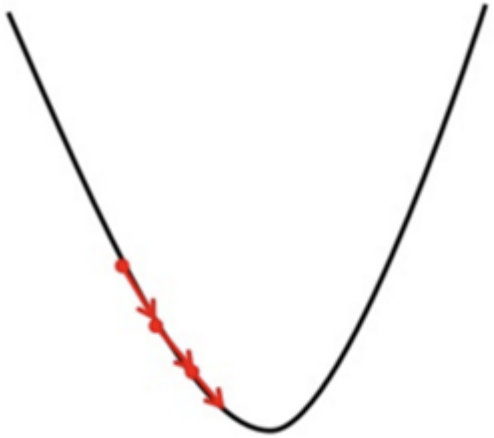


Gradient Descent

- In linear regression, targets is to get the best-fit regression line – with minimum Root Mean Squared error between the predicted value (pred) and true value (y).

- This is achieved using Gradient Descent algorithm.

- Initially chosen values of are refined in the direction of minima of the Root Mean Square error



Hands on Exercise

Regression

Dummy Variable Regression

- Independent variables can categorical variables, for example
  - Gender
  - Brand of laptop
  - Nationality

- Since algorithms expect numerical values in independent variables, these need to be encoded

- Discussion – what can be a problem if, for example, brand of laptop are coded as follows: HP=1, Dell=2, Lenovo=3, Asus=4, Acer=5

Dummy Variable Regression

- The correct way to encode the categorical variables is by using dummy variables
- A dummy variable takes on 1 and 0 only
- If a categorical variable has n possible values, then create n-1 dummy variables
- For example, if Laptop brand can take following 5 values HP, Dell, Lenovo, Asus, Acer; then create 4 dummy variable: Brad_HP, Brand_Dell, Brand_Lenovo, Brand_Asus. Each of these variable can take value 0 or 1