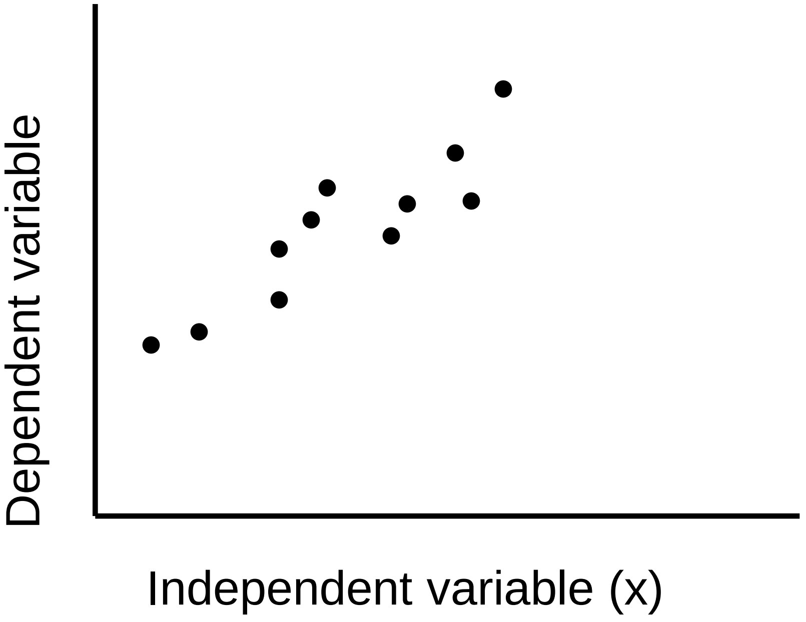
Machine Learning

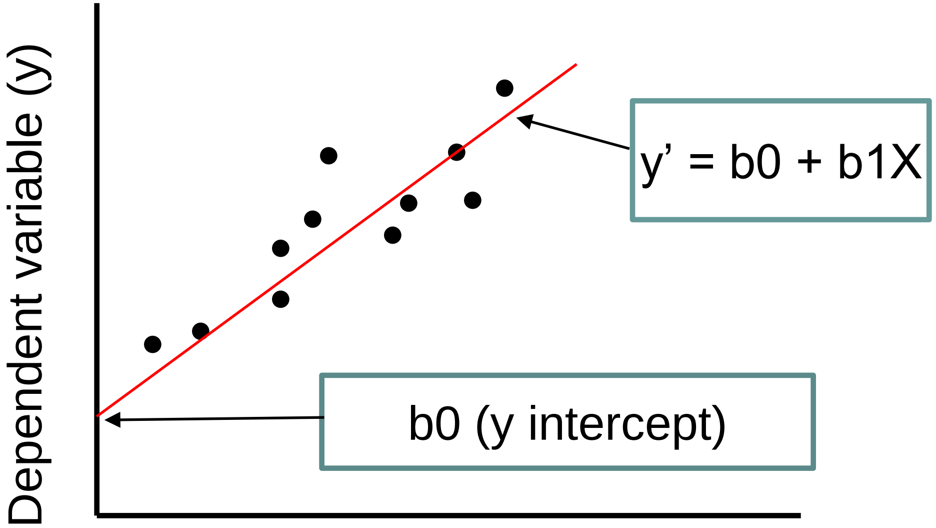
Linear Regression

Linear 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.





y’ = b0 + b1X

b0 (y intercept)

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)
  - $y = b_0 + b_1 * x_1 + \ldots\ldots.$
  - Predicted y is represented as $\hat{Y}$
  - Linear Regression in its **basic form** fits a straight line. The model is designed to fit a line that minimizes the squared errors (also called residuals.)

Correlation

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

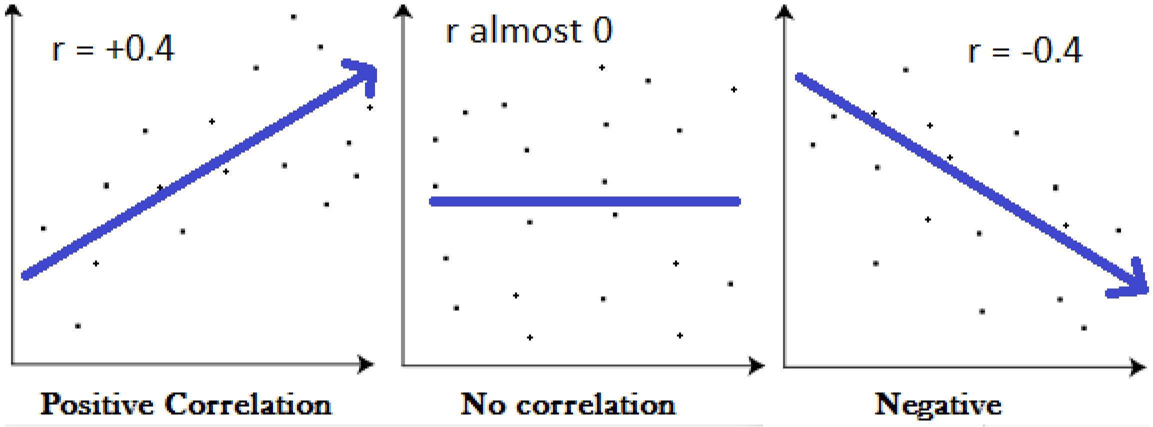
- Correlation between two variables indicates how closely their relationship follows a straight line. Pearson's correlation coefficient is commonly used to measure strength of linear relationship. It ranges between -1 and +1.

- 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

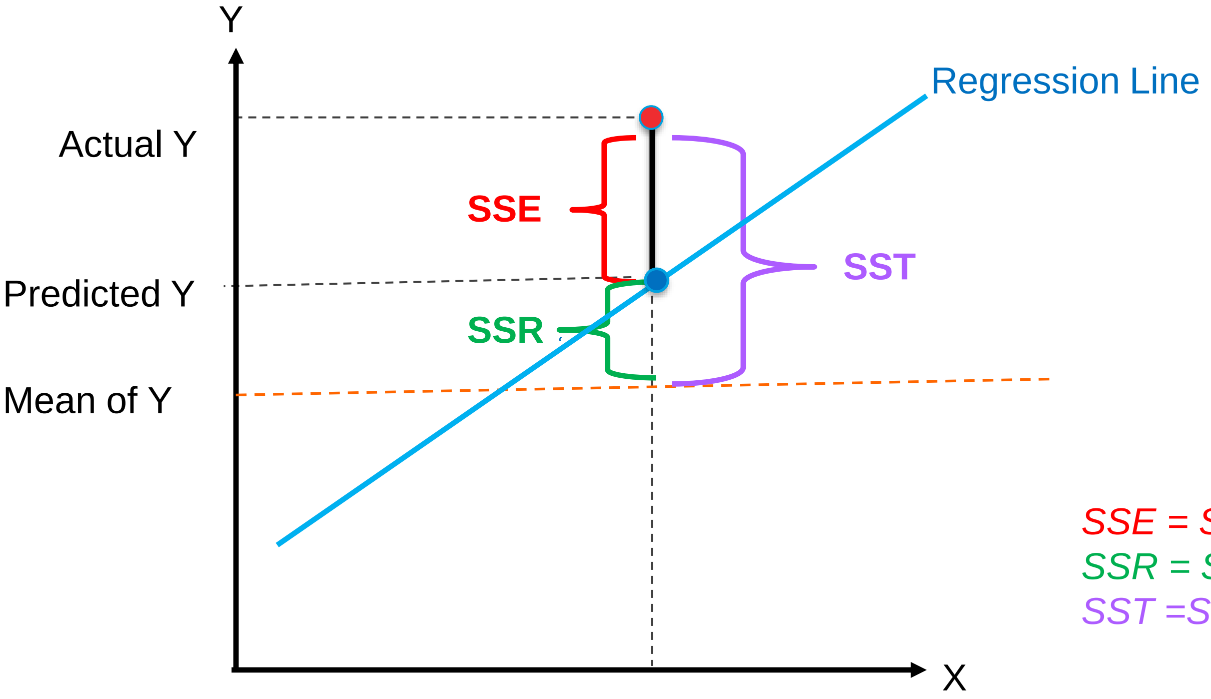
$$r = r_{xy} = \frac{\mathrm{Cov}\left(x, y\right)}{S_x \times S_y}$$

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



Linear Regression

- Linear regression fits a line to build a model. There are infinite number of lines that can be drawn through the training points. Which line should we consider as the model
- The line with the **least value total sum of squared prediction errors** (in diagram shown as SSE) is considered as best fit line. This is called **Cost Function** or **Loss Function**



SSE = Sum of Squares of Errors
SSR = Sum of Squares Regression
SST =Sum of Squares Total

Linear 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.

- Coefficient of Determination is denoted as R^2 .

- $R^2 = \frac{\mathbf{SSR}}{\mathbf{SST}} = 1 - \frac{\mathbf{SSE}}{\mathbf{SST}}$

- 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.

Model Performance Measurement

- Performance if Linear Regression Model is represented using
  - Coefficient of Determination (R^2). Higher value of is R^2 indicates better accuracy
.
  - Mean Squared Error (MSE). Lower value of MSE indicates better accuracy

  - The above are most important measures of regression performance. There are other measures such as RMSE, Mean Absolute Error

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:

$$y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

where k is the number of variables

Feature Selection

- Some of the predictor variables (independent variables) could be insignificant (not useful) for the regression model

- Such features can be detected using "*Statsmodel*" library

- "*OLS*" (Ordinary Least Squares) function performs Hypothesis testing for each predictor variable and provides 'P' value

- If the "P" value is greater or equal to 0.05, then the predictor variable is insignificant. Such predictor variables can be dropped from the model

- Drop the predictor variable with highest 'p' value and repeat the process to check whether all predictor variables are significant for the model

- This process is described in detail on the next slide

Feature Selection

- Multiple methods exist for feature selection. A common method is “Backward Selection” or “Backward Elimination”
  - *Step-1: In this method, build a model with all X variables included in the model*
  - *Step-**2A**: P-value of each X variable is checked. P-value >= 0.05 indicates that the variable is not significant.*
    - *If there are Xs with p-value >= 0.05, then drop the X which has **highest p-value** (i.e. drop only one X variable which is least significant). Perform Step-2B*
    - *If all Xs have p-value < 0.05 then feature selection is complete*
  - *Step-**2B**: After dropping the least significant X, build the model again and Perform Step-2A.*
- Other methods include “Forward Selection”, “Stepwise Selection”
- These methods face challenge regarding reliability of p-value, when **multicollinearity** exists among Xs (Refer Appendix - VIF for additional information)

Linear Regression

- Advantages –
  - Very intuitive and easy to understand method
  - Regression equation coefficient provide insight about impact of predictors on Target variable

- Disadvantages -
  - Linear regression models relationships between dependent and independent variables that are linear
  - Outliers can have significant effect of regression model
  - Assumes no multicollinearity, i.e. there is no strong correlation between independence variables (Xs)
  - All the assumptions of Linear Regression are often not satisfied

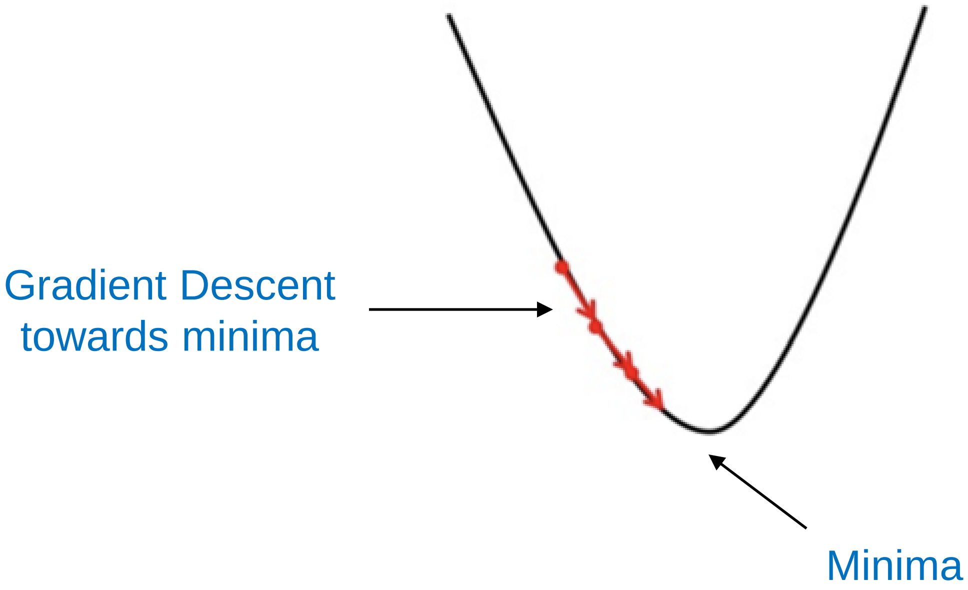
(***further study – Assumption of Linear Regression***)

Gradient Descent

- In linear regression, targets is to get the best-fit regression line – with minimum Mean Squared error

- This is achieved using Gradient Descent algorithm.

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



Gradient Descent towards minima

Minima

Hands on Exercise

Linear 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

Appendix

(For Reference)

Assumptions in Linear Regression

- Some of the key assumptions are:
  - Linear relationship between the features and target variable
  - Little or no Multicollinearity between the features
  - Homoscedasticity Assumption: .A scatter plot of residual values vs predicted values is a good way to check.
  - Normal distribution of error terms
  - Little or No autocorrelation in the residuals

VIF - Variable inflation Factor

- Useful in Linear Regression to check multicollinearity among independent variables
- Multicollinearity results in coefficients becoming unstable
- Multicollinearity identifies independent variables that can be predicted using linear regression of other independent variables
- VIF = $1/(1-R^2)$
- Independent variables with VIF >= 5 (common cut off) are candiated to be dropped.
- Drop only one variable which highest p-value (Do take into consideration the domain knowledge while deciding which column to drop) and check VIF values again. Repeat until all VIF values are < 5

Adjusted R-squared

- Adjusted R-squared is useful to overcome a limitation of R-Squared

- Adjusted R-squared = $1 - (1 - R^2) * \dfrac{n - 1}{n - k - 1}$

n = number of row in the training data

k = number of independent variables

- R-squared increases when we add new predictors to the model, even if the newly added predictors are not significant and don’t add any value to the predicting power of the model

- Adjusted R-squared increases if the newly added predictor improves the predicting power of the model.

Thank you

- Prashant Koparkar