Linear Regression

Statistical Learning Theory

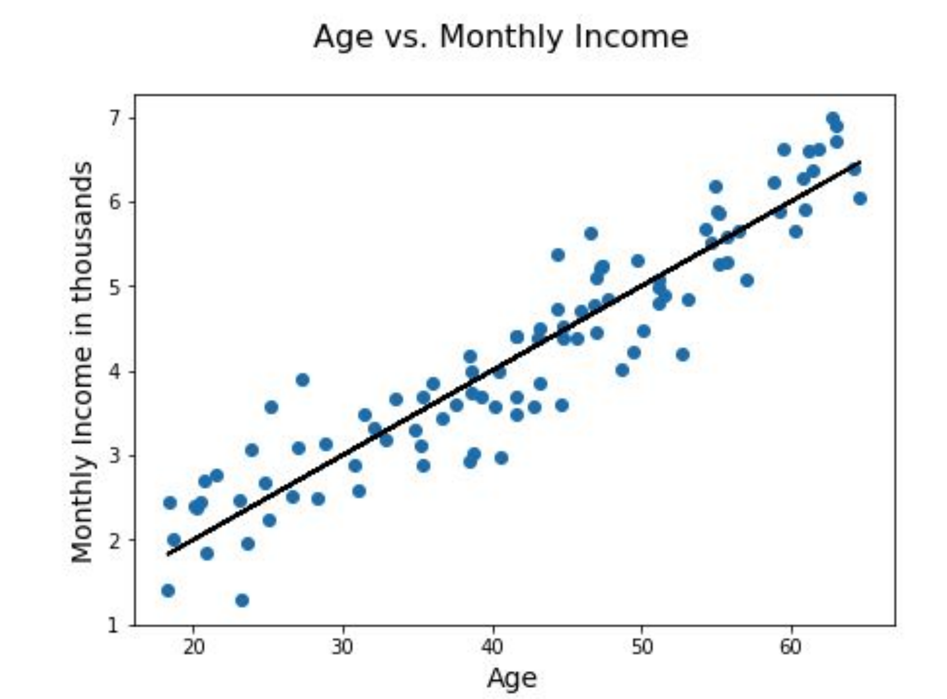
- Statistical Learning Theory is based on the idea of using data along with statistics to provide a framework for learning.
- In Statistical Learning Theory, the main idea is to construct a model to draw certain conclusions from data, and next, to use this model to make predictions.

Different types of data variables

- Dependent variables: data that can be controlled directly (other names: outcome variables, target variables, response variables)
- Common examples include income, price of a home, temperature, etc.
- Independent variables: data that cannot be controlled directly (other names: predictor variables, input variables, explanatory variables, features)
- Common examples are age, day of the week etc.

Relationships between variables

- As the dependent variable depends on the independent variable, when plotting data it is generally shown on the y-axis.
- The independent variable is shown on the x-axis



Statistical Modeling

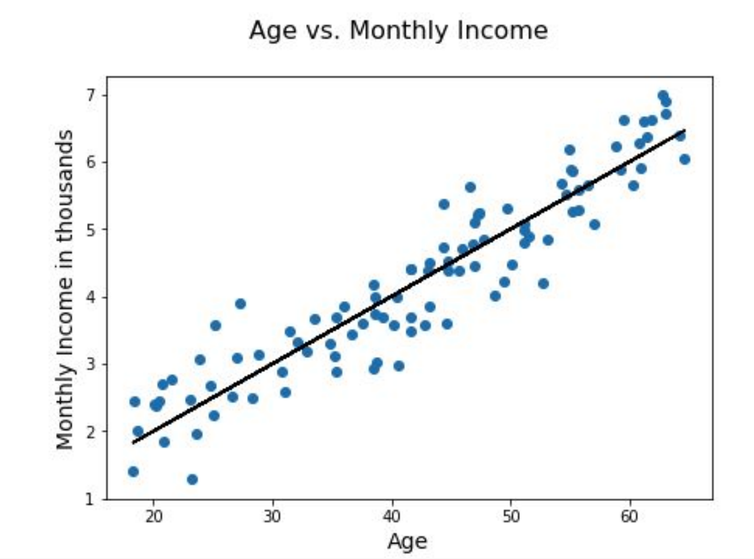
- A statistical model can be thought of as some kind of a transformation that helps us express dependent variables as a function of one or more independent variables.
- A statistical model defines a relationship between a dependent and an independent variable

Linear Regression

- Regression analysis is a way to estimate the relationship between a dependent variable (eg income) and one or more independent variables (eg age, etc.)
- The most common form of regression analysis is linear regression in which we only explore linear relationships between the dependent and independent variables.

Simple Linear Regression

- Simple Linear Regression uses a single feature (one independent variable) to model a linear relationship with a target (the dependent variable) by fitting an optimal model (i.e. the best straight line) to describe this relationship.
- Looking at the same plot from slide 4, we can see the linear relationship denoted by the line drawn which shows that as age increases, the income tends to increase



Simple Linear Regression

- A simple linear regression model would fit a line to the data points as shown in the previous slide. This line can then be used to describe the data and conduct further experiments using this fitted model. A straight line can be written as : ***y = mx + c.***
- There are 4 main components here:
    - A dependent variable that needs to estimated and predicted (here: y)
    - An independent variable, the input variable (here: x)
    - The slope which determines the angle of the line. Here, the slope is denoted as m.
    - The intercept which is the constant determining the value of y when x is 0.

Limitations of Simple Linear Regression

- Examines only the relationship between one independent variable and the dependent variable
- In many instances, we believe there to more than one independent variable to influence the dependent variable.
- For example, income is not only dependent on age but other factors like education level, seniority in a company, years of experience in the industry as well which simple linear regression fails to account for.

Multiple Linear Regression

- Multiple Linear Regression uses more than one feature to predict a target variable by fitting the best linear relationship.
- Going back to our previous example, we know that age is not the only factor that drives income. Other factors that can play a role are, among others, the number of hours worked, years of education, and the city of employment

Multiple Linear Regression

- Here is what our previous model will look like after adding in years of education:

  estimated monthly income=slope_sen*seniority+slope_ed*years_of_education+intercept

- When thinking of lines and slopes statistically, slope parameters associated with a particular predictor *xi* are often denoted by βi. Extending this example mathematically, you would write a multiple linear regression model as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_n x_n$$

  Where n is the number of predictors, β0 is the intercept and y is the predicted value of the dependent variable

Building linear models

- Lets now switch over to the jupyter notebook to talk more about how linear regression models are built and how to interpret them.