Machine Learning

Learning with Regression and Trees

Satishkumar L. Varma

Department of Information Technology
SVKM's Dwarkadas J. Sanghvi College of Engineering, Vile Parle, Mumbai.
ORCID | Scopus | Google Scholar | Google Site | Website

Outline

- Learning with Regression and Trees
  - Learning with Regression
    - Simple Linear Regression
    - Multiple Linear Regression
    - Logistic Regression
  - Learning with Trees
    - Decision Trees
    - Constructing Decision Trees using Gini Index
    - Classification and Regression Trees (CART)

Types of Regression

- Regression models used to find the relationship between a DV and IV.
- Simple linear regression
  - To models the relationship between a DV and a single IV.
- Multiple linear regression
  - If you have more than one independent variable.
- Multiple Regression vs. Multivariate Regression
- Multiple Regression:
  - The influence of several IVs on a DV is examined.
  - One DV is taken into account to analyzed.
- Multivariate Regression:
  - Several regression models are calculated to allow conclusions to be drawn about several DV.
  - Several dependent variables are analyzed.

Simple Linear Regression

$$\hat{y} = b \cdot x + a$$

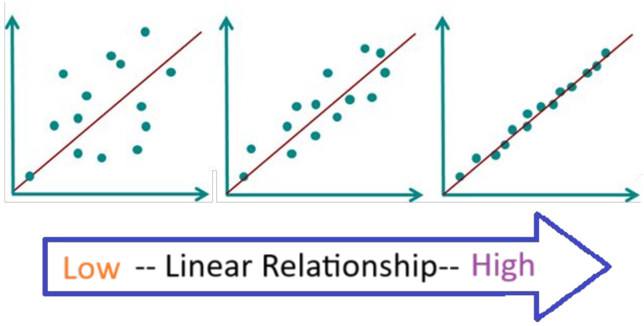
➡

Multiple Linear Regression

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

Simple Linear Regression

- Simple linear regression is used to estimate the relationship between two quantitative variables.
  - Models the relationship between a DV and a single IV.
  - DV must be a continuous/real value.
  - IV can be measured on continuous or categorical values.
- Simple linear regression is used:
  - To predict the value of a DV based on an IV.
  - To know, How strong the relationship is between two variables
  - (e.g., the relationship between rainfall and soil erosion).



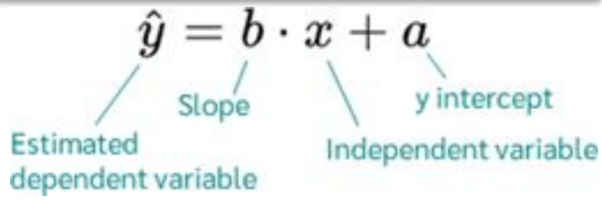
- The greater the linear relationship between the IV and the DV, the more accurate is the prediction.
- The greater the linear relationship between the DV and IVs, the more the data points lie on a straight line.
- In linear regression analysis, a straight line is drawn in the scatter plot.
- Visually, the relationship between the variables can be shown in a scatter plot.
- To determine this straight line, linear regression uses the method of least squares.

Simple Linear Regression

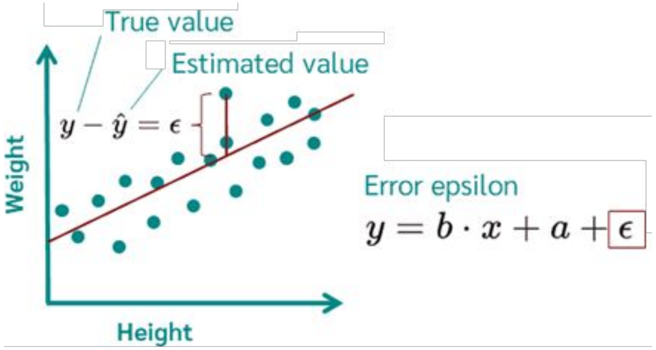
- Regression models describe the relationship between variables by fitting a line.
- It uses a straight line
  - while logistic and nonlinear regression models use a curved line.
- It allows you to estimate how a dependent variable changes as the independent variable(s) change.
- If you have more than one independent variable, use multiple linear regression.
- **Example**: Simple Linear Regression
  - To know the relationship between income and happiness.
  - Let us survey 500 people whose incomes range from 15k to 75k and
  - Ask them to rank their happiness on a scale from 1 to 10.
    - Independent variable: income
    - Dependent variable: happiness
  - The value of the dependent variable at a certain value of the independent variable
    - (e.g., the amount of soil erosion at a certain level of rainfall).
- As both variables are quantitative, we can perform a regression analysis
  - to see if there is a linear relationship between them.

Regression Line

- The regression line can be described by the following equation:



- Definition of "Regression coefficients":
  - a : point of intersection with the y-axis
  - b : gradient of the straight line
  - ŷ is the respective estimate of the y-value.
- This means that for each x-value the corresponding y-value is estimated.
- In our example, this means that the height of people is used to estimate their weight.
- y= a0 + a1x + ε
  - a0= It is the intercept of the Regression line (can be obtained putting x=0)
  - a1= It is the slope of the regression line, which tells whether the line is increasing or decreasing.
  - ε = The error term. (For a good model it will be negligible)



Error Estimation: Fitting Regression Line

- Error in the estimation while fitting a straight line.
  - No Error (perfect estimate)
    - If all points (measured values) were exactly on one straight line.
    - However, this is almost never the case;
  - Error (distance between the estimated value and the true value)
    - Need to find a straight line by keeping the error as small as possible
    - This distance or error is called the "residual",
    - Error is abbreviated as "e" (error)
    - Error is also represented by the greek letter epsilon (ϵ)

Error Estimation: Fitting Regression Line

- Calculate Regression line
  - To determine the regression coefficients (a and b)
  - so that the sum of the squared residuals is minimal.
  - OLS – "Ordinary Least Squares"
- The regression coefficient b can now have different signs, which can be interpreted as follows
  - $b > 0$: there is a positive correlation between x and y (the greater x, the greater y)
  - $b < 0$: there is a negative correlation between x and y (the greater x, the smaller y)
  - $b = 0$: there is no correlation between x and y
- Standardized regression coefficients are usually designated by the letter "beta" **β**.
- These are values that are comparable with each other.
- Here the unit of measurement of the variable is no longer important.

Ordinary Least Squares

- Ordinary Least Squares regression (OLS), often called linear regression.
- OLS is a technique for estimating coefficients of linear regression equations.
- Equation describe the relationship between one or more IV and a DV.
- OLS is often evaluated using r-squared (R^2).
- Least squares stand for the minimum squares error (SSE).
- **Example:**
- It takes into account the sum of squared errors instead of the errors
- As it sometimes can be -ve or +ve leads to nearly null value.
  - x = {2, 3, 5, 2, 4 } and predicted values, y = {3, 2, 5, 1, 5}
  - Total error = (3-2)+(2-3)+(5-5)+(1-2)+(5-4) = 1-1+0-1+1=0
  - Average error = 0/5 = 0 (it could lead to false conclusions)
- So, use the mean squared error
  - Total error = (3-2)^2+(2-3)^2+(5-5)^2+(1-2)^2+(5-4)^2 = 4
  - Average error = 4/5 = 0.8
  - Scaling the error back to the data; sqrt(0.8) = 0.89
  - That is the predictions differ by 0.89 from the real value.

Ordinary Least Squares

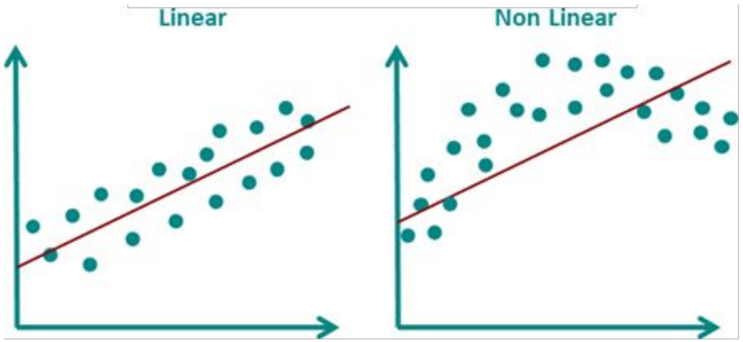
- OLS is often evaluated using r-squared (R^2).
- R-Squared (R^2 or the coefficient of determination)
- It is a statistical measure in a regression model.
- Regression model determines the proportion of variance in the DV that can be explained by the IV.
- That is r-squared shows how well the data fit the regression model (the goodness of fit).

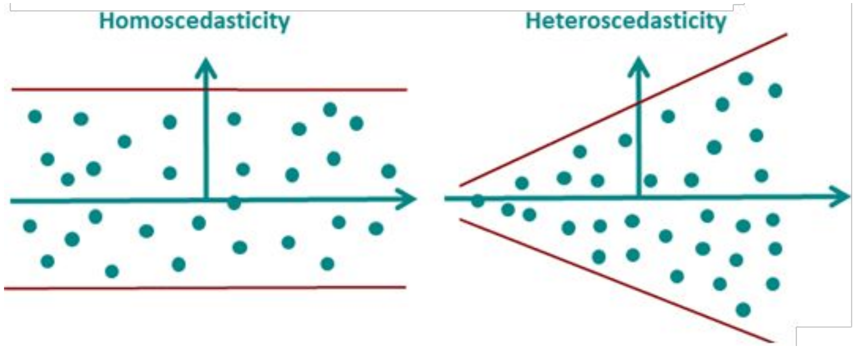
Assumptions of Linear Regression

- Assumptions of Linear Regression
- In order to interpret the results of the regression analysis meaningfully, certain conditions must be met.
  - Linearity: There must be a linear relationship between the dependent and independent variables.
  - Homoscedasticity: The residuals must have a constant variance.
  - Normality: Normally distributed error
  - No multicollinearity: No high correlation between the independent variables
  - No auto-correlation: The error component should have no auto correlation

Assumptions of Linear Regression

- Linearity
  - In linear regression, a straight line is drawn through the data.
  - This straight line should represent all points as good as possible.
  - If the points are distributed in a non-linear way, the straight line cannot fulfill this task.
- Homoscedasticity
  - Since in practice the regression model never exactly predicts the DV, there is always an error.
  - This very error must have a constant variance over the predicted range.





Assumptions of Linear Regression

- Normal distribution of the error
- The next requirement of linear regression is that the error epsilon must be normally distributed.
- There are two ways to find it out: One is the analytical way and the other is the graphical way.
- Analytical way (We can use either the Kolmogorov-Smirnov test or the Shapiro-Wilk test.
  - Eg. Kolmogorov-Smirnov (Statistics=0.16; df=12; p-value=0.873)
  - Eg. Shapiro-Wilk (Statistics=0.973; df=12; p-value=0.936)
  - If the p-value is greater than 0.05,
    - there is no deviation of the data from the normal distribution and
  - one can assume that the data are normally distributed.
- Graphical variant
  - Looked at the histogram or better to use QQ-plot (Quantile-Quantile-plot).
  - The more the data lie on the line, the better the normal distribution.

Assumptions of Linear Regression

- Multicollinearity
- Multicollinearity means that two or more independent variables are strongly correlated with one another.
- Problem with multicollinearity: The effects of each IV cannot be clearly separated from one another.
- Auto-correlation
  - If there is a high correlation between x1 and x2, then it is difficult to determine b1 and b2.
  - If both are completely equal, the regression model does not know how large b1 and b2 should be, becoming unstable.

$$\hat{y} = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

$$\hat{x}_1 = b_2 \cdot x_2 + \ldots + b_k \cdot x_k + a$$

$$\hat{x}_2 = b_1 \cdot x_1 + \ldots + b_k \cdot x_k + a$$

$$\hat{x}_k = b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + a$$

Toleranz

$$T = 1 - R^2$$

Coefficient of determination

Warning:

$$T < 0.1$$

VIF (Variance Inflation Factor)

$$VIF = \frac{1}{1 - R^2}$$

Coefficient of determination

Warning:

$$VIF > 10$$

Measures of Regression Model

- Two main **measures** are used to find out how well the regression model can predict or explain the DV
- Coefficient of determination R^2 (Also known as the variance explanation)
  - indicates how large the portion of the variance is that can be explained by the IVs.
  - The more variance can be explained, the better the regression model is.
  - In order to calculate R^2,
    - the variance of the estimated value is related to the variance in the observed values:

$$R^2 = \frac{\overbrace{s_{\hat{y}}^2}^{\text{Variance of the predicted values}}}{\underbrace{s_y^2}_{\text{Variance of the observed values}}}$$

- Adjusted R^2
  - The coefficient of determination R^2 is influenced by the number of IVs used.
  - The more IVs are included in the regression model, the greater the variance resolution R^2.
  - To take this into account, the adjusted R2 is used.

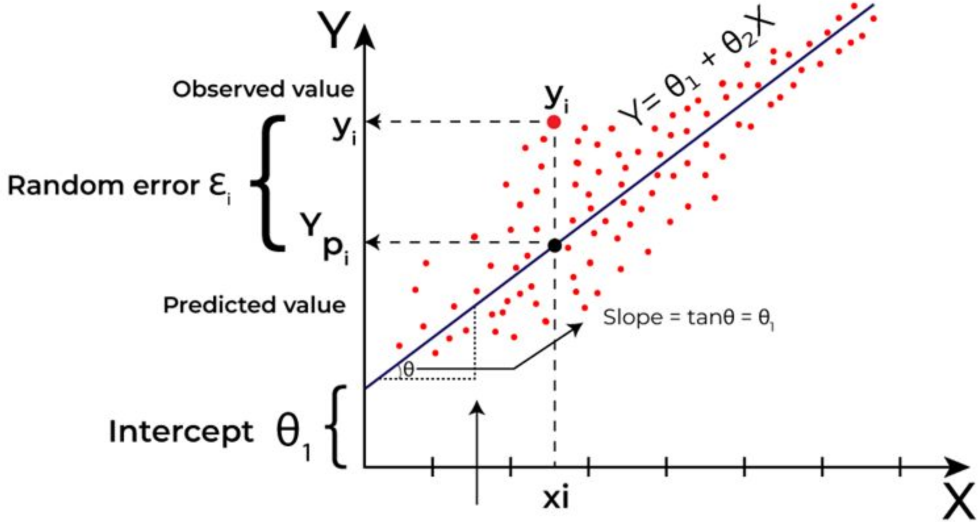
$$R^2_{adj} = 1 - (1 - R^2) \cdot \frac{n-1}{n-p-1}$$

Measures of Regression Model

- Standard estimation error
- The standard error of the estimate is the estimation of the accuracy of any predictions.
- It is denoted as SEE.
- It is the standard deviation of the estimation error.
- It gives an impression of how much the prediction differs from the correct value.
- It is the dispersion of the observed values around the regression line.
- The coefficient of determination R^2 and the SEE are used for simple and multiple linear regression.
- The regression line depreciates the sum of squared deviations of prediction.
- It is also known as the sum of squares error.

The best Fit Line equation

- The best Fit Line equation provides a straight line
- This line represents the relationship between the DV and IVs.
- The slope of the line indicates how much the DV changes for a unit change in the IV(s).



The best Fit Line equation

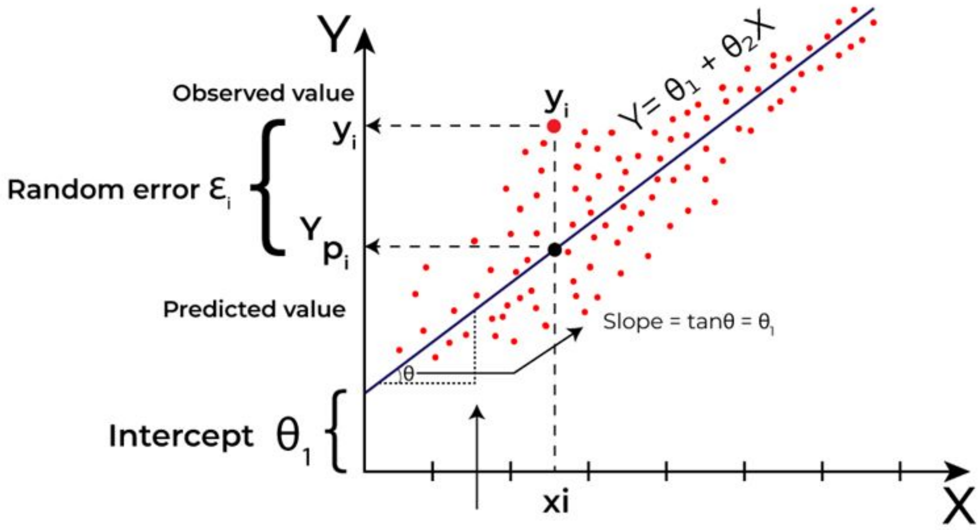
- Here Y is called a dependent or target variable and X is called an IV also known as the predictor of Y.
- The independent feature is the experience i.e X and
- the respective salary Y is the dependent variable.
- The model gets the best regression fit line by finding the best θ1 and θ2 values.
  - θ1: intercept
  - θ2: coefficient of x
- Once we find the best θ1 and θ2 values, we get the best-fit line.
- So when we are finally using our model for prediction, it will predict the value of y for the input value of x.
- Let's assume there is a linear relationship between X and Y then the salary can be predicted using:

$$\hat{Y} = \theta_1 + \theta_2 X \quad \text{OR} \quad \hat{y}_i = \theta_1 + \theta_2 x_i$$

- $y_i \epsilon Y \quad (i = 1, 2, \cdots, n)$ are labels to data (Supervised learning)
- $x_i \epsilon X \quad (i = 1, 2, \cdots, n)$ are the input independent training data (univariate – one input variable(parameter))
- $\hat{y}_i \epsilon \hat{Y} \quad (i = 1, 2, \cdots, n)$ are the predicted values.

The best Fit Line equation

- How to update θ1 and θ2 values to get the best-fit line?
- it is very important to update the θ1 and θ2 values,
- to reach the best value that minimizes the error between the predicted y value (pred) and the true y value (y).



$$minimize \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Simple Linear Regression

- Example

Linear regression equation

y=a + bx

x and y are two variables on the regression line.

b = Slope of the line.

a = y-intercept of the line.

x = Values of the first data set.

y = Values of the second data set.

$$a\,(intercept) = \frac{\sum y \sum x^2 - \sum x \sum xy}{(\sum x^2) - (\sum x)^2}$$

$$b\,(slope) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

Question: Find linear regression equation for the following two sets of data:

x	y
2	4
4	9
6	3
8	12

Solution: n = 4

x	y	x^2	xy
2	4	4	8
4	9	16	36
6	3	36	18
8	12	64	96
$\sum x = 20$	$\sum y = 28$	$\sum x^2 = 120$	$\sum xy = 158$

Caclualte; a(intercept) = 2.5

Caclualte; b(islope) = 0.9

Linear regression is given by: y = a + bx

y = 2.5 + 0.9x

References

Text books:

1. Ethem Alpaydin, “Introduction to Machine Learning , 4th Edition, The MIT Press, 2020.
2. Peter Harrington, “Machine Learning in Action”, 1st Edition, Dreamtech Press, 2012."
3. Tom Mitchell, “Machine Learning”, 1st Edition, McGraw Hill, 2017.
4. Andreas C, Müller and Sarah Guido, “Introduction to Machine Learning with Python: A Guide for Data Scientists”, 1ed, O'reilly, 2016.
5. Kevin P. Murphy, “Machine Learning: A Probabilistic Perspective”, 1st Edition, MIT Press, 2012."

Reference Books:

6. Aurélien Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”, 2nd Edition, Shroff/O'Reilly, 2019.
7. Witten Ian H., Eibe Frank, Mark A. Hall, and Christopher J. Pal., “Data Mining: Practical machine learning tools and techniques”, 1st Edition, Morgan Kaufmann, 2016.
8. Han, Kamber, “Data Mining Concepts and Techniques”, 3rd Edition, Morgan Kaufmann, 2012.
9. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar, “Foundations of Machine Learning”, 1ed, MIT Press, 2012.
10. H. Dunham, “Data Mining: Introductory and Advanced Topics”, 1st Edition, Pearson Education, 2006.

Thank You.