Linear Regression

Linear regression

- Linear regression is one of the most well known algorithm in statistics and machine learning.
- linear regression was developed in the field of statistics
- It is studied as a model for understanding the relationship between input and output numerical variables
- It is both a statistical algorithm and a machine learning algorithm.

Linear regression

- Linear regression is a linear model
- Ex: a model that assumes a linear relationship between the input variables (x) and the single output variable (y).
- More specifically, y can be calculated from a linear combination of the input variables (x).
- When there is a single input variable (x), the method is referred to as simple linear regression.
- When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

Linear regression

- Different techniques can be used to prepare or train the linear regression equation from data
- the most common of which is called Ordinary Least Squares.
- It is common to refer to a model prepared this way as Ordinary Least Squares Linear Regression or Least Squares Regression.

Linear Regression Model Representation

- The representation is a linear equation that combines a set of input values (x) along with its solution which is the predicted output for that set of input values (y).
- Both the input values (x) and the output value are numeric.
- in a simple regression problem (a single x and a single y), the form of the model would be:
- y = B0 + B1 * x
- B1 - scale factor to each input value or column, called a coefficient
- B0 – giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is called the intercept or the bias coefficient
- In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane.

Linear Regression Model Representation

- Complexity of a regression model (like linear regression)
- This refers to the number of coefficients used in the model.
- When a coefficient becomes zero, it removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0).

Linear Regression Learning the Model

- Learning a linear regression model requires estimating the values of the coefficients used in the representation
- Techniques to prepare a linear regression model.

Linear Regression Learning the Model

1. Simple Linear Regression:

- With simple linear regression when we have a single input, we can use statistics to estimate the coefficients.
- This requires that we calculate statistical properties from the data such as mean, standard deviations, correlations and covariance.

Linear Regression Learning the Model

2. Ordinary Least Squares

- When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients.
- The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals.
- This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together.
- This is the quantity that ordinary least squares seeks to minimize.
- This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients.

Linear Regression Learning the Model

2. Ordinary Least Squares

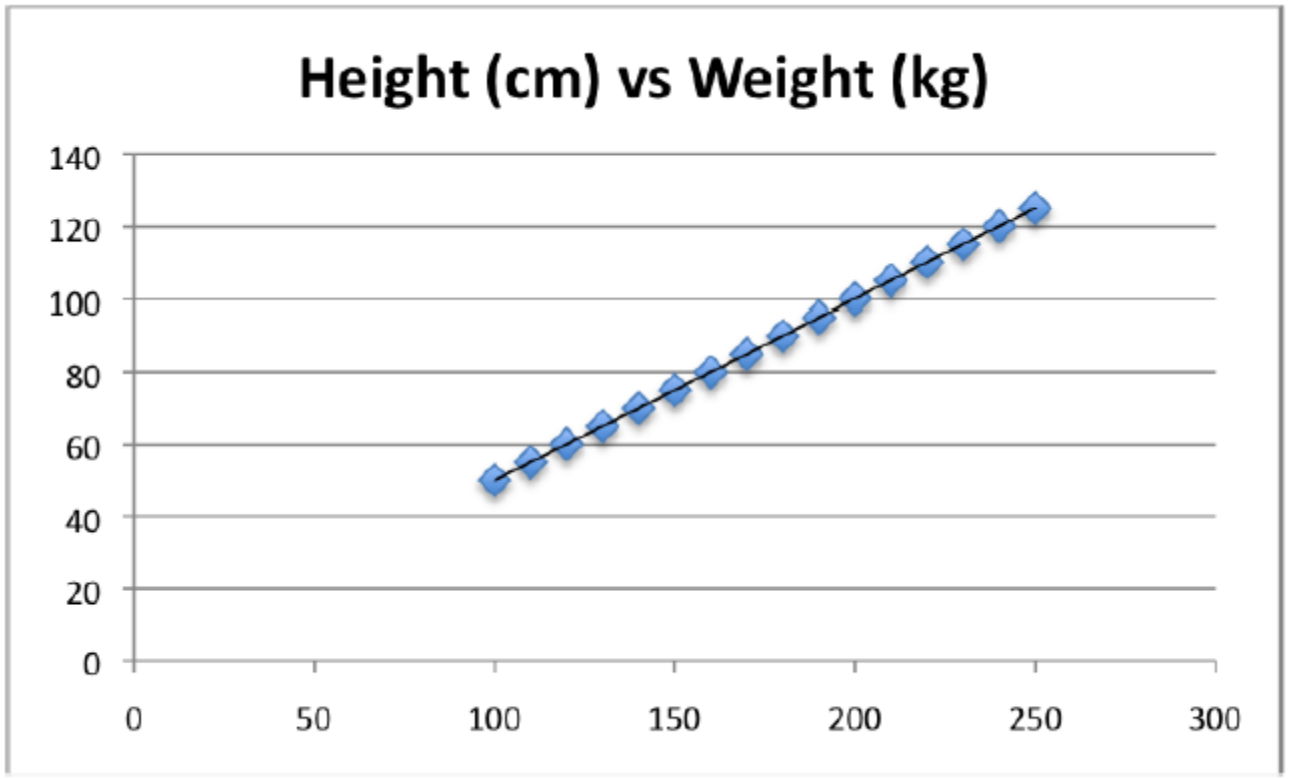
- When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients.
- The Ordinary Least Squares procedure seeks to minimize the sum of the squared residuals.
- This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together.
- This is the quantity that ordinary least squares seeks to minimize.
- This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients.

Making Predictions with Linear Regression

- Given the representation is a linear equation, making predictions
- involves solving the equation for a specific set of inputs
- Ex: Imagine we are predicting weight (y) from height (x).
- Our linear regression model representation for this problem:
- y = B0 + B1 * X1
- weight = B0 + B1 * height
- B0 is the bias coefficient and B1 is the coefficient for the height column.
- We use a learning technique to find a good set of coefficient values. Once found, we can plug in different height values to predict the weight.
- Ex: let's use B0 = 0.1 and B1 = 0.5. Calculate the weight (in kilograms) for a person with the height of 182 centimeters.
- weight = 0.1 + 0.5* 182=91.1
- Plot the above equation as a line in two-dimensions.
- The B0 is our starting point regardless of height.

Making Predictions with Linear Regression

Sample Height vs Weight Linear Regression



Run through heights from 100 to 250 centimeters and get weight values, creating our line

Preparing Data For Linear Regression

- How our data must be structured to make best use of the model?
- In practice, we can uses these rules as heuristics when using Ordinary Least Squares Regression, the most common implementation of linear regression.

1. Linear Assumption: Linear regression assumes that the relationship between our input and output is linear.

- When we have a lot of attributes. we may need to transform data to make the relationship linear (e.g. log transform for an exponential relationship).

2. Remove Noise: Linear regression assumes that our input and output variables are not noisy.

- Consider using data cleaning operations that better clarifies our data.
- This is most important for the output variable and we want to remove outliers in the output variable (y) if possible.

Preparing Data For Linear Regression

- How our data must be structured to make best use of the model?

3. Remove Collinearity: Linear regression will overfit our data when we have highly correlated input variables. Consider calculating pairwise correlations for our input data and removing the most correlated.

4. Gaussian Distributions. Linear regression will make more reliable predictions if our input and output variables have a Gaussian distribution.

- We may get benefit using transforms (e.g. log or BoxCox) on our variables to make their distribution more Gaussian looking.

5. Rescale Inputs: Linear regression will often make more reliable predictions if you rescale

- input variables using standardization or normalization.

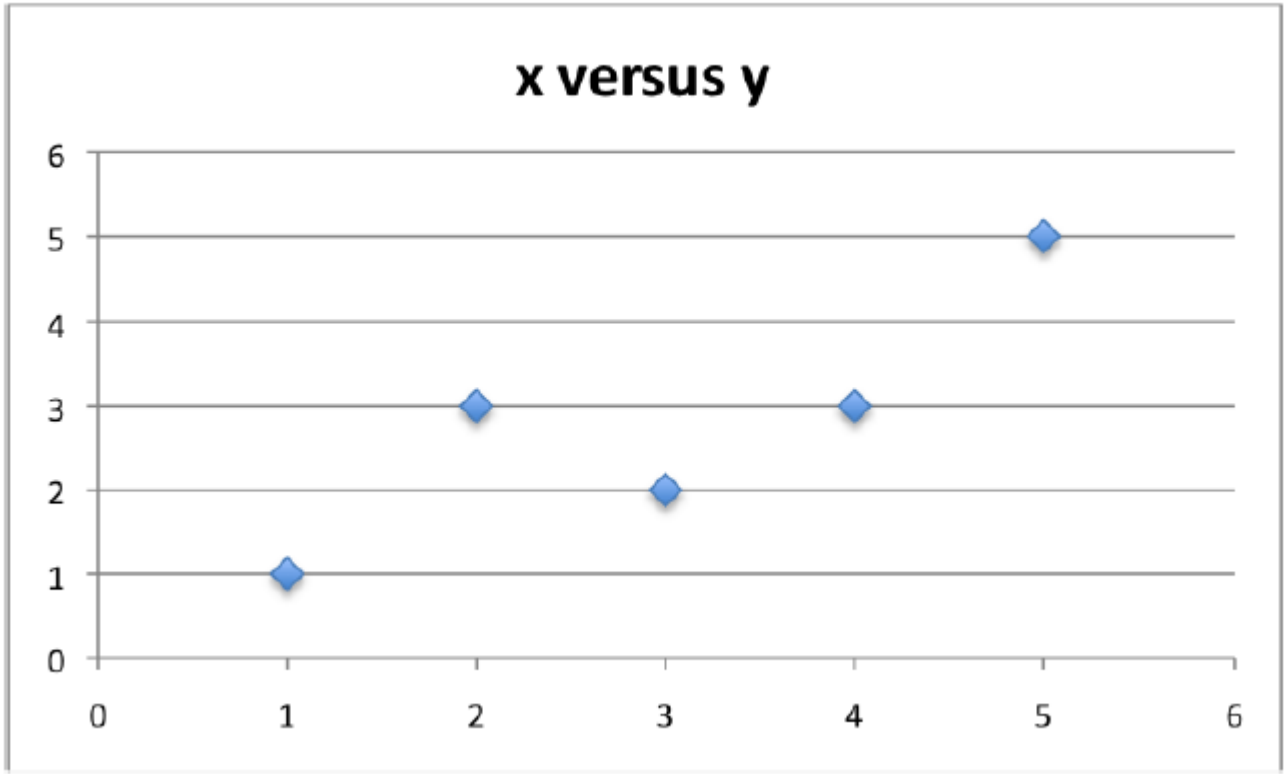
Simple Linear Regression- Example

x	y
1	1
2	3
4	3
3	2
5	5

- Data Set
- The attribute x is the input variable
- y is the output variable we are trying to predict.
- We can see the relationship between x and y looks kind-of linear.
- Plot the line
- to generally describe the relationship between the data.
- This is a good indication that using linear regression might be appropriate for this dataset.

Simple Linear Regression- Example

- Simple Linear Regression Dataset: Scatter plot of x versus y.



Simple Linear Regression- Example

- Simple linear regression: When we have a single input attribute (x) and we use linear regression
- Multiple linear regression: we have multiple input attributes (e.g. X1, X2, X3, etc.)
- Problem Statement: Create a simple linear regression model from our training data, then make predictions for our training data and identify how well the model learned the relationship in the data.

Simple Linear Regression- Example

- With simple linear regression we model our data as follows:
- y = B0 + B1 * x
- This is a line where
  - y is the output variable we want to predict,
  - x is the input variable
  - B0 and B1 are coefficients
  - B0 is called the intercept because it determines where the line intercepts the y-axis.
  - In machine learning we can call this the bias, because it is added to offset all predictions that we make.
  - The B1 term is called the slope because it defines the slope of the line or how x translates into a y value before we add our bias.

Simple Linear Regression- Example

- The goal is to find the best estimates for the coefficients to minimize the errors in predicting y from x
- In Simple regression we can estimate coefficients directly from our data.
- We can start off by estimating the value for B1 as:

$$B1 = \frac{\sum_{i=1}^{n}(x_i - mean(x)) \times (y_i - mean(y))}{\sum_{i=1}^{n}(x_i - mean(x))^2}$$

- Where mean() is the average value for the variable in our dataset.
- The xi and yi refer to the fact that we need to repeat these calculations across all values in our dataset
- i refers to the ith value of x or y.
- We can calculate B0 using B1 and some statistics from our dataset, as follows:
- B0 = mean(y) - B1 * mean(x)

Simple Linear Regression- Estimating The Slope (B1)

- Let's start with the top part of the equation, the numerator. First we need to calculate the
- mean value of x and y. The mean is calculated as:

$$\frac{1}{n} \times \sum_{i=1}^{n} x_i$$

- Where n is the number of values (5 in this case).
- calculate the mean value of our x and y variables:
- mean(x) = 3
- mean(y) = 2.8

Simple Linear Regression- Example

- Calculate the error of each variable from the mean

Residual of each x value from the mean.

x	mean(x)	x - mean(x)
1	3	-2
2		-1
4		1
3		0
5		2

Residual of each y value from the mean.

y	mean(y)	y - mean(y)
1	2.8	-1.8
3		0.2
3		0.2
2		-0.8
5		2.2

Simple Linear Regression- Example

- multiply the error for each x with the error for each y and calculate the sum of these multiplications.

x - mean(x)	y - mean(y)	Multiplication
-2	-1.8	3.6
-1	0.2	-0.2
1	0.2	0.2
0	-0.8	0
2	2.2	4.4

Simple Linear Regression- Example

- Summing the final column we have calculated our numerator as 8.
- Now we need to calculate the denominator of the equation for calculating B1
- This is calculated as the sum of the squared differences of each x value from the mean.
- We have already calculated the difference of each x value from the mean, so we square each value and calculate the sum.
- Squared residual of each x value from the mean.

x - mean(x)	squared
-2	4
-1	1
1	1
0	0
2	4

Simple Linear Regression- Example

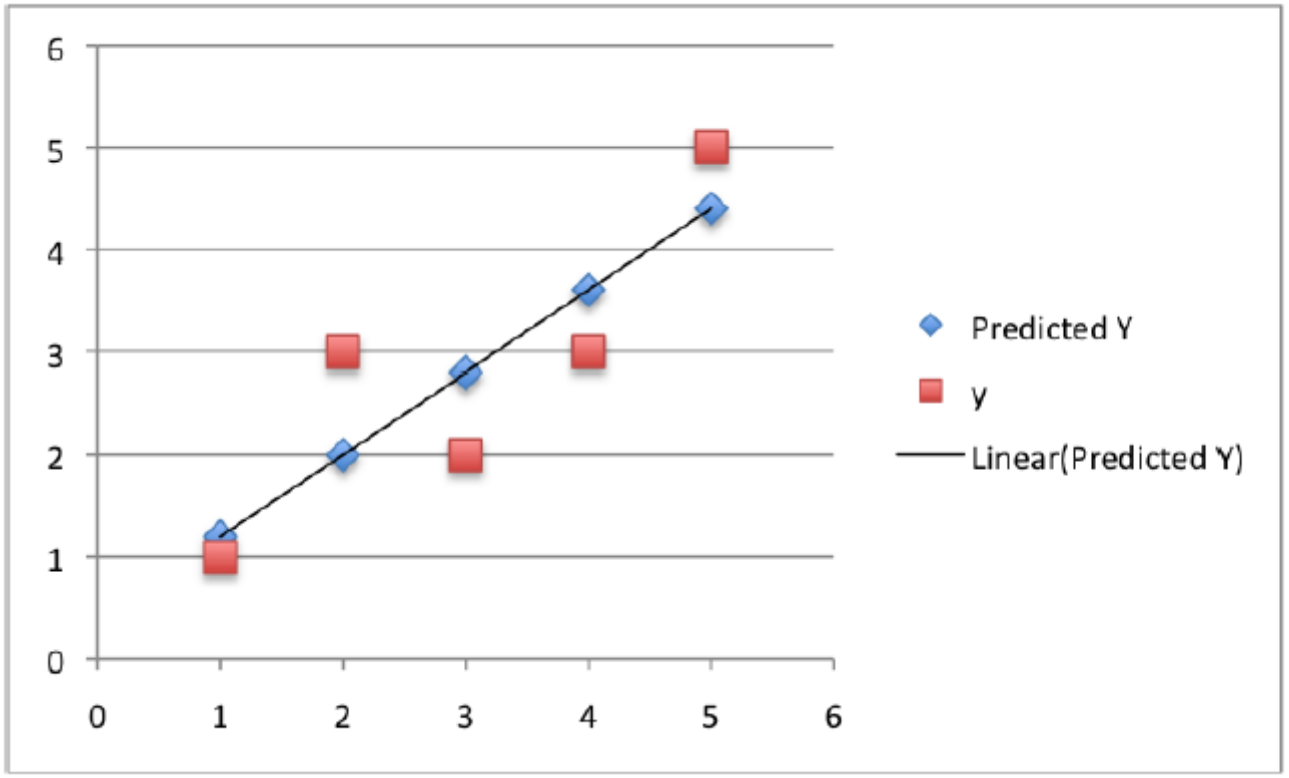
- Calculating the sum of these squared values gives the denominator of 10.
- Now we can calculate the value of our slope.
- B1 = 8/10=0.8

Simple Linear Regression- Example

- Estimating the Intercept (B0)
- B0 = mean(y) - B1 * mean(x)
- B0 = 2.8 – 0.8 * 3 = 0.4
- Making Predictions
- We now have the coefficients for our simple linear regression equation.
- y = B0 + B1 * x
- y = 0.4 + 0.8 * x
- Problem: Try out the model by making predictions for our training data and plot these predictions as a line with our data. This gives a visual idea of how well the line models our data.

Simple Linear Regression- Example

Simple Linear Regression Predictions (x vs y in red and x vs prediction in blue)



Simple Linear Regression- Example

- Estimating Error
- We can calculate an error score for our predictions called the Root Mean Squared Error or RMSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - y_i)^2}{n}}$$

- Error for predicted values

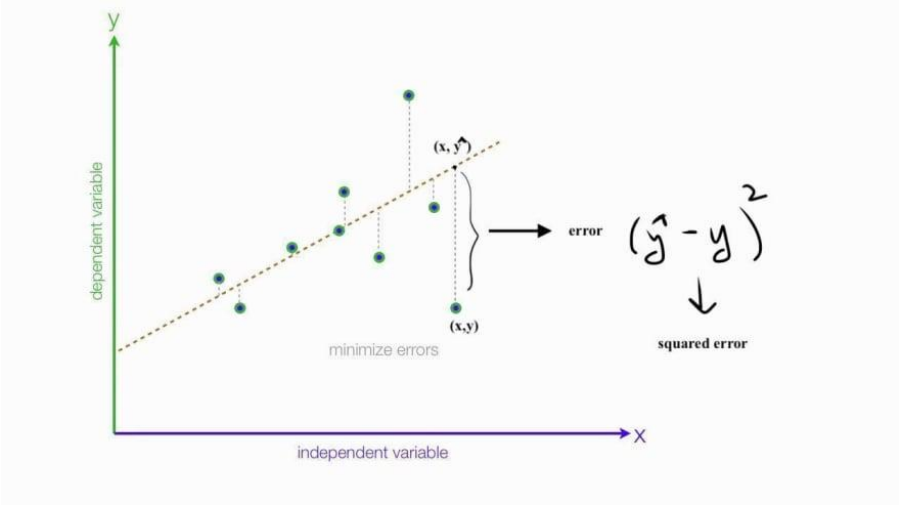
Predicted	y	Predicted - y
1.2	1	0.2
2	3	-1
3.6	3	0.6
2.8	2	0.8
4.4	5	-0.6

Squared error for predicted values

Predicted - y	squared error
0.2	0.04
-1	1
0.6	0.36
0.8	0.64
-0.6	0.36

Simple Linear Regression- Example

- The sum of these errors is 2.4 units, dividing by 5 and taking the square root gives us:
- RMSE = 0.692820323
- Each prediction is on average wrong by about 0.692 units.



Intuitive Formula

- If there is no relationship between X and Y, the best guess for all values of X is the mean of Y.
- If there is a relationship (slope is not zero), the best guess for the mean of X is still the mean of Y
- As X departs from the mean, so does Y.
- At any rate, the regression line always passes through the means of X and Y.
- This means that, regardless of the value of the slope, when X is at its mean, so is Y

Intuitive Formula

- To find the slope, Pedhazur formula is:
- This says that the slope B1(here b) is the sum of deviation cross products divided by the sum of squares for X

$$b = \frac{\sum xy}{\sum x^2}$$

- To find B0(here a), subtract and rearrange to find

$$\bar{Y} = a + b\bar{X}$$

- B0= y- B1 * x

$$\bar{Y} - b\bar{X} = a + b\bar{X} - b\bar{X}$$

$$a = \bar{Y} - b\bar{X}$$