Spearman's Rank-Order Correlation

Pearson’s correlation works well if the relationship between variables is linear and if the variables are roughly normal. But it is not robust in the presence of outliers. Spearman’s rank correlation is an alternative that mitigates the effect of outliers and skewed distributions.

To compute Spearman’s cor-relation, we have to compute the rank of each value, which is its index in the sorted sample.

Find Spearman's Rank-Order Correlation between English and Math of a calss.

English	56	75	45	71	61	64	58	80	76	61
Sort Order	2	8	1	7	4.5	6	3	10	9	4.5

Math	66	70	40	60	65	56	59	77	67	63
Sort Order	7	9	1	4	6	2	3	10	8	5

```
eng = [56, 75, 45, 71, 61, 64, 58, 80, 76, 61]
Mat = [66, 70, 40, 60, 65, 56, 59, 77, 67, 63]

ser_eng_rank = pd.Series(eng).rank()
ser_mat_rank = pd.Series(mat).rank()

ser_eng_rank.corr(ser_mat_rank, method='spearman')
```

0.6686960980480710 6

Is this Correlation by chance ?

Hacker's Way : To check, if the correlation is by chance?, we have to rerandamize one series and calculate the Correlation. We have to perform this test around 10,000 times, record the correlation. Now check the percentage of number of times the correlation >= 0.669. If it is more than 5% we can say that the correlation is by chance.

```
def isCorrelationByChance(ser1, ser2, num_iters):
    orig_corr = ser1.corr(ser2)
    corr_dict = dict()
    rand_corr_more_than_original = 0
    for i in range(num_iters):
        sample_ser1 = ser1.sample(n=len(ser1))
        sample_ser1.reset_index(drop=True, inplace=True)
        corr_dict[i] = sample_ser1.corr(ser2)
        if(corr_dict[i]  >= orig_corr):
            rand_corr_more_than_original += 1
        percentage = rand_corr_more_than_original*100/num_iters
    return percentage


print(isCorrelationByChance(ser_eng, ser_mat, 10000))
```

The percentage is oscillating between 3 and 4, less than 5, hence the correlation is not by change. In other words for any sample this relation would hold true.

Simple Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation(Y = **theta$_0$** + **theta$_1$** X) to observed data. One variable is considered to be an Independent variable(X), and the other is considered to be a dependent variable (Y).

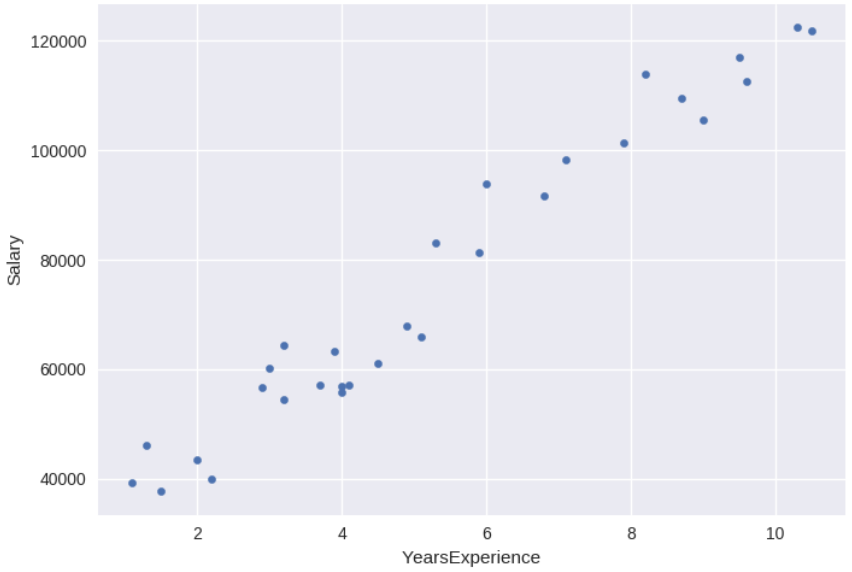
Linear Equation $h(X) = \theta_0 + \theta_1 * X$

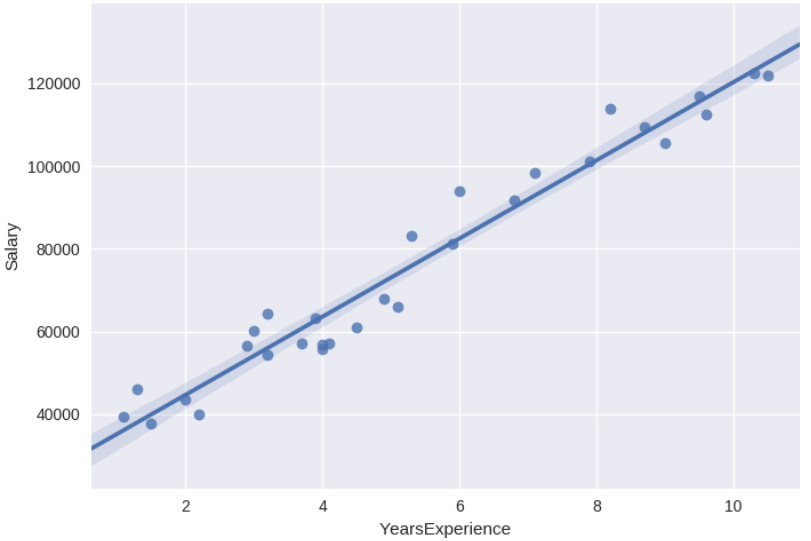
The co-coefficients or parameters

Θ_0 – intercept

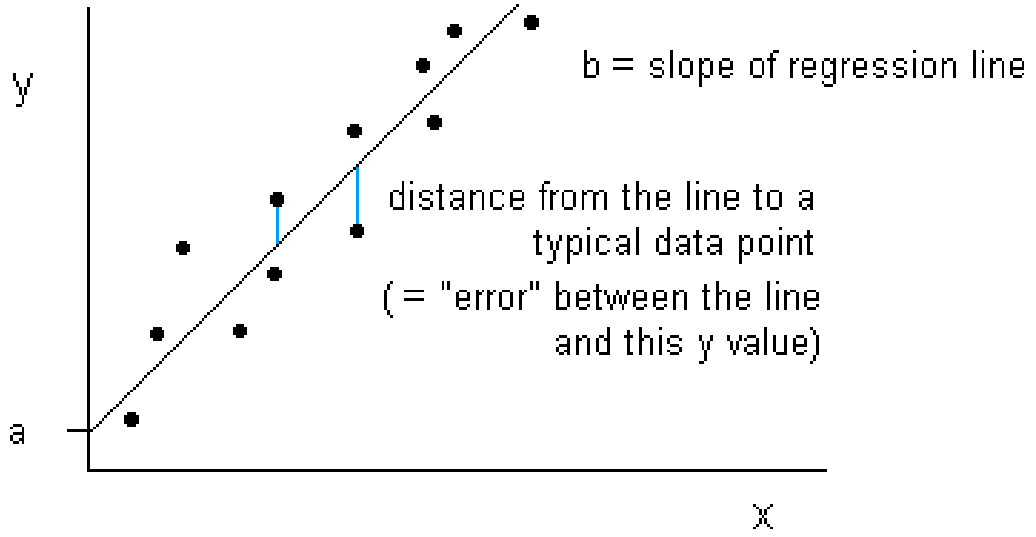
Θ_1 – slope

* Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This can be done by looking at the scatter plot between independent and dependent variable.





Least-Squares Regression: (ordinary least squares - OLS)The most common method for fitting a regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). Because the deviations are first squared, then summed, there are no cancellations between positive and negative values.



$e_i = \hat{y} - y$ Least-Squares Regression objective is to reduce over all error**, that is minimize**

Residual Sum of Squares (RSS) = $\sum_{i=1}^{n} \left(\hat{y} - y_i\right)^2$ n – number of samples