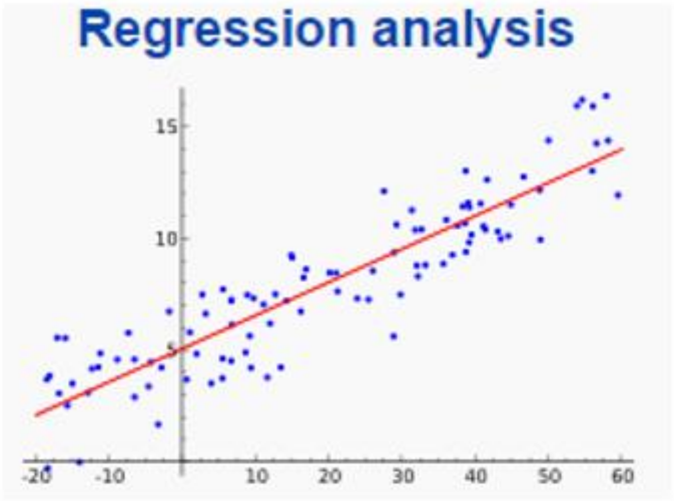
K2 Analytics
Building Skills, Building Individuals

Linear Regression

Linear Regression

- In statistics, **linear regression** is an approach for modeling the relationship between **a scalar dependent variable y** and **one or more explanatory variables (or independent variables) denoted X.**
  - The case of one explanatory variable is called **simple linear regression.**
  - For more than one explanatory variable, the process is called **multiple linear regression.**



https://en.Wikipedia.org/wiki/Linear_regression

Linear Regression… e.g.

Ohm’s Law:

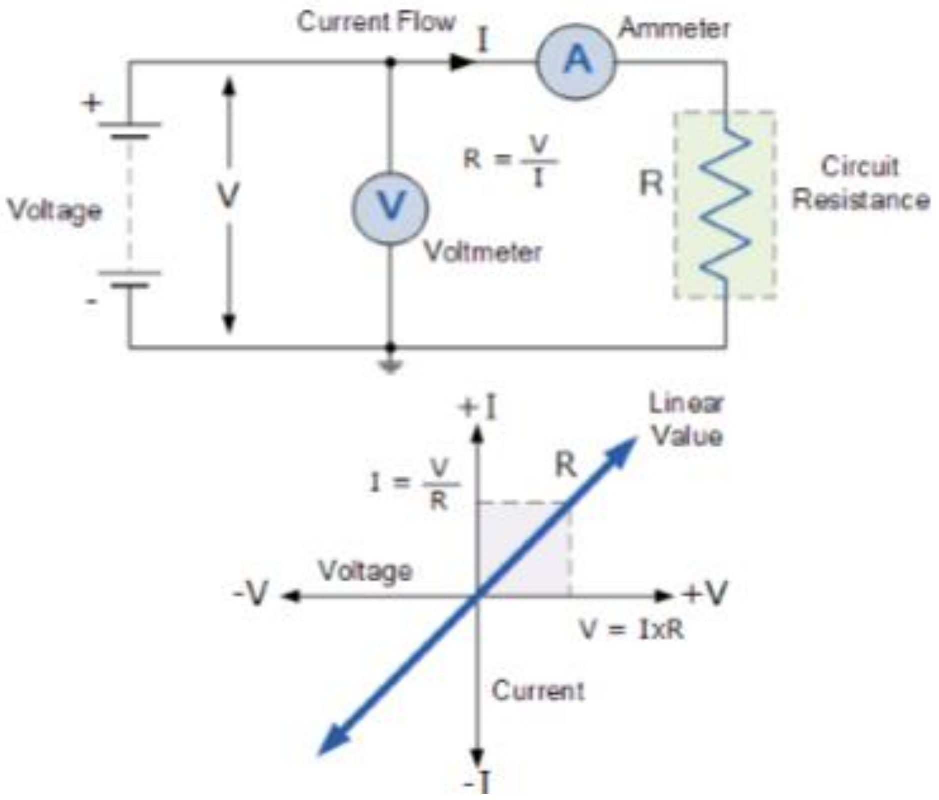
- In physics, it is observed that the relationship between Voltage (V), Current (I) and Resistance (R) is a linear relationship expressed as

  V = I * R

  I = V / R

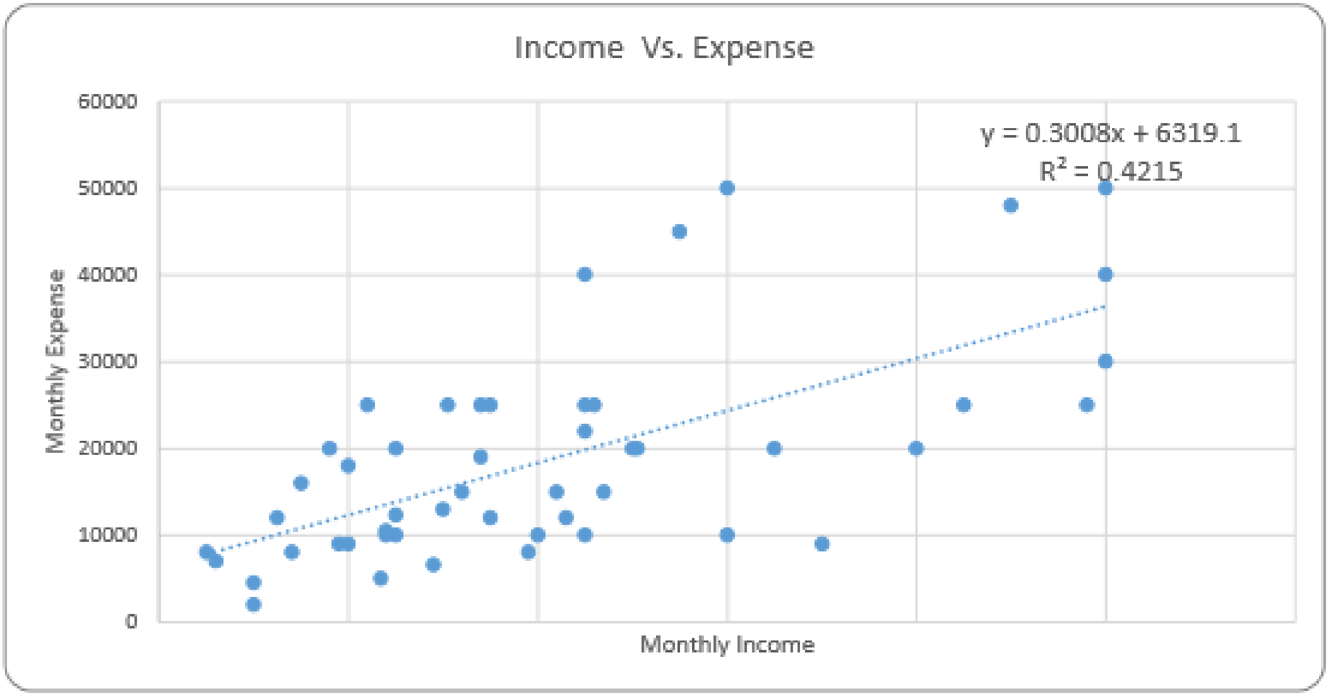
- In a circuit board for a given Resistance R, as you increase the Voltage V, the current I increases proportionately

https://www.electronics-tutorials.ws/dccircuits/dcp_1.html



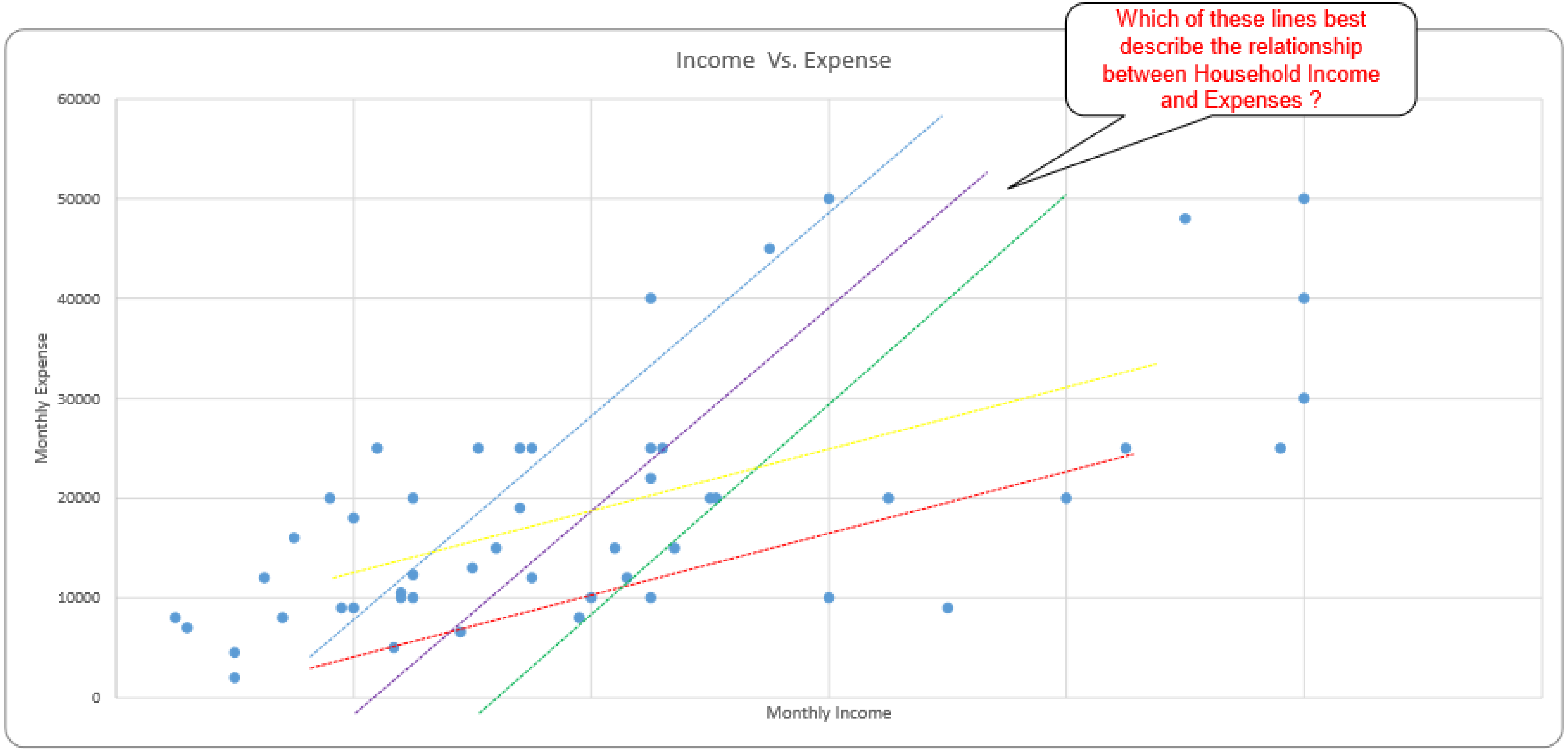
Sample Monthly Income-Expense Data of a Household

Monthly Income (in Rs.)	Monthly Expense (in Rs.)
5,000	8,000
6,000	7,000
10,000	4,500
10,000	2,000
12,500	12,000
14,000	8,000
15,000	16,000
18,000	20,000
19,000	9,000
20,000	9,000
20,000	18,000
22,000	25,000
23,400	5,000
24,000	10,500
24,000	10,000

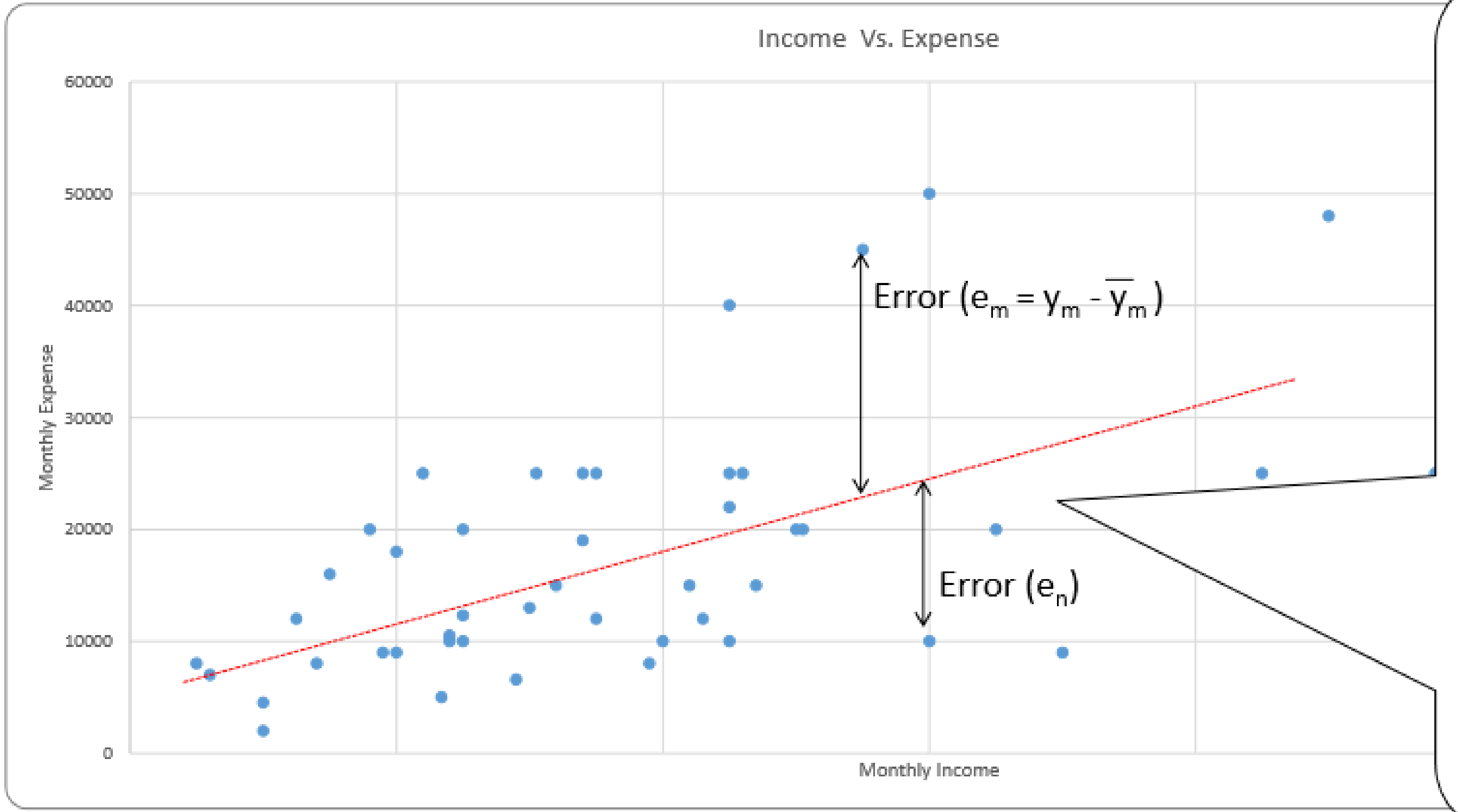


We have to find the relationship between Income and Expenses of a household

Line of Best fit



Line of Best fit…



The Line of Best Fit will be the one where Sum of Square of Error (SSE) term will be minimum (OLS Technique)

$Y_{i(hat)} = b_o + b_1X_i$ is the sample regression equation

$$\begin{aligned} SSE &= \Sigma\ e_{i(hat)}^2 \quad (1) \\ &= \Sigma\ (Y_i - Y(_{i(hat)})^2 \quad (2) \\ &= \Sigma\ (Y_i - b_o - b_1X_i)^2 \quad (3) \end{aligned}$$

Using calculus we get

$$b_o = \frac{\Sigma Y_i - b_1 \Sigma X_i}{n}$$

$$b_1 = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

Importing Required Packages

- Import Packages

```
## Import Packages
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import os
```

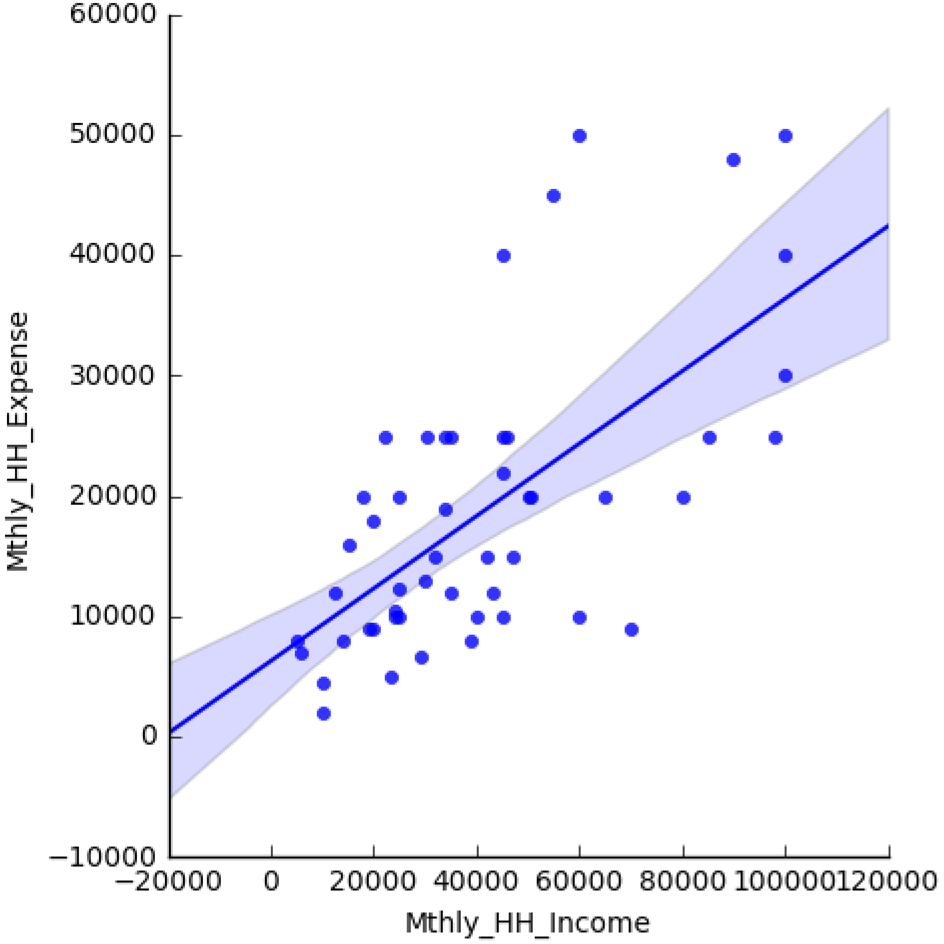
Import Datafile

```
## Set the working directory and import data
os.chdir("D:/K2Analytics/datafile")
inc_exp = pd.read_csv("Inc_Exp_Data.csv")
inc_exp.head()
```

Index	Mthly_HH_Income	Mthly_HH_Expense	No_of_Fly_Members	Emi_or_Rent_Amt	Annual_HH_Income	Highest_Qualified_Member	No_of_Earning_Members
0	5000	8000	3	2000	64200	Under-Graduate	1
1	6000	7000	2	3000	79920	Illiterate	1
2	10000	4500	2	0	112800	Under-Graduate	1
3	10000	2000	1	0	97200	Illiterate	1
4	12500	12000	2	3000	147000	Graduate	1
5	14000	8000	2	0	196560	Graduate	1

Scatter Plot

```
## Scatter Plot
import seaborn as sns
%matplotlib inline
sns.lmplot(x ="Mthly_HH_Income",
           y = "Mthly_HH_Expense", data = inc_exp)
```



Simple Linear Regression

```python
## Simple Linear Regression Model
import statsmodels.formula.api as sm
linear_mod = sm.ols(formula ="Mthly_HH_Expense ~ Mthly_HH_Income" ,
                    data = inc_exp).fit()

#Get the model summary
linear_mod.summary()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Mthly_HH_Expense   R-squared:                       0.421
Model:                            OLS   Adj. R-squared:                  0.409
Method:                 Least Squares   F-statistic:                     34.97
Date:                Sun, 25 Nov 2018   Prob (F-statistic):           3.40e-07
Time:                        21:58:20   Log-Likelihood:                -526.77
No. Observations:                  50   AIC:                             1058.
Df Residuals:                      48   BIC:                             1061.
Df Model:                           1
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [95.0% Conf. Int.]
-----------------------------------------------------------------------------------
Intercept        6319.1018   2488.733      2.539      0.014      1315.168  1.13e+04
Mthly_HH_Income     0.3008      0.051      5.914      0.000         0.198     0.403
==============================================================================
Omnibus:                        6.455   Durbin-Watson:                   2.417
Prob(Omnibus):                  0.040   Jarque-Bera (JB):                5.471
Skew:                           0.774   Prob(JB):                       0.0649
Kurtosis:                       3.479   Cond. No.                     9.27e+04
==============================================================================
```

Coefficient of Determination

- In statistics, the coefficient of determination, denoted R2 or r2 and pronounced "R squared", is a number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variable(s).
  - The total sum of squares (proportional to the variance of the data):

    $$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2,$$

  - The regression sum of squares, also called the explained sum of squares:

    $$SS_{\text{reg}} = \sum_i (f_i - \bar{y})^2,$$

  - The sum of squares of residuals, also called the residual sum of squares:

    $$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

  The most general definition of the coefficient of determination is

  $$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

https://en.wikipedia.org/wiki/Coefficient_of_determination

ANOVA – F test for Linear Regression

Analysis of Variance (ANOVA) consists of calculations that provide information about levels of variability within a regression model and form a basis for tests of significance. The basic regression line concept, DATA = FIT + RESIDUAL, is rewritten as follows:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i).$$

The first term is the total variation in the response y, the second term is the variation in mean response, and the third term is the residual value. Squaring each of these terms and adding over all of the n observations gives the equation

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2.$$

This equation may also be written as **SST = SSM + SSE**, where SS is notation for ***sum of squares*** and T, M, and E are notation for ***total***, ***model***, and ***error***, respectively.

The square of the sample correlation is equal to the ratio of the model sum of squares to the total sum of squares: r^2 **= SSM/SST**.
This formalizes the interpretation of r^2 as explaining the fraction of variability in the data explained by the regression model.

The sample variance s_y^2 is equal to $\sum(y_i - \bar{y})^2/(n - 1)$ **= SST/DFT**, the total sum of squares divided by the total degrees of freedom (DFT).
For simple linear regression, the MSM (mean square model) = $\sum(\hat{y}_i - \bar{y})^2/(1)$ **= SSM/DFM**, since the simple linear regression model has one explanatory variable x.
The corresponding MSE (mean square error) = $\sum(y_i - \hat{y}_i)^2/(n - 2)$ **= SSE/DFE**, the estimate of the variance about the population regression line (σ^2).

…contd

ANOVA calculations are displayed in an ***analysis of variance table***, which has the following format for simple linear regression:

Source	Degrees of Freedom	Sum of squares	Mean Square	F
Model	1	$\sum(\hat{y}_i - \bar{y})^2$	SSM/DFM	MSM/MSE
Error	n - 2	$\sum(y_i - \hat{y}_i)^2$	SSE/DFE	
Total	n - 1	$\sum(y_i - \bar{y})^2$	SST/DFT	

The "F" column provides a statistic for testing the hypothesis that
$\beta_1 \neq 0$
against the null hypothesis that $\beta_1 = 0$.

The test statistic is the ratio MSM/MSE, the mean square model term divided by the mean square error term. When the MSM term is large relative to the MSE term, then the ratio is large and there is evidence against the null hypothesis.

For simple linear regression, the statistic MSM/MSE has an *F* distribution with degrees of freedom (DFM, DFE) = (1, *n* - 2).

Multiple Linear Regression

- Multiple linear regression is the most common form of linear regression analysis.
- Multiple linear regression is used to explain the relationship between one continuous depended variable with two or more independent variables.
- The independent variables can be continuous or categorical (dummy coded as appropriate)
- Independent variables should not be multi-collinear

Correlation Check

```
## Correlation check
inc_exp.corr()
```

	Mthly_HH_Income	Mthly_HH_Expense	No_of_Fly_Members	Emi_or_Rent_Amt	Annual_HH_Income	No_of_Earning_M
Mthly_HH_Income	1.000000	0.649215	0.448317	0.036976	0.970315	0.347883
Mthly_HH_Expense	0.649215	1.000000	0.639702	0.405280	0.591222	0.311915
No_of_Fly_Members	0.448317	0.639702	1.000000	0.085808	0.430868	0.597482
Emi_or_Rent_Amt	0.036976	0.405280	0.085808	1.000000	0.002716	-0.097431
Annual_HH_Income	0.970315	0.591222	0.430868	0.002716	1.000000	0.296679
No_of_Earning_Members	0.347883	0.311915	0.597482	-0.097431	0.296679	1.000000

Multiple Linear Regression

```
## Multiple Linear Regression
m_linear_mod = sm.ols(formula ="Mthly_HH_Expense ~ Mthly_HH_Income+\
                       No_of_Fly_Members+ Emi_or_Rent_Amt+\
                       Annual_HH_Income" ,data = inc_exp).fit()

m_linear_mod.summary()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       Mthly_HH_Expense   R-squared:                       0.709
Model:                            OLS   Adj. R-squared:                  0.683
Method:                 Least Squares   F-statistic:                     27.40
Date:                Tue, 27 Nov 2018   Prob (F-statistic):           1.48e-11
Time:                        13:22:56   Log-Likelihood:                -509.59
No. Observations:                  50   AIC:                             1029.
Df Residuals:                      45   BIC:                             1039.
Df Model:                           4
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [95.0% Conf. Int.]
-------------------------------------------------------------------------------------
Intercept         -5124.8763   2818.362     -1.818      0.076     -1.08e+04   551.597
Mthly_HH_Income       0.4092      0.157      2.608      0.012         0.093     0.725
No_of_Fly_Members  3224.4195    719.071      4.484      0.000      1776.136  4672.703
Emi_or_Rent_Amt       0.6569      0.158      4.162      0.000         0.339     0.975
Annual_HH_Income     -0.0167      0.013     -1.314      0.196        -0.042     0.009
==============================================================================
Omnibus:                        0.142   Durbin-Watson:                   2.377
Prob(Omnibus):                  0.932   Jarque-Bera (JB):                0.013
Skew:                           0.035   Prob(JB):                        0.993
Kurtosis:                       2.963   Cond. No.                     1.75e+06
==============================================================================
```

Note : The Beta of Mthly_HH_Income is **Positive** and Beta of Annual_HH_Income is Negative.

Both are Collinear with each other and is leading to Multi-Collinearity Problem

Multiple Linear Regression…

```
In [42]: m_linear_mod = sm.ols(formula ="Mthly_HH_Expense ~ Mthly_HH_Income+\
    ...:                       No_of_Fly_Members+ Emi_or_Rent_Amt+\
    ...:                       Annual_HH_Income" ,data = inc_exp).fit()

In [43]: m_linear_mod.summary()
Out[43]:
<class 'statsmodels.iolib.summary.Summary'>
"""
```

Variance Inflation Factor - VIF

- Multi-collinearity is typically checked using VIF
- **Varaince inflation factors (VIF)** measure how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related.
- **VIF = 1 / (1 – R^2)**
- (1 – R^2) for each independent Variable is computed by Regressing that Variable w.r.t all other Independent Variable. For e.g.

  Mthly_HH_Income = f (No_of_Fly_Members, Emi_or_Rent_Amt , Annual_HH_Income)
  No_of_Fly_Members = f (Mthly_HH_Income, Emi_or_Rent_Amt , Annual_HH_Income)
  Annual_HH_Income = f (Mthly_HH_Income, Emi_or_Rent_Amt , No_of_Fly_Members)
  Emi_or_Rent_Amt = f (Mthly_HH_Income, No_of_Fly_Members , Annual_HH_Income)

- By regressing each variable with other we trying to find how much of variance of a variable can be explained by all other variables taken together

VIF check in Python

```
In [63]: def VIF(formula,data):
    ...:     import pip #To install packages
    ...:     #pip.main(["install","dmatrices"])
    ...:     #pip.main(["install","statsmodels"])
    ...:     from patsy import dmatrices
    ...:     from statsmodels.stats.outliers_influence import variance_inflation_factor
    ...:     y , X = dmatrices(formula,data = data,return_type="dataframe")
    ...:     vif = pd.DataFrame()
    ...:     vif["VIF Factor"] = [variance_inflation_factor(X.values, i) \
    ...:        for i in range(X.shape[1])]
    ...:     vif["features"] = X.columns
    ...:     return(vif.round(1))

In [64]: VIF=VIF("Mthly_HH_Expense ~ Mthly_HH_Income+\
    ...:         No_of_Fly_Members+ Emi_or_Rent_Amt+\
    ...:         Annual_HH_Income" ,data = inc_exp)

In [65]: VIF
Out[65]:
   VIF Factor           features
0         8.6          Intercept
1        17.7     Mthly_HH_Income
2         1.3   No_of_Fly_Members
3         1.0     Emi_or_Rent_Amt
4        17.4    Annual_HH_Income
```

VIF	Status of predictors
VIF = 1	Not correlated
1 < VIF < 5	Moderately correlated
VIF > 5 to 10	Highly correlated

Multiple Linear Regression

- Multiple linear Regression Model

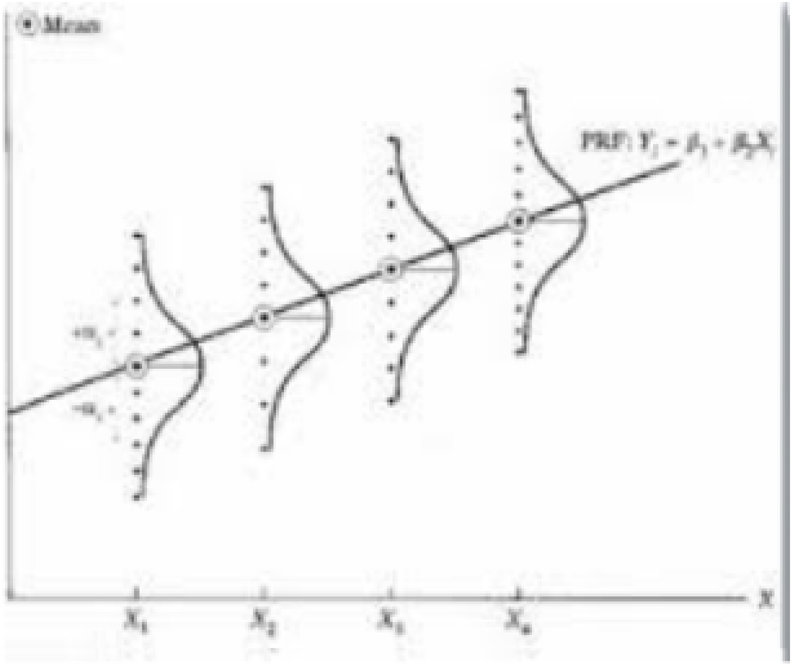
```
In [71]: m_linear_mod = sm.ols(formula = "Mthly_HH_Expense ~ Mthly_HH_Income+\
    ...:                               No_of_Fly_Members+ Emi_or_Rent_Amt",
    ...:                               data = inc_exp).fit()

In [72]: m_linear_mod.params
Out[72]:
Intercept           -5148.070385
Mthly_HH_Income         0.210439
No_of_Fly_Members    3232.573874
Emi_or_Rent_Amt         0.685093
dtype: float64
```

Summary of Multiple Linear Regression Model

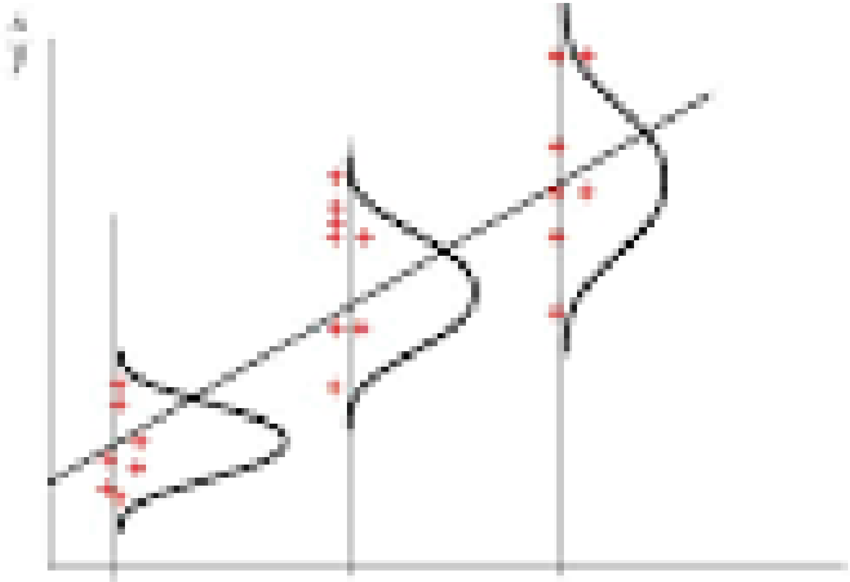
```
In [73]: m_linear_mod.summary()
Out[73]:
<class 'statsmodels.iolib.summary.Summary'>
"""
                            OLS Regression Results
==============================================================================
Dep. Variable:        Mthly_HH_Expense   R-squared:                       0.698
Model:                             OLS   Adj. R-squared:                  0.678
Method:                  Least Squares   F-statistic:                     35.40
Date:                 Tue, 08 Aug 2017   Prob (F-statistic):           5.17e-12
Time:                         11:55:12   Log-Likelihood:                -510.53
No. Observations:                   50   AIC:                             1029.
Df Residuals:                       46   BIC:                             1037.
Df Model:                            3
Covariance Type:             nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
Intercept         -5148.0704   2840.472     -1.812      0.076   -1.09e+04     569.503
Mthly_HH_Income       0.2104      0.042      5.009      0.000       0.126       0.295
No_of_Fly_Members  3232.5739    724.699      4.461      0.000    1773.830    4691.318
Emi_or_Rent_Amt       0.6851      0.158      4.347      0.000       0.368       1.002
==============================================================================
Omnibus:                        0.916   Durbin-Watson:                   2.326
Prob(Omnibus):                  0.633   Jarque-Bera (JB):                0.560
Skew:                           0.258   Prob(JB):                        0.756
Kurtosis:                       3.041   Cond. No.                     1.46e+05
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.46e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
"""
```

Homoscedasticity Vs Heteroscedasticity

Homoscedasticity simply stated means variance of the error term across observations is same (homogeneous)



Heterocedasticity occurs when the variance of the error term differs across observations



Thank you

Contact us:
ar.jakhotia@k2analytics.co.in

Earning is in Learning
- Rajesh Jakhotia