K2 Analytics
Building Skills, Building Individuals

Basic Data Mining Techniques
Linear Regression

- Rajesh Jakhotia

26-Oct-2015

Earning is in Learning
- Rajesh Jakhotia

About K2 Analytics

At K2 Analytics, we believe that skill development is very important for the growth of an individual, which in turn leads to the growth of Society & Industry and ultimately the Nation as a whole. For this it is important that access to knowledge and skill development trainings should be made available easily and economically to every individual.

Our Vision: *"To be the preferred partner for training and skill development"*

Our Mission: *"To provide training and skill development training to individuals, make them skilled & industry ready and create a pool of skilled resources readily available for the industry"*

*We have chosen Business Intelligence and Analytics as our focus area. With this endeavour we make this presentation on "**Basic Data Mining Techniques**" accessible to all those who wish to learn Analytics. We hope it is of help to you. For any feedback / suggestion or if you are looking for job in analytics then feel free to write back to us at ar.jakhotia@k2analytics.co.in*

K2 Analytics
Building Skills, Building Individuals

Linear Regression

Introduction to Linear Regression

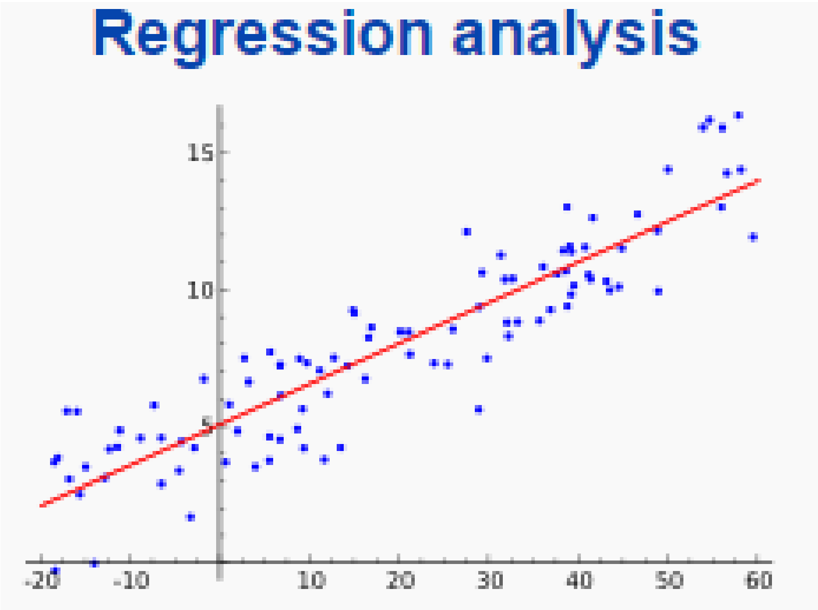
Ordinary Least Square

Simple Linear Regression

Multiple 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 Relationship … e.g.

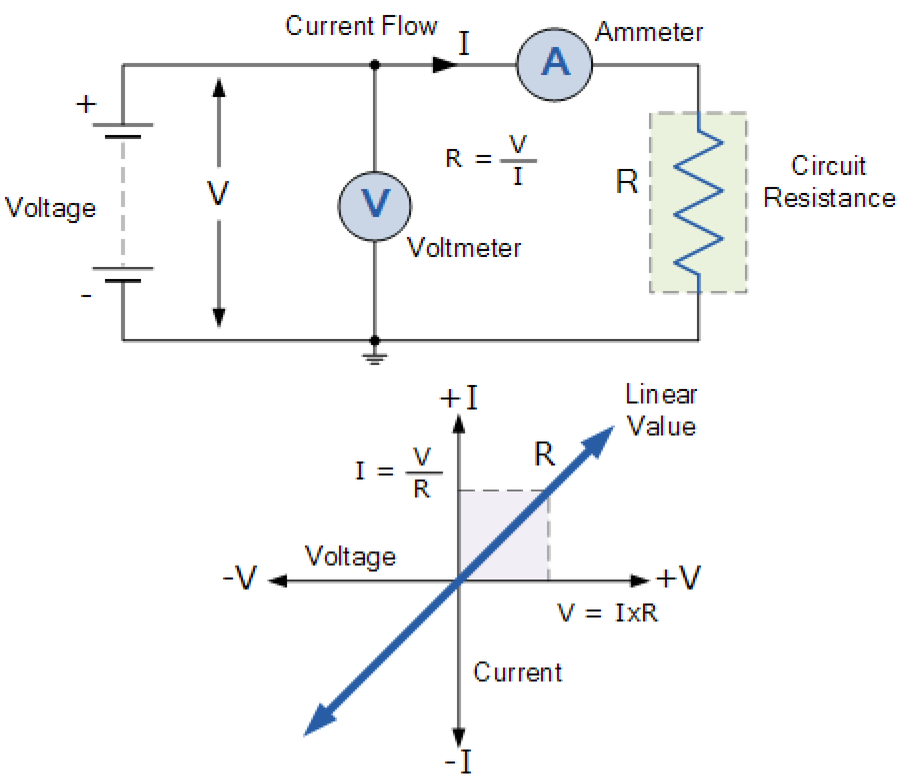
Ohm's Law:

- 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

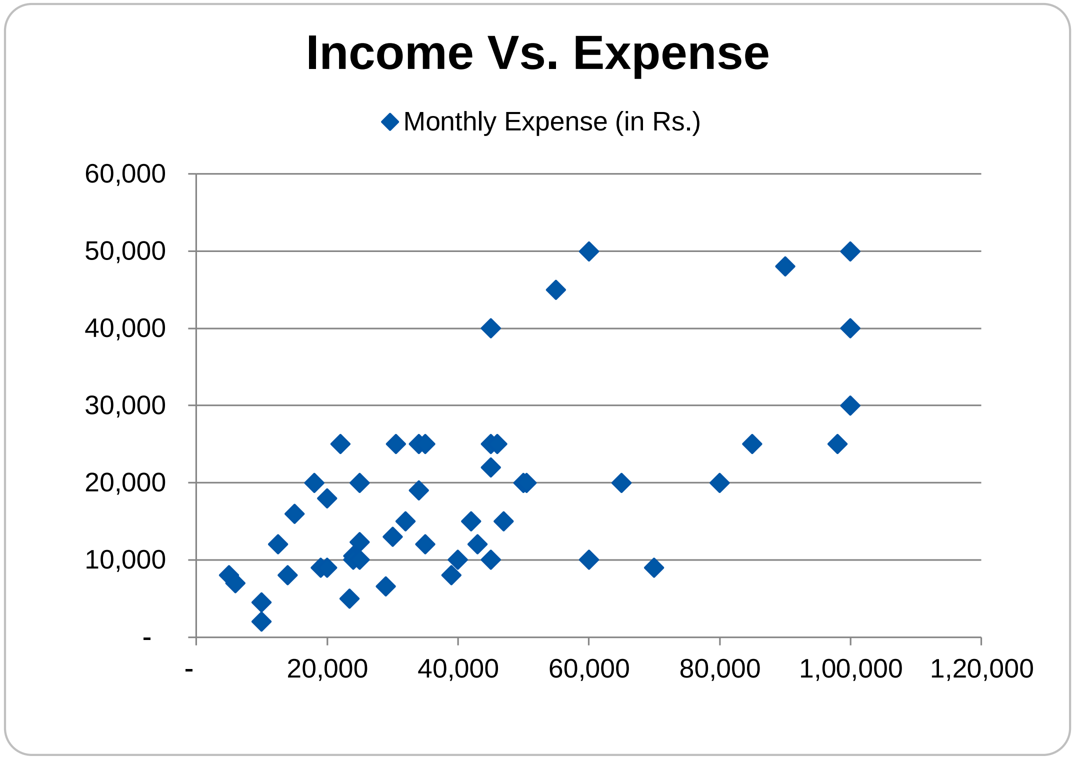
- In a circuit board for a given Resistance R, as you increase the Voltage V, the Current I increases proprotionately



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

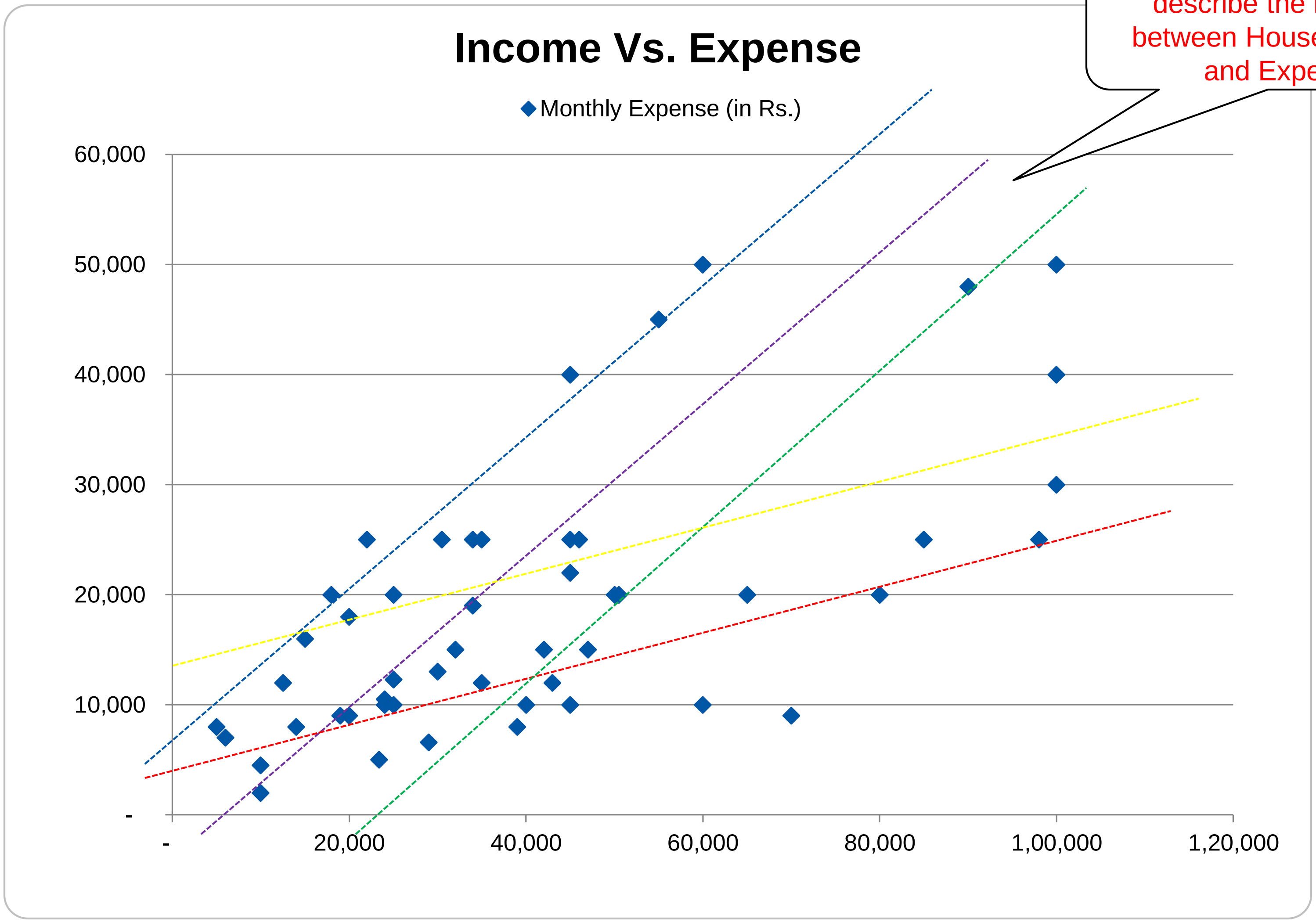
Sample Monthly Income-Expense Data of a Household

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



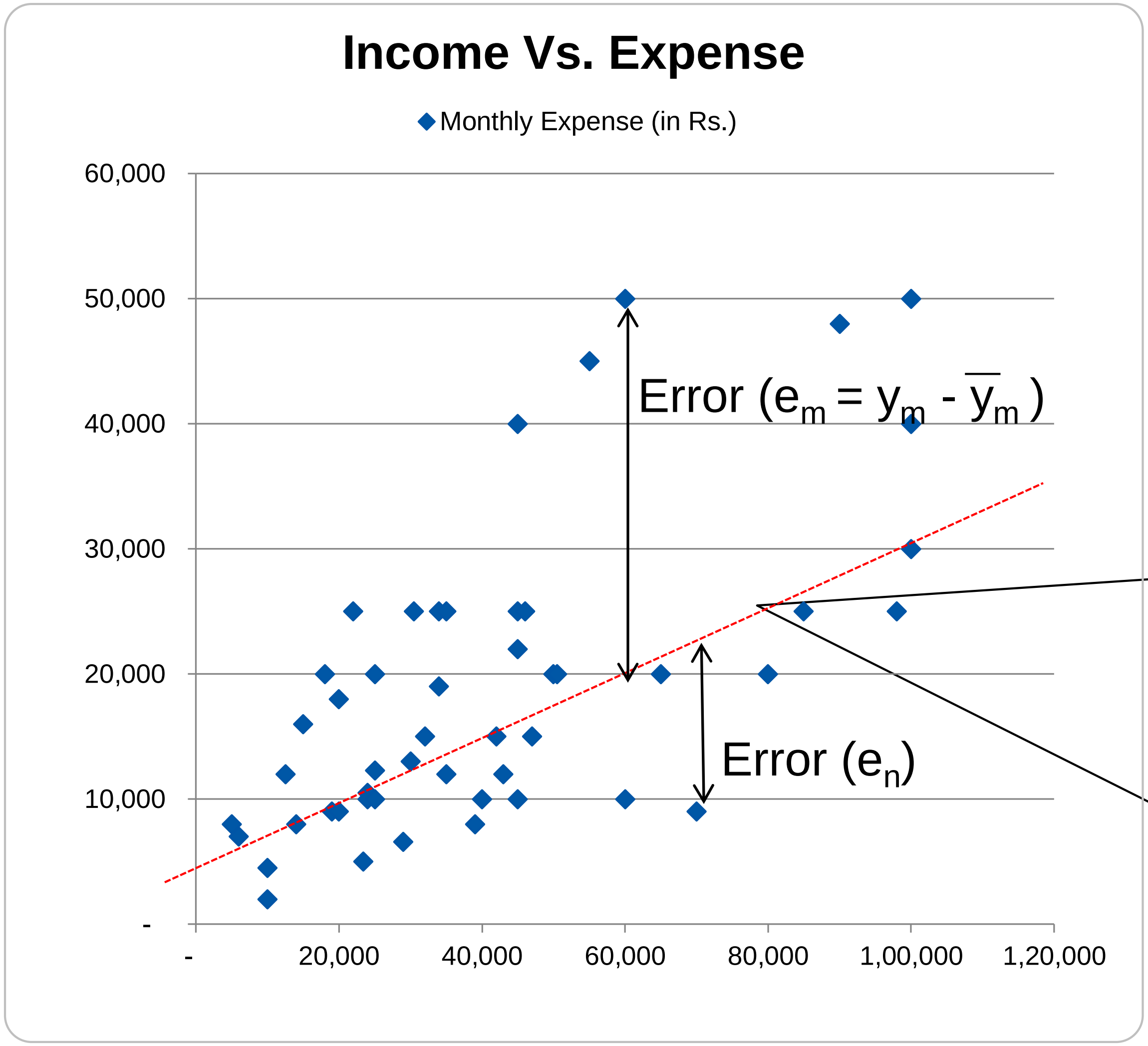
We have to find the relationship between Income and Expenses of a household

Line of Best Fit



Which of these lines best describe the relationship between Household Income and Expenses ?

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)

$Y_{i(hat)} = b_o + b_1 X_i$ is the **sample regression equation**

SSE = $\Sigma\ e_{i(hat)}^2$ (1)
= $\Sigma\ (Y_i - Y(_{i(hat)})^2$ (2)
= $\Sigma\ (Y_i - b_o - b_1 X_i)^2$ (3)

Using calculus we get

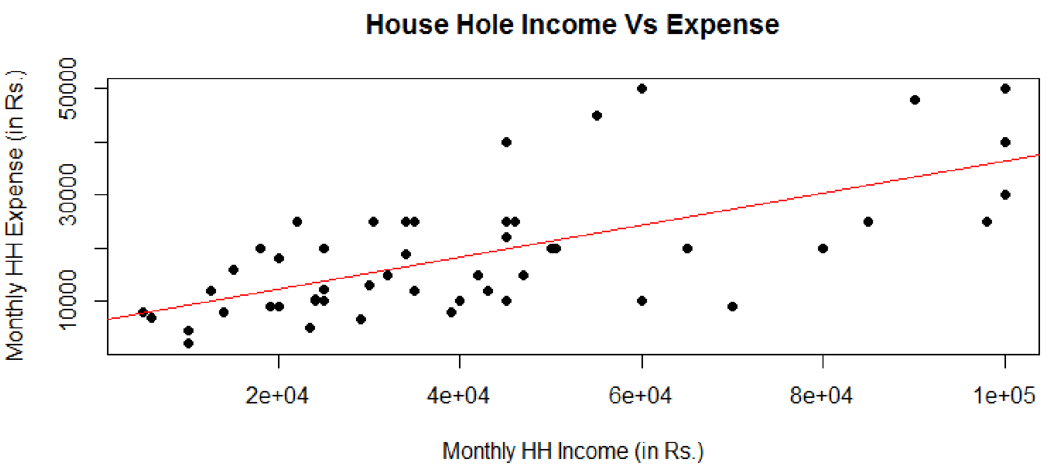
$$b_o = \frac{\Sigma Y_i - b_1 \Sigma X_i}{n}$$

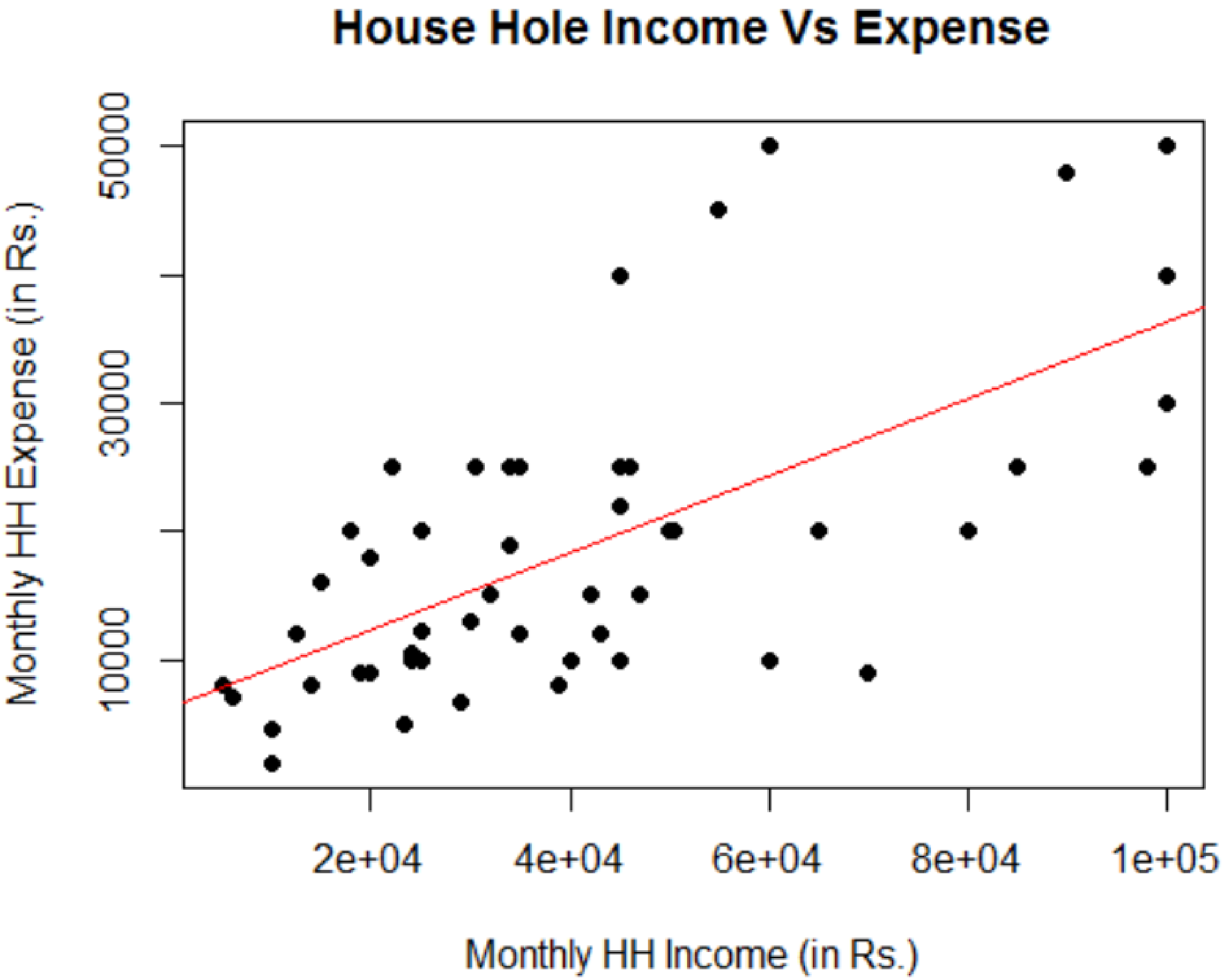
$$b_1 = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

OLS - Ordinary Least Square

Simple Linear Regression in R

```
setwd("D:/K2Analytics/Datafile")
inc_exp <- read.csv("Inc_Exp_Data.csv", header=T)
View(inc_exp)
## Scatter plot
plot ( inc_exp$Mthly_HH_Income, inc_exp$Mthly_HH_Expense,
    main=" House Hole Income Vs Expense ",
    xlab="Monthly HH Income (in Rs.)", ylab="Monthly HH Expense (in Rs.)",
    pch=19 )
# Add fit lines; regression line (y~x)
abline(lm(inc_exp$Mthly_HH_Expense ~
    inc_exp$Mthly_HH_Income),
    col="red")
```




House Hole Income Vs Expense
Monthly HH Expense (in Rs.)
10000
30000
50000
2e+04
4e+04
6e+04
8e+04
1e+05
Monthly HH Income (in Rs.)

Simple Linear Regression in R

Linear Regression Model

linear_mod <- lm(Mthly_HH_Expense ~ Mthly_HH_Income, data = inc_exp)

Get the coefficient and intercept

linear_mod

```
Call:
lm(formula = Mthly_HH_Expense ~ Mthly_HH_Income, data = inc_exp)

Coefficients:
    (Intercept)  Mthly_HH_Income
      6319.1018           0.3008
```

Get the R-Squared (Coefficient of Determination)

Coefficient of Determination is how much of the total variance in Y is explained by the model, i.e. variance in X

summary(linear_mod)$r.squared

```
[1] 0.4214804
```

Linear Relationship significance test

```
> summary(linear_mod)

Call:
lm(formula = Mthly_HH_Expense ~ Mthly_HH_Income, data = inc_exp)

Residuals:
   Min     1Q Median     3Q    Max 
-18372  -6263  -1940   5164  25635 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     6.319e+03  2.489e+03   2.539   0.0144 *  
Mthly_HH_Income 3.008e-01  5.086e-02   5.914  3.4e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9291 on 48 degrees of freedom
Multiple R-squared:  0.4215,	Adjusted R-squared:  0.4094 
F-statistic: 34.97 on 1 and 48 DF,  p-value: 3.397e-07
```

p-value suggests that the **linear relationship** between expense and income is significant

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 dependent variable from two or more independent variables.
- The independent variables can be continuous or categorical (dummy coded as appropriate)
- Independent variables should not be multi-collinear

```
setwd("D:/K2Analytics/Datafile")

inc_exp <- read.csv("Inc_Exp_Data.csv", header=T)

## Correlation between Indpendent Variables

cor(inc_exp)
```

	Mthly_HH_Income	Mthly_HH_Expense	No_of_Fly_Members	Emi_or_Rent_Amt
Mthly_HH_Income	1.00000000	0.6492153	0.44831731	0.03697611
Mthly_HH_Expense	0.64921525	1.0000000	0.63970156	0.40528027
No_of_Fly_Members	0.44831731	0.6397016	1.00000000	0.08580759
Emi_or_Rent_Amt	0.03697611	0.4052803	0.08580759	1.00000000

Multiple Linear Regression...contd

Multiple Linear Regression Model

m_linear_mod <- lm (Mthly_HH_Expense ~ Mthly_HH_Income + No_of_Fly_Members
+ Emi_or_Rent_Amt + Annual_HH_Income,
data = inc_exp
)

summary(m_linear_mod)

```
Call:
lm(formula = Mthly_HH_Expense ~ Mthly_HH_Income + No_of_Fly_Members +
    Emi_or_Rent_Amt + Annual_HH_Income, data = inc_exp)

Residuals:
     Min       1Q   Median       3Q      Max
-14887.4  -3455.9    588.8   3955.7  14494.0

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5.125e+03  2.818e+03  -1.818 0.075664 .
Mthly_HH_Income    4.092e-01  1.569e-01   2.608 0.012318 *
No_of_Fly_Members  3.224e+03  7.191e+02   4.484 5.01e-05 ***
Emi_or_Rent_Amt    6.569e-01  1.578e-01   4.162 0.000141 ***
Annual_HH_Income  -1.666e-02  1.268e-02  -1.314 0.195533
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6806 on 45 degrees of freedom
Multiple R-squared:  0.7089,	Adjusted R-squared:  0.6831
F-statistic:  27.4 on 4 and 45 DF,  p-value: 1.475e-11
```

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

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

Multi-collinearity

- Multicollinearity (also collinearity) is a phenomenon in which two or more predictor variables in a multiple regression model are highly correlated, meaning that one can be linearly predicted from the others with a substantial degree of accuracy.
  - In this situation the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.
  - Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors.
  - That is, a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.
- E.g. Monthly Income and Annual Income Variables

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

Variance Inflation Factor (VIF)

- Multi-collinearity is typically checked using VIF
- VIF = 1 / (1 - R^2)
- (1 - R^2) is also called Tolerance and it is opposite of Coefficient of Determination
- How is R^2 for each Indpendent Variable computed?
  - R^2 for each Indpendent Variable is computed by Regressing that Variable w.r.t all other Indepnedent 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 are trying to find how much of variance of a variable can be explained by all other variables taken together

Variance Inflation Factor

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

library(car)

vif (linear_mod)

```
Mthly_HH_Income  Annual_HH_Income No_of_Fly_Members    Emi_or_Rent_Amt
      17.735808         17.426934          1.259203           1.026453
```

Multiple Linear Regression

Multiple Linear Regression Model

```
m_linear_mod <- lm ( Mthly_HH_Expense ~
                     Mthly_HH_Income + No_of_Fly_Members + Emi_or_Rent_Amt,
                     data = inc_exp
                     )
```

Display the Multiple Linear Regression Model

```
m_linear_mod
```

```
Call:
lm(formula = Mthly_HH_Expense ~ Mthly_HH_Income + No_of_Fly_Members +
    Emi_or_Rent_Amt, data = inc_exp)

Coefficients:
      (Intercept)    Mthly_HH_Income  No_of_Fly_Members    Emi_or_Rent_Amt
       -5148.0704             0.2104          3232.5739             0.6851
```

Summary of Multiple Linear Regression Model

summary(m_linear_mod)

```
Call:
lm(formula = Mthly_HH_Expense ~ Mthly_HH_Income + No_of_Fly_Members +
    Emi_or_Rent_Amt, data = inc_exp)

Residuals:
     Min       1Q   Median       3Q      Max
-15684.5  -4581.5    -99.2   3522.3  16275.3

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5.148e+03  2.840e+03  -1.812   0.0765 .
Mthly_HH_Income    2.104e-01  4.201e-02   5.009 8.52e-06 ***
No_of_Fly_Members  3.233e+03  7.247e+02   4.461 5.23e-05 ***
Emi_or_Rent_Amt    6.851e-01  1.576e-01   4.347 7.56e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6860 on 46 degrees of freedom
Multiple R-squared:  0.6978,    Adjusted R-squared:  0.6781
F-statistic:  35.4 on 3 and 46 DF,  p-value: 5.172e-12
```

Note the improvement in R Squared value in Multiple Linear Model as compared to Simple Linear Model

K2 Analytics
Building Skills, Building Individuals

Thank you

Email Id : ar.jakhotia@k2analytics.co.in

Earning is in Learning
- Rajesh Jakhotia