

# Simple Linear Regression

---

UTKARSH GAIKWAD

CLASS STARTING SHARP AT 4:03 PM

# Checking relationships between 2 variables

---

Income per month	Spending
100000	50000
80000	35000
60000	29000
45000	20000
30000	15000

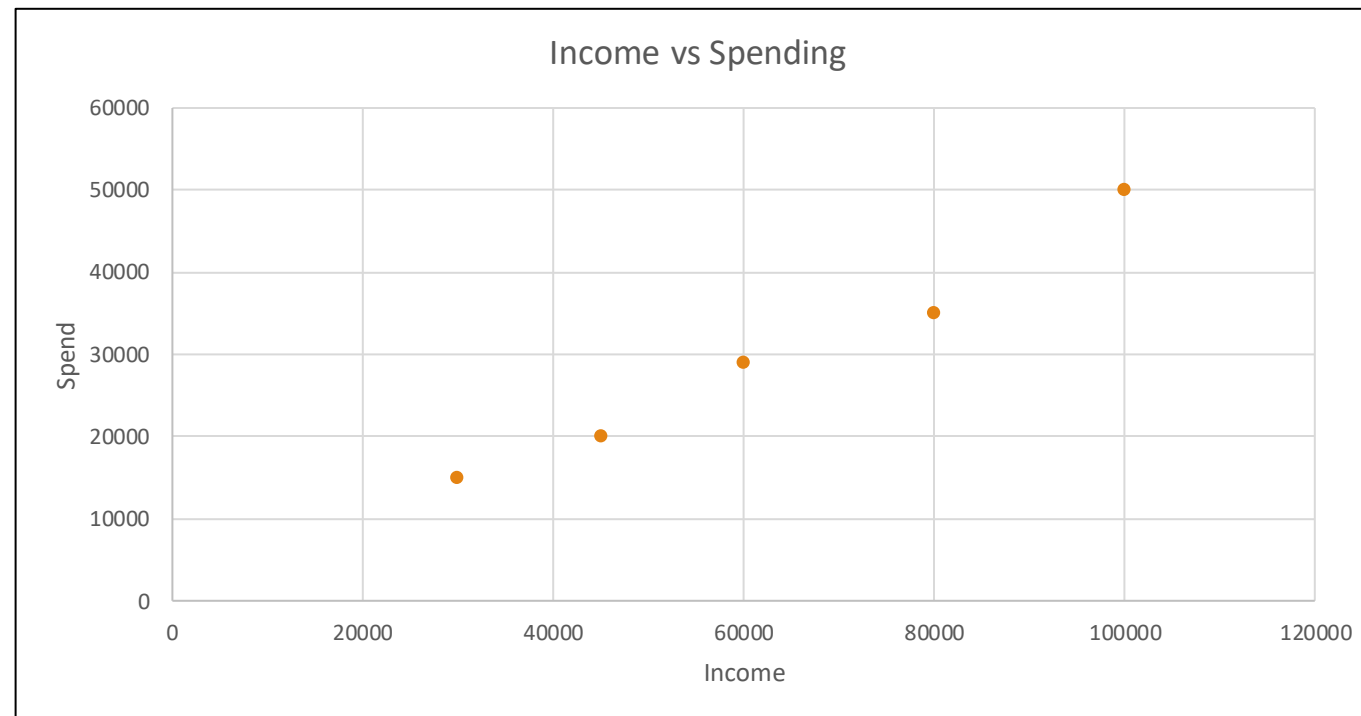
Spending ~ Income

Independent Feature

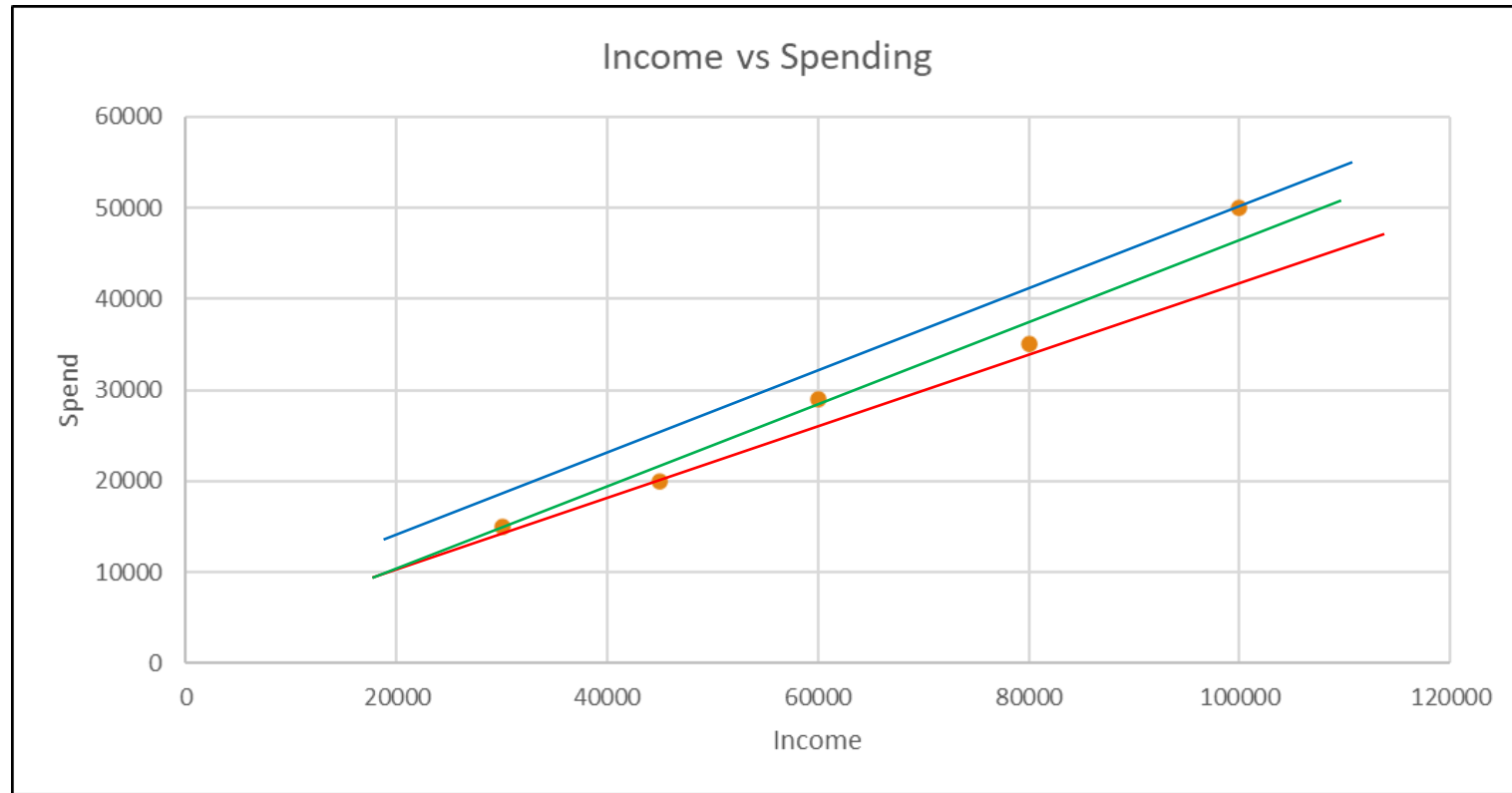
Dependent Feature / Target Feature

# Visualizing the Relationship between Income vs Spend

---

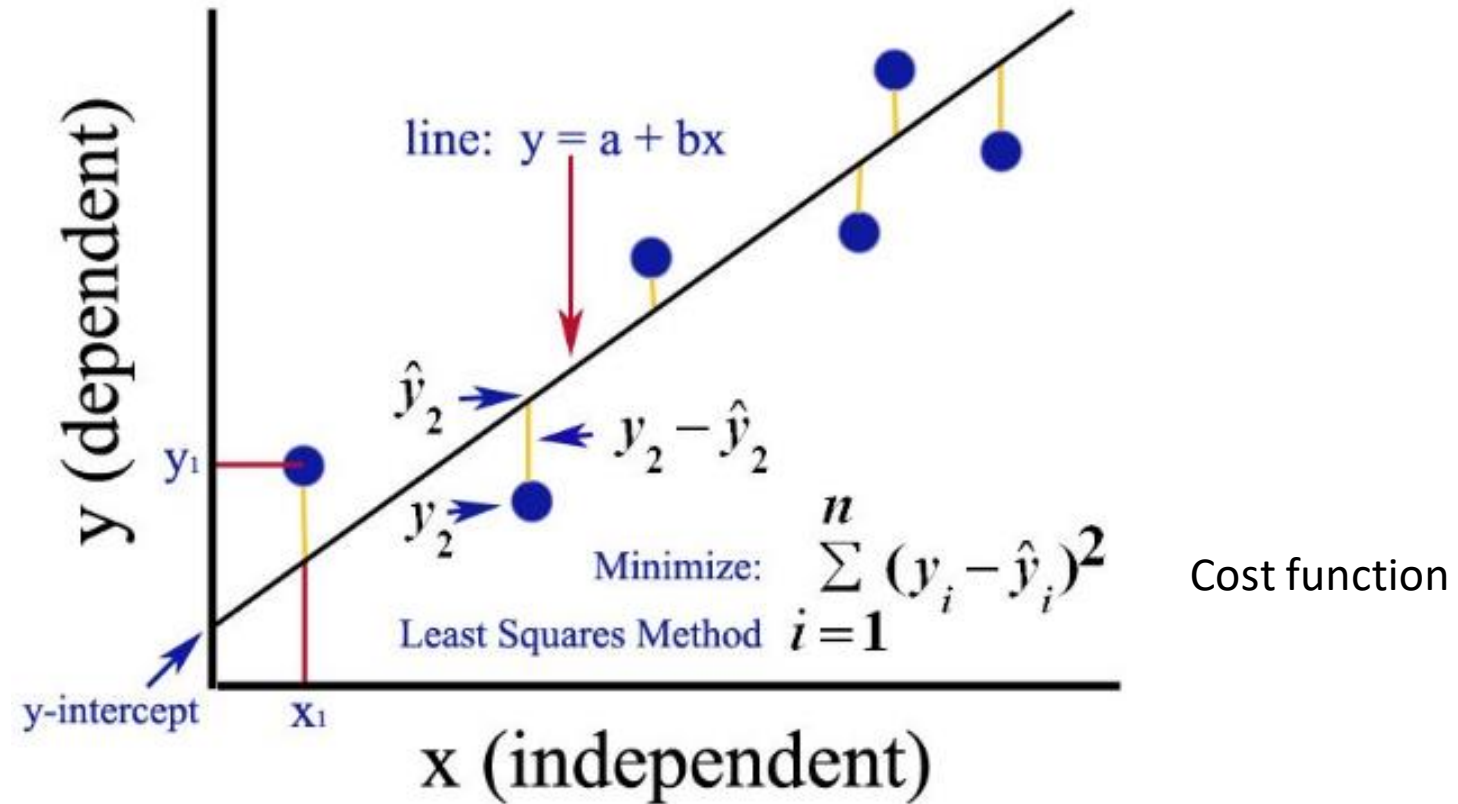


# Predicting Spend based on income



Which line  
to fit for  
this  
problem?

# Least Squared Error model



# Equation Of Line

---

$$y = mx + c$$

y : Dependent Feature(Target)

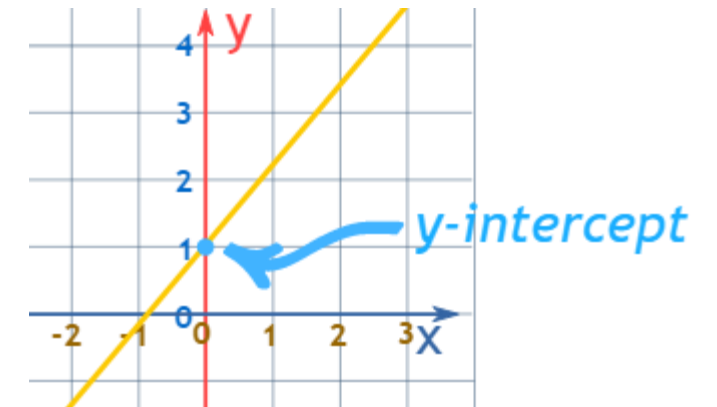
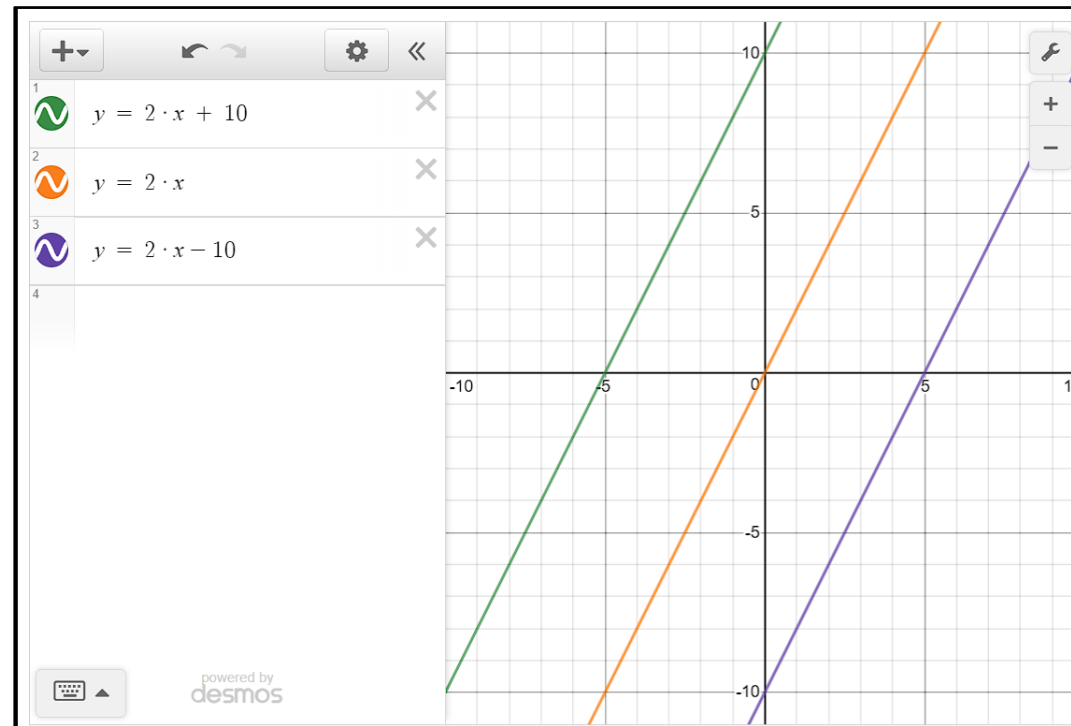
x : Independent Feature

m : Slope of a line

c : y intercept of a line

# What happens if we change Line intercept ?

At  $x = 0$  ,  $y = c$  (intercept)



# What is slope?

If  $x$  increases by 1 or unit value , how much will  $y$  change?

Eg.  $y = 2x + 3$

$x_0 = 3$  increases by one,  $x_1 = 4$

$y_0 = 2 \cdot 3 + 3 = 9$

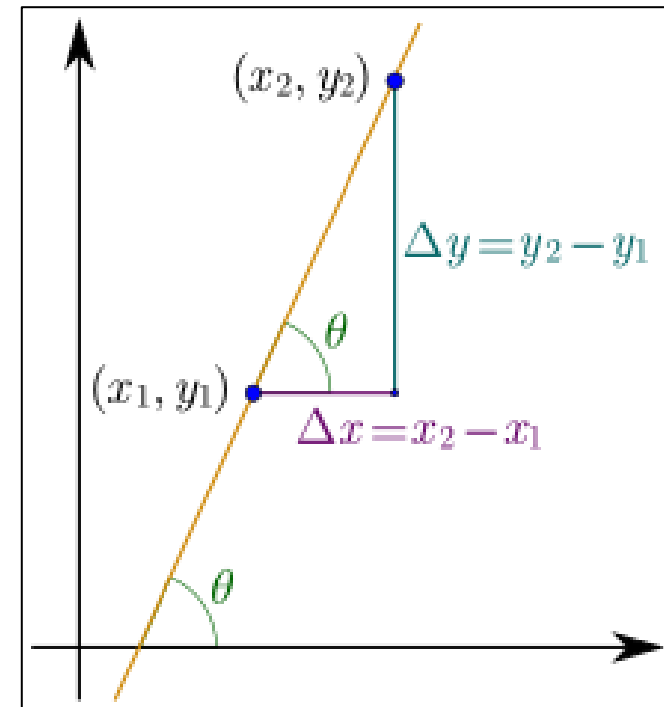
$y_1 = 2 \cdot 4 + 3 = 11$

$y_1 - y_0 = 11 - 9 = 2$

Slope =  $(y_2 - y_1) / (x_2 - x_1)$

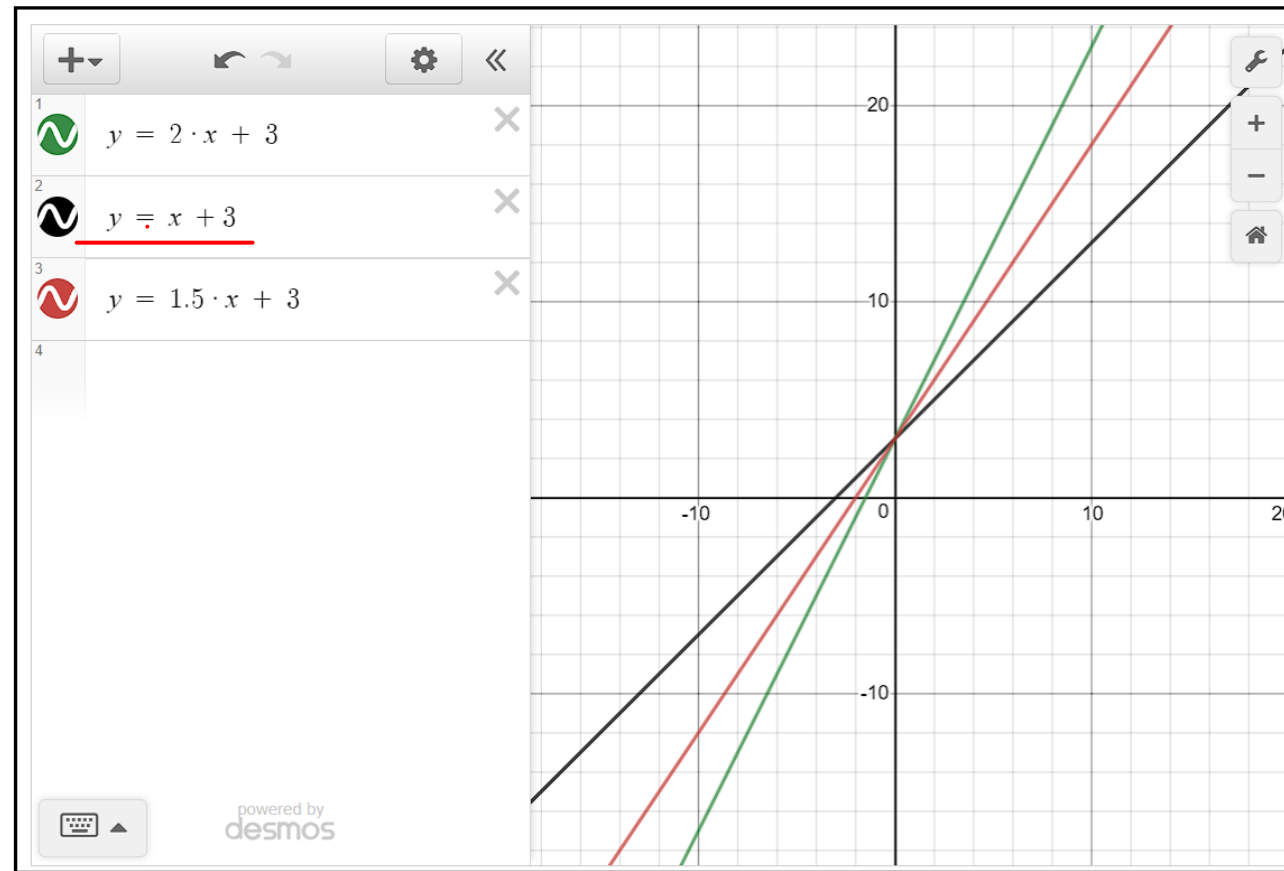
When  $x$  increased by 1 ,  $y$  increased by 2

$y = 2x + 3$  , Slope = 2





# Lines with different slopes



# Simple linear Regression Objective

---

- fit a line  $y_{actual} = \beta_0 + \beta_1 \cdot x + \varepsilon$
- $y_{pred} = \beta_0 + \beta_1 \cdot x$
- Minimise the Squared error for given relationships
- Least Squares error method
- Formula for slope :  $\beta_1 = \frac{cov(x,y)}{var(x)} = \frac{\Sigma(x-\bar{x})(y-\bar{y})/n}{\Sigma(x-\bar{x})^2/n}$
- Formula for Intercept :  $\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$
- $\bar{x}$ : **Mean of all  $x$  values**
- $\bar{y}$ : **Mean of all  $y$  values**

# Solving income vs Spend problem

---

Income per month (x)	Spending (y)
100000	50000
80000	35000
60000	29000
45000	20000
30000	15000

$$Spending = \beta_0 + \beta_1 \cdot Income + \varepsilon$$

# B0 and B1 Calculation

Income per month (x)	Spending (y)	X mean	Y mean	x-xmean	y-ymean	prod	(x-xmean)^2
100000	50000	63000	29800	37000	20200	747400000	1369000000
80000	35000	63000	29800	17000	5200	88400000	289000000
60000	29000	63000	29800	-3000	-800	2400000	9000000
45000	20000	63000	29800	-18000	-9800	176400000	324000000
30000	15000	63000	29800	-33000	-14800	488400000	1089000000

300600000

616000000

$$\text{Formula for slope : } \beta_1 = \frac{\text{cov}(x,y)}{\text{var}(x)} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

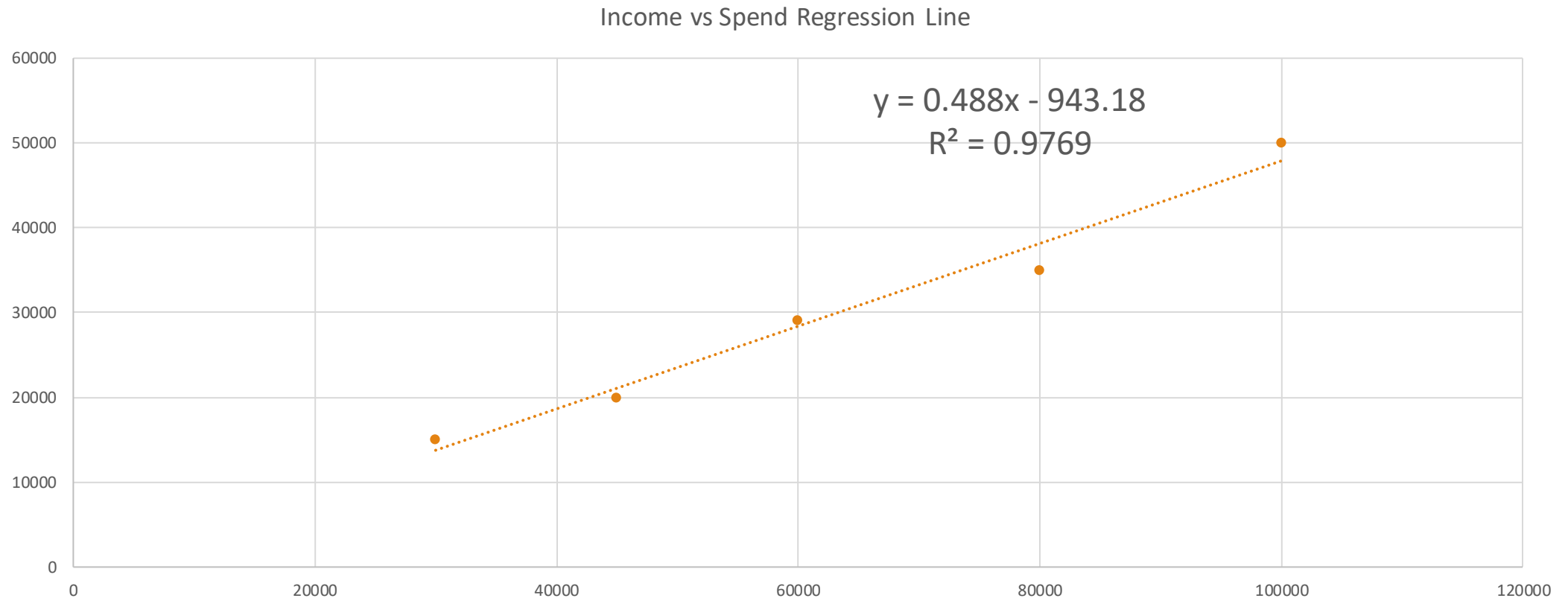
$$\text{Formula for Intercept : } \beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$$

SUM	1503000000	3080000000
COV	300600000	616000000

<b>B0</b>	<b>0.4880</b>
<b>B1</b>	<b>-943.1818</b>

# Regression Line

---



# Metrics to evaluate Model (Regression only)

---

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)
- R squared

# Mean Squared Error

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Mean      Error      Squared

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

Income per month (x)	Spending (y)	Ycap	Error	Squared Error
100000	50000	47855.52	2144.48	4598796.70
80000	35000	38095.78	-3095.78	9583848.98
60000	29000	28336.04	663.96	440844.26
45000	20000	21016.23	-1016.23	1032731.07
30000	15000	13696.43	1303.57	1699298.47

B1	0.488
B0	-943.182

Sum	17355519.48
Count	5
Average	3471103.896
MSE	3471103.896
RMSE	1863.089879

# Mean Absolute Error

---

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Income per month (x)	Spending (y)	Ycap	Error	absolute error
100000	50000	47855.52	2144.48	2144.48
80000	35000	38095.78	-3095.78	3095.78
60000	29000	28336.04	663.96	663.96
45000	20000	21016.23	-1016.23	1016.23
30000	15000	13696.43	1303.57	1303.57

Sum	8224.03
Count	5
MAE	1644.81



# Mean Absolute Percentage Error

---

$$MAPE = \frac{1}{N} \sum_{t=1}^N \left| \frac{E_t - A_t}{A_t} \right|$$

Income per month	Spending (y)	Ycap	Error	absolute error	Abs Percentage error
100000	50000	47855.5	2144	2144.5	4.29%
80000	35000	38095.8	-3096	3095.8	8.85%
60000	29000	28336	664	663.96	2.29%
45000	20000	21016.2	-1016	1016.2	5.08%
30000	15000	13696.4	1304	1303.6	8.69%
				MAPE	5.84%

# R squared metric

---

Formula

$$R^2 = 1 - \frac{RSS}{TSS}$$

$R^2$  = coefficient of determination

$RSS$  = sum of squares of residuals

$TSS$  = total sum of squares

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

# R squared metric Calculation

---

Income per month (x)	Spending (y)	Ycap	Error	Error^2	yi - ymean	(yi - ymean)^2
100000	50000	47855.52	2144.48	4598796.70	20200	408040000
80000	35000	38095.78	-3095.78	9583848.98	5200	27040000
60000	29000	28336.04	663.96	440844.26	-800	640000
45000	20000	21016.23	-1016.23	1032731.07	-9800	96040000
30000	15000	13696.43	1303.57	1699298.47	-14800	219040000

ymean	29800
-------	-------

RSS	17355519.48	TSS	750800000
R2	0.9769		

# Thank You

---

PING ME ON SKYPE GROUP FOR ANY QUERIES

PERFORM THE PRACTICAL AND YOU CAN LEAVE FOR THE DAY