Linear Regression model

Dr. Pritam Anand.
Assistant Professor,
DA-IICT, Gandhinagar.

Credit Card Dataset

Income	Limit	Rating	Cards	Age	Balance
14.891	3606	283	2	34	333
106.025	6645	483	3	82	903
104.593	7075	514	4	71	580
148.924	9504	681	3	36	964
55.882	4897	357	2	68	331
80.18	8047	569	4	77	1151
20.996	3388	259	2	37	203
71.408	7114	512	2	87	872
15.125	3300	266	5	66	279
71.061	6819	491	3	41	1350
63.095	8117	589	4	30	1407

Task

Income (hundred thousand dollar) x	Balance (thousand dollar) y
0.550798	5.651202
0.708148	7.321263
0.290905	5.167304
0.510828	5.609367
0.892947	9.406379
0.896293	9.379439
0.125585	2.734997
0.207243	4.876649
0.051467	3.584138
0.44081	5.437239

Training data

Task

Income (hundred thousand dollar)	Balance (thousand dollar)
0.550798	5.651202
0.708148	7.321263
0.290905	5.167304
0.510828	5.609367
0.892947	9.406379
0.896293	9.379439
0.125585	2.734997
0.207243	4.876649
0.051467	3.584138
0.44081	5.437239

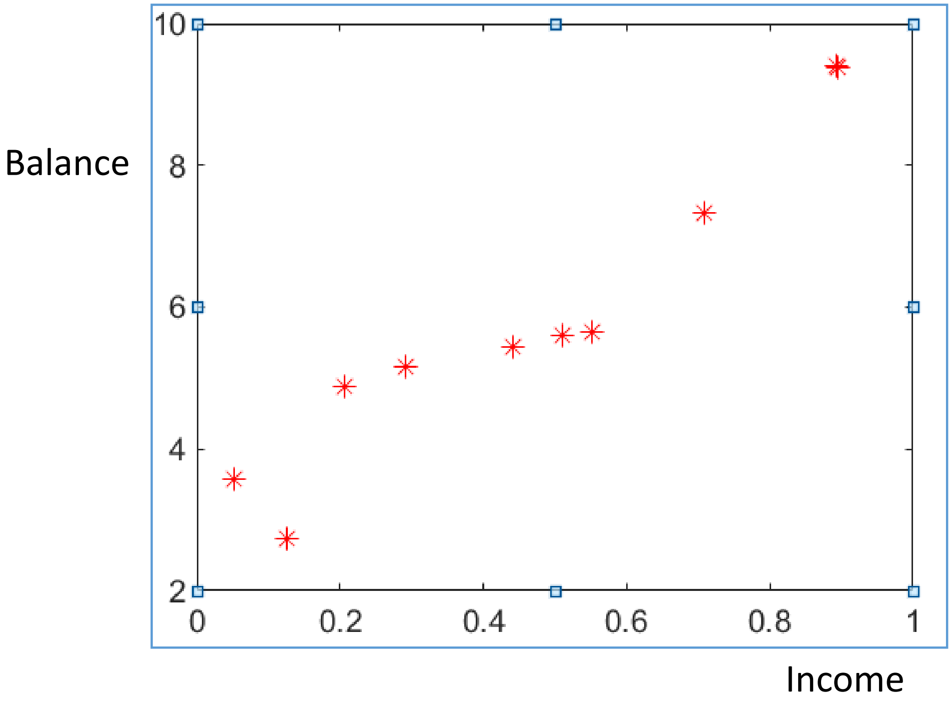
Training data

Income (x) (hundred thousand dollar)	Balance (y) (thousand dollar)
0.96703	[redacted]
0.547232	[redacted]
0.972684	[redacted]
0.714816	[redacted]
0.697729	[redacted]
0.216089	[redacted]
0.976274	[redacted]
0.00623	[redacted]
0.252982	[redacted]
0.434792	[redacted]
0.779383	[redacted]
0.197685	[redacted]
0.862993	[redacted]
0.983401	[redacted]
0.163842	[redacted]
0.597334	[redacted]
0.008986	[redacted]
0.386571	[redacted]
0.04416	[redacted]
0.956653	[redacted]

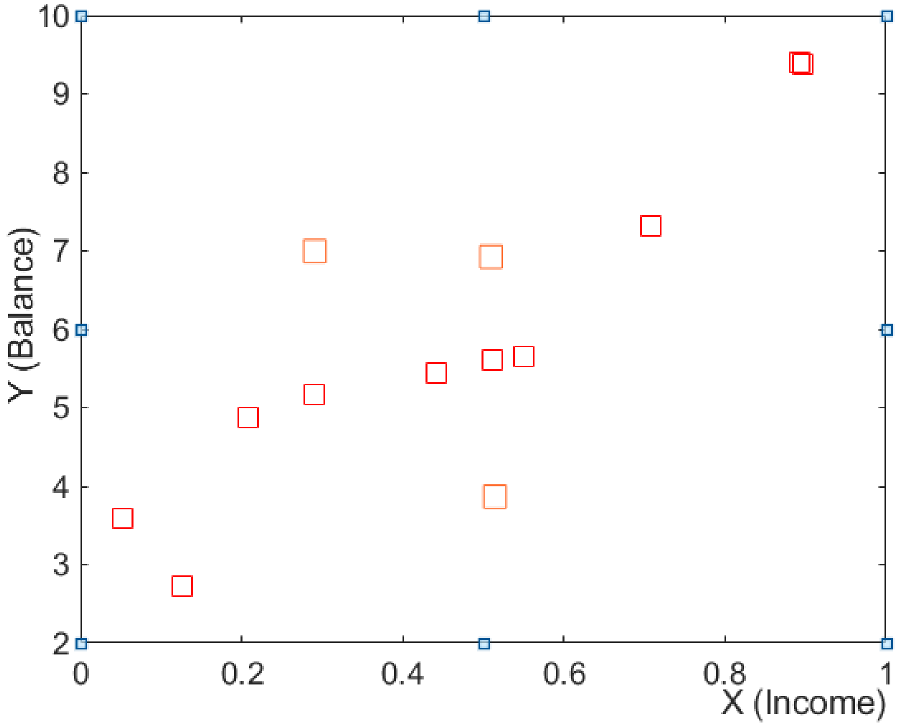
Testing Data

Task

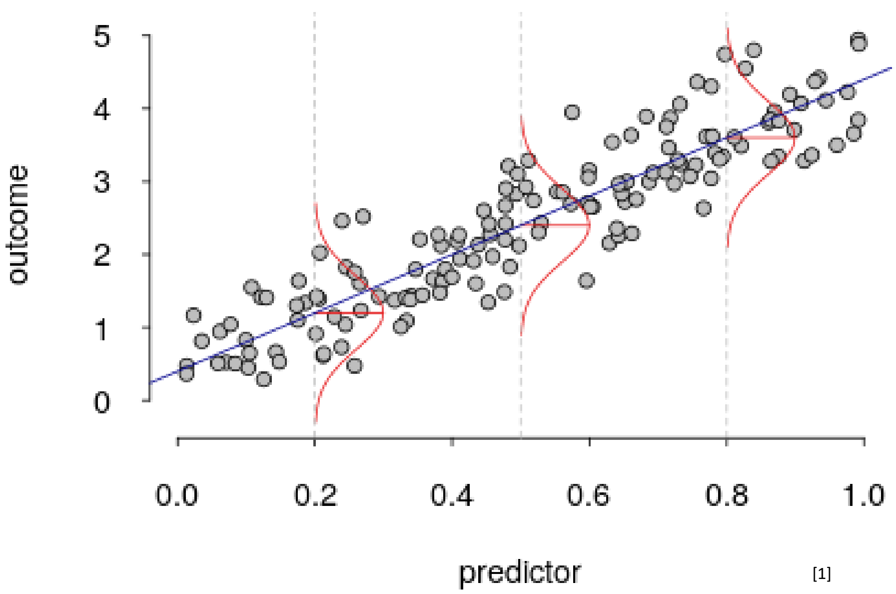
Income (hundred thousand dollar)	Balance (thousand dollar)
x_1= 0.550798	y_1= 5.651202
x_2 =0.708148	y_2= 7.321263
x_3 =0.290905	y_3= 5.167304
x_4 =0.510828	y_4= 5.609367
x_5 =0.892947	y_5= 9.406379
x_6 =0.896293	y_6= 9.379439
x_7 =0.125585	y_7= 2.734997
x_8 =0.207243	y_8= 4.876649
x_9 =0.051467	y_9= 3.584138
x_{10} =0.44081	y_{10}= 5.437239



Random Relation



Random Relation



[1]

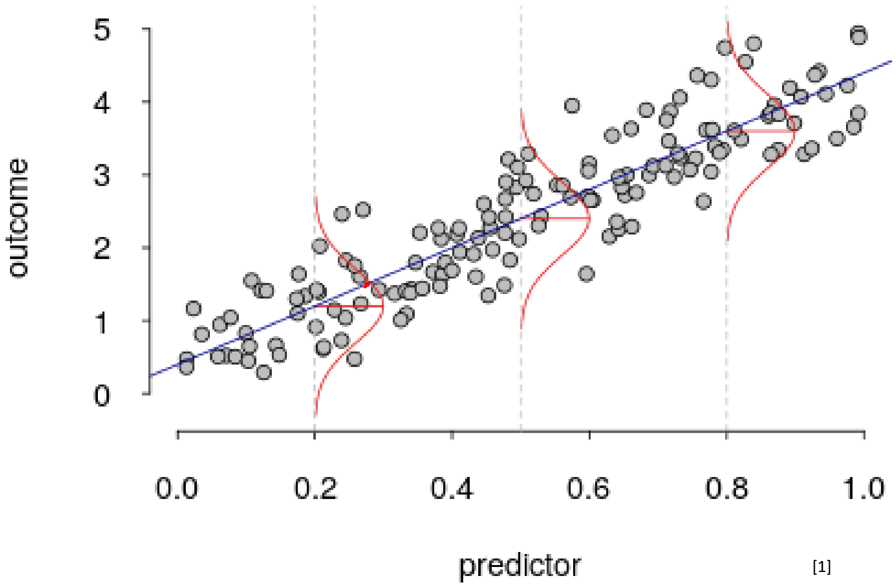
[1] Image Source:- https://www.google.com/url?sa=i&url=https%3A%2F%2Fjanhove.github.io%2Fanalysis%2F2019%2F04%2F11%2Fassumptionsrelevance&psig=AOvVaw000ZtKPluw1mRonOwupIf&ust=1692611152164000&source=images&cd=vfe&opi=89978449&ved=0CBAQjhxqFwoTCKCJ8f_56oADFQAAAAAdAAAAABAD

Regression model Assumptions

- (x_i, y_i) are iid (identically and independently distributed) random variables.

- $y_i = \quad E(y_i | x) + \epsilon_i$

- $E(\epsilon_i) = 0$

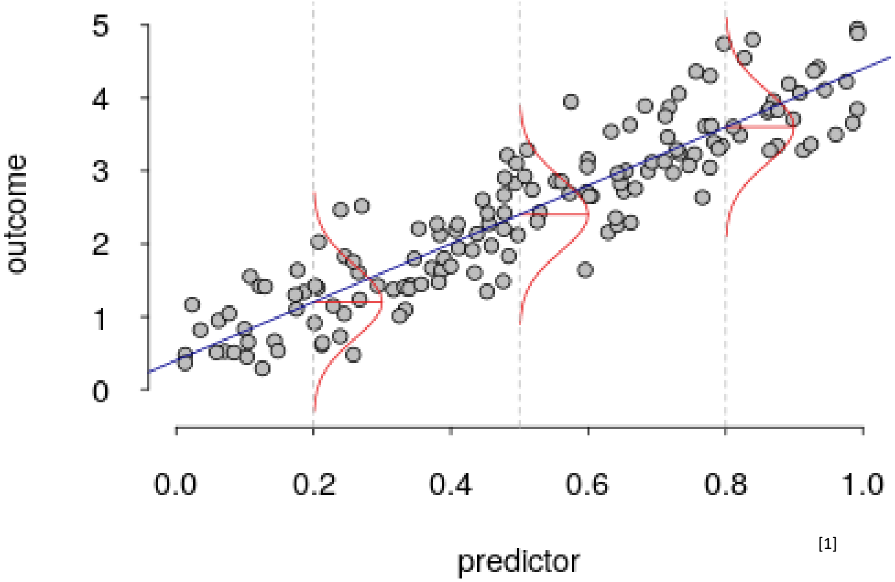


[1]

---

[1] Image Source:-
https://www.google.com/url?sa=i&url=https%3A%2F%2Fjanhove.github.io%2Fanalysis%2F2019%2F04%2F11%2Fassumptionsrelevance&psig=AOvVaw000ZtKPluw1mRonOwupIf&ust=1692611152164000&source=images&cd=vfe&opi=89978449&ved=0CBAQjhxqFwoTCKCJ8f_56oADFQAAAAAdAAAAABAD

$$y_i = \underset{\overset{\uparrow}{f(x)}}{E(y_i|x)} + \epsilon_i$$

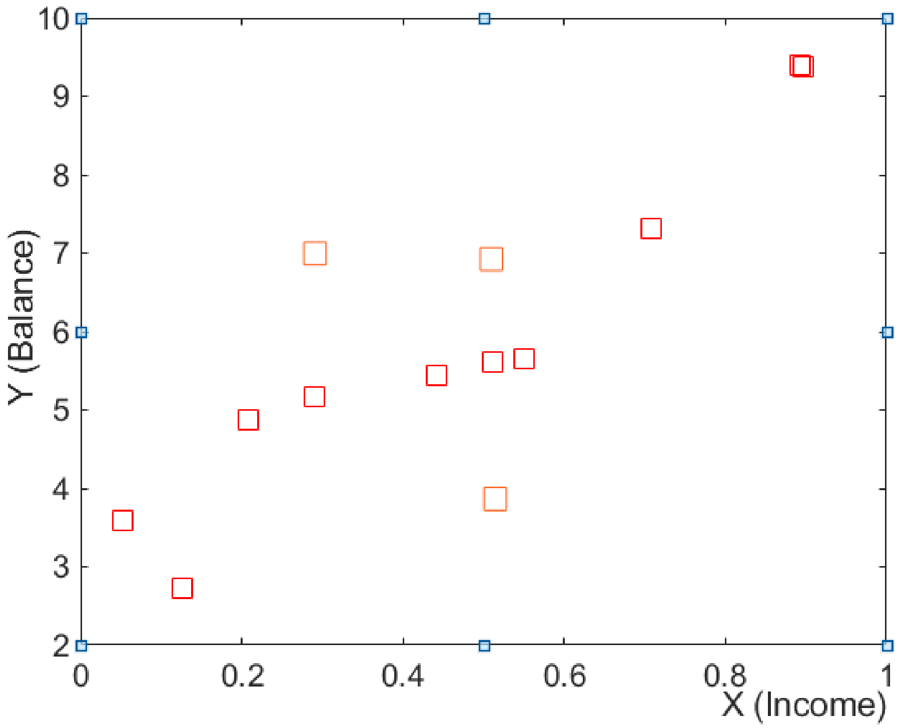


[1]

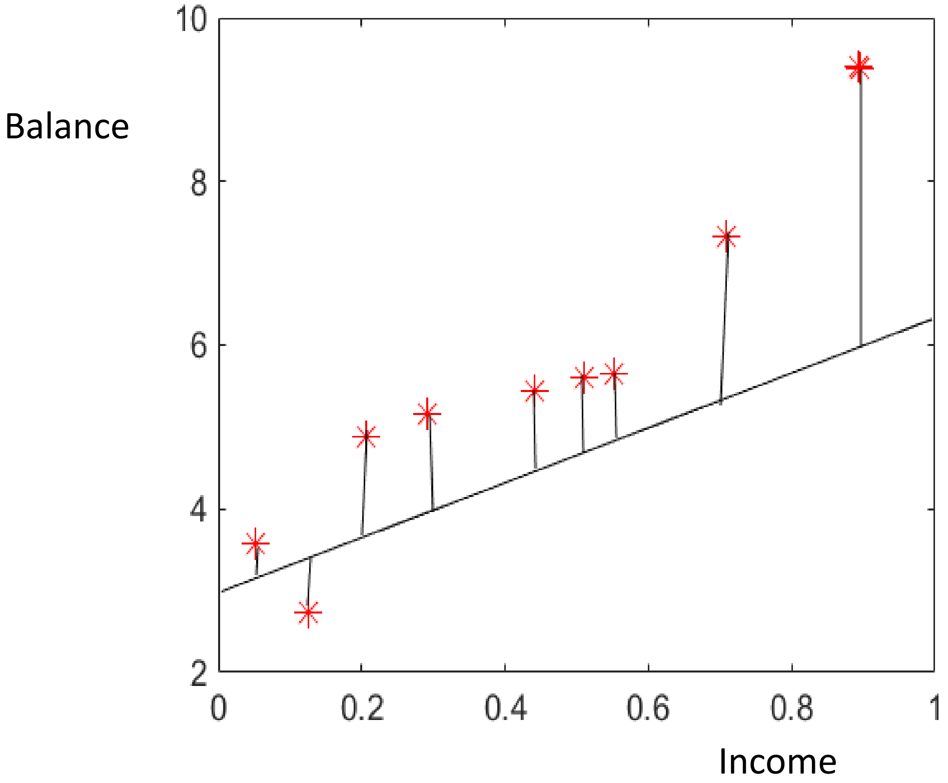
[1] Image Source:-
https://www.google.com/url?sa=i&url=https%3A%2F%2Fjanhove.github.io%2Fanalysis%2F2019%2F04%2F11%2Fassumptionsrelevance&psig=AOvVaw000ZtKPluw1mRonOwupIf&ust=1692611152164000&source=images&cd=vfe&opi=89978449&ved=0CBAQjhxqFwoTCKCJ8f_56oADFQAAAAAdAAAAABAD

$$y_i = \underset{\underset{f(x)}{\uparrow}}{E(y_i|x)} + \epsilon_i$$

$$E(\epsilon_i) = 0$$



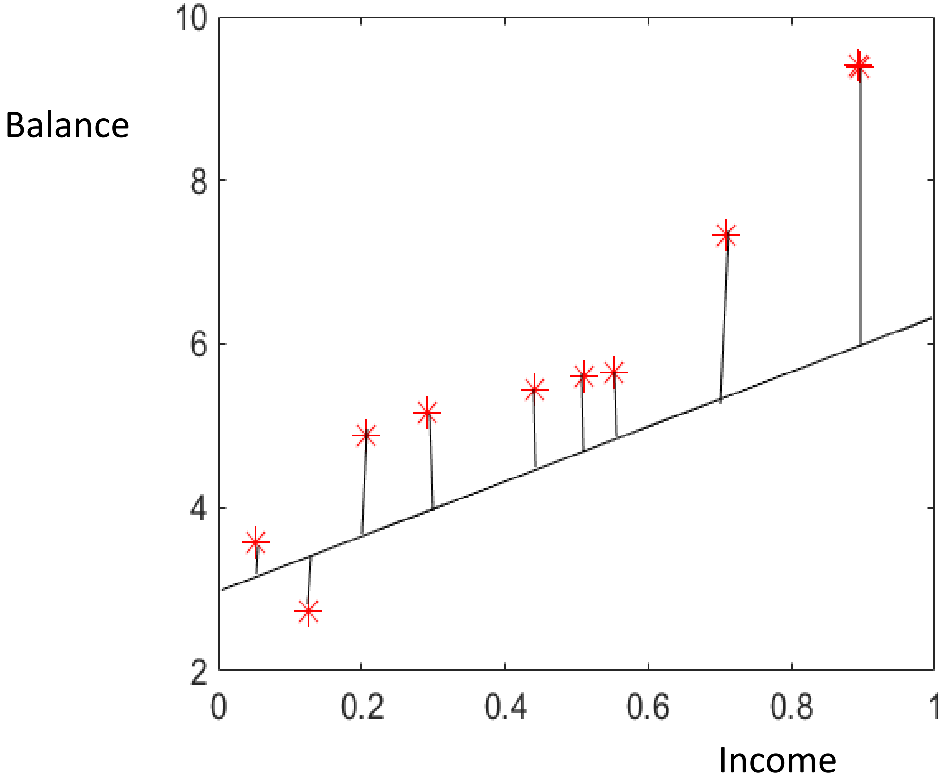
Income (hundred thousand dollar)	Balance (thousand dollar)
x_1= 0.550798	y_1= 5.651202
x_2 =0.708148	y_2= 7.321263
x_3 =0.290905	y_3= 5.167304
x_4 =0.510828	y_4= 5.609367
x_5 =0.892947	y_5= 9.406379
x_6 =0.896293	y_6= 9.379439
x_7 =0.125585	y_7= 2.734997
x_8 =0.207243	y_8= 4.876649
x_9 =0.051467	y_9= 3.584138
x_{10} =0.44081	y_{10}= 5.437239



$f(x) = \beta_0 + \beta_1 x$

Income (hundred thousand dollar)	Balance (thousand dollar)
x_1= 0.550798	y_1= 5.651202
x_2 =0.708148	y_2= 7.321263
x_3 =0.290905	y_3= 5.167304
x_4 =0.510828	y_4= 5.609367
x_5 =0.892947	y_5= 9.406379
x_6 =0.896293	y_6= 9.379439
x_7 =0.125585	y_7= 2.734997
x_8 =0.207243	y_8= 4.876649
x_9 =0.051467	y_9= 3.584138
x_{10} =0.44081	y_{10}= 5.437239

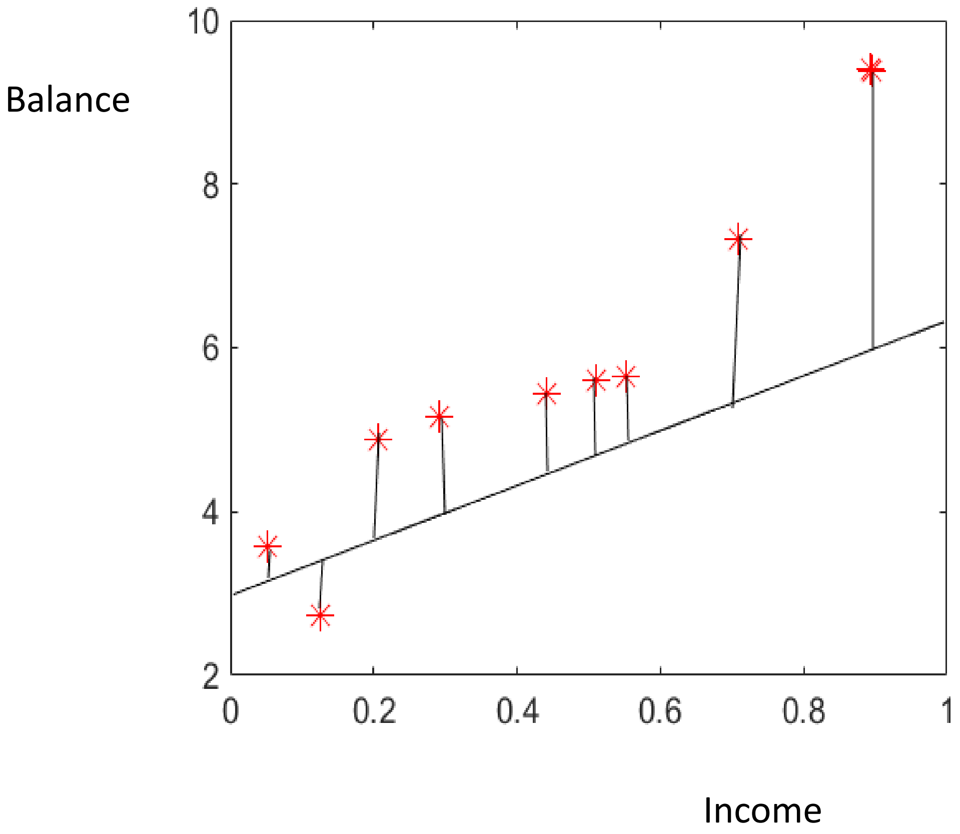
Error (thousand dollar)
$(y_1-(\beta_0 + \beta_1 x_1))^2$
$(y_2-(\beta_0 + \beta_1 x_2))^2$
$(y_3-(\beta_0 + \beta_1 x_3))^2$
$(y_4-(\beta_0 + \beta_1 x_4))^2$
$(y_5-(\beta_0 + \beta_1 x_5))^2$
$(y_6-(\beta_0 + \beta_1 x_6))^2$
$(y_7-(\beta_0 + \beta_1 x_7))^2$
$(y_8-(\beta_0 + \beta_1 x_8))^2$
$(y_9-(\beta_0 + \beta_1 x_9))^2$
$(y_{10}-(\beta_0 + \beta_1 x_{10}))^2$



$f(x) = \beta_0 + \beta_1 x$

Income (hundred thousand dollar)	Balance (thousand dollar)
$x_1 = 0.550798$	$y_1 = 5.651202$
$x_2 = 0.708148$	$y_2 = 7.321263$
$x_3 = 0.290905$	$y_3 = 5.167304$
$x_4 = 0.510828$	$y_4 = 5.609367$
$x_5 = 0.892947$	$y_5 = 9.406379$
$x_6 = 0.896293$	$y_6 = 9.379439$
$x_7 = 0.125585$	$y_7 = 2.734997$
$x_8 = 0.207243$	$y_8 = 4.876649$
$x_9 = 0.051467$	$y_9 = 3.584138$
$x_{10} = 0.44081$	$y_{10} = 5.437239$

Error (thousand dollar)
$(y_1-(\beta_0 + \beta_1 x_1))^2$
$(y_2-(\beta_0 + \beta_1 x_2))^2$
$(y_3-(\beta_0 + \beta_1 x_3))^2$
$(y_4-(\beta_0 + \beta_1 x_4))^2$
$(y_5-(\beta_0 + \beta_1 x_5))^2$
$(y_6-(\beta_0 + \beta_1 x_6))^2$
$(y_7-(\beta_0 + \beta_1 x_7))^2$
$(y_8-(\beta_0 + \beta_1 x_8))^2$
$(y_9-(\beta_0 + \beta_1 x_9))^2$
$(y_{10}-(\beta_0 + \beta_1 x_{10}))^2$



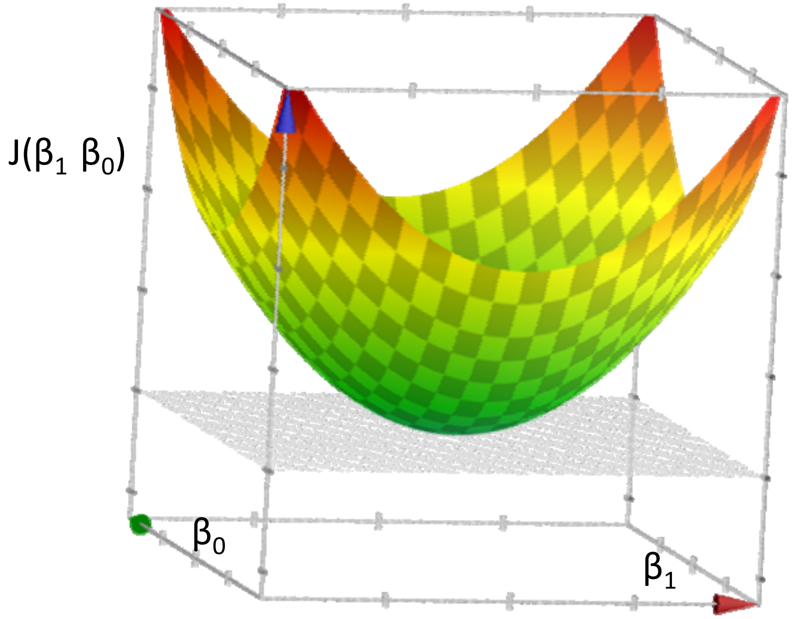
$f(x) = \beta_0 + \beta_1 x$

$$\frac{1}{10}\sum_{i=1}^{10}(y_i-(\beta_0 + \beta_1 x_i))^2$$

For given Training Set T = { (x_1,y_1), (x_2,y_2),….., (x_n,y_n)}, we need to solve

$$Min_{(\beta_1,\beta_0)} \quad J(\beta_1,\beta_0) = \quad \sum_{i=1}^{n} (yi-(\beta_0+\beta_1 x_i))^2$$

[1]



[1] https://www.google.com/search?q=plot+z%3Dx%5E2%2By%5E2&oq=plot+z%3D&aqs=chrome.1.69i57j0i512l4j0i22i30l5.6542j0j4&sourceid=chrome&ie=UTF-8

$$\min_{x} \; x^T x$$

$$\min_{u} \; \boxed{(y - Ay)^T (y - Ay)}$$

For given Training Set T = { (x_1,y_1), (x_2,y_2),....., (x_n,y_n)}, we need to solve

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \\ x_n \end{bmatrix}$$

$$Min_{(\beta_1,\beta_0)} \quad J(\beta_1,\beta_0) = \quad \sum_{i=1}^{n} (yi-(\beta_0 + \beta_1 x_i))^2$$

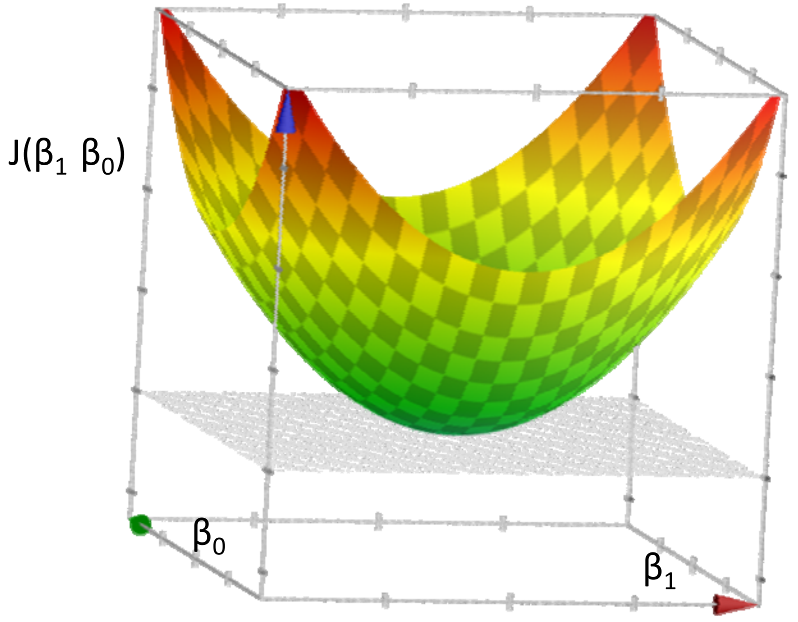
$$u = \begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix}$$

$$A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ x_4 & 1 \\ x_5 & 1 \\ x_6 & 1 \\ x_7 & 1 \\ & \\ x_n & 1 \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ \\ y_n \end{bmatrix}$$

$$f = x_1^2 + x_2^2 + - - + x_n^2$$

$$\frac{\partial f}{\partial x_1} = \begin{bmatrix} 2x_1 \\ \\ \end{bmatrix}$$

$$\partial P$$



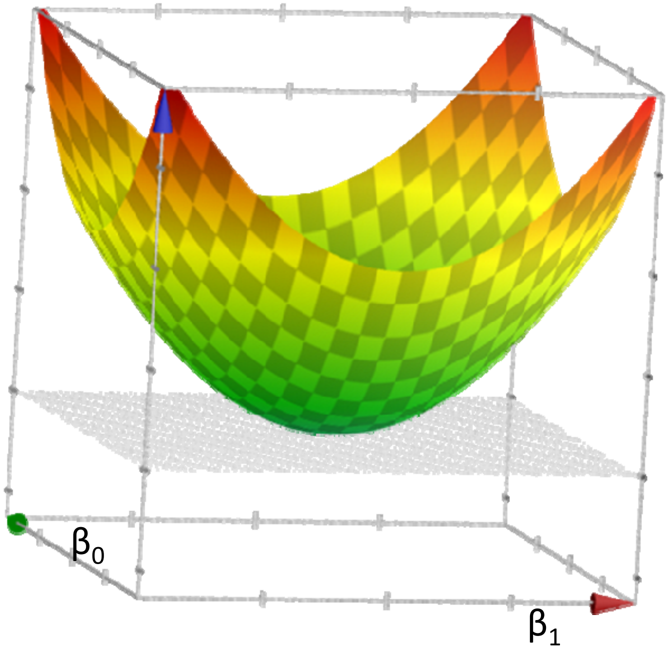
Ax $\quad m \times n$

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{1n} \\ & & \\ a_{m1} & a_{m2} & a_{mn} \end{bmatrix}$$

For given Training Set T = { $(x_1,y_1), (x_2,y_2),\ldots\ldots, (x_n,y_n)$}, we need to solve

$$Min_{(\beta_1,\beta_0)} \quad J(\beta_1,\beta_0) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \qquad \ldots\ldots(1)$$

$$u = \begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix} \qquad A = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ x_3 & 1 \\ x_4 & 1 \\ x_5 & 1 \\ x_6 & 1 \\ x_7 & 1 \\ & \\ x_n & 1 \end{bmatrix} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ \\ y_n \end{bmatrix}$$



The Least Square problem reduces to

$$Min_{(u)} \quad J(u) = \boxed{(Y - Au)^T (Y - Au)}$$

$$AX = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n \end{bmatrix}$$

$$\begin{bmatrix} a_{11} & a_{21} & a_{n1} \\ a_{12} & a_{22} & a_{n2} \\ \\ a_{1n} & a_{2n} & a_{nn} \end{bmatrix} = A^T$$

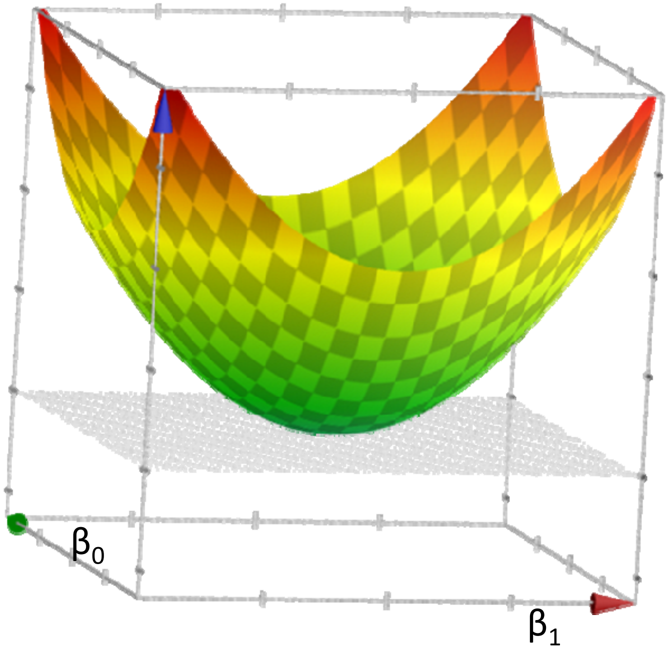
For given Training Set $T = \{ (x_1,y_1), (x_2,y_2),\ldots\ldots, (x_n,y_n)\}$, we need to solve

$$Min_{(\beta_1,\beta_0)} \quad J(\beta_1,\beta_0) = \quad \sum_{i=1}^{n} (yi-(\beta_0 + \beta_1 x_i))^2 \qquad \ldots\ldots(1)$$

The Least Square problem reduces to

$$Min_{(u)} \quad J(u) = \quad (Y - Au)^T (Y - Au)$$

$$\nabla u \, J(u) = A^T (Y - Au)$$



$$\underline{\nabla_u (y - Au)^T (y - Au)} \qquad \underline{u = (A^T A)^{-1} A^T y}$$

$$2 \cdot A^T \cancel{2} \boxed{(y - Au)} = 0$$

$$u =$$

$$-\cancel{2} A^T (y - Au) = 0$$

$$A^T (y - Au) = 0 \qquad \Rightarrow$$

$$A^T A u = A^T y$$

$$(A^T A)^{-1} (A^T A) u = (A^T A)^{-1} A^T y$$

For given Training Set T = { $(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)$}, we need to solve

$$Min_{(\beta_1,\beta_0)} \quad J(\beta_1,\beta_0) = \sum_{i=1}^{n} (yi-(\beta_0 + \beta_1 x_i))^2 \qquad .....(1)$$

The Least Square problem reduces to

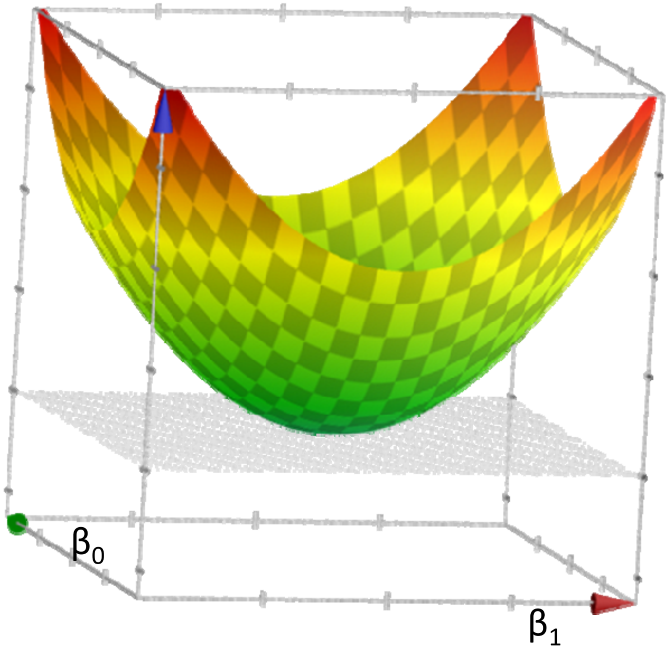
$$Min_{(u)} \quad J(u) = (Y - Au)^T (Y - Au)$$

Setting the gradient for J(u) = 0

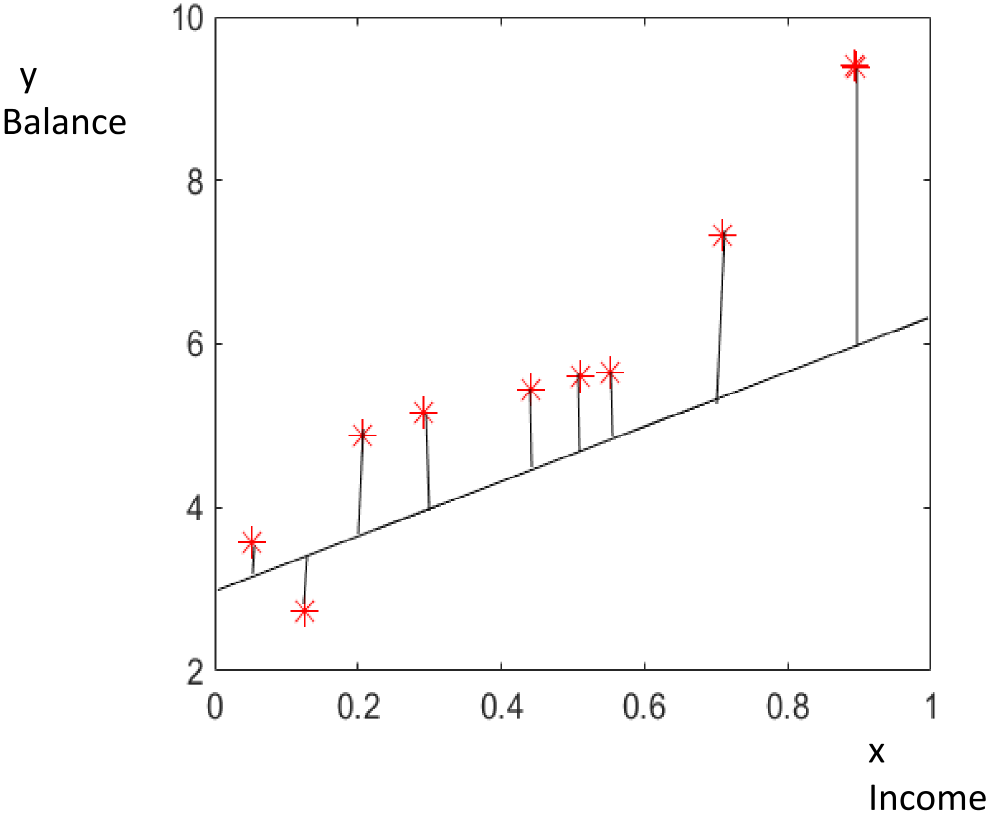
$$\nabla u \, J(u) = 0$$

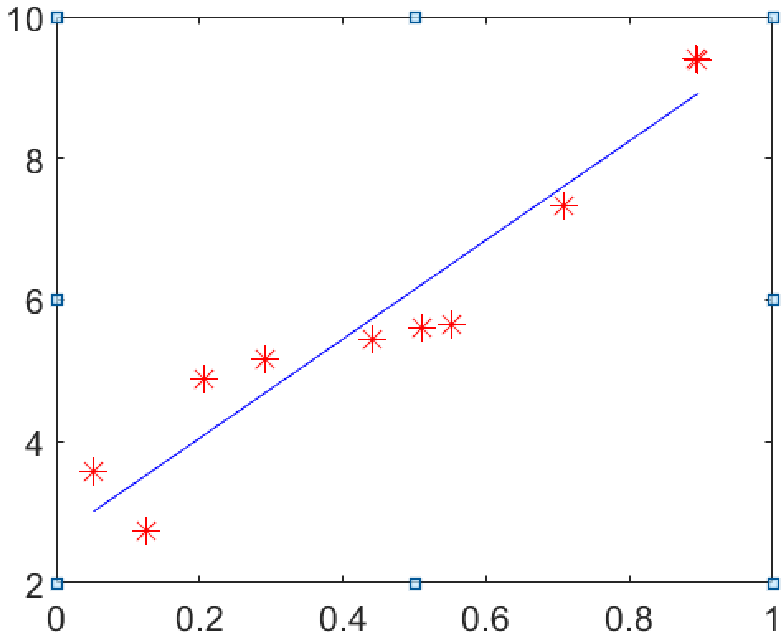
$$A^T (Y - Au) = 0$$

$$\boxed{u = \begin{bmatrix} \beta_1 \\ \beta_0 \end{bmatrix} = (A^T A)^{-1} \; A^T Y}$$



Best Fitting Line





Best Fitting Line

$$f(x) = 2.6 + 6.9x$$

Best Fitting Line

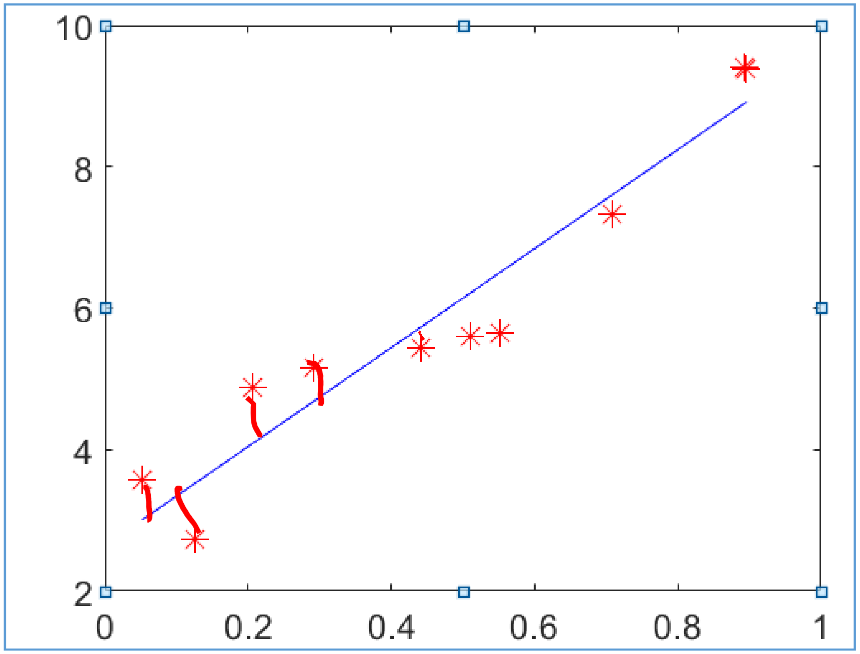
Training Error:-

$$\frac{1}{10}\sum_{i=1}^{10}(yi-(\beta_0+\beta_1 x_i))^2 = 0.3537$$

Training RMSE =

$$\sqrt{\frac{1}{10}\sum_{i=1}^{10}(yi-(\beta_0+\beta_1 x_i))^2}$$

= 0.5947



Best Fitting Line

$f(x) = 2.6 + 6.9x$

Income (x) (thousand dollar)	Balance (y) (thousand dollar)	Estimated f(x) (thousand dollar)
0.96703	9.675083	9.41205399
0.547232	6.293266	6.47467789
0.972684	9.730614	9.45161938
0.714816	7.474346	7.64728226
0.697729	7.342933	7.5277212
0.216089	4.619033	4.15763074
0.976274	9.765597	9.47673972
0.00623	4.012784	2.68921946
0.252982	4.762698	4.41577473
0.434792	5.626166	5.68791617
0.779383	7.989045	8.09906511
0.197685	4.552625	4.02885271
0.862993	8.705537	8.6840969
0.983401	9.835217	9.52660279
0.163842	4.43522	3.79205019
0.597334	6.622444	6.8252457
0.008986	4.01882	2.70850244
0.386571	5.37153	5.35051307
0.04416	4.096522	2.95461902
0.956653	9.574695	9.33944573

Income (x) (thousand dollar)	Balance (y) (thousand dollar)	Estimated f(x) (thousand dollar)
0.96703	9.675083	9.41205399
0.547232	6.293266	6.47467789
0.972684	9.730614	9.45161938
0.714816	7.474346	7.64728226
0.697729	7.342933	7.5277212
0.216089	4.619033	4.15763074
0.976274	9.765597	9.47673972
0.00623	4.012784	2.68921946
0.252982	4.762698	4.41577473
0.434792	5.626166	5.68791617
0.779383	7.989045	8.09906511
0.197685	4.552625	4.02885271
0.862993	8.705537	8.6840969
0.983401	9.835217	9.52660279
0.163842	4.43522	3.79205019
0.597334	6.622444	6.8252457
0.008986	4.01882	2.70850244
0.386571	5.37153	5.35051307
0.04416	4.096522	2.95461902
0.956653	9.574695	9.33944573

Test $RMSE = \underline{\sqrt{\frac{1}{k}\sum_{i=1}^{k}(yi - f(xi))^2}} =$

0.9426

Income (x) (thousand dollar)	Balance (y) (thousand dollar)	Estimated f(x) (thousand dollar)
0.96703	9.675083	9.41205399
0.547232	6.293266	6.47467789
0.972684	9.730614	9.45161938
0.714816	7.474346	7.64728226
0.697729	7.342933	7.5277212
0.216089	4.619033	4.15763074
0.976274	9.765597	9.47673972
0.00623	4.012784	2.68921946
0.252982	4.762698	4.41577473
0.434792	5.626166	5.68791617
0.779383	7.989045	8.09906511
0.197685	4.552625	4.02885271
0.862993	8.705537	8.6840969
0.983401	9.835217	9.52660279
0.163842	4.43522	3.79205019
0.597334	6.622444	6.8252457
0.008986	4.01882	2.70850244
0.386571	5.37153	5.35051307
0.04416	4.096522	2.95461902
0.956653	9.574695	9.33944573

Training $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(yi - f(xi))^2} = 0.5947$

Testing $RMSE = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(yi - f(xi))^2} = 0.9426$

Evaluation Criteria

Testing set is $\{(x_1,y_1), (x_2,y_2),.......,(x_k,y_k)\}$, y_i is estimated by $f(x_i)$, $\bar{y}$ is the mean of y_i and f_x is the mean of $f(x_i)$

fx

- Sum of Squared Error (SSE) $= \sum_{i=1}^{k}(yi - f(x_i))^2$
- Root Mean Square of Error ($RMSE$) $= \sqrt{\frac{1}{k}\sum_{i=1}^{k}(yi - f(xi))^2}$
- Mean Absolute of Error (MAE) $= \frac{1}{k}\sum_{i=1}^{k} \underline{|y_i - f(x_i)|}$
- Mean Absolute Percentage of Error ($MAPE$) $= \frac{1}{k}\sum_{i=1}^{k} | {}^{y_i - f(x_i)}/_{y_i} | \times 100$
- Normalized Mean Square of Error (NMSE) $= {}^{\sum_{i=1}^{k}(yi - f(x_i))^2}/_{\sum_{i=1}^{k}(yi - \bar{y})^2}$ $= \frac{SSE}{SST}$
- $R^2 = {}^{\boxed{\sum_{i=1}^{k}(f(x_i) - fx)^2}}/_{\sum_{i=1}^{k}(yi - \bar{y})^2}$