Linear Regression - 2

Cost function/SSE

- S(a_0, a_1)= $\sum_{i=1}^{m}\left(y_i - (a_0 + a_1 x_i)\right)^2$
- S is a function of $a_0 \; and \; a_1$. As any of the values of $a_0 \; and \; a_1$ is changed , S changes.
- We want to determine those values of parameters $a_0 \; and \; a_1$; for which S is minimum.
- S ≥ 0. (Always)
- S will be minimum when partial derivatives with respect to $a_0 \; and \; a_1$ become zero.

- S(a_0, a_1)= $\sum_{i=1}^{m}\left(y_i-(a_0+a_1x_i)\right)^2$
- $\frac{\partial S}{\partial a_0}=\ \frac{\partial}{\partial a_0}\left(\sum_{i=1}^{m}\left(y_i-(a_0+a_1x_i)\right)^2\right)$
- $\frac{\partial S}{\partial a_0}=\ \ \sum_{i=1}^{m}\frac{\partial}{\partial a_0}\left(y_i-(a_0+a_1x_i)\right)^2$
- $\frac{\partial S}{\partial a_0}=\sum_{i=1}^{m}2\left(y_i-(a_0+a_1x_i)\right)\left(\frac{\partial}{\partial a_0}(y_i-(a_0+a_1x_i))\right)$
- $=\sum_{i=1}^{m}2\left(y_i-(a_0+a_1x_i)\right)$(-1)
-
- $\frac{\partial S}{\partial a_0}$ **= -2** $\sum_{i=1}^{m}\left(y_i-(a_0+a_1x_i)\right)$

-

- S(a_0, a_1)= $\sum_{i=1}^{m}\left(y_i - (a_0 + a_1 x_i)\right)^2$
- $\frac{\partial S}{\partial a_1} = \frac{\partial}{\partial a_1}\left(\sum_{i=1}^{m}\left(y_i - (a_0 + a_1 x_i)\right)^2\right)$
- $\frac{\partial S}{\partial a_1} = \sum_{i=1}^{m} \frac{\partial}{\partial a_1}\left(y_i - (a_0 + a_1 x_i)\right)^2$
- $\frac{\partial S}{\partial a_1} = \sum_{i=1}^{m} 2\left(y_i - (a_0 + a_1 x_i)\right)\left(\frac{\partial}{\partial a_1}(y_i - (a_0 + a_1 x_i))\right)$
- $= \sum_{i=1}^{m} 2\left(y_i - (a_0 + a_1 x_i)\right)(-x_i)$
- $\frac{\partial S}{\partial a_1}$ **= -2** $\boldsymbol{\sum_{i=1}^{m} x_i\left(y_i - (a_0 + a_1 x_i)\right)}$
-

- $\frac{\partial S}{\partial a_0} = 0, \frac{\partial S}{\partial a_1} = 0$
- $\frac{\partial S}{\partial a_0}$ = -2 $\sum_{i=1}^{m}\left(y_i - (a_0 + a_1 x_i)\right)$ = 0
- $\sum_{i=1}^{m}\left(y_i - (a_0 + a_1 x_i)\right)$ = 0
- $\sum_{i=1}^{m} y_i = \sum_{i=1}^{m} a_0 + \sum_{i=1}^{m} a_1 x_i$
- $\boldsymbol{\sum_{i=1}^{m} y_i = m a_0 + a_1 \sum_{i=1}^{m} x_i}$

- $\frac{\partial S}{\partial a_1}$ = -2 $\sum_{i=1}^{m} x_i\left(y_i - (a_0 + a_1 x_i)\right)$ =0 gives
- $\sum_{i=1}^{m} x_i\left(y_i - (a_0 + a_1 x_i)\right)$=0
- $\sum_{i=1}^{m} x_i y_i = \sum_{i=1}^{m} a_0\, x_i + \sum_{i=1}^{m} a_1 x_i^2$
- $\boldsymbol{\sum_{i=1}^{m} x_i y_i = a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2}$

Normal Equations

The above equations are simplified to

$$ma_0 + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i$$

and

$$ma_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i.$$

Since the x_i and y_i are known quantities, the above two equations (called the **normal equations**), can be solved for the two unknown a_0 and a_1. Differentiating $\frac{\partial S}{\partial a_0}$ and $\frac{\partial S}{\partial a_1}$ with respect to a_0 to a_1 respectively, we find

$$\frac{\partial^2 S}{\partial a_0^2} \quad \text{and} \quad \frac{\partial^2 S}{\partial a_1^2}$$

and both will be positive at the points. Hence these values provide a minimum of S.

Computation Table

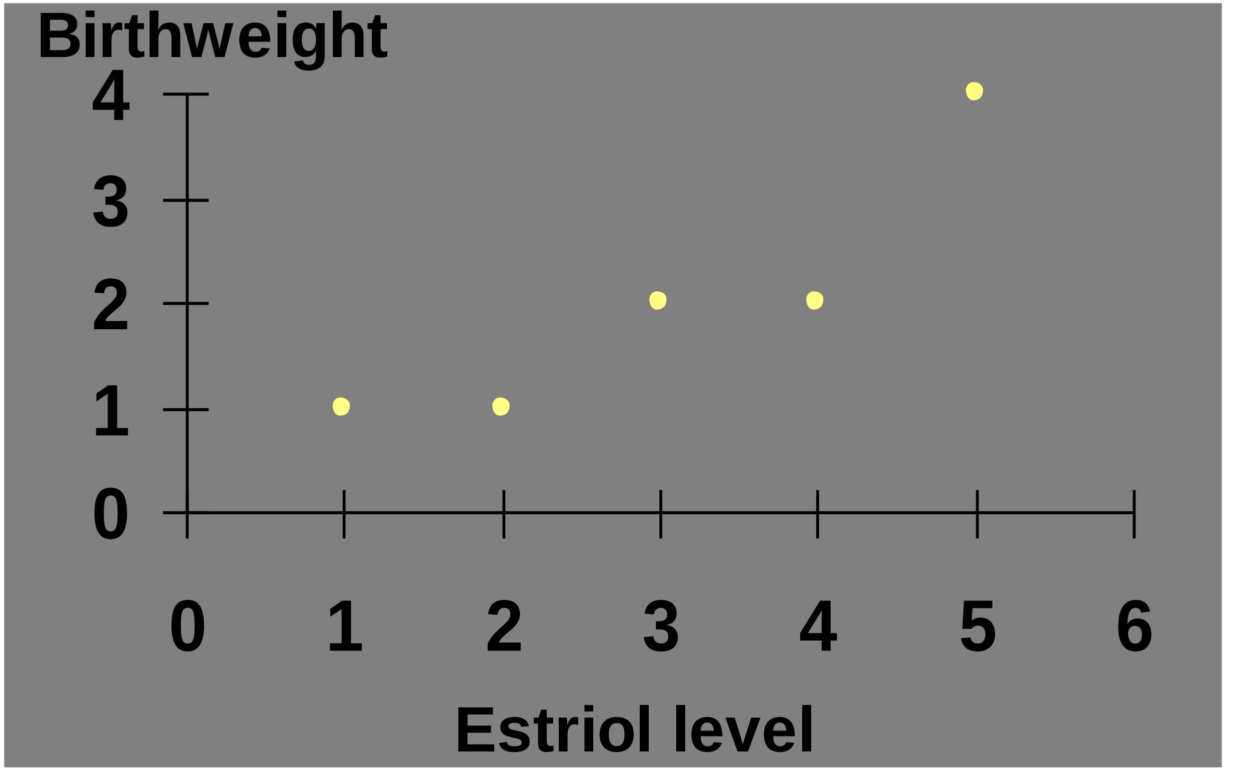
X_i	Y_i	X_i^2	Y_i^2	X_iY_i
X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	X_nY_n
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	ΣX_iY_i

Parameter Estimation Example

- Obstetrics: What is the **relationship** between Mother's Estriol level & Birthweight using the following data?

Estriol (mg/24h)	**Birthweight** (g/1000)
1	1
2	1
3	2
4	2
5	4

Scatterplot
Birthweight vs. Estriol level



Parameter Estimation Solution Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	**10**	**55**	**26**	**37**

Formation of Normal Equations from Data

- $ma_0 + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i$
- $a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i$

i	x_i^2	x_i	y_i	$x_i y_i$
1	1	1	1	1
2	4	2	1	2
3	9	3	2	6
4	16	4	2	8
5	25	5	4	20
Total	55	15	10	37

Solution of Normal Equatiuons –Method 1

- $5\ a_0\ \ + 15\ a_1 = 10$
- 15 a_0 + 55 a_1 = 37
- **Solving by elimination:**
- $5\ a_0\ \ + 15\ \ = 10$ } X 3 $\Longrightarrow$ 15 a_0 + 45 a_1 = 30

  15 a_0 + 55a_1 = 37

  ----

  10 a_1 = 7 gives a_1 = 0.7
- $5\ a_0\ \ + 15a_1 = 10$ } /5 $\Longrightarrow$ a_0 + 3 a_1 = 2
- a_0 + 2.1 = 2 $\Longrightarrow$ a_0 = -0.1
- Line fitted is $y = a_0 + a_1 x$
- $y = -0.1 +\ .7\ x$

Home Work

- Calculate $(y_i - \hat{y}_i)$ and S=$\sum(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\boldsymbol{2}}$
- Do the same by changing a_0 and a_1

i	$\boldsymbol{x_i}$	$\boldsymbol{y_i}$	$\boldsymbol{\hat{y}_i = -0.1 + .7\,x_i}$	$(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})$	$(\boldsymbol{y_i} - \boldsymbol{\hat{y}_i})^{\boldsymbol{2}}$
Total					

Interpretation of Coefficients

- 1. Slope (a_1)

  $y_2 - y_1 = a_1(x_2 - x_1)$; change in y = slope times change in x ; $\Delta\, y = a_1\, \Delta x$

  - If $a_1 = 0.7$, then *Y* Is Expected to Increase by 0.7 for Each 1 Unit Increase in *X*

- Y-Intercept (a_0)

$$m a_0 + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i$$

$$a_0 + a_1 X = Y$$

  Value of *Y* when *X* = 0;

  If a_0 = -0.1 , then Average *Y* Is Expected to Be -0.1 When *X* Is 0

Solution of Normal Equations: Method 2

- $ma_0 + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i$
- $a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i$

- $ma_0 + a_1 \sum_{i=1}^{m} x_i - \sum_{i=1}^{m} y_i = 0$
- $a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} x_i y_i = 0$

$$\frac{a_0}{\begin{vmatrix} \sum_{i=1}^{m} x_i & -\sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i^2 & -\sum_{i=1}^{m} x_i y_i \end{vmatrix}} = -\frac{a_1}{\begin{vmatrix} m & -\sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} x_i & -\sum_{i=1}^{m} x_i y_i \end{vmatrix}} = \frac{1}{\begin{vmatrix} m & \sum_{i=1}^{m} x_i \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 \end{vmatrix}}$$

$$a_0 = \frac{-(\sum_{i=1}^{m} x_i)(\sum_{i=1}^{m} x_i y_i) + \sum_{i=1}^{m} y_i \sum_{i=1}^{m} x_i^2}{m \sum_{i=1}^{m} x_i^2 - (\sum_{i=1}^{m} x_i)^2}$$

$-$

Formulas of parameters

- $a_1 = -\frac{-m\sum x_i y_i + \sum x_i \sum y_i}{m\sum {x_i}^2 - (\sum x_i)^2}$
- $a_1 = \frac{m\sum x_i y_i - \sum x_i \sum y_i}{m\sum {x_i}^2 - (\sum x_i)^2}$

Data

- $ma_0 + a_1 \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i$
- $a_0 \sum_{i=1}^{m} x_i + a_1 \sum_{i=1}^{m} x_i^2 = \sum_{i=1}^{m} x_i y_i$

i	x_i^2	x_i	y_i	$x_i y_i$
1	1	1	1	1
2	4	2	1	2
3	9	3	2	6
4	16	4	2	8
5	25	5	4	20
Total	55	15	10	37

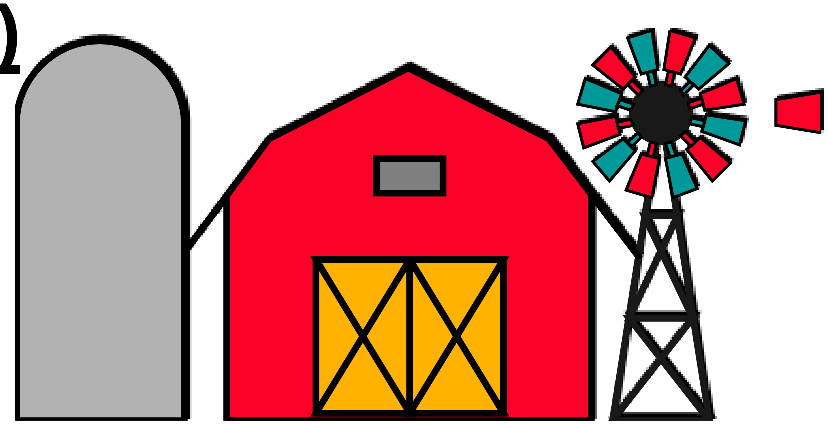
- $\sum x_i = 15$; $\sum y_i = 10$; $\sum {x_i}^2 = 55$; $\sum x_i\, y_i = 37$

- $a_0 = \frac{55x10-15x37}{5x55-15^2} = \frac{550-555}{275-225} = \frac{-5}{50}$ = - 0.1
- $a_1 = \frac{m\sum x_i y_i - \sum x_i \sum y_i}{m\sum {x_i}^2 - (\sum x_i)^2}$
- $a_1 = \frac{5X55-15X10}{5X55-225} = \frac{185-150}{50} = \frac{35}{50}$ = 0.7

Parameter Estimation Thinking Challenge

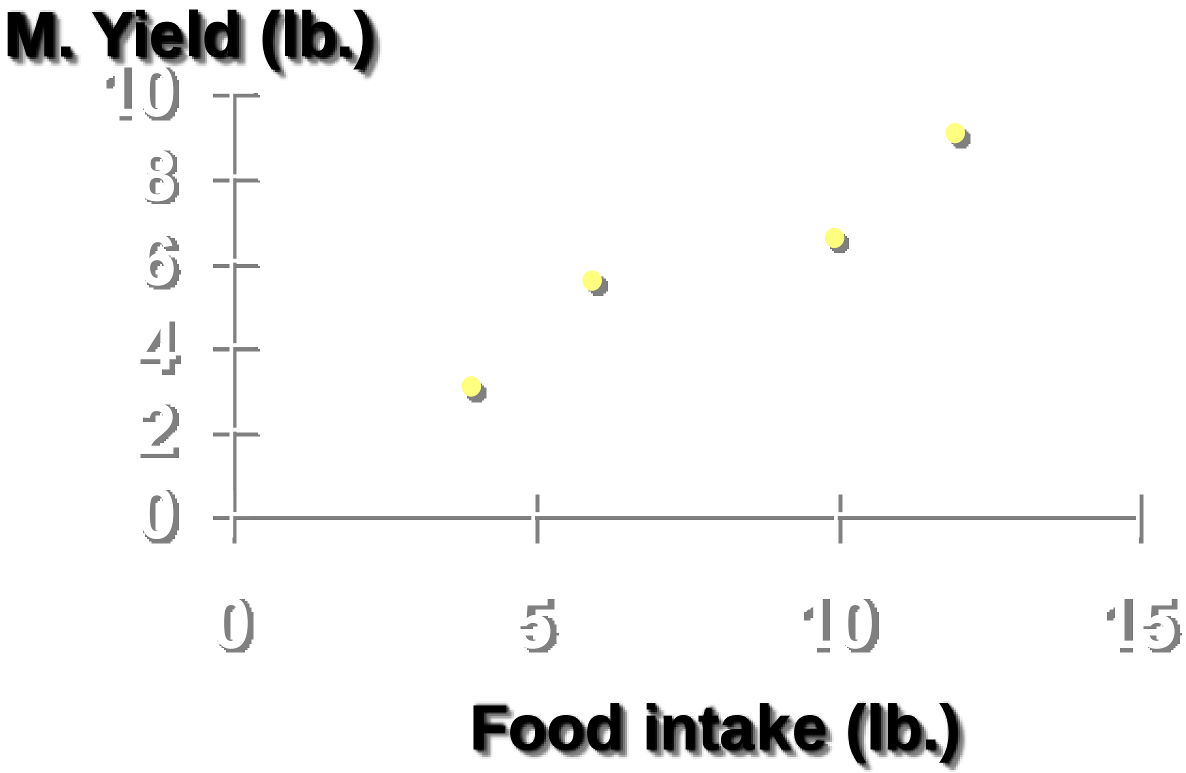
- You're a Vet epidemiologist for the county cooperative. You gather the following data:
- **Food (lb.)** / **Milk yield (lb.)**

Food (lb.)	Milk yield (lb.)
4	3.0
6	5.5
10	6.5
12	9.0



- What is the **relationship** between cows' food intake and milk yield?

Scattergram
Milk Yield vs. Food intake*



Parameter Estimation Solution Table*

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
4	3.0	16	9.00	12
6	5.5	36	30.25	33
10	6.5	100	42.25	65
12	9.0	144	81.00	108
32	**24.0**	**296**	**162.50**	**218**

Parameter Estimation Solution*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{218 - \frac{(32)(24)}{4}}{296 - \frac{(32)^2}{4}} = 0.65$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 6 - (0.65)(8) = 0.80$$

- http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights
- height weight data set

Sr	Height	Weight	xy	
iii	(cms) = x	(Kg)= y		
1	135	57	7695	18225
2	165	70	11550	27225
3	155	63	9765	24025
4	160	65	10400	25600
5	150	62	9300	22500
m= 5	$\sum x_i$ 765	317	48710	117575

Income	hrs/week
8000	38
6400	50
2500	15
3000	30
6000	50
5000	38
8000	50
4000	20
11000	45
25000	50
4000	20
8800	35
5000	30
7000	43

Income	hrs/week
8000	35
18000	37.5
5400	37
15000	35
3500	30
24000	45
1000	4
8000	37.5
2100	25
8000	46
4000	30
1000	200
2000	200
4800	30