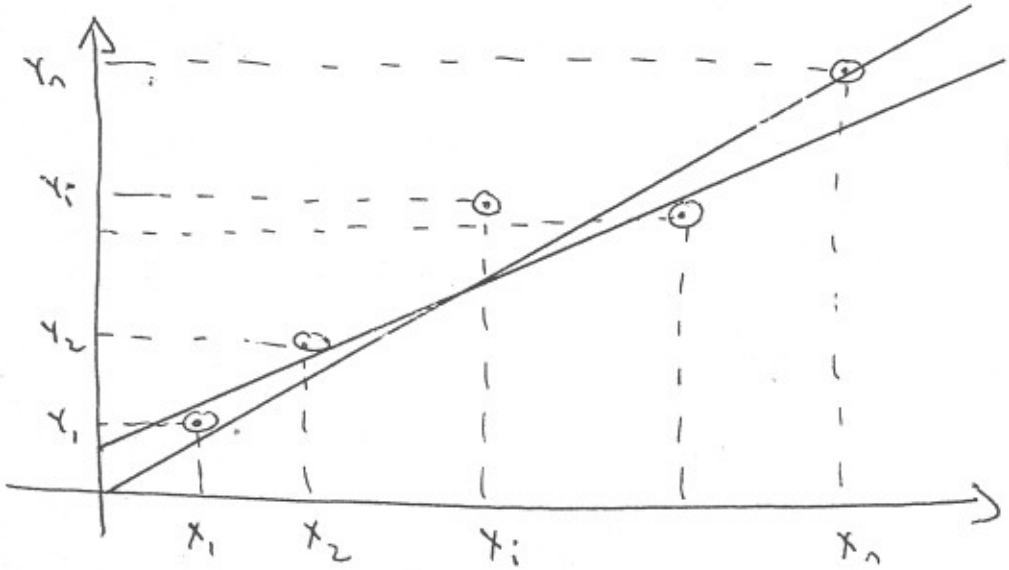
Chap 11 Least Squares Regression

11.1 Linear Regression

Given data pts (X_i, Y_i) $i = 1, 2, \cdots, n$

What is the best line that can be drawn for representing the relationship between X and Y?



2 of many possible lines that can be "fit" to the data.

Let the equation of the line be $\hat{Y} = a_0 + a_1 X$

Then for each X_i, $i = 1, 2, \cdots, n$ the predicted value of Y using the line is $\hat{Y}_i = a_0 + a_1 X_i$. The difference between the actual (observed, measured, data) value Y_i and the predicted value $\hat{Y}_i$ is called the residual (error, deviation) and denoted e_i.

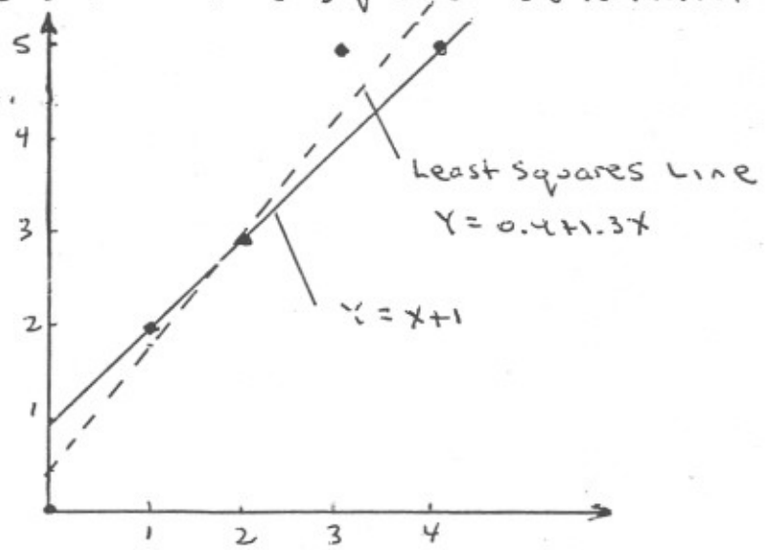
$$e_i = Y_i - \hat{Y}_i$$

$$= Y_i - (a_0 + a_1 X_i) \qquad i = 1, 2, \cdots, n$$

The "best" fit in the Least Squares Sense is the one line for which the sum of the squares of the deviations is a minimum. Therefore a_0 & a_1 are chosen to minimize $\sum_{i=1}^{n} e_i^2$.

example - Given the data pts below, fit a line thru them and calculate the sum of the squared deviations.

i	X_i	Y_i
1	0	0
2	1	2
3	2	3
4	3	5
5	4	5



Suppose we use the line thru (1,2) & (4,5)

$$Y - Y_0 = m(X - X_0)$$

$$m = \frac{5-2}{4-1} = 1$$

$$\Rightarrow Y - 2 = (X - 1)$$

$$Y = X + 1$$

Computing $\sum_{i=1}^{5} e_i^2$ for the line $Y = X + 1$,

i	X_i	Y_i	$\hat{Y}_i$	$e_i = Y_i - \hat{Y}_i$	e_i^2
1	0	0	1	-1	1
2	1	2	2	0	0
3	2	3	3	0	0
4	3	5	4	1	1
5	4	5	5	0	0

$$\sum_{i=1}^{5} e_i^2 = 2$$

It can be shown that the "best" of all possible lines in the Least Squares sense is the one whose coefficients a_0 & a_1 satisfy

$$n a_0 + \left(\sum_{i=1}^{n} x_i\right) a_1 = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_i\right) a_0 + \left(\sum_{i=1}^{n} x_i^2\right) a_1 = \sum_{i=1}^{n} x_i y_i$$

NORMAL EQS

In this example,

i	x_i	y_i	x_i^2	$x_i y_i$
1	0	0	0	0
2	1	2	1	2
3	2	3	4	6
4	3	5	9	15
5	4	5	16	20
	10	15	30	43

$$\Rightarrow \quad 5a_0 + 10a_1 = 15$$
$$10a_0 + 30a_1 = 43$$

$$a_0 + 2a_1 = 3$$
$$10a_0 + 30a_1 = 43$$

$$a_0 = \frac{\begin{vmatrix} 3 & 2 \\ 43 & 30 \end{vmatrix}}{\begin{vmatrix} 1 & 2 \\ 10 & 30 \end{vmatrix}} = \frac{4}{10}, \qquad a_1 = \frac{\begin{vmatrix} 1 & 3 \\ 10 & 43 \end{vmatrix}}{\begin{vmatrix} 1 & 2 \\ 10 & 30 \end{vmatrix}} = \frac{13}{10}$$

The Least Squares line is $\hat{Y} = 0.4 + 1.3X$.

The sum of squared deviations is computed as follows:

i	X_i	Y_i	$\hat{Y}_i$	e_i	e_i^2
1	0	0	0.4	-0.4	0.16
2	1	2	1.7	0.3	0.09
3	2	3	3.0	0	0
4	3	5	4.3	0.7	0.49
5	4	5	5.6	-0.6	0.36
					1.0 1

Simple Statistics

Given a set of data pts (X_i, Y_i), $i = 1, 2, \ldots, n$

the mean of the observed values Y_i, $i = 1, 2, \ldots, n$

is $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$

The residual in each Y_i value about the mean $\bar{Y}$ is given by

$$\epsilon_i = Y_i - \bar{Y}, \quad i = 1, 2, \ldots, n$$

The total variation in the set of Y_i values measured with respect to the mean $\bar{Y}$ is

$$SST = \sum_{i=1}^{n} \epsilon_i^2$$

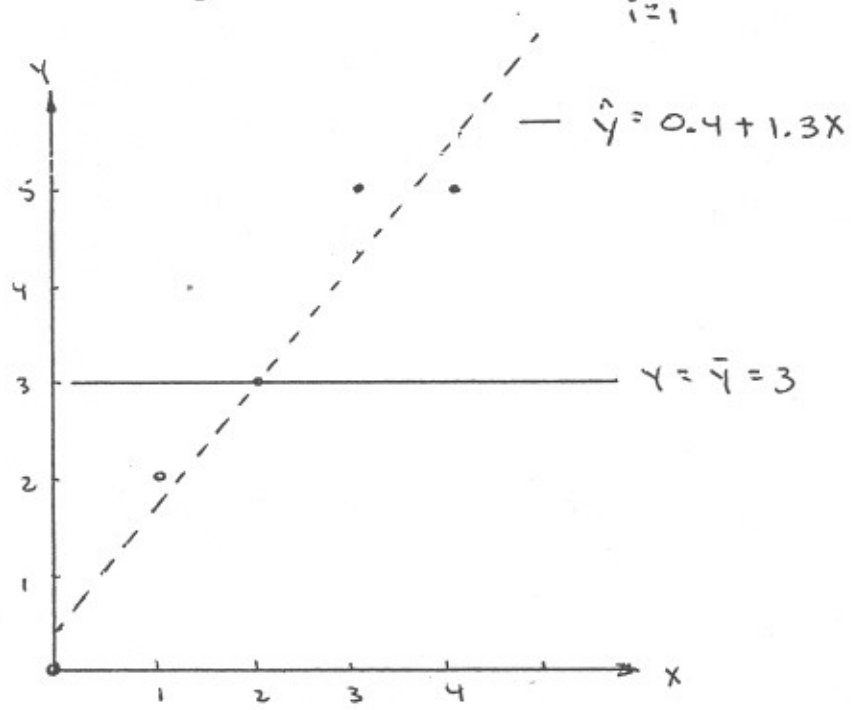
$$= \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$= (Y_1 - \bar{Y})^2 + (Y_2 - \bar{Y})^2 + \cdots + (Y_n - \bar{Y})^2$$

In the previous example,

i	x_i	y_i	$\epsilon_i = y_i - \bar{y}$	ϵ_i^2
1	0	0	-3	9
2	1	2	-1	1
3	2	3	0	0
4	3	5	2	4
5	4	5	2	4

$\bar{y} = \frac{1}{5}(15) = 3$ $\qquad \sum_{i=1}^{5} \epsilon_i^2 = 18$ (SST)



It can be shown that, $SST = SSE + SSR$

where $SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ Residual (Error) Sum of Squares (Unexplained Variatio

$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ Regression Sum of Squares (Explained Variation)

$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ Total Sum of Squares

Continuing with this example,

$SST = 18$

$SSE = 1.1$

$\Rightarrow SSR = SST - SSE$

$= 18 - 1.1$

$= 16.9$

Note, the residual in each Y_i value about the regression line is $e_i = Y_i - \hat{Y}_i$, $i = 1, 2, \ldots, n$

Therefore, SSE is a measure of the total variation in the set of Y_i values measured with respect to the least squares regression line.

There is a significant reduction in the total variation of the Y_i values about the mean, i.e. from SST to SSE

$$\frac{SST}{18} \quad : \quad \frac{SSE}{1.1}$$

Hence, $\hat{Y}_i$, $i = 1, 2, \ldots, n$ fit the given data better than $\bar{Y}$, $i = 1, 2, \ldots, n$

A measure of how good the regression line fit is given by

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} \qquad (P\ 337)$$

Coefficient of determination

$r = \sqrt{\frac{SSR}{SST}}$ is the correlation coefficient

r^2 is called the <u>coefficient of determination</u>

In the previous example,

$$r^2 = \frac{SST - SSE}{SST} = \frac{18 - 1.1}{18} = 0.939$$

$$r = \sqrt{\frac{SST - SSE}{SST}} = 0.969 \quad \underline{\text{Correlation Coefficient}}$$

Before regression, another measure of total variation in the data is given by

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}} = \sqrt{\frac{SST}{n-1}}$$

Standard Deviation

In the previous example, $s_Y = \sqrt{\frac{18}{5-1}} = 2.12$

After regression, an equivalent measure of variation in the data about the regression line is

$$s_{Y|X} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-2}} = \sqrt{\frac{SSE}{n-2}}$$

Standard Error of the Estimate

In the previous example,

$$s_{Y|X} = \sqrt{\frac{1.1}{5-2}} = 0.606$$

For a perfect fit, i.e. $\hat{Y}_i = Y_i$, $i = 1, 2, \ldots, n$

$$\Rightarrow \quad \hat{Y}_i = a_0 + a_1 \hat{X}_i = Y_i$$
$$\Rightarrow \quad e_i = Y_i - \hat{Y}_i = 0 \qquad \left.\right\} \; i = 1, 2, \ldots, n$$

$$SSE = 0$$

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2 = \sum (Y_i - \bar{Y})^2 = SST$$

$$r^2 = \frac{SST - 0}{SST} = 1 \quad \Rightarrow \quad r = 1$$

$$s_{Y|X} = 0$$

11.2 Polynomial Regression

The least squares quadratic thru (X_i, Y_i), $i = 1, 2, \ldots, n$ is

$$\hat{Y} = a_0 + a_1 X + a_2 X^2$$

where a_0, a_1, a_2 are chosen to minimize the sum of the squared residuals

$$\Rightarrow \quad \underset{a_0, a_1, a_2}{\text{Min}} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\Rightarrow \quad \underset{a_0, a_1, a_2}{\text{Min}} \sum_{i=1}^{n} \left[Y_i - (a_0 + a_1 X_i + a_2 X_i^2) \right]^2$$

The resulting coefficients a_0, a_1, a_2 satisfy

$$n a_0 + (\sum X_i) a_1 + (\sum X_i^2) a_2 = \sum Y_i$$

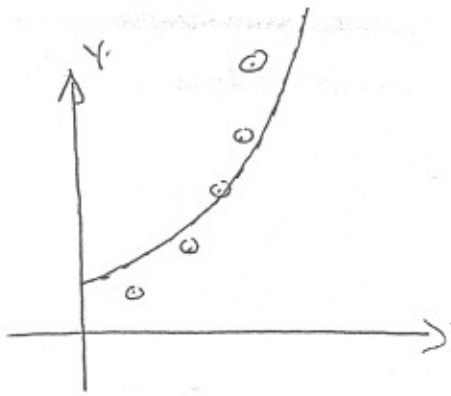
$$(\sum X_i) a_0 + (\sum X_i^2) a_1 + (\sum X_i^3) a_2 = \sum X_i Y_i$$

$$(\sum X_i^2) a_0 + (\sum X_i^3) a_1 + (\sum X_i^4) a_2 = \sum X_i^2 Y_i$$

11.1.5 Applications of Linear Regression - Linearization of Nonlinear Relationships

p342

Consider the exponential model $Y = ae^{bX}$



$\Rightarrow \ln Y = \ln a + bX$

Let $z = \ln Y$

$\Rightarrow z = a_0 + a_1 X$

where $a_0 = \ln a$

$a_1 = b$

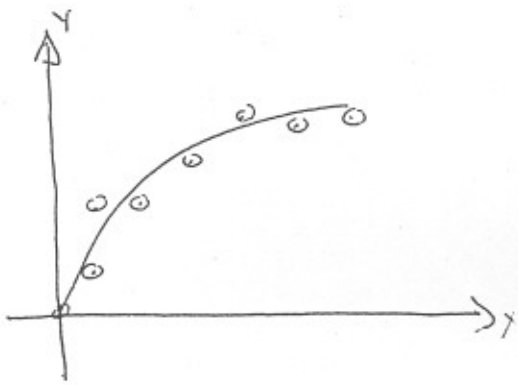
X	Y	$z = \ln Y$
X_1	Y_1	$z_1 = \ln Y_1$
X_2	Y_2	$z_2 = \ln Y_2$
⋮		
X_n	Y_n	$z_n = \ln Y_n$

$\Rightarrow a_0 \& a_1 \Rightarrow a = e^{a_0}$

$b = a_1$

Consider a saturation-growth-rate equation $Y = a\dfrac{X}{b+X}$

$\Rightarrow \dfrac{1}{Y} = \dfrac{b+X}{aX} = \dfrac{1}{a}\left[1 + \dfrac{b}{X}\right] = \dfrac{1}{a} + \dfrac{b}{a}\left(\dfrac{1}{X}\right)$



Let $z = \dfrac{1}{Y}$, $u = \dfrac{1}{X}$

$\Rightarrow z = a_0 + a_1 u$

$z = \dfrac{1}{a} + \dfrac{b}{a} u$

$a_0 = \dfrac{1}{a}$

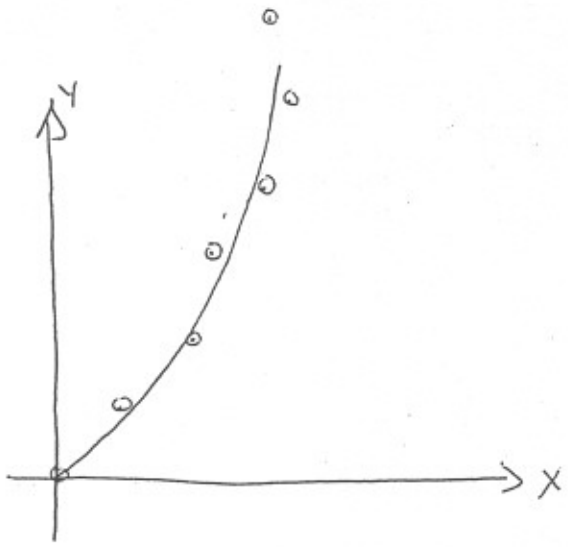
$a_1 = \dfrac{b}{a}$

X_i	Y_i	$u_i = \frac{1}{X_i}$	$z_i = \frac{1}{Y_i}$
X_1	Y_1	u_1	z_1
X_n	Y_n	u_n	z_n

$a = \dfrac{1}{a_0}$

$b = aa_1 = \dfrac{a_1}{a_0}$

Consider a power equation $Y = aX^b$



$\Rightarrow \log Y = \log a + b \log X$

Let $z = \log Y$ & $u = \log X$

$\Rightarrow \quad z = \log a + bu$

$= a_0 + a_1 u$

where $a_0 = \log a$

$a_1 = b$

X_i	Y_i	$u_i = \log X_i$	$z_i = \log Y_i$
X_1	Y_1	u_1	z_1
X_n	Y_n	u_n	z_n

a_0

$a = 10$

$b = a_1$

Fa 00
EGN 3420 Exam 2 Name _____

SHOW ALL WORK!

Problem 1 (35 pts)

For the table of data points given below

x_i	y_i	$u_i=\log x_i$	$z_i=\log y_i$	u_i^2	$u_i z_i$	$\hat{Y}_i=aX_i^b$	$e_i=Y_i-\hat{Y}_i$	e_i^2
1	100	0	2.0000	0	0	102.0426	-2.0426	4.1723
2	53	0.3010	1.7243	0.0906	0.5191	51.3134	1.6866	2.8445
3	35	0.4771	1.5441	0.2276	0.7367	34.3234	0.6766	0.4578
4	25	0.6021	1.3979	0.3625	0.8416	25.8036	-0.8036	0.6458
		1.3802	6.6663	0.6807	2.0974			8.1204

A) Fit a power equation model $y = ax^b$ to the data in the table.

~~B) Find the total sum of squares SST~~

C) Calculate $SSE = \sum_i (y_i - \hat{y}_i)^2$

You may use the remaining columns in the table to help solve the problem. Round all answers to 4 places after the decimal point.

A) $z = a_0 + a_1 u$

$n a_0 + \Sigma u_i a_1 = \Sigma z_i$

$\Sigma u_i a_0 + \Sigma u_i^2 a_1 = \Sigma u_i z_i$

$4a_0 + 1.3802 a_1 = 6.6663$

$1.3802 a_0 + 0.6807 a_1 = 2.0974$

$a_0 = 2.0088$

$a_1 = -0.9918$

$a = 10^{a_0} = 102.0426$

$b = a_1 = -0.9918$

B)

C) $SSE = 8.1204$

Fa 00
EGN 3420 Exam 2 Name _____

SHOW ALL WORK!

Problem 2 (35 pts)

A Least Squares Line was fit through the 4 data points in the table below. The Sum of the squares of the errors, SSE = 27/4.

x_i	y_i	x_i^2	$x_i y_i$	$\hat{y}_i$	e_i	e_i^2
-5	-10	25	50	$\frac{y_2}{4} - 10$	$-\frac{y_2}{4}$	$\frac{y_2^2}{16}$
0	y_2	0	0	$\frac{y_2}{4}$	$\frac{3}{4}y_2$	$\frac{9y_2^2}{16}$
1	2	1	2	$\frac{y_2}{4} + 2$	$-\frac{y_2}{4}$	$\frac{y_2^2}{16}$
4	8	16	32	$\frac{y_2}{4} + 8$	$-\frac{y_2}{4}$	$\frac{y_2^2}{16}$
0	y_2	42	84			

a) Find the missing value for y_2.

b) Find the equation of the Least Squares Line.

You may use the remaining columns in the table to help solve the problem. Round all answers to 4 places after the decimal point.

$n a_0 + \Sigma x_i a_1 = \Sigma y_i$

$\Sigma x_i a_0 + \Sigma x_i^2 a_1 = \Sigma x_i y_i$

$4 a_0 + 0(a_1) = y_2$

$0(a_0) + 42 a_1 = 84$

$a_0 = \frac{y_2}{4}$

$a_1 = 2$

$\hat{y}_i = \frac{y_2}{4} + 2x_i$

$e_i = y_i - \hat{y}_i$

$\Sigma e_i^2 = \frac{12 y_2^2}{16}$

$\frac{27}{4} = \frac{3}{4} y_2^2$, $y_2^2 = 9$, $y_2 = 3$

b) $a_0 = \frac{y_2}{4} = \frac{3}{4}$

$\hat{y} = a_0 + a_1 x$

$\hat{y} = \frac{3}{4} + 2x$