Simple least squares

- Summary
  - Model form: $y = a_0 + a_1 x + e$
  - $S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$ becomes minimizes where $\frac{\partial S_r}{\partial a_0} = 0$ & $\frac{\partial S_r}{\partial a_1} = 0.$
  - Rearranging and solving for a_0 and a_1

$$n a_0 + \left(\sum x_i\right) a_1 = \sum y_i \qquad \left(\sum x_i\right) a_0 + \left(\sum x_i^2\right) a_1 = \sum x_i y_i$$

$$\Rightarrow \quad a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2} \qquad a_0 = \bar{y} - a_1 \bar{x}$$

- Question: what if our model we want to find is non-linear?

Ex. Activation energy in rate constant

$$k = k_0 e^{-E/RT}$$

→ Linearize !

Linearization

- Want to model non-linear relationships between independent (x) and dependent (y) variables.
  1. Make a simple linear model through a suitable transformation.

$$y = f(x) + e \quad \rightarrow \quad y = a_0 + a_1 x + e$$

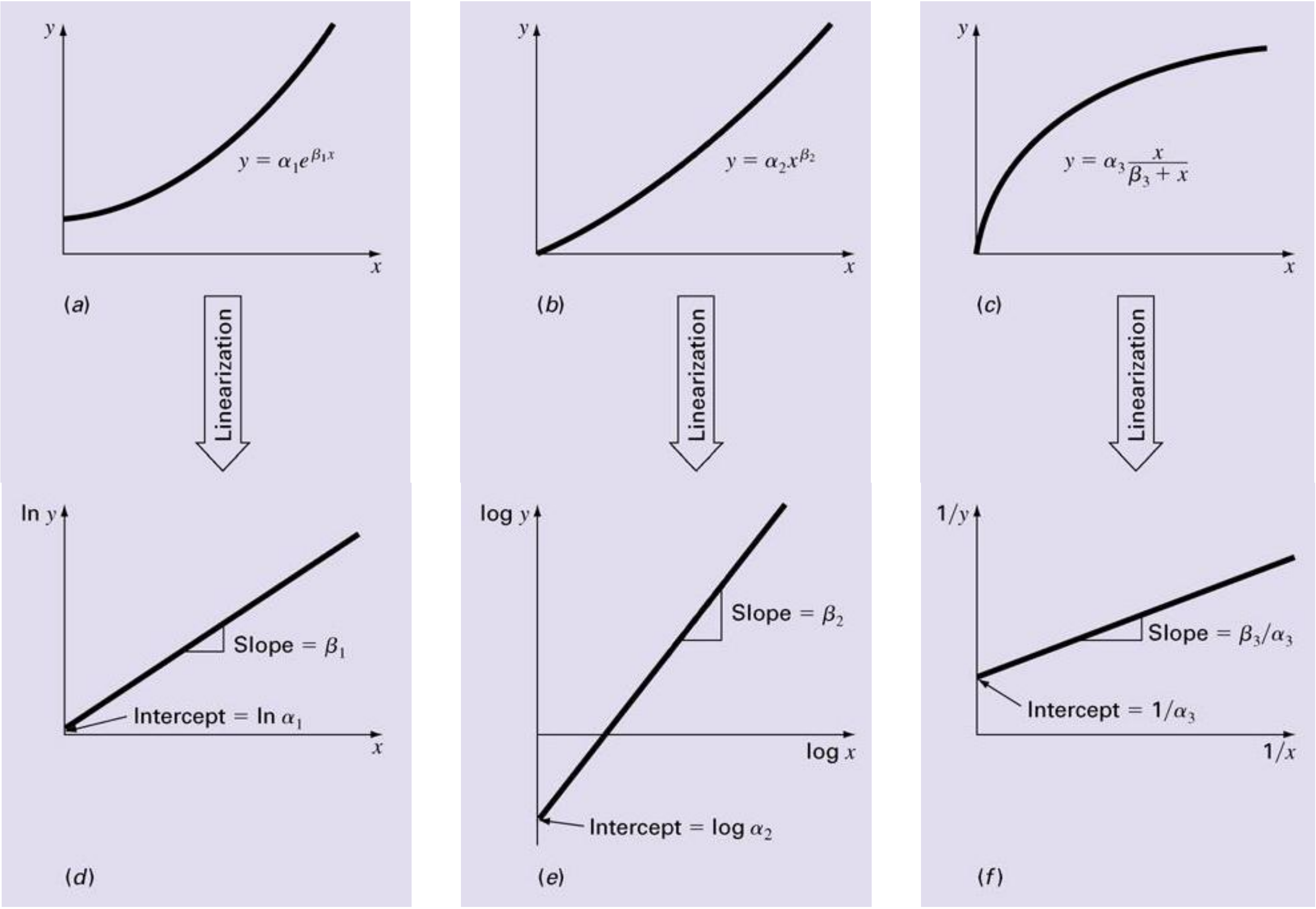
  2. Use previous results (simple least squares)

$$a_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \qquad a_0 = \bar{y} - a_1 \bar{x}$$

※Caution: transformation also changes P.D.F of variables (and errors)

We will discuss about this in model assessment.

Linearization (Cont.)



Polynomial regression

- For quadratic form

$$y = a_0 + a_1 x + a_2 x^2 + e$$

  - Sum of squares

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2 \right)^2$$

Again, S_r has a parabolic shape w.r.t a_0, a_1, and a_2. with plus signs of a_0^2, a_1^2, and a_2^2.

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2) = 0$$

Polynomial regression (Cont.)

- Rearranging the previous equations gives

$$(n)a_0 + \left(\sum x_i\right)a_1 + \left(\sum x_i^2\right)a_2 = \sum y_i$$
$$\left(\sum x_i\right)a_0 + \left(\sum x_i^2\right)a_1 + \left(\sum x_i^3\right)a_2 = \sum x_i y_i$$
$$\left(\sum x_i^2\right)a_0 + \left(\sum x_i^3\right)a_1 + \left(\sum x_i^4\right)a_2 = \sum x_i^2 y_i$$

$$\Rightarrow \begin{pmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{pmatrix}$$

the above equations can be solved easily. (three unknowns and three equations.)

- For general polynomials

$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_m x^m + e$$

- From the results of two cases ($y = a_0 + a_1x$ & $y = a_0 + a_1x + a_2x^2$)

$$\frac{\partial S_r}{\partial a_0} = \frac{\partial S_r}{\partial a_1} = \cdots = \frac{\partial S_r}{\partial a_m} = 0$$

we need to solve (m+1) linear algebraic equations for (m+1) parameters.

Multiple least squares

- Consider when there are more than two independent variables, x_1, x_2, …, x_m. ➔ regression plane.

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_m x_m + e$$

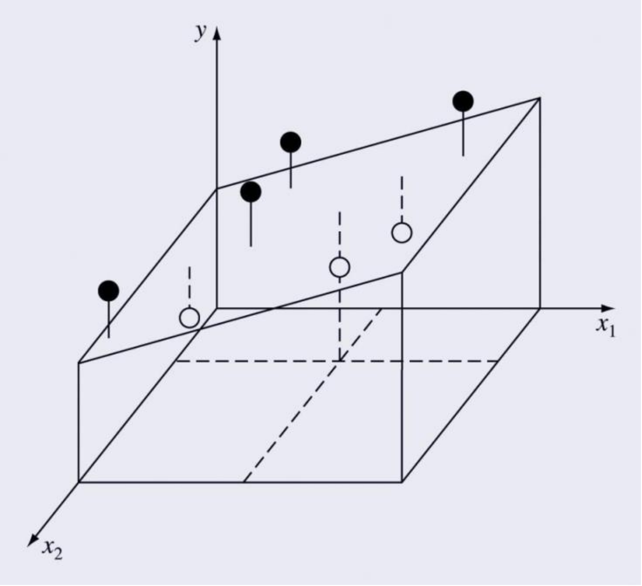
- For 2-D case, $y = a_0 + a_1 x_1 + a_2 x_2$.
  - Again, S_r has a parabolic shape w.r.t a_0, a_1.

$$S_r = \sum (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i})^2$$

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_{1,i} (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}) = 0$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_{2,i} (y_i - a_0 - a_1 x_{1,i} - a_2 x_{2,i}) = 0$$



Multiple least squares (Cont.)

- Rearranging and solve for a_0, a_1 and a_2 gives

$$\begin{pmatrix} n & \sum x_{1,i} & \sum x_{2,i} \\ \sum x_{1,i} & \sum x_{1,i}^2 & \sum x_{1,i}x_{2,i} \\ \sum x_{2,i} & \sum x_{1,i}x_{2,i} & \sum x_{2,i}^2 \end{pmatrix} \begin{Bmatrix} a_1 \\ a_2 \\ a_3 \end{Bmatrix} = \begin{Bmatrix} \sum y_i \\ \sum x_{1,i}y_i \\ \sum x_{2,i}y_i \end{Bmatrix}$$

- For an m-dimensional plane,

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_mx_m + e$$

- Same as in general polynomials,

$$\frac{\partial S_r}{\partial a_0} = \frac{\partial S_r}{\partial a_1} = \cdots = \frac{\partial S_r}{\partial a_m} = 0$$

we need to solve (m+1) linear algebraic equations for (m+1) parameters.

General least squares

- The following form includes all cases (simple least squares, polynomial regression, multiple regression)

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

$$\text{where } z_0, z_1, \ldots, z_m \; : m+1 \text{ different functions}$$

Ex. Simple and multiple least squares

$$Z_0 = 1, Z_1 = x_1, Z_2 = x_2, \cdots, Z_m = x_m$$

polynomial regression

$$Z_0 = x^0 = 1, Z_1 = x^1, Z_2 = x^2, \cdots, Z_m = x^m$$

- Same as before,

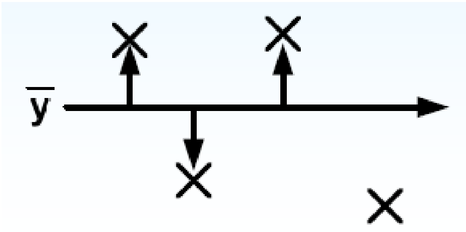
$$\frac{\partial S_r}{\partial a_0} = \frac{\partial S_r}{\partial a_1} = \cdots = \frac{\partial S_r}{\partial a_m} = 0$$

we need to solve (m+1) linear algebraic equations for (m+1) parameters.

Quantification of errors

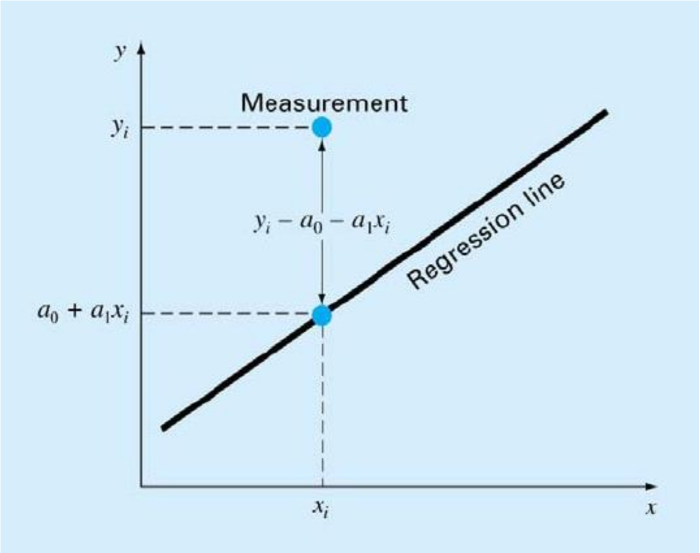
$$S_t = \sum (y_i - \bar{y})^2$$

Total sum of squares around the mean for the response variable, y



$$S_r = \sum e_i^2 = \sum (y_i - a_0 z_{0,i} - a_1 z_{1,i} - \cdots - a_m z_{m,i})^2$$

Sum of squares of residuals around the regression line



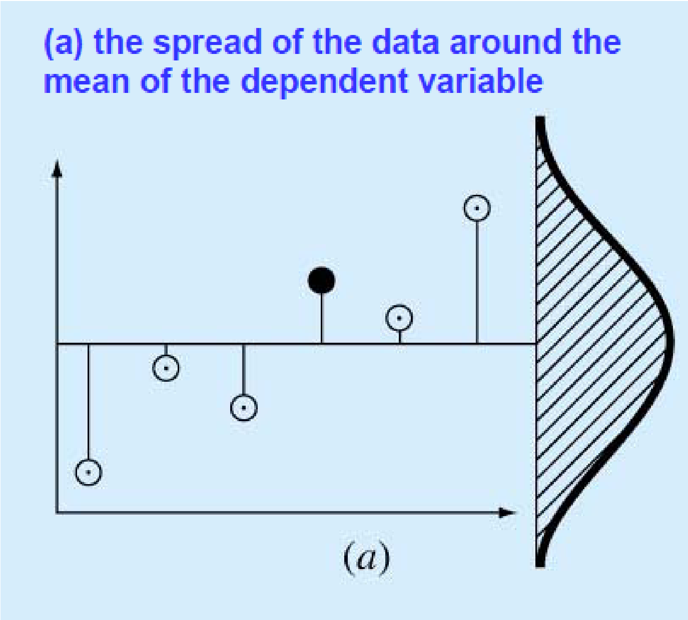
Quantification of errors (Cont.)

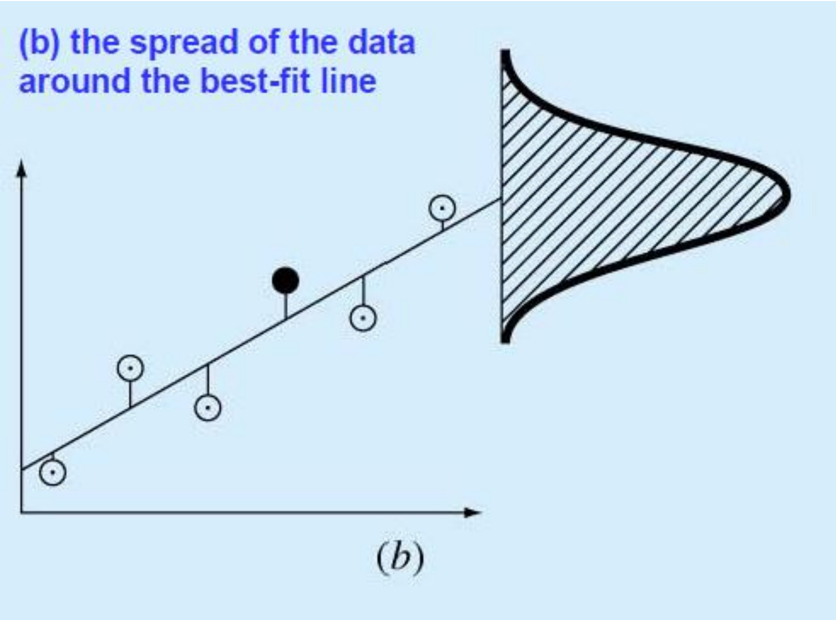
$$S_y = \sqrt{\frac{1}{n-1}\sum(y_i - \bar{y})^2} = \sqrt{\frac{S_t}{n-1}}$$

Standard deviation of *y*

$$S_{y/x} = \sqrt{\frac{S_r}{n-(m+1)}}$$

Standard error of predicted *y*
→ quantify appropriateness of regression





Quantification of errors (Cont.)

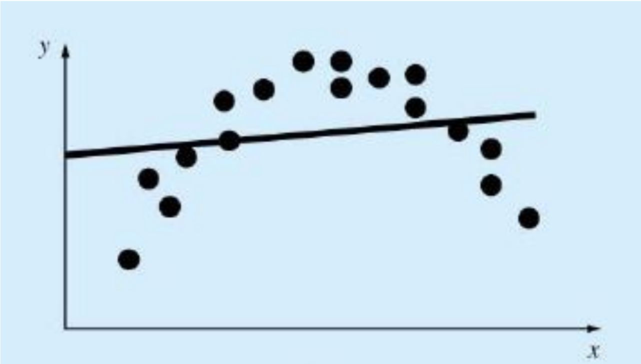
- Coefficients of determination, R^2

$$R^2 = \sqrt{\frac{S_t - S_r}{S_t}}$$

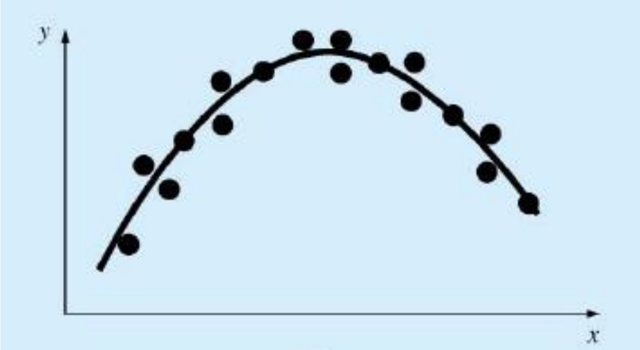
The amount of variability in the data explained by the regression model.

$R^2 = 1$ when $S_r = 0$: perfect fit (a regression curve passes through data points)

$R^2 = 0$ when $S_r = S_t$: as bad as doing nothing



(a)



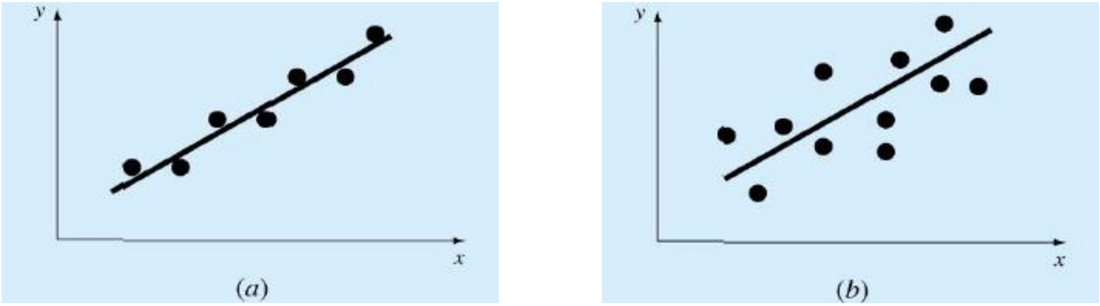
(b)

It is evident from the figures that a parabola is adequate.
R^2 of (b) is higher than that of (a)

Quantification of errors (Cont.)

- **Warning!** : $R^2 \approx 1$ **does not guarantee** that the model is adequate, nor the model will predict new data well.
  - It is possible to force R^2 to be one by adding as many terms as there are observations.
  - S_r can be big when variance of random error is large.

(Usual assumption on error is that error is random is unpredictable)



Practice using Minitab

(1) Wind tunnel example with higher polynomials

(2) Simple regression with increasing random noise

Confidence intervals - coefficients

- Coefficients in the regression model have confidence interval.

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

  - Why? They are also statistics like $\bar{x}$ & s. That is, they are numerical quantities calculated in a sample (not entire population). They are estimated values of parameters.

Statistic that we want to find its confidence interval

$$statistic \pm A \times \sigma_{statistic}$$

Value that depends on P.D.F of the statistic & confidence level α

Standard error of the statistic

statistic	A	$\sigma_{statistic}$
$\bar{x}$	$z_{\alpha/2}$	$\sigma_x / \sqrt{n}$
$\bar{x}$	$t_{\nu,\alpha/2}$	$s_x / \sqrt{n}$

※The standard error of a statistic is the standard deviation of the sampling distribution of that statistic

Confidence intervals – coefficients (cont.)

- Matrix representation of GLS

$$y = a_0 z_0 + a_1 z_1 + a_2 z_2 + \cdots + a_m z_m + e$$

$$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{e}$$

– matrix of the calculated values of the basis functions at the measured values of the independent variable

– observed valued of the dependent variable

– unknown coefficients

– residuals

$$\mathbf{Z} = \begin{bmatrix} Z_{01} & Z_{11} & \cdots & Z_{m1} \\ Z_{02} & Z_{12} & \cdots & Z_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{0n} & Z_{1n} & \cdots & Z_{mn} \end{bmatrix}$$

$$\mathbf{y}^T = \lfloor y_1 \ \ y_2 \ \ \cdots \ \ y_n \rfloor$$

$$\mathbf{a}^T = \lfloor a_0 \ \ a_1 \ \ \cdots \ \ a_m \rfloor$$

$$\mathbf{e}^T = \lfloor e_1 \ \ e_2 \ \ \cdots \ \ e_n \rfloor$$

m+1: number of coefficients
n: number of data points

Confidence intervals – coefficients (Cont.)

- Example

Fitting quadratic polynomials to five data points

x	-1.0	-0.5	0.0	0.5	1.0
y	1.0	0.5	0.0	0.5	2.0

$$y = a_0 + a_1 x + a_2 x^2 + e$$

$$\mathbf{y} = \mathbf{Za} + \mathbf{e}$$

$$\begin{bmatrix} 1.0 \\ 0.5 \\ 0.0 \\ 0.5 \\ 2.0 \end{bmatrix} = \begin{bmatrix} 1 & -1.0 & 1.0 \\ 1 & -0.5 & 0.25 \\ 1 & 0.0 & 0.0 \\ 1 & 0.5 & 0.25 \\ 1 & 1.0 & 1.0 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Three unknowns
Five equations

Can you solve this problem?

Confidence intervals – coefficients (Cont.)

- Solutions

$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{e}$

Sum of squares of errors

$$S_r = \sum e_i^2 = \mathbf{e}^T\mathbf{e} = (\mathbf{y} - \mathbf{Z}\mathbf{a})^T(\mathbf{y} - \mathbf{Z}\mathbf{a})$$

$$\frac{\partial S_r}{\partial \mathbf{a}} = 0 \quad \longrightarrow \quad (\mathbf{Z}^T\mathbf{Z})\mathbf{a} = \mathbf{Z}^T\mathbf{y}$$

Called "normal equations"

1. LU decomposition or other methods to solve L.A.E

$$(\mathbf{Z}^T\mathbf{Z})\mathbf{a} = \mathbf{Z}^T\mathbf{y} \qquad \Rightarrow \text{"}\mathbf{A}\mathbf{x} = \mathbf{b}\text{"}$$

2. Matrix inversion

$$(\mathbf{Z}^T\mathbf{Z})\mathbf{a} = \mathbf{Z}^T\mathbf{y} \qquad \Rightarrow \mathbf{a} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}$$

computationally not efficient, but statistically useful

Confidence intervals – coefficients (Cont.)

- Matrix inversion approach

$$\mathbf{a} = \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{y}$$

Denote Z_{ii}^{-1} as the diagonal element of $\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$

Confidence interval of estimated coefficients

$$a_{i-1} \pm t_{n-(m+1),\alpha/2}\sqrt{S_{y/x}^2 Z_{ii}^{-1}}$$

$t_{n-(m+1),\alpha/2}$ Student t statistics

$S_{y/x} = \sqrt{\dfrac{S_r}{n-(m+1)}}$ Standard error of estimate

What if confidence intervals contain zero?