Least-square Linear Regression

Qin Songmei (秦松梅) & Zeng Qi (曾琪)

Linear Regression

- A set of N data points (x_i, y_i)
- A straight-line model:

$$y(x) = y(x; a, b) = a + bx$$

y: dependent variable
x: independent variable
a, b: regression coefficients

- Measure how well the model agrees with the data

Famous linear relations in astronomy

- Period -luminosity relation of Cepheids
- Absolute magnitude-Metallicity relation of RR Lyrae
- M_{BH}-σ relation
- Tully-Fisher (L - Vmax) relation
- Faber-Jackson (L~σ) relation

……

All are statistical scaling relations, none of them are first principle like F=ma

Six methods for linear regression

- OLS(Y/X) (standard ordinary least squares)---a

Minimizes sum of square of vertical distances

- OLS(X/Y) (inverse line)----------b

Minimizes sum of square of horizontal distances

- OLS bisector

Bisects the angle formed by the two lines

- OR (Orthogonal regression)----------c

Minimizes sum of squares of perpendicular distances

- RMA (Reduced major-axis)----------d

Geometric mean of the two slopes

- Mean OLS

Arithmetic mean of the two slopes

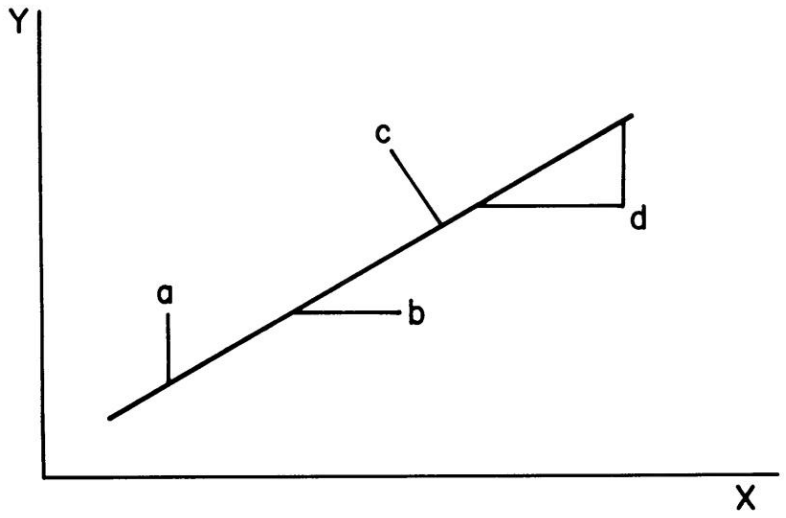


FIG. 1.—Illustration of the different methods for minimizing the distance of the data from a fitted line: (*a*) OLS($Y \mid X$), where the distance is measured vertically; (*b*) OLS($X \mid Y$), where the distance is taken horizontally; (*c*) OR, where the distance is measured vertically to the line; and (*d*) RMA, where the distances are measured both perpendicularly and horizontally. No illustration of the OLS bisector is drawn in this figure.

Isobe, Feigelson, Akritas & Babu, ApJ 364, 105 1990
Feigelson & Babu, ApJ 397, 55 1992

Formulae for six Linear Regression Slopes

Method	Expression for Slope	The Estimate of the Variance of the Slope $\widehat{\mathrm{Var}}(\hat{\beta}_i)$
OLS(Y|X)	$\hat{\beta}_1 = S_{xy}/S_{xx}$	$[\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \hat{\beta}_1 x_i - \overline{y} + \hat{\beta}_1\overline{x})^2]/S_{xx}^2$
OLS(X|Y)	$\hat{\beta}_2 = S_{yy}/S_{xy}$	$[\sum_{i=1}^{n}(y_i - \overline{y})^2(y_i - \hat{\beta}_2 x_i - \overline{y} + \hat{\beta}_2\overline{x})^2]/S_{xy}^2$
OLS-bisector	$\hat{\beta}_3 = (\hat{\beta}_1 + \hat{\beta}_2)^{-1}[\hat{\beta}_1\hat{\beta}_2 - 1 + [(1+\hat{\beta}_1^2)(1+\hat{\beta}_2^2)]^{\frac{1}{2}}]$	$[\hat{\beta}_3^2/(\hat{\beta}_1 + \hat{\beta}_2)^2(1+\hat{\beta}_1^2)(1+\hat{\beta}_2^2)]\ [(1+\hat{\beta}_2^2)^2\widehat{\mathrm{Var}}(\hat{\beta}_1)) + 2(1+\hat{\beta}_1^2)(1+\hat{\beta}_2^2)\widehat{\mathrm{Cov}}(\hat{\beta}_1,\hat{\beta}_2) + (1+\hat{\beta}_1^2)^2\widehat{\mathrm{Var}}(\hat{\beta}_2)]$
Orthogonal regression	$\hat{\beta}_4 = \frac{1}{2}[(\hat{\beta}_2 - \hat{\beta}_1^{-1}) + sign(S_{xy})[4 + (\hat{\beta}_2 - \hat{\beta}_1^{-1})^2]^{\frac{1}{2}}]$	$\hat{\beta}_4^2[\hat{\beta}_1^{-2}\widehat{\mathrm{Var}}(\hat{\beta}_1) + 2\widehat{\mathrm{Cov}}(\hat{\beta}_1,\hat{\beta}_2) + \hat{\beta}_1^2\widehat{\mathrm{Var}}(\hat{\beta}_2)]/[4\hat{\beta}_1^2 + (\hat{\beta}_1\hat{\beta}_2 - 1)^2]$
Reduced major axis	$\hat{\beta}_5 = sign(S_{xy})(\hat{\beta}_1\hat{\beta}_2)^{\frac{1}{2}}$	$\frac{1}{4}[(\hat{\beta}_2/\hat{\beta}_1)\widehat{\mathrm{Var}}(\hat{\beta}_1) + 2\widehat{\mathrm{Cov}}(\hat{\beta}_1,\hat{\beta}_2) + (\hat{\beta}_1/\hat{\beta}_2)\widehat{\mathrm{Var}}(\hat{\beta}_2)]$
OLS-mean	$\hat{\beta}_6 = \frac{1}{2}(\hat{\beta}_1 + \hat{\beta}_2)$	$\frac{1}{4}[\widehat{\mathrm{Var}}(\hat{\beta}_1) + \widehat{\mathrm{Var}}(\hat{\beta}_2) + 2\widehat{\mathrm{Cov}}(\hat{\beta}_1,\hat{\beta}_2)]$

Isobe, Feigelson, Akritas & Babu, ApJ 364, 105 1990
Feigelson & Babu, ApJ 397, 55 1992

Formulae for six Linear Regression Intercept

$$\hat{\alpha}_j = \bar{y} - \hat{\beta}_j \bar{x}$$

$$\widehat{\mathrm{Var}}\,(\hat{\alpha}_j) = \frac{1}{n^2} \sum_i^n \left\{ y_i^0 - \hat{\beta}_j x_i^0 - n\bar{x} \right.$$

$$\left. \times \left[\frac{\gamma_{1j}}{S_{xx}} x_i^0 (y_i^0 - \hat{\beta}_1 x_i^0) + \frac{\gamma_{2j}}{S_{xy}} y_i^0 (y_i^0 - \hat{\beta}_2 x_i^0) \right] \right\}^2 , \quad (9)$$

where $\hat{\gamma}_{ij}$ are given by

$$\gamma_{11} = 1 , \quad (10)$$

$$\gamma_{12} = 0 , \quad (11)$$

$$\gamma_{13} = \gamma_1 (1 + \hat{\beta}_2^2) , \quad (12)$$

$$\gamma_{14} = \gamma_2 |\hat{\beta}_1|^{-1} , \quad (13)$$

$$\gamma_{15} = \tfrac{1}{2} \sqrt{\hat{\beta}_2 / \hat{\beta}_1} , \quad (14)$$

$$\gamma_{21} = 0 , \quad (15)$$

$$\gamma_{22} = 1 , \quad (16)$$

$$\gamma_{23} = \gamma_1 (1 + \hat{\beta}_1^2) , \quad (17)$$

$$\gamma_{24} = |\hat{\beta}_1| \gamma_2 , \quad (18)$$

$$\gamma_{25} = \tfrac{1}{2} \sqrt{\hat{\beta}_1 / \hat{\beta}_2} , \quad (19)$$

with

$$\gamma_1 = \hat{\beta}_3 [(\hat{\beta}_1 + \hat{\beta}_2) \sqrt{(1 + \hat{\beta}_1^2)(1 + \hat{\beta}_2^2)}]^{-1} , \quad (20)$$

$$\gamma_2 = \hat{\beta}_4 [4\hat{\beta}_1^2 + (\hat{\beta}_1 \hat{\beta}_2 - 1)^2]^{-1/2} . \quad (21)$$

Isobe, Feigelson, Akritas & Babu, ApJ 364, 105 1990

Structural regression models

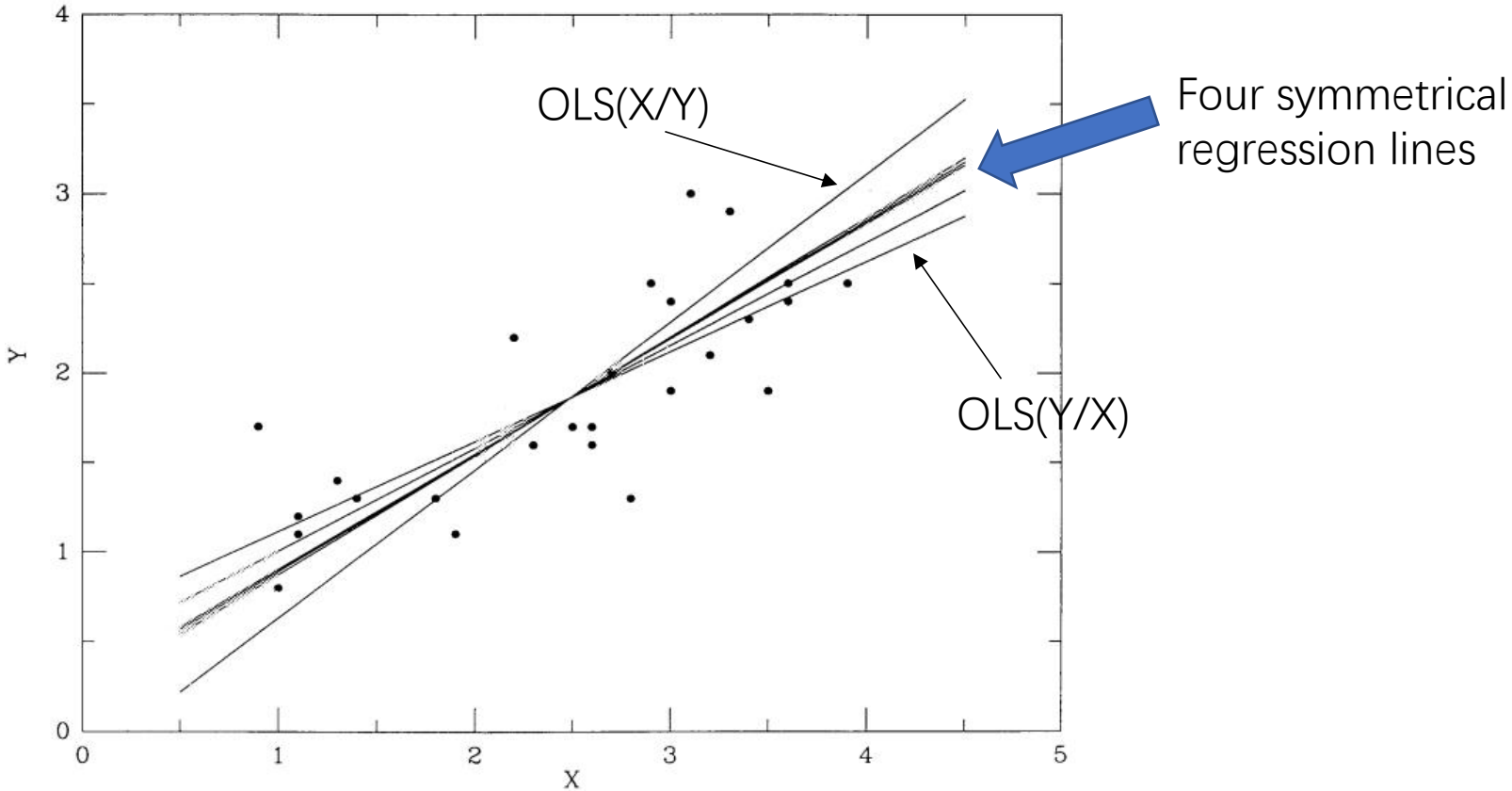


FIG. 1.—[Figs. 1–6 are diagrams of the bivariate regressions discussed in the text, using a hypothetical data set with 26 points.] Unweighted least-squares models (§ 2). Starting with the steepest line, they are OLS($X \mid Y$), OLS mean, OLS bisector, reduced major axis, orthogonal, and OLS($Y \mid X$) regressions. The lines are calculated using the program SLOPES.

Some Guidelines:

1. Avoid specifying variables:

- OLS(Y/X)
- OLS(X/Y)

2. Treat variables symmetrically:

- OLS bisector
- OR (Orthogonal regression)
- RMA (Reduced major-axis)
- Mean OLS

Models

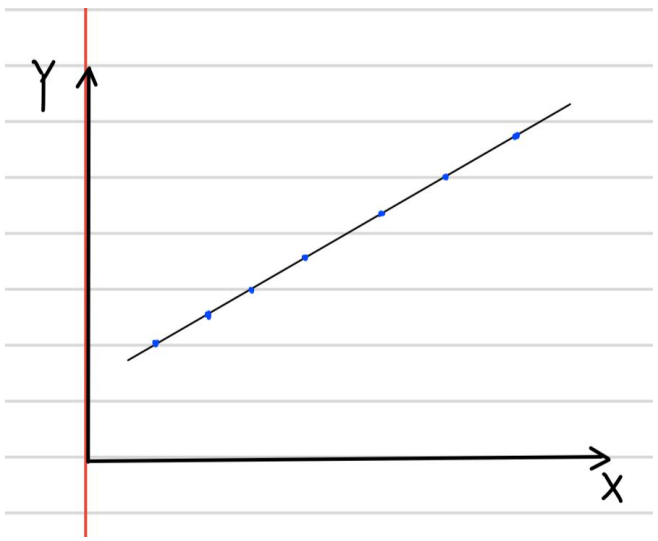
1. Unweighted linear regression models

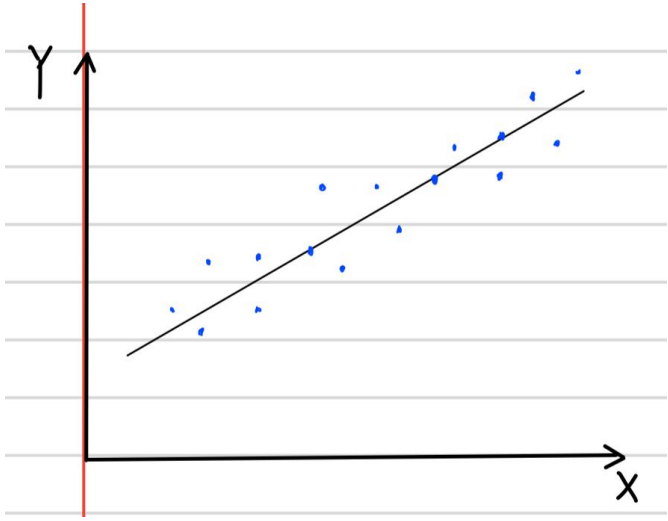
- **Functional regression**:

The true points lie precisely on the line (x_i, y_i)

- **Structural regression** :

The true points are scattered about the line
$(x_i, y_i, \mathrm{X}, \mathrm{Y}, \sigma_x, \sigma_y)$
(Isobe, Feigelson, Akritas & Babu, ApJ 364, 105 1990)





2. Measurement error and unknown intrinsic scatter

- **Direct generalization of the OLS**
- **Weighted least-squares (WLS)**

(Akritas, 1996, apj, 470, 706)

Data with Errors in Y

- chi-square merit function:

$$\chi^2(a,b) = \sum_{i=1}^{N} \left(\frac{y_i - a - bx_i}{\sigma_i} \right)^2$$

- derivatives of $\chi^2(a, b)$:

$$0 = \frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^{N} \frac{y_i - a - bx_i}{\sigma_i^2}$$

$$0 = \frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^{N} \frac{x_i(y_i - a - bx_i)}{\sigma_i^2}$$

- The solution for the best-fit model parameters a and b:

$$\Delta \equiv SS_{xx} - (S_x)^2$$

$$a = \frac{S_{xx}S_y - S_x S_{xy}}{\Delta}$$

$$b = \frac{SS_{xy} - S_x S_y}{\Delta}$$

Some notations:

$$S \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \quad S_x \equiv \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \quad S_y \equiv \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}$$

$$S_{xx} \equiv \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \quad S_{xy} \equiv \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2}$$

Data with Errors in Y

- With the propagation of errors: $\sigma_f^2 = \sum_{i=1}^{N} \sigma_i^2 \left(\frac{\partial f}{\partial y_i}\right)^2$

- The derivatives of a and b with respect to y_i: $\frac{\partial a}{\partial y_i} = \frac{S_{xx} - S_x x_i}{\sigma_i^2 \Delta}$ $\frac{\partial b}{\partial y_i} = \frac{S x_i - S_x}{\sigma_i^2 \Delta}$

- The variances in the estimates of a and b: $\sigma_a^2 = S_{xx}/\Delta$ $\sigma_b^2 = S/\Delta$

Data with Errors in Both Coordinates

- σ_{x_i}, σ_{yi} the x and y standard deviations for the ith point

$$\chi^2(a,b) = \sum_{i=1}^{N} \frac{(y_i - a - bx_i)^2}{\sigma_{y\,i}^2 + b^2\sigma_{x\,i}^2}$$

$$\mathrm{Var}(y_i - a - bx_i) = \mathrm{Var}(y_i) + b^2\mathrm{Var}(x_i) = \sigma_{y\,i}^2 + b^2\sigma_{x\,i}^2 \equiv 1/w_i$$

- derivatives of $\chi^2(a, b)$:

$$\partial\chi^2/\partial b = 0$$

$$\partial\chi^2/\partial a = 0$$

Data with Errors in Both Coordinates

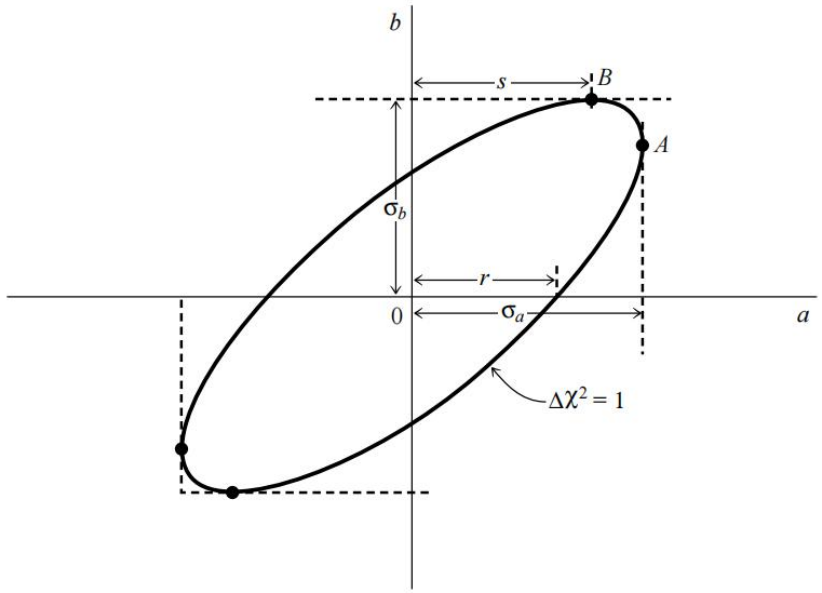
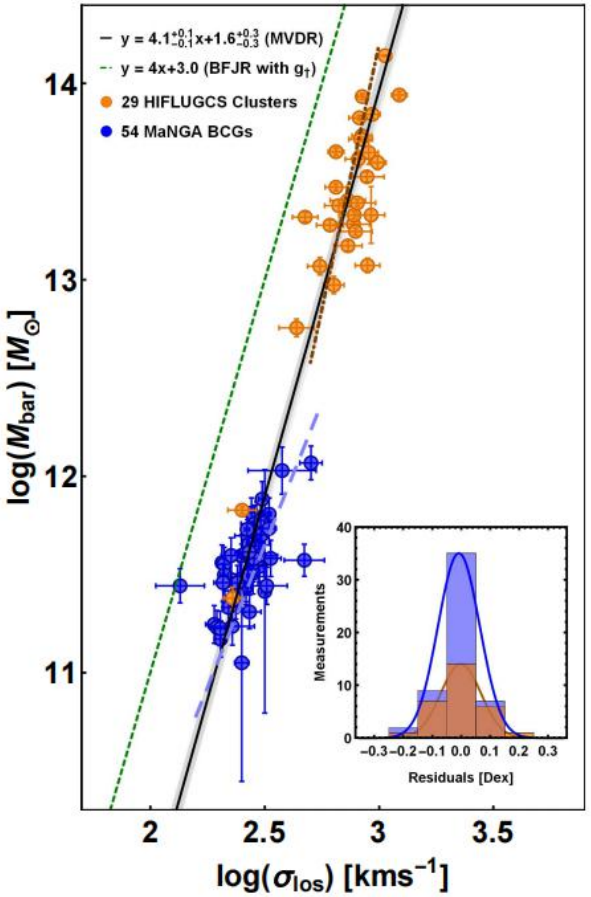


Figure 15.3.1. Standard errors for the parameters a and b. The point B can be found by varying the slope b while simultaneously minimizing the intercept a. This gives the standard error σ_b, and also the value s. The standard error σ_a can then be found by the geometric relation $\sigma_a^2 = s^2 + r^2$.

$$a = \left[\sum_i w_i(y_i - bx_i)\right] \Big/ \sum_i w_i$$

Data with intrinsic scatter and measurement error

Example:



The log-likelihood function is written as

$$-2\ln\mathcal{L} = \sum_i \ln\left(2\pi\sigma_i^2\right) + \sum_i \frac{\Delta_i^2}{\sigma_i^2}\,, \tag{4}$$

with

$$\Delta_i^2 = \frac{(y_i - m\,x_i - b)^2}{m^2+1}\,, \tag{5}$$

where i runs over all data points, and σ_i includes the observational uncertainties $(\sigma_{x_i}, \sigma_{y_i})$ and the lognormal intrinsic scatter $\sigma_{\rm int}$,

$$\sigma_i^2 = \frac{m^2\sigma_{x_i}^2}{m^2+1} + \frac{\sigma_{y_i}^2}{m^2+1} + \sigma_{\rm int}^2\,. \tag{6}$$

(Tian et al. 2021)

Code about least-square

- Linear regression for data with measurement errors and intrinsic scatter (BCES): https://github.com/rsnemmen/BCES

- An alternative approach using Stan (python version) : https://github.com/astrobayes/BMAD/blob/master/chapter_4/code_4.11.py