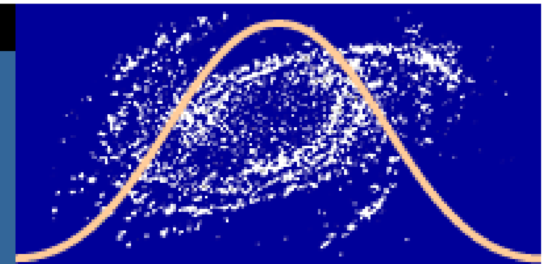
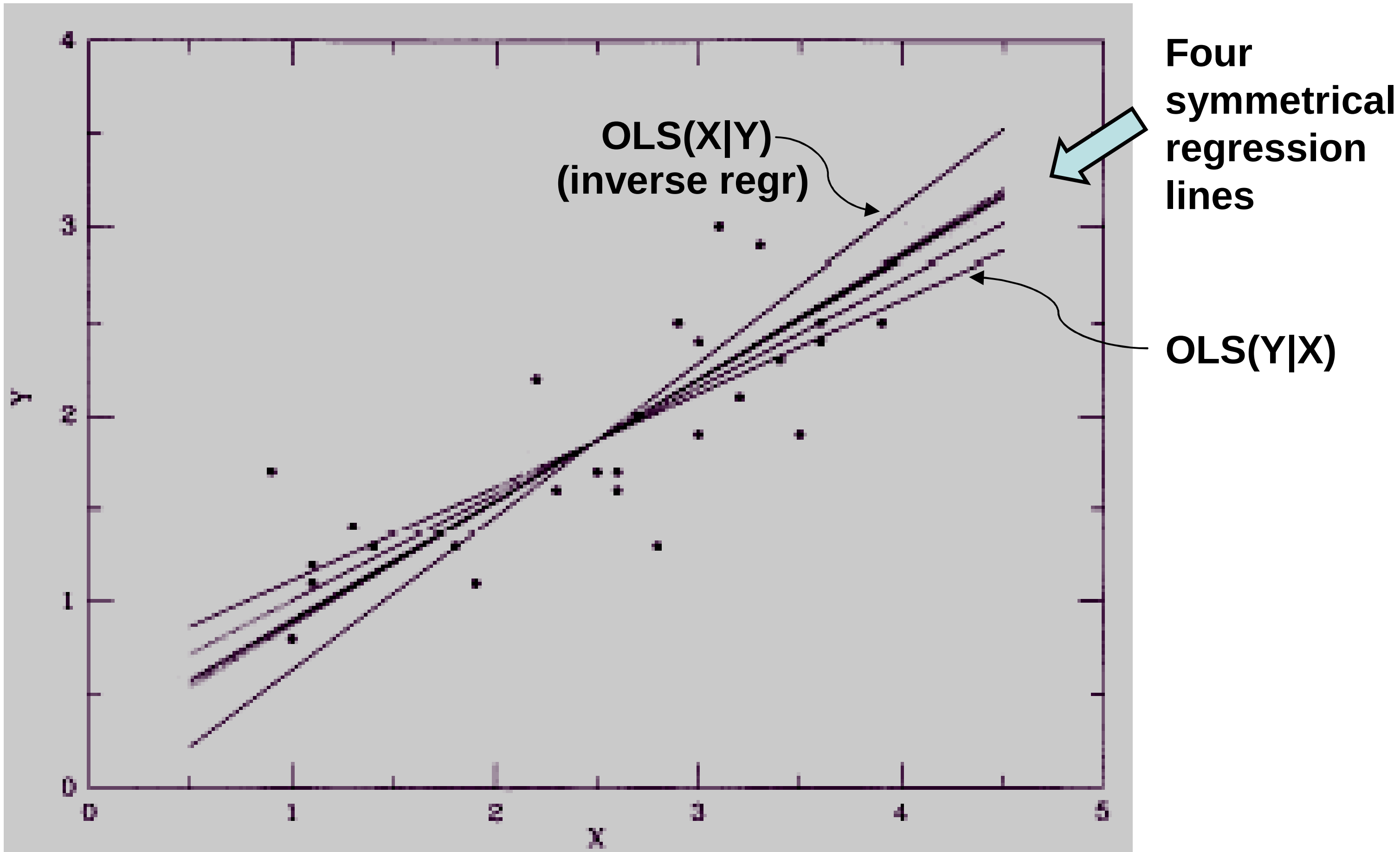


Linear regression issues in astronomy

G. Jogesh Babu
Center for Astrostatistics

Structural regression

Seeking the intrinsic relationship between two properties



$$E(Y)=\alpha+\beta E(X)$$

Analytical formulae for slopes of the 6 OLS lines

TABLE 1

LINEAR REGRESSION FORMULAE FOR SLOPES

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

NOTE.—An estimate of covariance term is given by:

$$\widehat{\mathrm{Cov}}(\hat{\beta}_1, \hat{\beta}_2) = (\hat{\beta}_1 S_{xx}^2)^{-1}\left\{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})[y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x})][y_i - \bar{y} - \hat{\beta}_2(x_i - \bar{x})]\right\}.$$

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

Comments

- Standard estimates of variances of slopes are valid strictly under a very restrictive assumption: errors are independent of X values
- The estimates are valid even when this condition is violated. These are derived using the so called `delta method’

Relations among the slopes

Suppose $S_{XY} > 0$

- If $\beta_5 < 1$, then

$$\beta_3 \leq 1 \text{ and } \beta_1 \leq \beta_4 \leq \beta_5 \leq \beta_3 \leq \beta_2 .$$

- If $\beta_5 > 1$, then

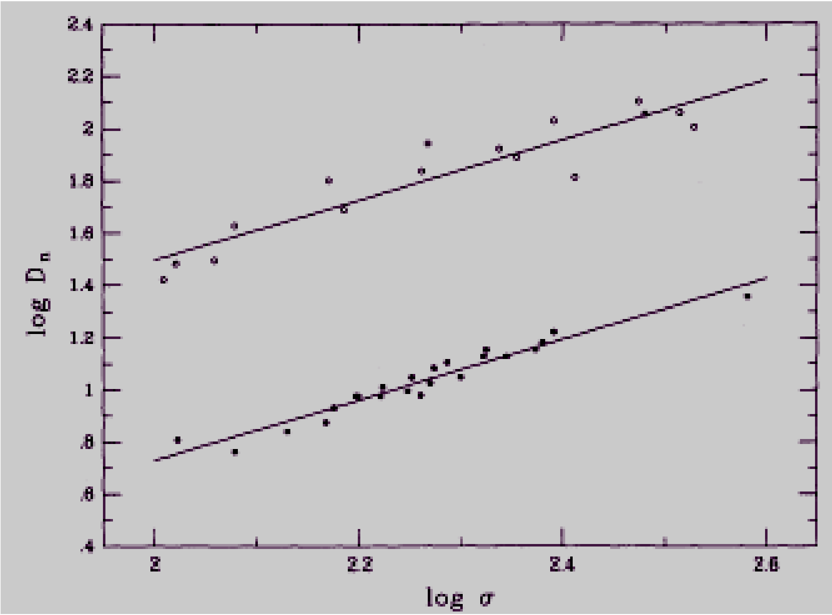
$$\beta_3 \geq 1 \text{ and } \beta_1 \leq \beta_3 \leq \beta_5 \leq \beta_4 \leq \beta_2 .$$

- If $\beta_5 = 1$, then

$$\beta_3 = \beta_4 = \beta_5 .$$

β_5 is the slope of the reduced major axis

- Feigelson & Babu, ApJ 397, p.55, 1992



Example: Faber-Jackson relation between diameter and stellar velocity dispersion of elliptical galaxies

TABLE 4

REGRESSIONS FOR COMA AND VIRGO log D_n' VERSUS log σ*

METHOD (1)	ASYMPTOTIC FORMULAE Intercept (2)	Slope (3)	BOOTSTRAP SLOPE (4)	JACKKNIFE SLOPE (5)
23 Coma Ellipticals				
OLS($Y \mid X$)	-1.595 ± 0.186	1.162 ± 0.082	1.186 ± 0.094	1.164 ± 0.111
OLS($X \mid Y$)	-1.765 ± 0.216	1.238 ± 0.096	1.261 ± 0.104	1.239 ± 0.128
OLS bisector	-1.678 ± 0.200	1.199 ± 0.088	1.223 ± 0.099	1.201 ± 0.119
Orthogonal	-1.694 ± 0.209	1.206 ± 0.092	1.231 ± 0.102	1.208 ± 0.124
Reduced major axis	-1.679 ± 0.200	1.199 ± 0.088	1.223 ± 0.099	1.201 ± 0.119
OLS mean	-1.680 ± 0.200	1.200 ± 0.088	1.224 ± 0.099	1.201 ± 0.119
16 Virgo Ellipticals				
OLS($Y \mid X$)	-0.790 ± 0.230	1.144 ± 0.101	1.143 ± 0.127	1.114 ± 0.118
OLS($X \mid Y$)	-1.183 ± 0.180	1.316 ± 0.082	1.322 ± 0.132	1.316 ± 0.093
OLS bisector	-0.978 ± 0.190	1.227 ± 0.085	1.227 ± 0.107	1.226 ± 0.099
Orthogonal	-1.021 ± 0.198	1.245 ± 0.089	1.246 ± 0.121	1.245 ± 0.104
Reduced major axis	-0.979 ± 0.190	1.227 ± 0.085	1.228 ± 0.108	1.227 ± 0.099
OLS mean	-0.986 ± 0.188	1.230 ± 0.084	1.233 ± 0.110	1.230 ± 0.098

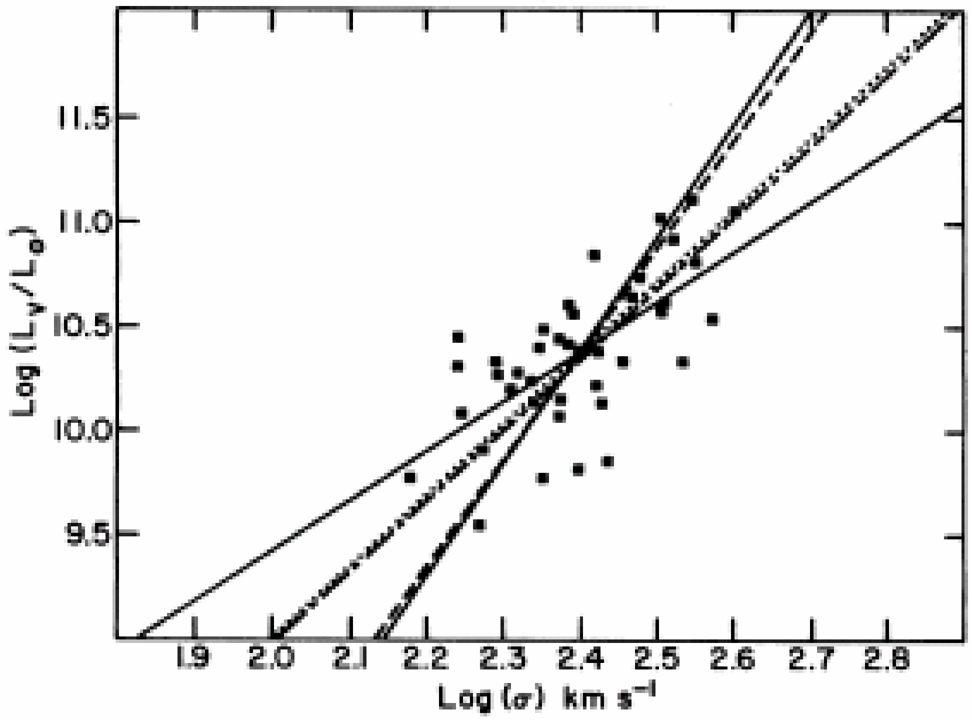
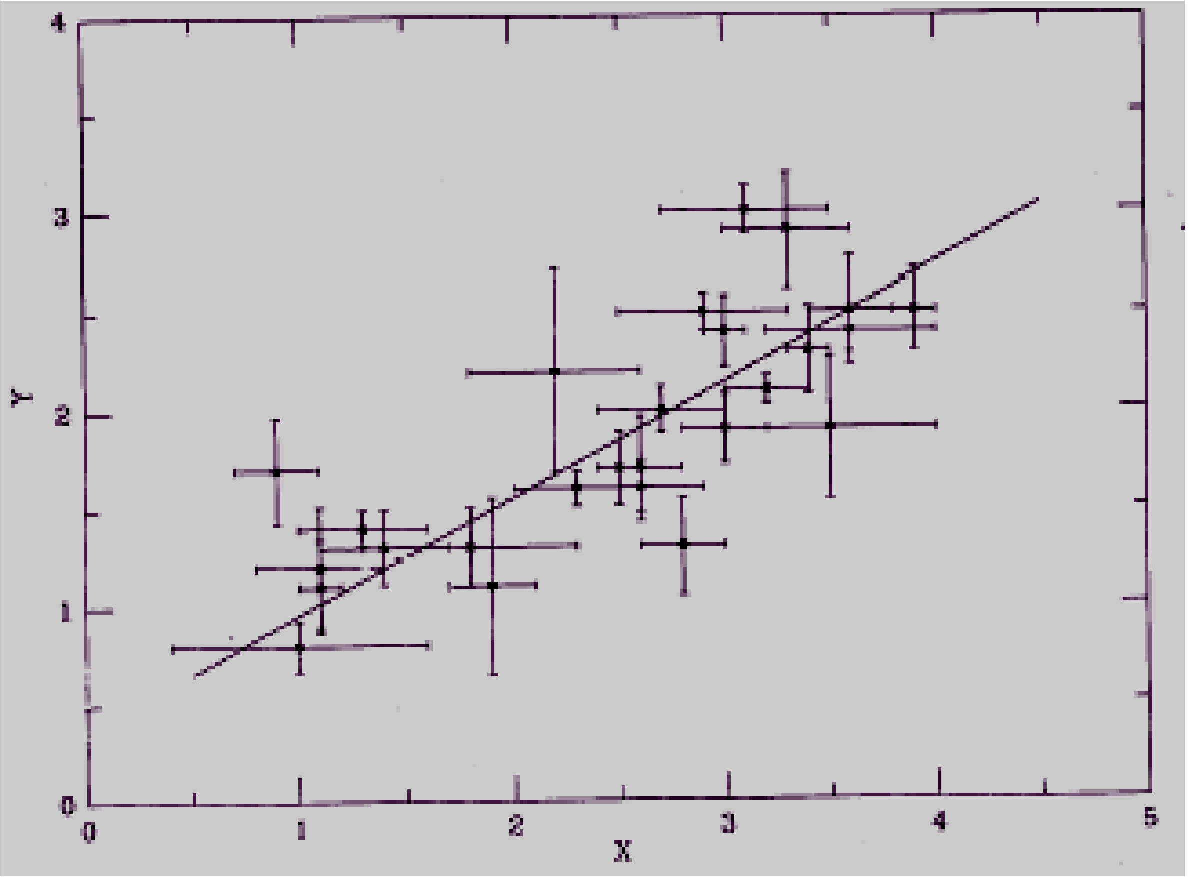


FIG. 2.—Example of a data set with large scatter obtained from Schechter's (1980) measurements of the Faber-Jackson relation in elliptical galaxies. The luminosity is in solar luminosity units. The two solid lines present OLS($Y \mid X$) (*shallowest line*) and OLS($X \mid Y$) (*steepest line*). The dot-dashed line, dashed line, and dotted line represent the OLS bisector, OR, and RMA, respectively.

The calculated slopes are 2.4 ± 0.4 and 5.4 ± 0.8 for the extrema OLS(L/σ) and OLS (σ/L), respectively, and 3.4 ± 0.4, 3.6 ± 0.4 and 5.2 ± 0.8 for the OLS bisector, reduced major axis, and orthogonal regression respectively. The scientific conclusions regarding distances and galaxy formation models obviously depend greatly on the regression method adopted. The dispersion of the five estimates is larger than the variance of any one estimate. The astronomer should calculate all the regression lines and be cautious about the confidence intervals and conclusions.

Heteroscedastic measurement errors in both variables



Homoscedastic functional
Deeming (Vistas Astr 1968)
Fuller "Measurement Error Models" (1987)

Heteroscedastic functional
York (Can J Phys 1966)
ODRPACK Boggs et al. (ACM Trans Math Soft 1990)

Heteroscedastic structural
BCES (Akritas & Bershady ApJ 1996)

Functional Regression

$$Y_i = y_i + \varepsilon_i$$

$$X_i = x_i + \tau_i$$

ε_i and τ_i are measurement errors

- We are interested in the real (regression) relation

$$y_i = bx_i + a$$

- x_i are fixed.

Fitting Power Law

- $f(z)= c\ z^{-\alpha}$ for $z > h > 0$ and for some $\alpha > 1$.
- $Y=\log (f(x))$, $X = \log z$
- $Y= a + b X$
- Fitting the curve is equivalent to estimating a and b by linear regression
- Clearly we use OLS(Y|X)
- X is independent variable and Y is dependent variable

Structural Regression

$$Y_i = y_i + \varepsilon_i$$

$$X_i = x_i + \tau_i$$

ε_i and τ_i are measurement errors

- We are interested in the real (regression) relation

$$y_i = bx_i + a$$

- For any i, x_i is a random variable, it has its own intrinsic variability

Regression with measurement errors and intrinsic scatter

Y = observed data
V = measurement errors

$$(Y_{1i}, Y_{2i}, V_i)\ , \quad i = 1, \ldots n$$

X = intrinsic variables
e = intrinsic scatter

$$Y_{1i} = X_{1i} + \epsilon_{1i} \quad \text{and} \quad Y_{2i} = X_{2i} + \epsilon_{2i}$$

Regression model

$$X_{2i} = \alpha_1 + \beta_1 X_{1i} + \epsilon_i$$

Slope estimator

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2) - \sum_{i=1}^{n} V_{12,i}}{\sum_{i=1}^{n} (Y_{1i} - \bar{Y}_1)^2 - \sum_{i=1}^{n} V_{11,i}}$$

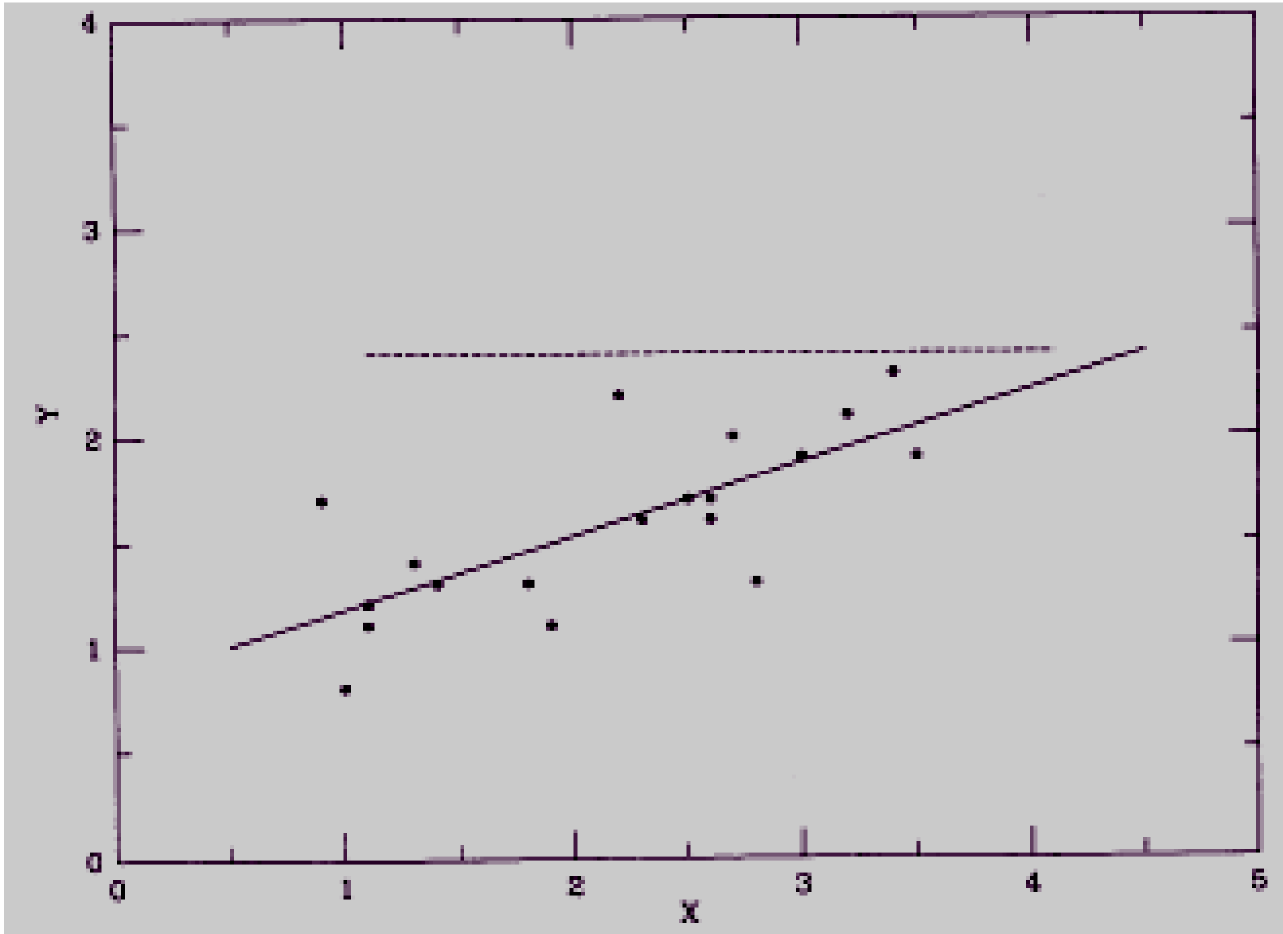
$$\hat{\alpha}_1 = \bar{Y}_2 - \beta_1 \bar{Y}_1 \ .$$

Slope variance

$$\hat{\sigma}^2_{\beta_1} = n^{-1} \sum_{i=1}^{n} (\xi_{1i} - \bar{\xi}_1)^2 \qquad \xi_{1i} = \frac{[Y_{1i} - E(Y_{1i})](Y_{2i} - \beta_1 Y_{1i} - \alpha_1) + \beta_1 V_{11,i} - V_{12,i}}{V(Y_{1i}) - E(V_{11,i})}$$

Akritas & Bershady, ApJ 470, 706 1996

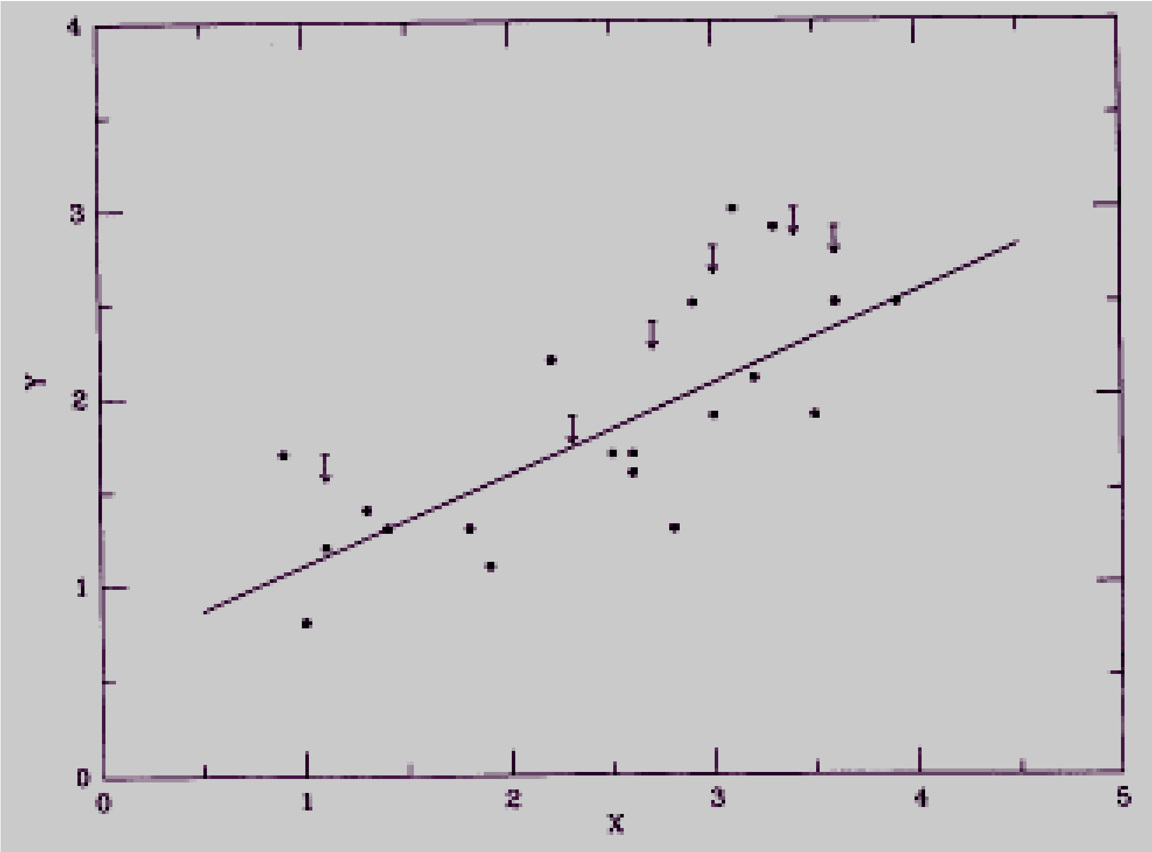
Truncation due to flux limits



Econometrics: Tobit & LIMDEP models (Amemiya, Advanced econometrics 1985; Maddala, Limited-dependent & Quantitative Variables in Econometrics 1983)

Astronomy: Malmquist bias in Hubble diagram (Deeming, Vistas Astr 1968, Segal, PNAS 1975)

Censoring due to non-detections



Correlation coefficients:
Generalized Kendall's τ (Brown, Hollander & Korwar 1974)

Linear regression with normal residuals:
EM Algorithm (Wolynetz Appl Stat 1979)

Linear regression with Kaplan-Meier residuals:
Buckley & James (Biometrika 1979) Schmitt (ApJ 1985)

Isobe, Feigelson & Nelson (ApJ 1986)
Implemented in Astronomy Survival Analysis (ASURV) package

Conclusions

Bivariate linear regression in astronomy can be surprisingly complex. Pay attention to precise question being asked, and details of situation. Several codes are available through http://astrostatistics.psu.edu/statcodes.

- **Functional vs. structural regression**
- **Symmetrical vs. dependent regression**
- **Weighting by measurement error**
- **Truncation & censoring due to flux limits**