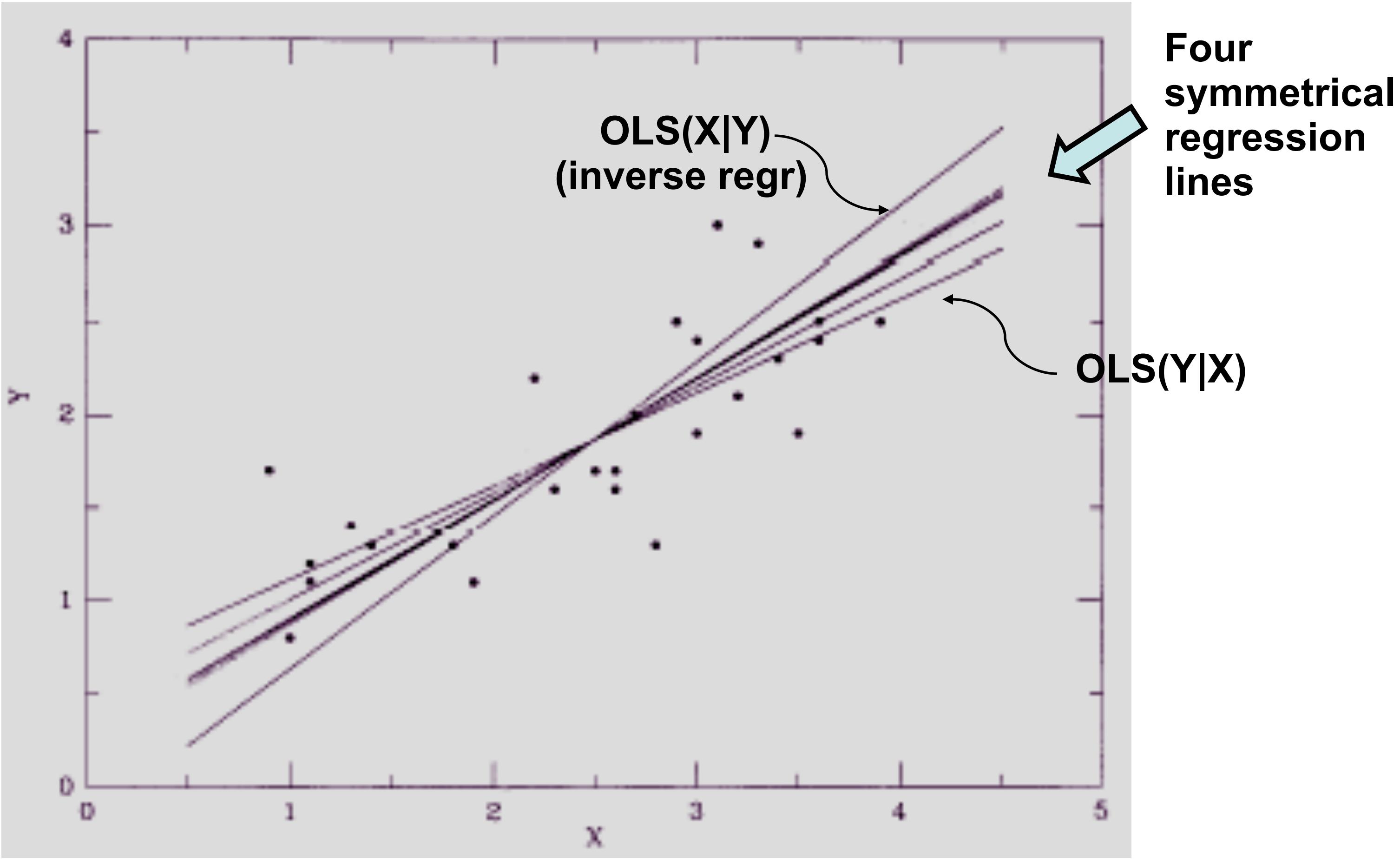
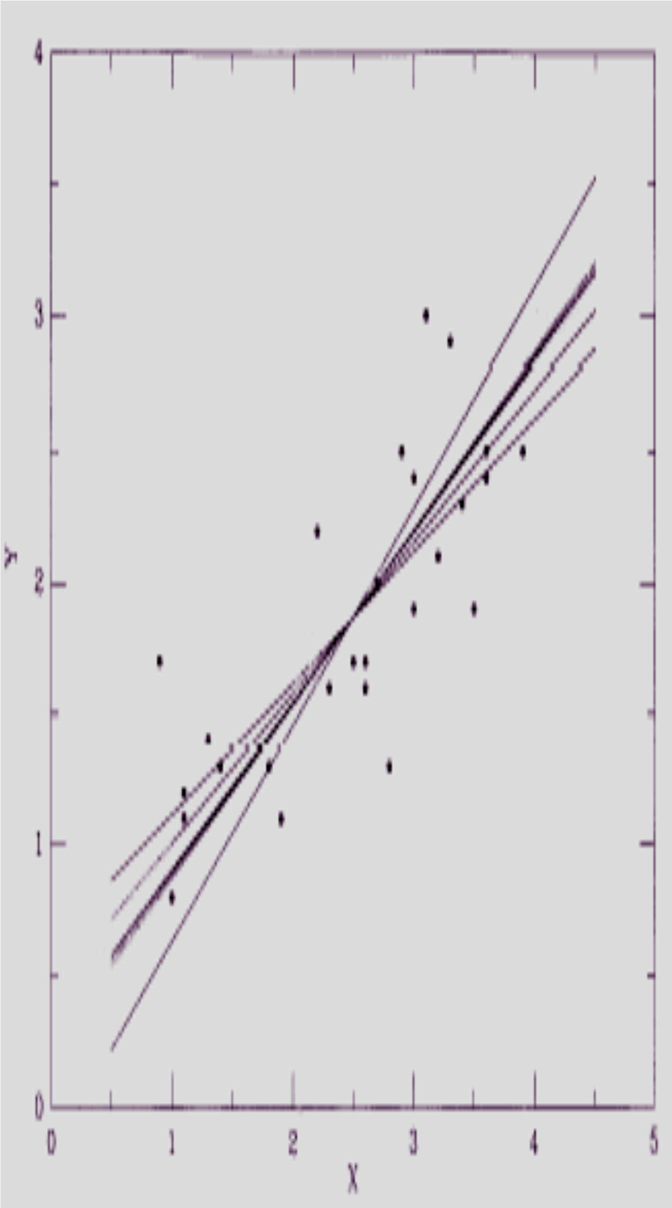
Linear regression issues in astronomy

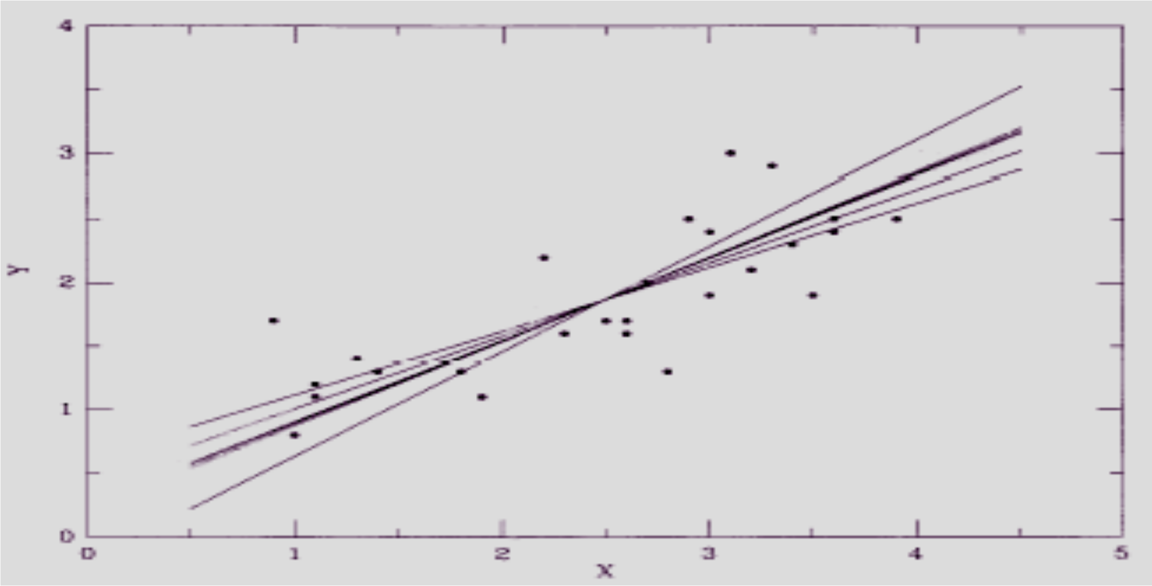
G. Jogesh Babu and Eric Feigelson
Center for Astrostatistics

Structural regression

Seeking the intrinsic relationship between two properties without specifying 'dependent' and 'independent' variables







Shrinking and Stretching

Slope depends on the units

Analytical formulae for slopes of the 6 OLS lines

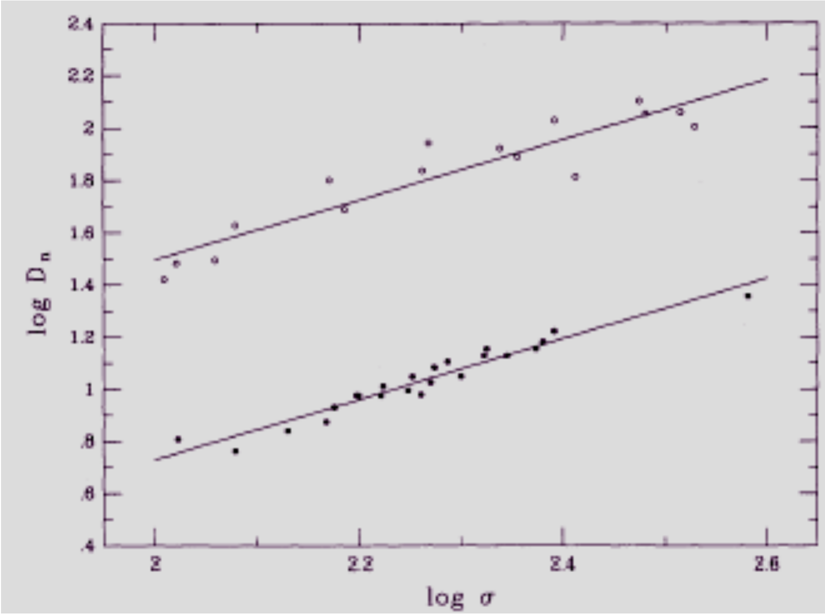
TABLE 1

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:

$$\widehat{\mathrm{Cov}}\,(\hat\beta_1,\hat\beta_2) = (\hat\beta_1 S_{xx}^2)^{-1}\left\{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})[y_i-\bar{y}-\hat\beta_1(x_i-\bar{x})][y_i-\bar{y}-\hat\beta_2(x_i-\bar{x})]\right\}.$$



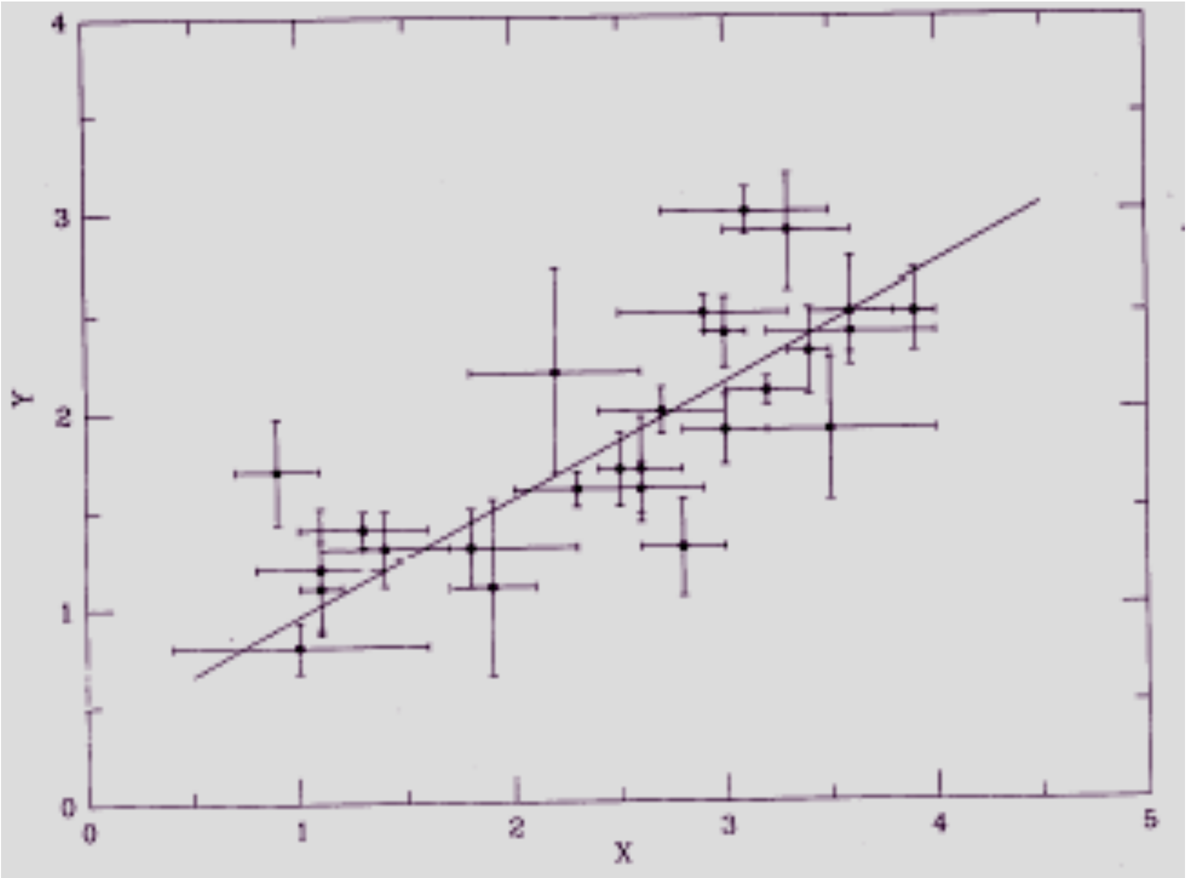
Example: Faber-Jackson relation between diameter and stellar velocity dispersion of elliptical galaxies

TABLE 4

REGRESSIONS FOR COMA AND VIRGO log D_n' VERSUS log σ[a]

METHOD (1)	ASYMPTOTIC FORMULAE Intercept (2)	ASYMPTOTIC FORMULAE 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

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)

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

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

BCES slope estimator

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

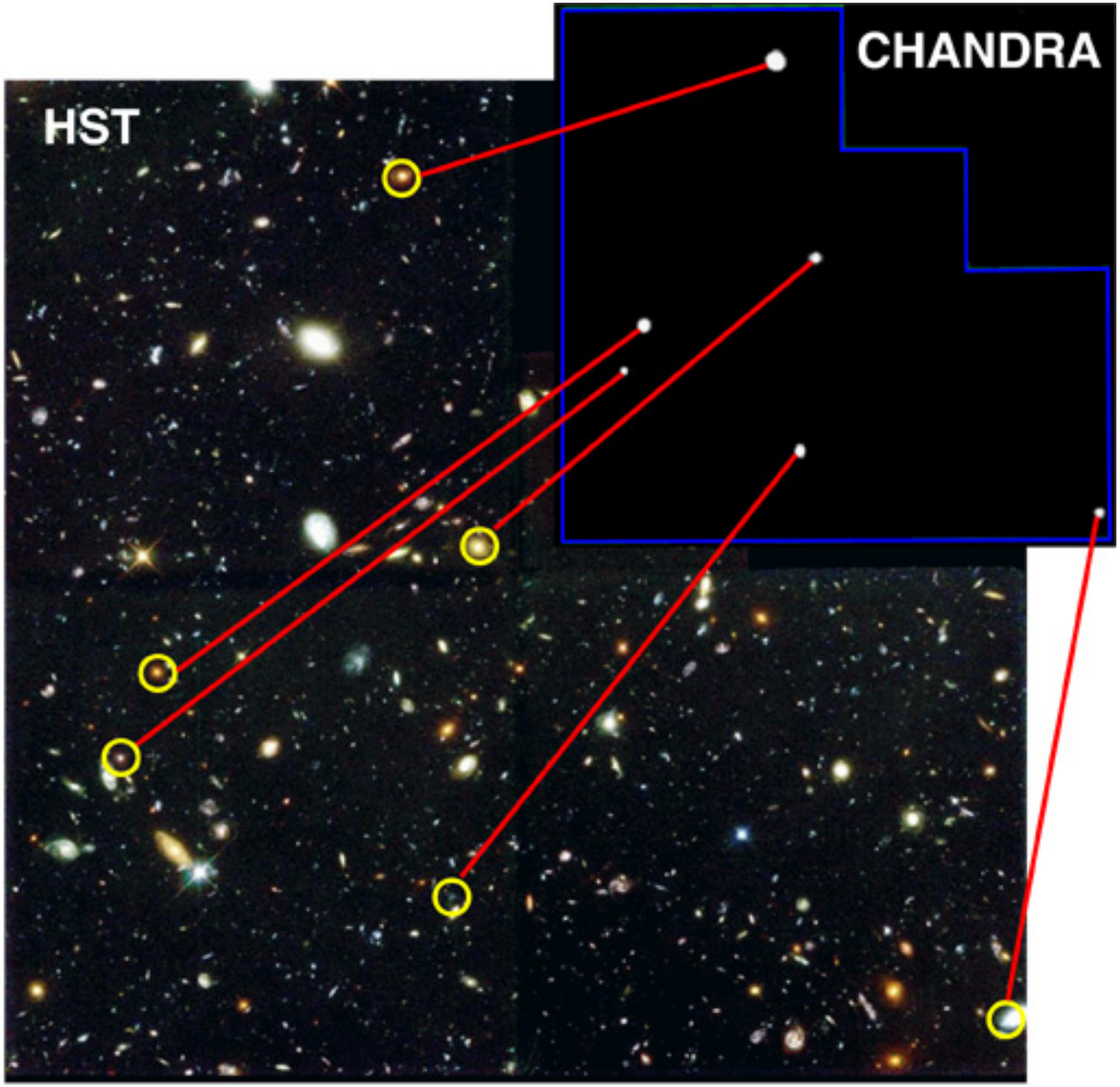
$$\hat{\alpha}_1 = \bar{Y}_2 - \beta_1 \bar{Y}_1 \, .$$

BCES slope variance

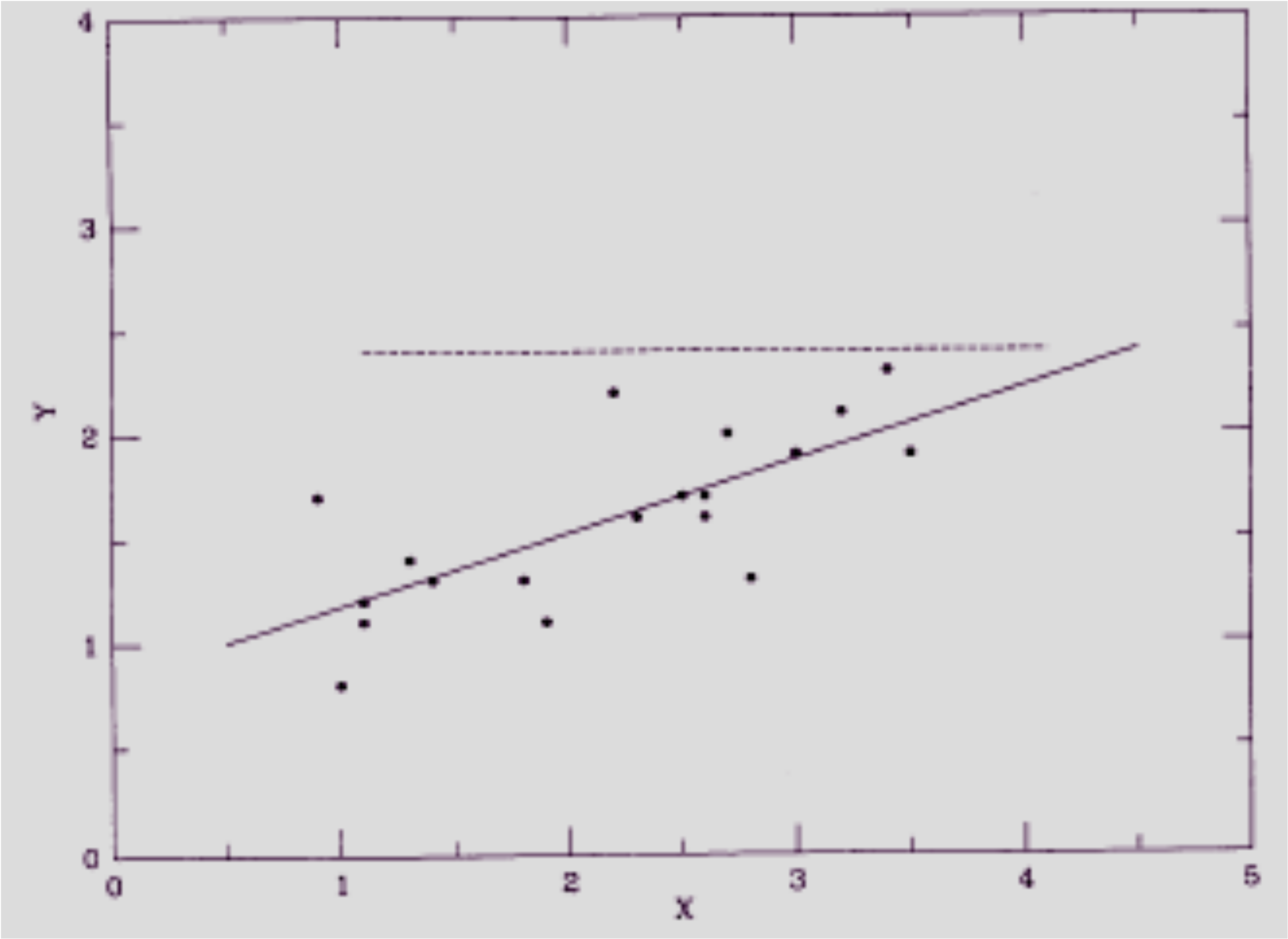
$$\hat{\sigma}^2_{\beta_1} = n^{-1} \sum_{i=1}^{n} (\xi_{1i} - \bar{\xi}_1)^2 \qquad \xi_{1i} = \frac{[Y_{1i} - E(Y_{1i})](Y_{2i} - \beta_1 Y_{1i} - \alpha_1) + \beta_1 V_{11,i} - V_{12,i}}{V(Y_{1i}) - E(V_{11,i})}$$

Akritas & Bershady, ApJ 470, 706 1996

X-ray sources in the HDFN



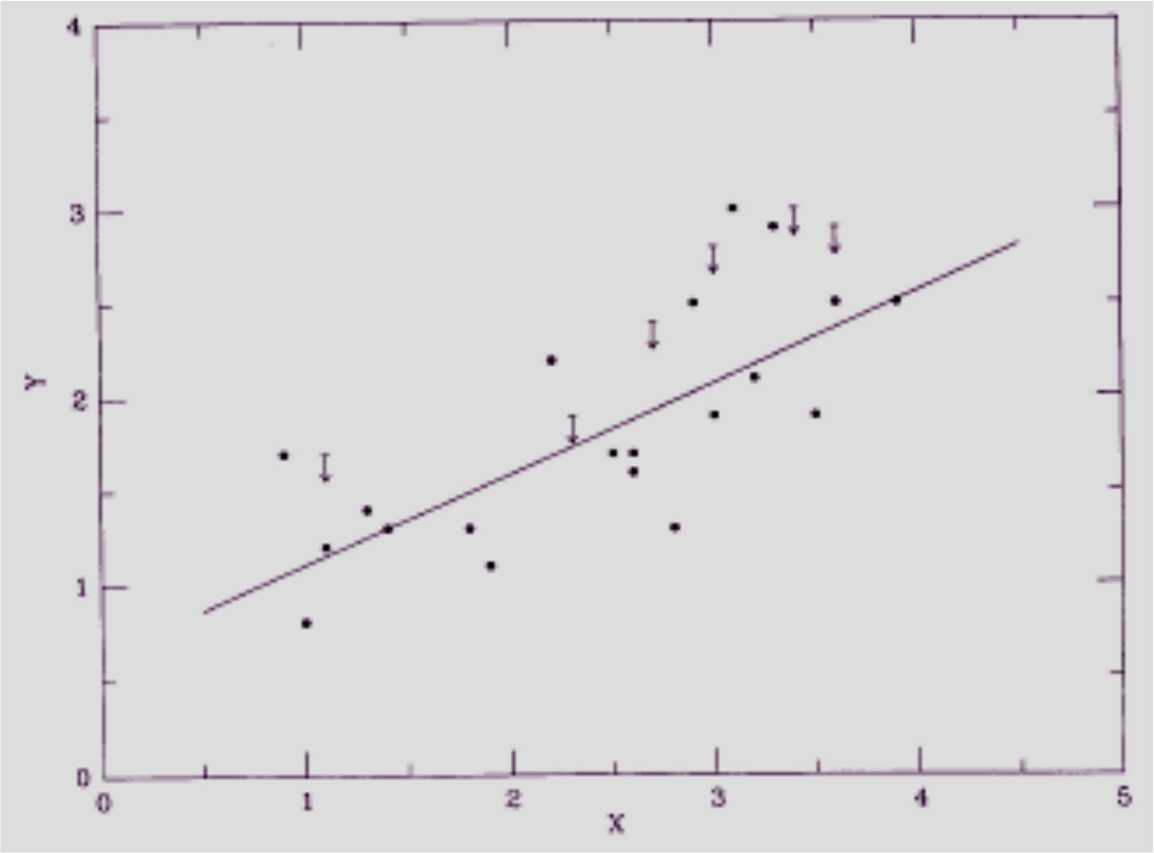
Truncation due to flux limits



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)

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

Bayesian Treatment of Measurement Errors in Linear Regression

- Errors-in-variables regression model (cf. monograph W. Fuller 1987)

  $\eta_i = \alpha + \beta\xi_i + \varepsilon_i$ (True relationship)
  $x_i = \xi_i + \varepsilon_{x,i}$ (True variables indirectly observed
  $y_i = \eta_i + \varepsilon_{y,i}$ with measurement errors)

- ξ_i is modeled as mixtures of Normals

Mixture of Normals Model

- Model the distribution of ξ as a mixture of K Gaussians, assume Gaussian intrinsic scatter and Gaussian measurement errors of known variance
- The model is hierarchically expressed as:

$$\xi_i \mid \pi, \mu, \tau^2 \sim \sum_{k=1}^{K} \pi_k N(\mu_k, \tau_k^2)$$

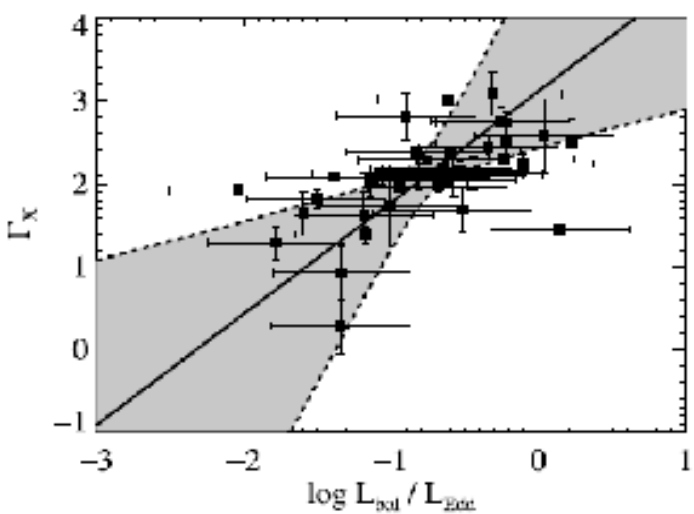
$$\eta_i \mid \xi_i, \alpha, \beta, \sigma^2 \sim N(\alpha + \beta \xi_i, \sigma^2)$$

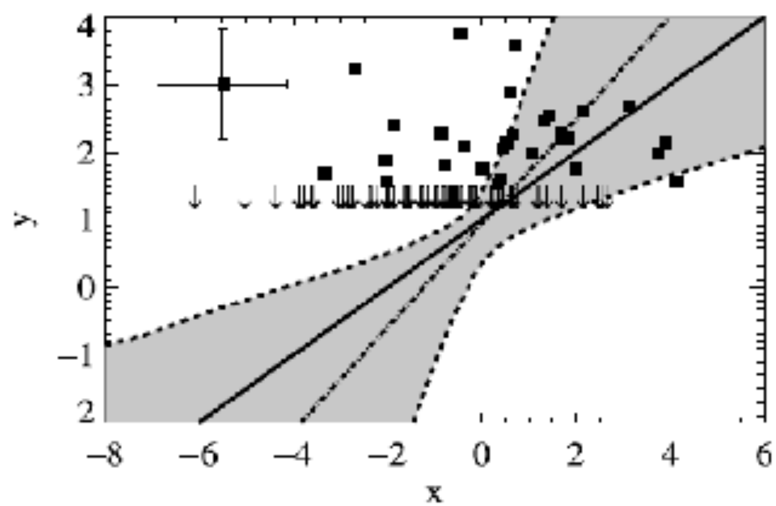
$$y_i, x_i \mid \eta_i, \xi_i \sim N([\eta_i, \xi_i], \Sigma_i)$$

$$\psi = (\pi, \mu, \tau^2), \quad \theta = (\alpha, \beta, \sigma^2), \quad \Sigma_i = \begin{pmatrix} \sigma_{y,i}^2 & \sigma_{xy,i} \\ \sigma_{xy,i} & \sigma_{x,i}^2 \end{pmatrix}$$

Prior distributions are assigned to the parameters ($\alpha,\beta,\sigma,\boldsymbol{\tau},\boldsymbol{\mu}$), Bayes' Theorem is applied, and posterior distributions are computed with Markov chain Monte Carlo techniques.

The method can be applied to censored and truncated regression problems, as well as measurement error problems. Performance is demonstrably better than earlier de-biased least-squares solutions (BCES, FITEXY). IDL code is available.





Conclusions

Bivariate linear regression in astronomy can be surprisingly complex. Pay attention to precise question being asked, and details of situation. Several codes 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

Other topics not considered here (some covered later in the Summer School):

- Robust & rank regression techniques to treat outliers
- Goodness-of-fit, model selection and parsimony
- Nonlinear regression
- Multivariate regression

References

- Isobe, Feigelson, Akritas & Babu, ApJ 364, 105, 1990
- Feigelson and Babu, ApJ 397, 55, 1992
- Brandon Kelly, ApJ 665, 1489, 2007
- W. Fuller 1987