

# Linear regression issues in astronomy

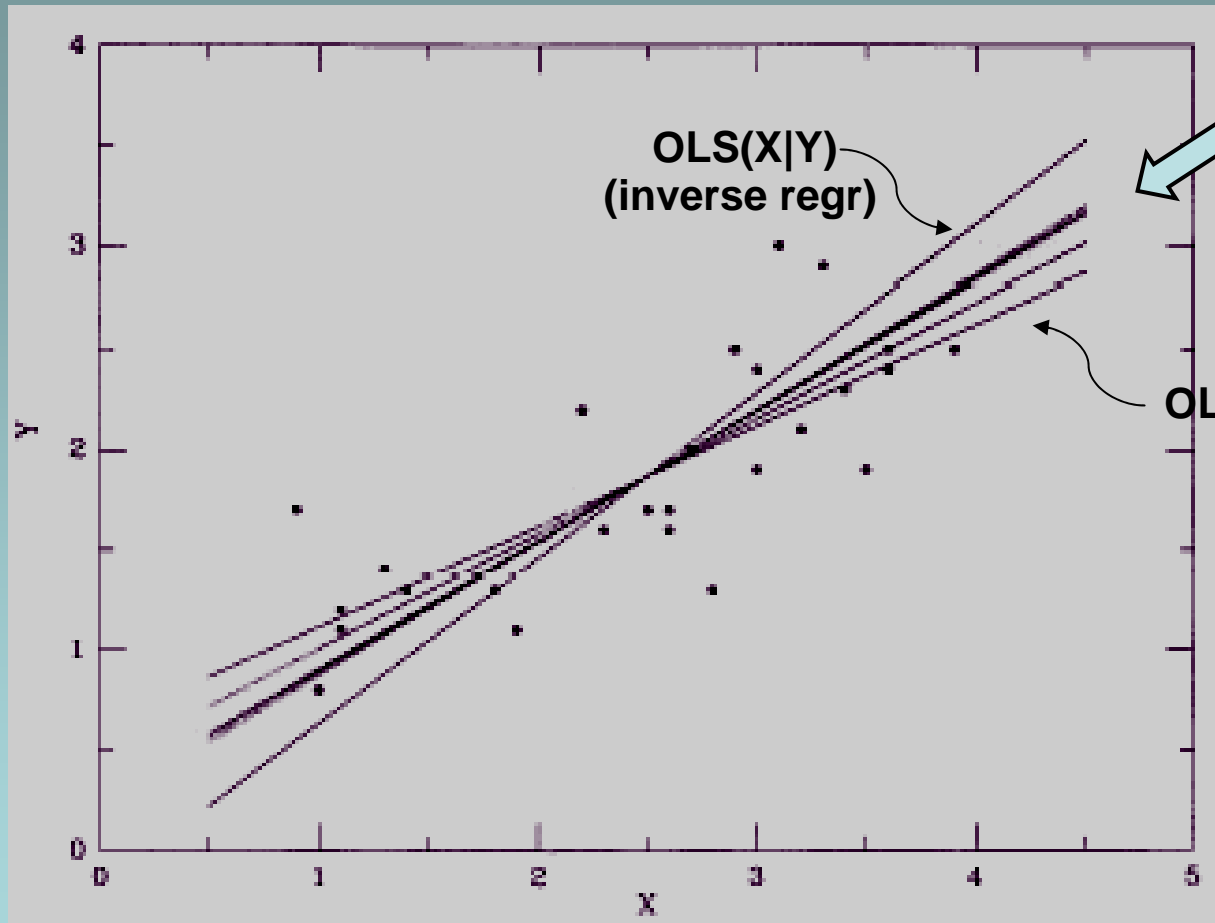
Eric Feigelson  
Summer School in astrostatistics

## References

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

# Structural regression

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



Four  
symmetrical  
regression  
lines

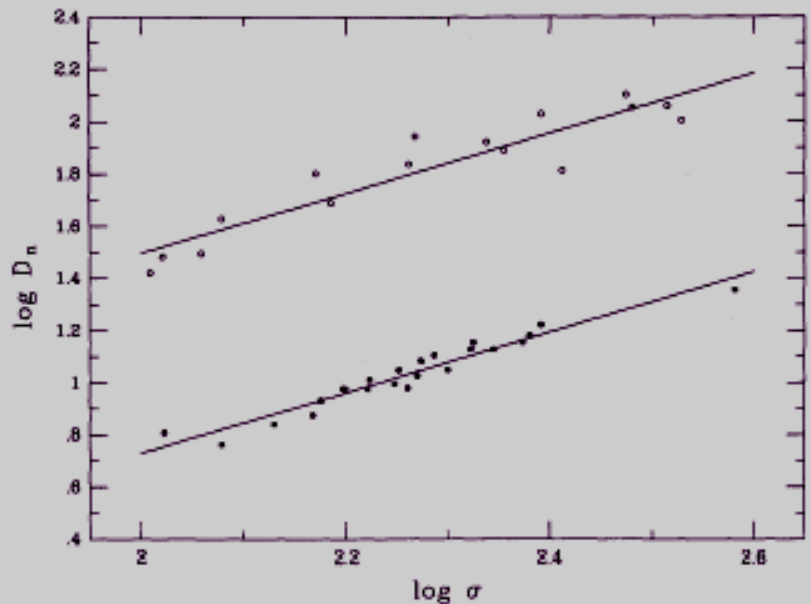
# Analytical formulae for slopes of the 6 OLS lines

TABLE 1  
LINEAR REGRESSION FORMULAE FOR SLOPES

Method	Expression for Slope	Estimate of the Variance of the Slope $\widehat{\text{Var}}(\beta_i)$
OLS(X Y) .....	$\beta_1 = \frac{S_{xy}}{S_{xx}}$	$\frac{1}{S_{xx}^2} \left[ \sum_{i=1}^n (x_i - \bar{x})^2 (y_i - \beta_1 x_i - \bar{y} + \beta_1 \bar{x})^2 \right]$
OLS(Y X) .....	$\beta_2 = \frac{S_{yy}}{S_{xy}}$	$\frac{1}{S_{xy}^2} \left[ \sum_{i=1}^n (y_i - \bar{y})^2 (y_i - \beta_2 x_i - \bar{y} + \beta_2 \bar{x})^2 \right]$
OLS bisector .....	$\beta_3 = (\beta_1 + \beta_2)^{-1} [\beta_1 \beta_2 - 1 + \sqrt{(1 + \beta_1^2)(1 + \beta_2^2)}]$	$\frac{\beta_3^2}{(\beta_1 + \beta_2)^2 (1 + \beta_1^2)(1 + \beta_2^2)} [(1 + \beta_2^2)^2 \widehat{\text{Var}}(\beta_1) + 2(1 + \beta_1^2)(1 + \beta_2^2) \widehat{\text{Cov}}(\beta_1, \beta_2) + (1 + \beta_1^2)^2 \widehat{\text{Var}}(\beta_2)]$
Orthogonal regression .....	$\beta_4 = \frac{1}{2} [(\beta_2 - \beta_1^{-1}) + \text{Sign}(S_{xy}) \sqrt{4 + (\beta_2 - \beta_1^{-1})^2}]$	$\frac{\beta_4^2}{4\beta_1^2 + (\beta_1 \beta_2 - 1)^2} [\beta_1^{-2} \widehat{\text{Var}}(\beta_1) + 2 \widehat{\text{Cov}}(\beta_1, \beta_2) + \beta_1^2 \widehat{\text{Var}}(\beta_2)]$
Reduced major-axis .....	$\beta_5 = \text{Sign}(S_{xy}) (\beta_1 \beta_2)^{1/2}$	$\frac{1}{4} \left[ \frac{\beta_2}{\beta_1} \widehat{\text{Var}}(\beta_1) + 2 \widehat{\text{Cov}}(\beta_1, \beta_2) + \frac{\beta_1}{\beta_2} \widehat{\text{Var}}(\beta_2) \right]$

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

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

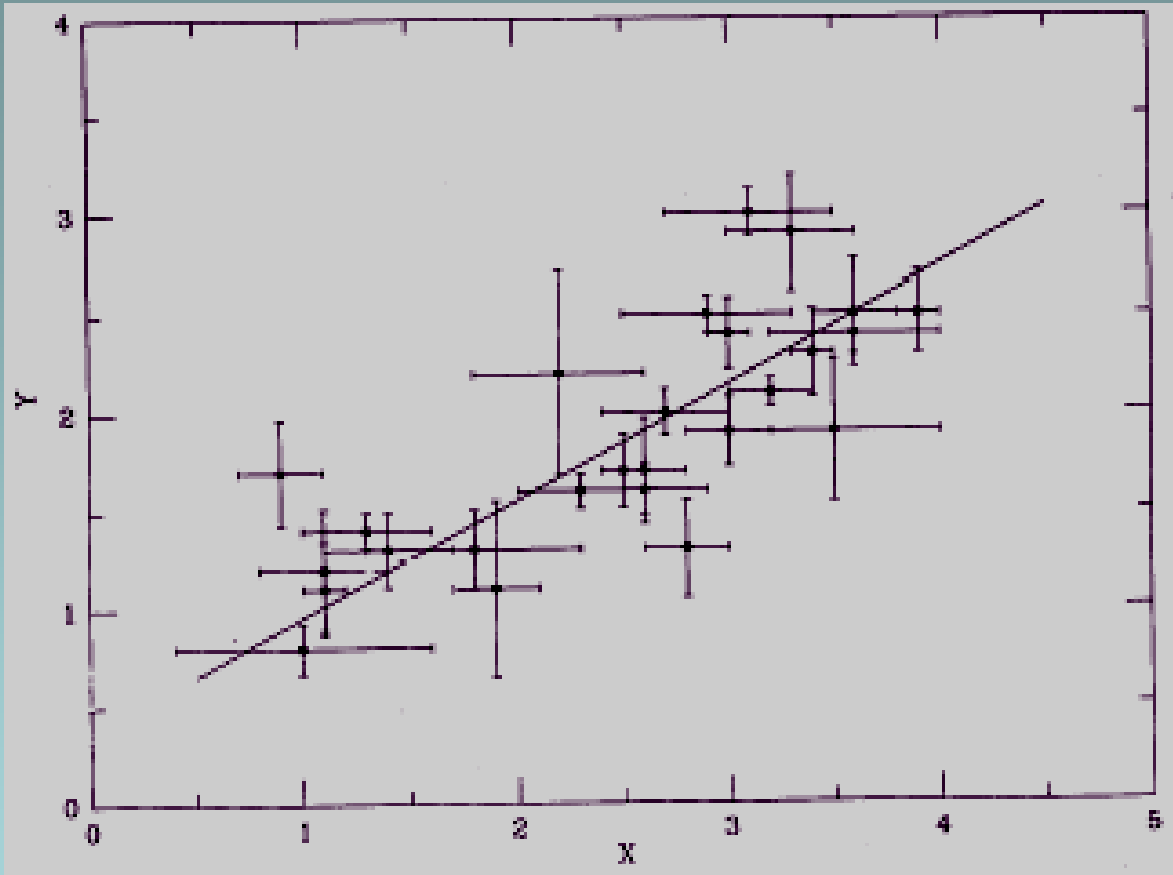


**Example: Faber-Jackson relation between diameter and stellar velocity dispersion of elliptical galaxies**

**TABLE 4**  
REGRESSIONS FOR COMA AND VIRGO  $\log D_n$  VERSUS  $\log \sigma^*$

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

# 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 & Bershadsky  
ApJ 1996)

# Regression with measurement errors and intrinsic scatter

$Y$  = observed data

$V$  = measurement errors

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

$X$  = 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

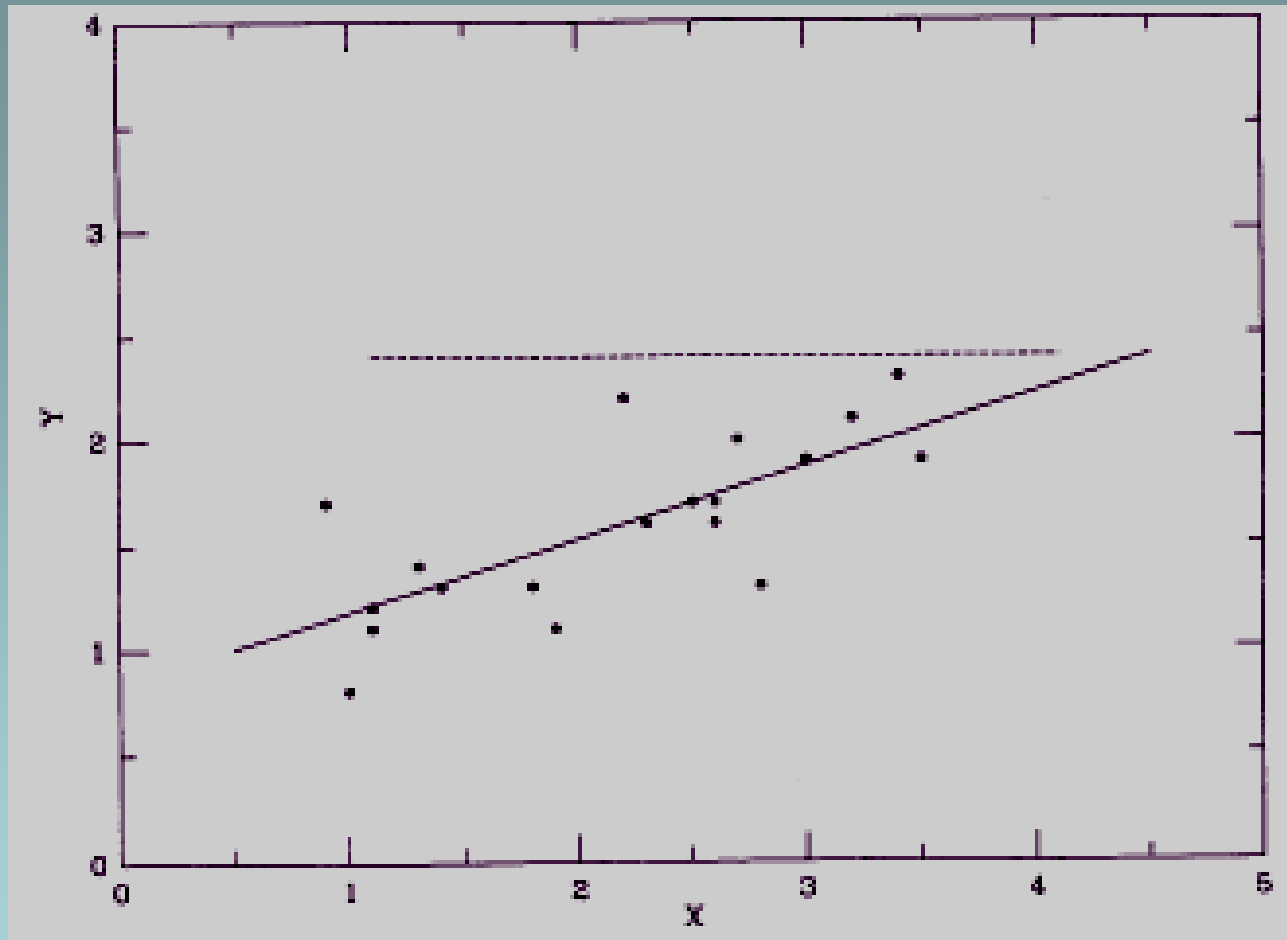
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2) - \sum_{i=1}^n V_{12,i}}{\sum_{i=1}^n (Y_{1i} - \bar{Y}_1)^2 - \sum_{i=1}^n V_{11,i}}$$
$$\hat{\alpha}_1 = \bar{Y}_2 - \beta_1 \bar{Y}_1.$$

BCES slope variance

$$\hat{\sigma}_{\beta_1}^2 = n^{-1} \sum_{i=1}^n (\xi_{1i} - \bar{\xi}_1)^2$$

$$\xi_{1i} = \frac{[Y_{1i} - E(Y_{1i})](Y_{2i} - \beta_1 Y_{1i} - \alpha_1) + \beta_1 V_{11,i} - V_{12,i}}{V(Y_{1i}) - E(V_{11,i})}$$

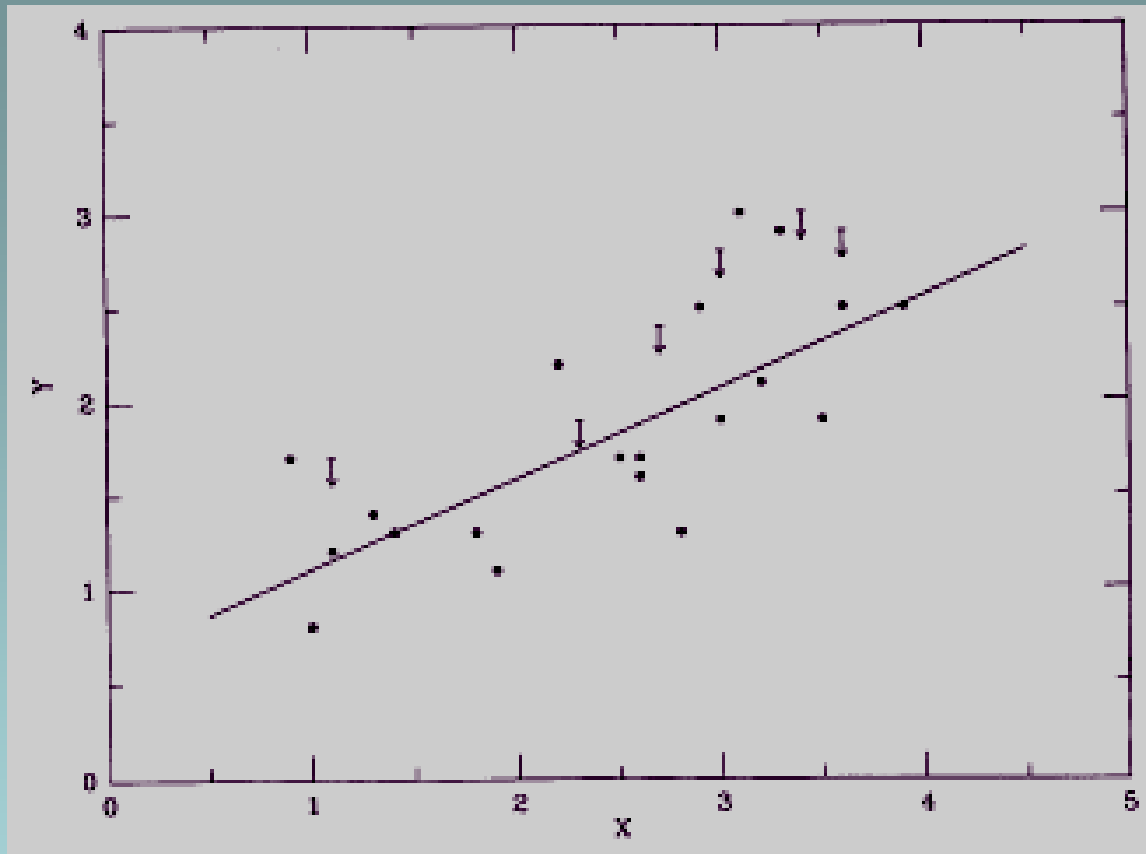
# Truncation due to flux limits



**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  $\tau$  (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



# Mixture of Normals Model

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

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

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

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

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

References: Kelly (2007, ApJ in press, arXiv:705.2774),  
Carroll et al.(1999, Biometrics, 55, 44)

Integrate complete data likelihood to obtain observed data likelihood:

$$\begin{aligned}
 p(x, y \mid \theta, \psi) &= \prod_{i=1}^n \iint p(x_i, y_i \mid \xi_i, \eta_i) p(\eta_i \mid \xi_i, \theta) p(\xi_i \mid \psi) d\xi_i d\eta_i \\
 &= \prod_{i=1}^n \sum_{k=1}^K \frac{\pi_k}{2\pi |V_{k,i}|^{1/2}} \exp\left\{-\frac{1}{2} (\mathbf{z}_i - \boldsymbol{\zeta}_k)^T V_{k,i}^{-1} (\mathbf{z}_i - \boldsymbol{\zeta}_k)\right\}
 \end{aligned}$$

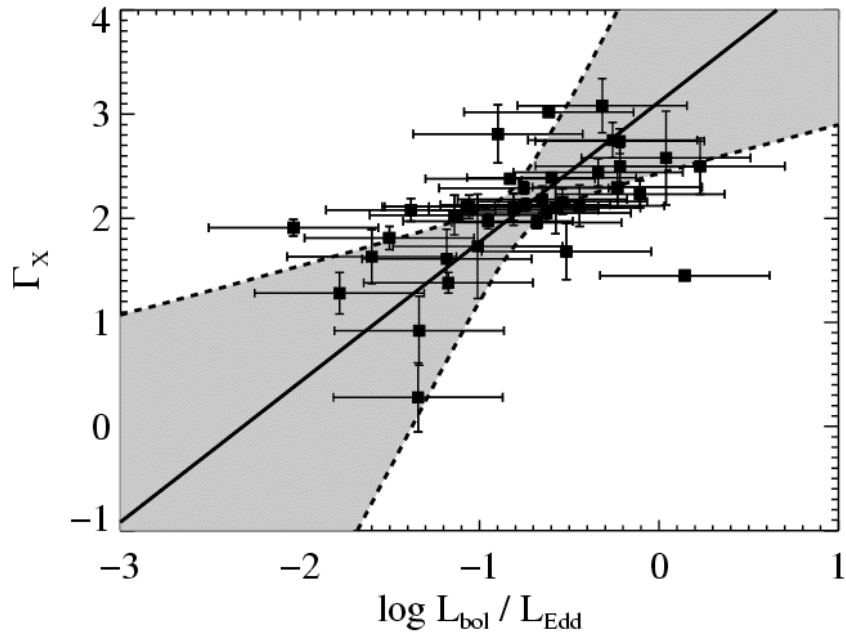
$$\mathbf{z}_i = (y_i \quad x_i)^T$$

$$\boldsymbol{\zeta}_k = (\alpha + \beta\mu_k \quad \mu_k)^T$$

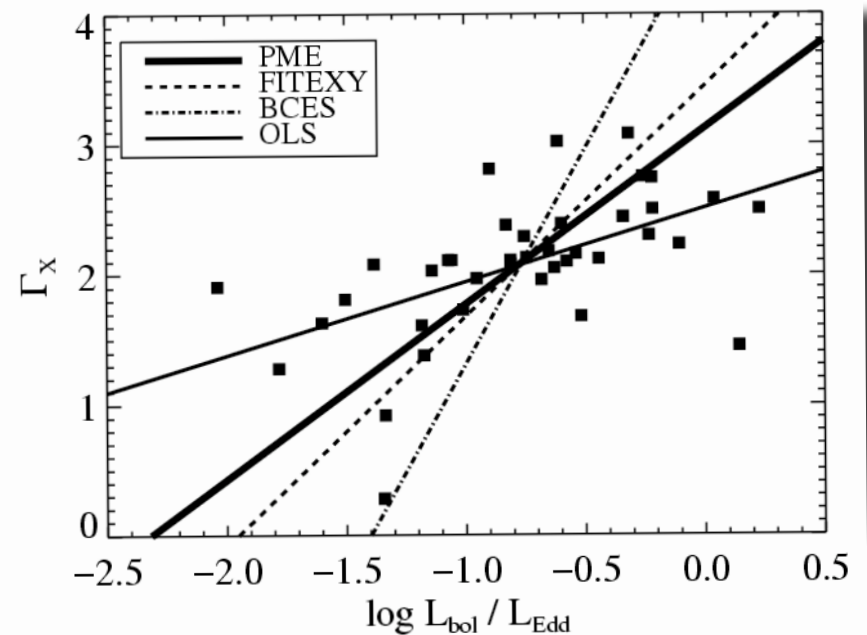
$$V_{k,i} = \begin{pmatrix} \beta^2 \tau_k^2 + \sigma^2 + \sigma_{y,i}^2 & \beta \tau_k^2 + \sigma_{xy,i} \\ \beta \tau_k^2 + \sigma_{xy,i} & \tau_k^2 + \sigma_{x,i}^2 \end{pmatrix}$$

Can be used to calculate a maximum-likelihood estimate (MLE), perform Bayesian inference. See Kelly (2007) for generalization to multiple covariates.

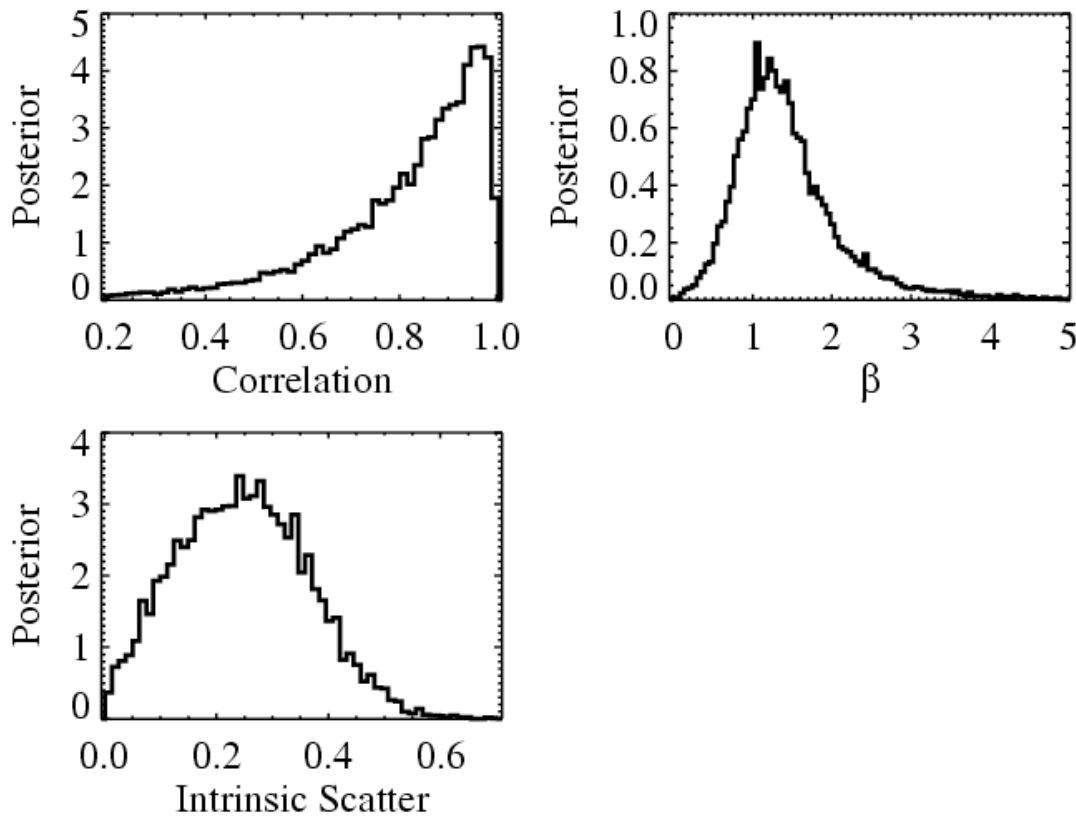
# Example: Dependence of Quasar X-ray Spectral Slope on Eddington Ratio



Solid line is posterior median,  
Shaded region contains 95%  
Of posterior probability.



# Posterior for Quasar Spectral Slope vs Eddington Ratio



For Comparison:

$$\hat{\beta}_{OLS} = 0.56 \pm 0.14$$

$$\hat{\beta}_{BCES} = 3.29 \pm 3.34$$

$$\hat{\beta}_{EXY} = 1.76 \pm 0.49$$

$$\hat{\sigma}_{OLS} = 0.41$$

$$\hat{\sigma}_{BCES} = 0.32$$

$$\hat{\sigma}_{EXY} = 0.0$$

## 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