Linear regression

Eric Feigelson

3rd INPE Advanced School in Astrophysics:
Astrostatistics 2009

Outline

1. Classical frequentist linear regression
2. Complications: Symmetry, censoring, truncation & measurement errors
3. Bayesian approaches

Concepts of regression

- Functional regression: Choose the dependent (response) variable. Model assumes points like exactly on the curve.
- Structural regression: Model accepts scatter about the curve.
- Choice of parametric model: Linear, polynomial or other simple family, complex model from astrophysical theory
- Assumption of i.i.d. (independently and identically distributed data points). Not consistent with different measurement errors for each point (heteroscedasticity), mixture of populations or outliers.
- Assumption of asymptotic normality. Generally valid via the Central Limit Theorem.

Functional linear regression model

The data: (x_i, y_i) for $i = 1, \ldots, N$, random i.i.d. sample of underlying population. x_i values are fixed. x can be a single variable or a vector of j variables x_{ij}.

The statistical model:

$$y = E(Y|X = x) = \alpha \ + \ \beta x \ + \ \epsilon$$

$$\epsilon = N(0, \sigma^2)$$

"The expected value of the dependent variable y given that the independent variable (or vector) x has value X is equal to the slope β times x plus the intercept α and a Gaussian noise value ϵ with zero mean and variance σ^2."

Ordinary least-squares linear regression

The least-squares method chooses estimates $\hat{\alpha}$ and $\hat{\beta}$ which minimized the sum of squared deviations

$$SS = \sum_{i}^{N} [y_i - (\hat{\alpha} + \hat{\beta} x_i)]^2$$

Setting $\delta SS/\delta\alpha = 0$ and $\delta SS/\delta\beta = 0$, the solution is

$$\hat{\beta} = \frac{N \sum x_i y_i - (\sum x_i)(\sum y_i)}{N \sum x_i^2 - (\sum x_i)^2}$$

$$\hat{\alpha} = \bar{y_i} - \hat{\beta} \bar{x_i}$$

$$Var(\hat{\beta}) = \frac{N\sigma^2}{N \sum x_i^2 - (\sum x_i)^2}$$

and similarly for $Var(\hat{\alpha})$ and $Cov(\hat{\alpha}, \hat{\beta})$. The estimators are unbiased and consistent if errors are uncorrelated with the regressor, $E[x_i \epsilon_i] = 0$, and are efficient if the errors are homoscedastic, $E[\epsilon^2 | x_i]$ are independent of i.

The OLS line is often the maximum-likelihood line

The probability density function (p.d.f.) of the i-th observation f_i is

$$f_i(y_i; x_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp -(y_i - \alpha - \beta x_i)^2/2\sigma^2$$

The log-likelihood $-lnL$ at the data locations $(x_i, y_i); i = 1, \ldots, N$ is

$$-lnL = \frac{N}{2} + \frac{N}{2}\log(2\pi) + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \alpha - \beta x_i)^2$$

If σ is known, the maximum-likelihood estimator (MLE) of α and β is the same as the ordinary least squares estimator obtained by minimizing $\sum(y_i - \alpha - \beta x_i)^2$.

Structural regression

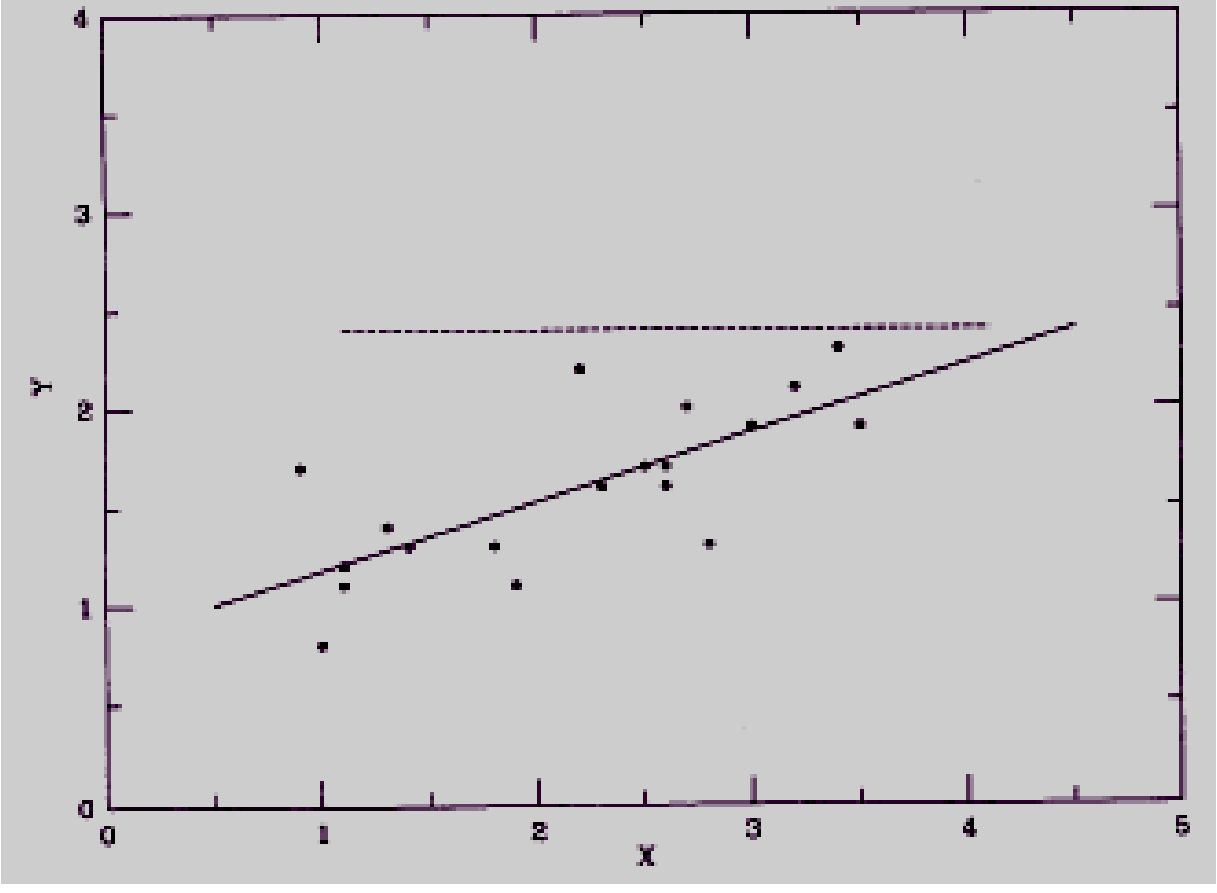
Seeking the intrinsic relationship between two properties without specifying dependent and independent variables. While functional regression is often used in human affairs to make decisions, structural regression is often needed in astronomy to understand the relationship between physical properties of planets, stars, galaxies, etc.

Generalizations and alternatives to OLS

- Symmetrical regression (Total LS, Principal Component Analysis)
- Robust regression (least absolute deviation, quantile, M-estimators, α-trimmed estimators, ...)
- Generalizations when errors are correlated with the data or the regressors (GLS, IRLS, instrumental variables, errors-in-variables, see below, ...)
- Collinearity (ridge regression, lasso and other penalized likelihoods)
- High dimensional problems (PCA, least angle regression)
- Censored regression (Buckley/James, Schmitt)
- Truncated regression (Tobit, Deeming, Teerikorpi)

Truncated regression

Truncation due to flux limits in astronomical surveys



Univariate distribution: Maximum-likelihood distribution function estimator (Lynden-Bell 1972; Woodroofe 1985).

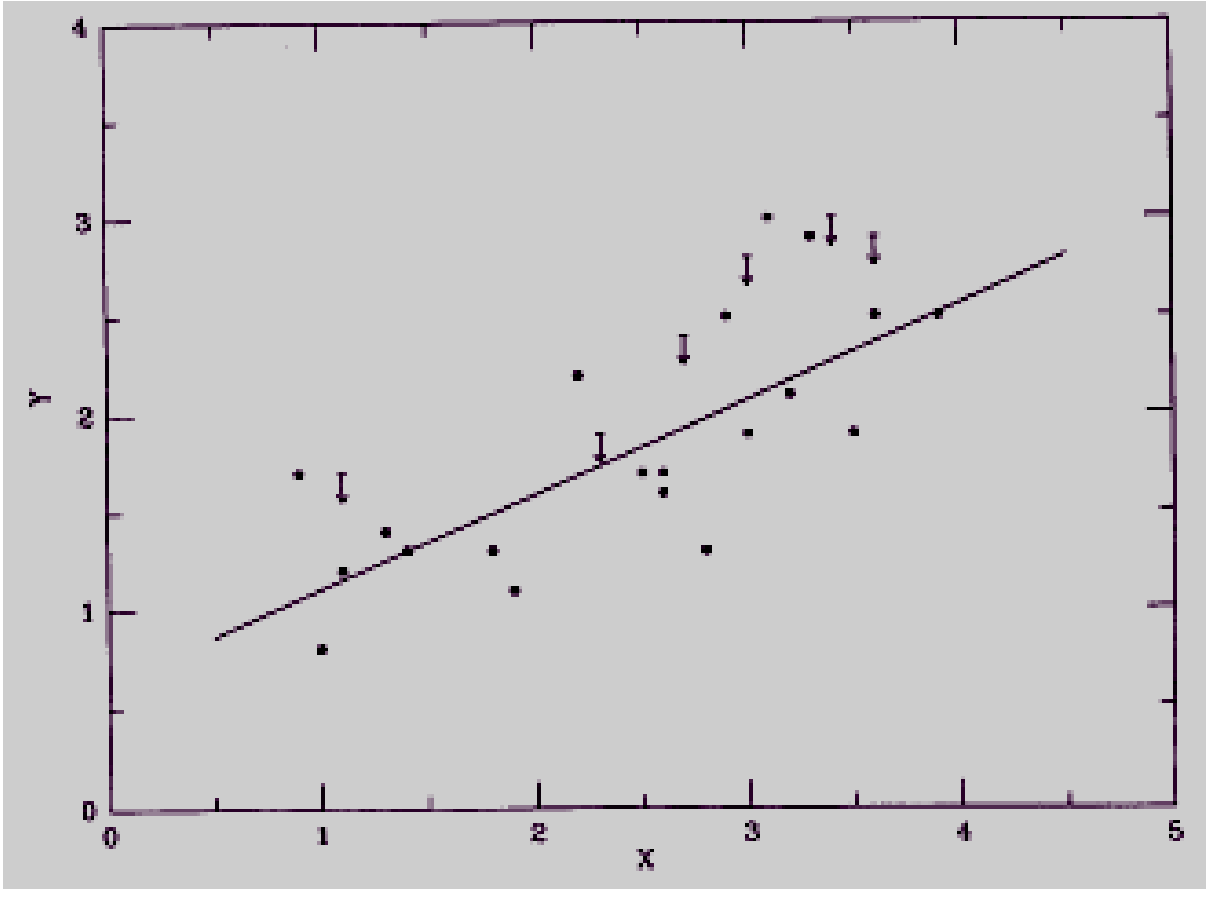
Correlation coefficient: Efron & Petrosian (1992)

OLS linear regression: Tobit and LIMDEP regression in econometrics; Malmquist bias correction in astronomy (Deeming 1968; Segal 1975; Teerikorpi 1984)

Not much development of statistical techniques for truncation, as information on the distribution of faint objects is missing

Censored regression

Censoring (or upper limits) due to non-detections of previously identified sample. Extensive field of 'survival analysis' treating censoring.



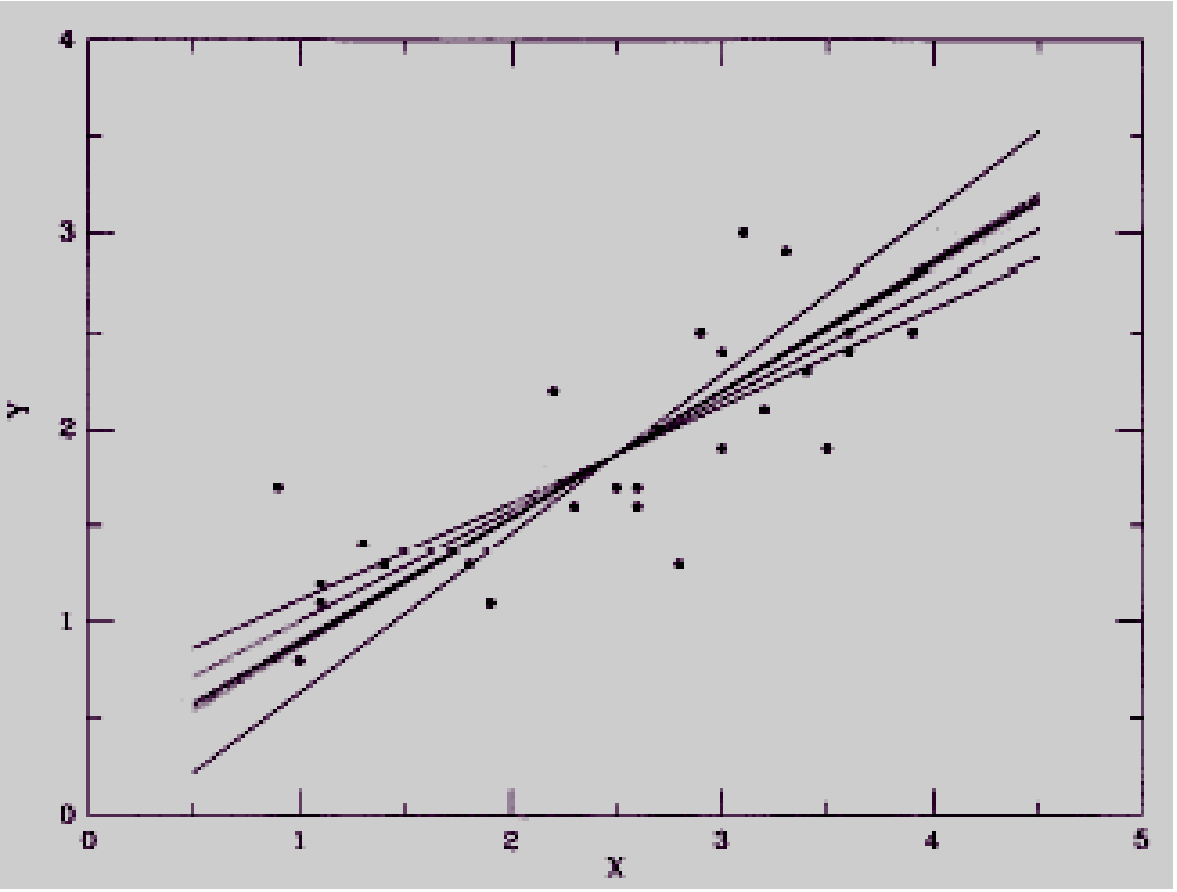
Univariate distribution Kaplan-Meier maximum likelihood distribution function estimator

2-sample tests: Gehan, logrank, Peto-Prentice

Correlation coefficients: Generalized Kendall's τ
Linear regression: Cox regression, EM algorithm with normal residuals, Buckley-James line with Kaplan-Meier residuals

Survival methods for astronomy implemented in ASURV package (Feigelson & Nelson ApJ 1986, http://astrostatistics.psu.edu/statcodes)

Structural OLS regression



Astronomers use up to six OLS lines: Standard OLS($Y|X$), the inverse line OLS($X|Y$), and four symmetrical lines. Orthogonal regression (Pearson 1901, first principal component), reduced major axis, and OLS bisector.

TABLE 1

LINEAR REGRESSION FORMULAE FOR SLOPES

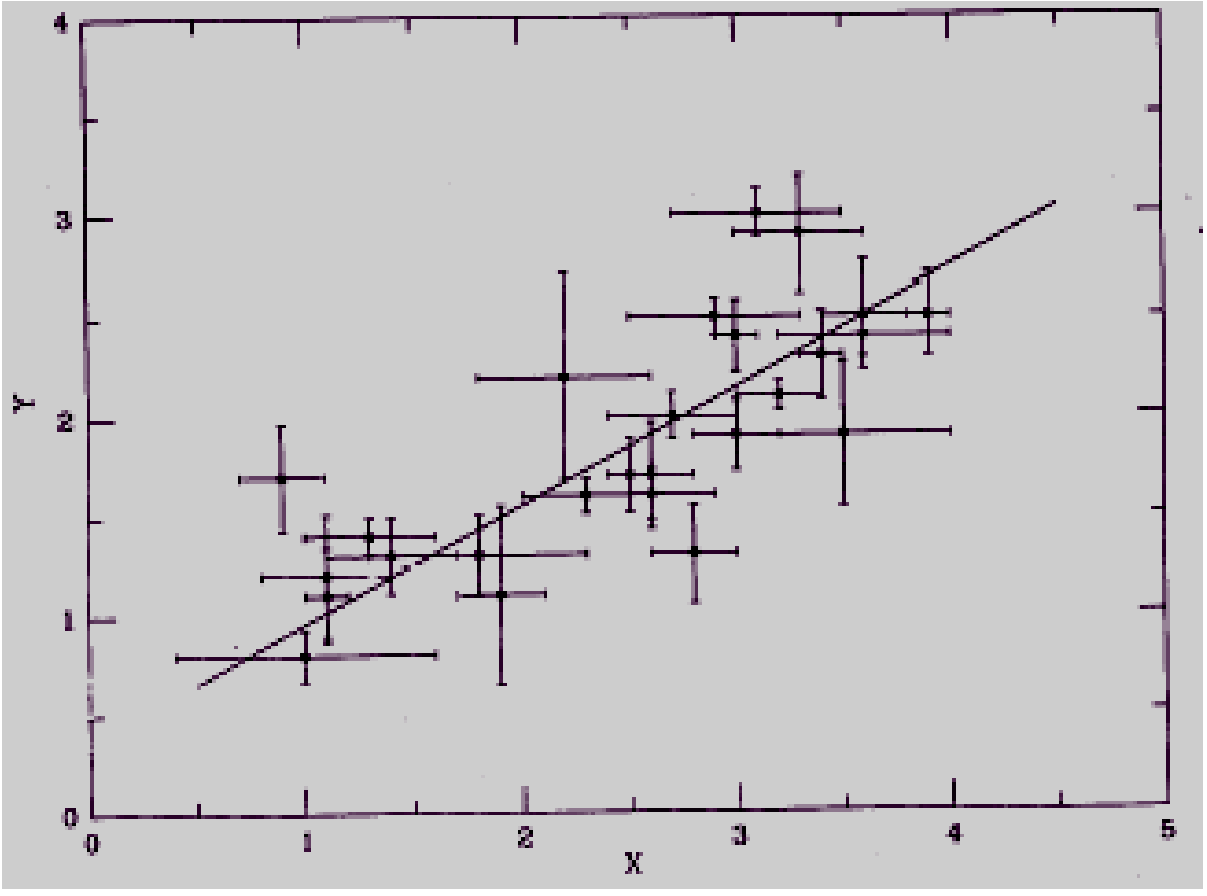
Method	Expression for Slope	Estimate of the Variance of the Slope $\widehat{\mathrm{Var}}\,(\hat{\beta}_i)$
OLS($X \mid Y$)	$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$	$\frac{1}{S_{xx}^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2(y_i-\hat{\beta}_1 x_i-\bar{y}+\hat{\beta}_1\bar{x})^2\right]$
OLS($Y \mid X$)	$\hat{\beta}_2 = \frac{S_{yy}}{S_{xy}}$	$\frac{1}{S_{xy}^2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2(y_i-\hat{\beta}_2 x_i-\bar{y}+\hat{\beta}_2\bar{x})^2\right]$
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:

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

These lines are not the same function and should not be used interchangeably. Both the values and the variances of the regression parameters are mathematically different. The scientific question is often too vague to specify a single method.

Regression with measurement errors



Homoscedastic functional model Deeming (Vistas Astro 1968), Fuller *Measurement Error Models* (1987), Carroll/Rupert/Stefanski *Measurement Errors in Non-Linear Models* (2nd ed, 2006)

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

Heteroscedastic structural model Press et al. (*Numerical Recipes* 1992), Akritas/Bershady (ApJ 1996), Tremaine et al. (ApJ 2002), Kelly (ApJ 2007)

An OLS regression with errors and scatter

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

(Y_1, Y_2) are the observed data, V_{12} are the measurement errors

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

(X_1, X_2) are the intrinsic variables, ϵ_{12} is the intrinsic scatter

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

Regression estimators & slope variance:

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

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

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

Akritas & Bershady, ApJ 1996

Bayesian approach of Zellner and Jaynes I

"Fitting the 'best' straight line to a scatter plot of data ... is undoubtedly the most common problem of inference faced by scientists ... But from the view point of orthodox statistics the problem turned out to be a horrendous can of worms; generations of efforts led only to a long list of false starts, and no satisfactory solution." (Jaynes 1990)

Model: $Y = \alpha + \beta X$ where the physical variables are indirectly measured through data (x_i, y_i) with (unknown) measurement errors: $x_i = X_i + e_i$, $y_i = Y_i + f_i$. For prior information I, Bayes' Theorem gives:

$$p(\alpha, \beta | x, y, I) = p(\alpha, \beta | I) \frac{p(x, y | \alpha, \beta, i)}{p(x, y | I)}$$

Bayesian approach of Zellner and Jaynes II

For the special case of Gaussian errors and no prior information, Bayes' Theorem gives the ordinary least squares solution: $\hat{\beta} = (\bar{xy} - \bar{x}\bar{y})/(\bar{x^2} - \bar{x}^2)$ and $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

Other solutions arise for different priors. For instance, for Gaussian errors, uniform priors on the X_i and Jeffreys priors to σ_x and σ_y, the posterior marginal p.d.f. for β is

$$p(\beta, x, y, I) \propto [\gamma^2 + (\beta - \hat{\beta})^2]^{-(n-1)/2)}$$

where $\gamma^2 = (\bar{y^2} - \bar{y}^2)(1 - r^2)/(\bar{x^2} - \bar{x}^2)$ and $r^2 = (\bar{xy} - \bar{x}\bar{y}/(\bar{x^2} - \bar{x}^2)(\bar{y^2} - \bar{y}^2)$. This is the least-squares solution for $e_i = 0$ (no errors in the x variable) and f_i errors are unknown.

The integrated Bayesian approach of Kelly

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:

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

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

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

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

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

Slides courtesy Brandon Kelly

Integrate complete data likelihood to obtain observed data likelihood:

$$p(x,y\,|\,\theta,\psi) = \prod_{i=1}^{n} \iint p(x_i,y_i\,|\,\xi_i,\eta_i)\,p(\eta_i\,|\,\xi_i,\theta)\,p(\xi_i\,|\,\psi)\,d\xi_i\,d\eta_i$$

$$= \prod_{i=1}^{n} \sum_{k=1}^{K} \frac{\pi_k}{2\pi\left|V_{k,i}\right|^{1/2}} \exp\left\{-\tfrac{1}{2}(\mathbf{z}_i-\zeta_k)^T V_{k,i}^{-1}(\mathbf{z}_i-\zeta_k)\right\}$$

$$\mathbf{z}_i = \begin{pmatrix} y_i & x_i \end{pmatrix}^T$$

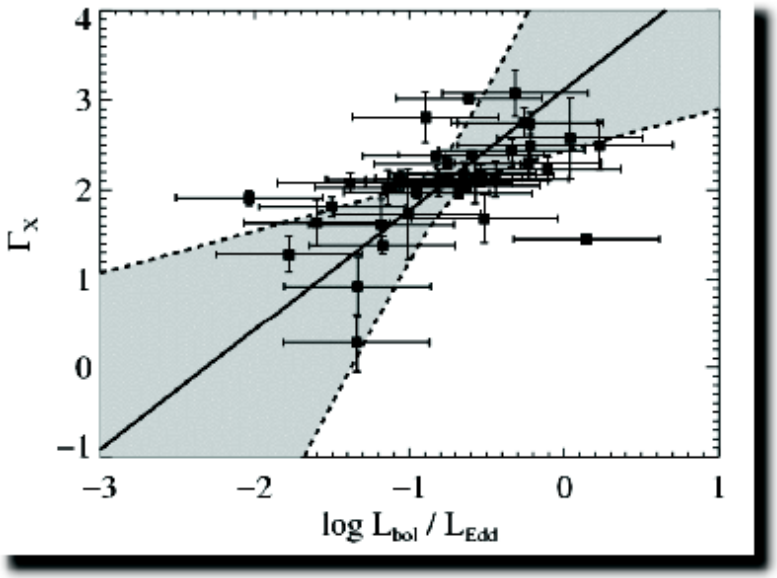
$$\zeta_k = \begin{pmatrix} \alpha+\beta\mu_k & \mu_k \end{pmatrix}^T$$

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

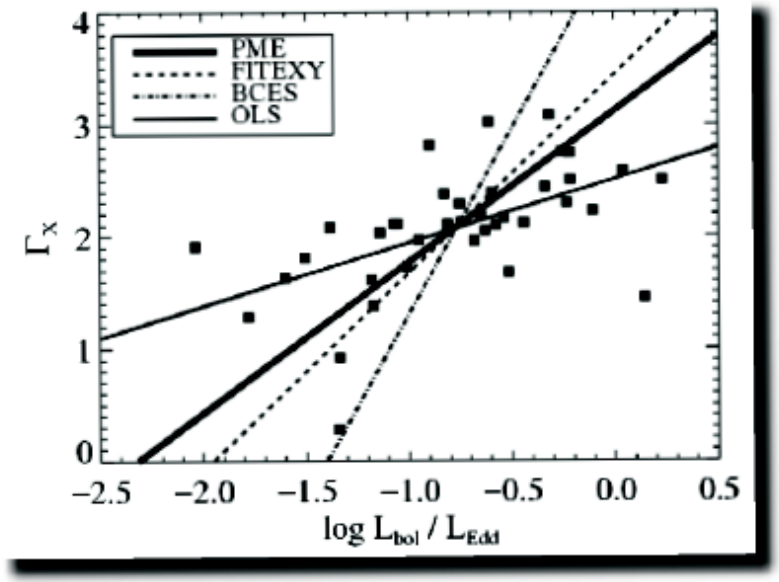
Can be used to calculate a maximum-likelihood estimate (MLE), perform Bayesian inference. See Kelly (2007) for generalization to multiple covariates.

Slides courtesy Brandon Kelly

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

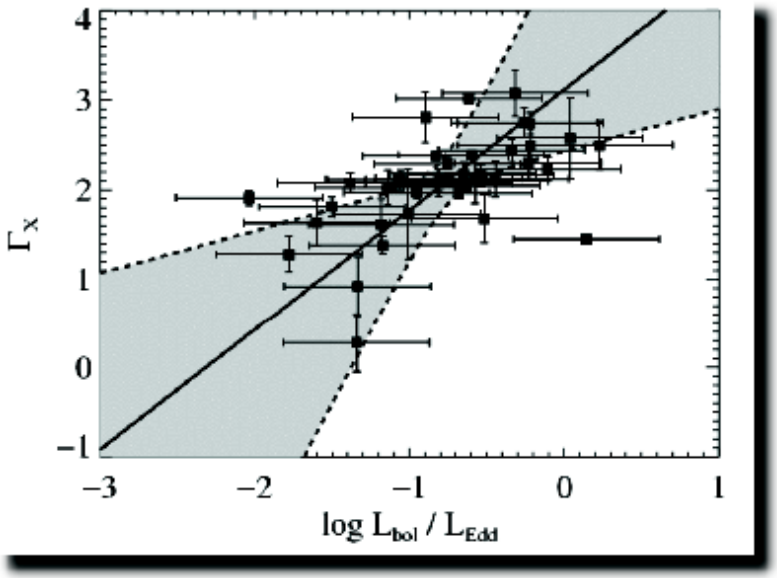


Solid line is posterior median,
Shaded region contains 95%
Of posterior probability.

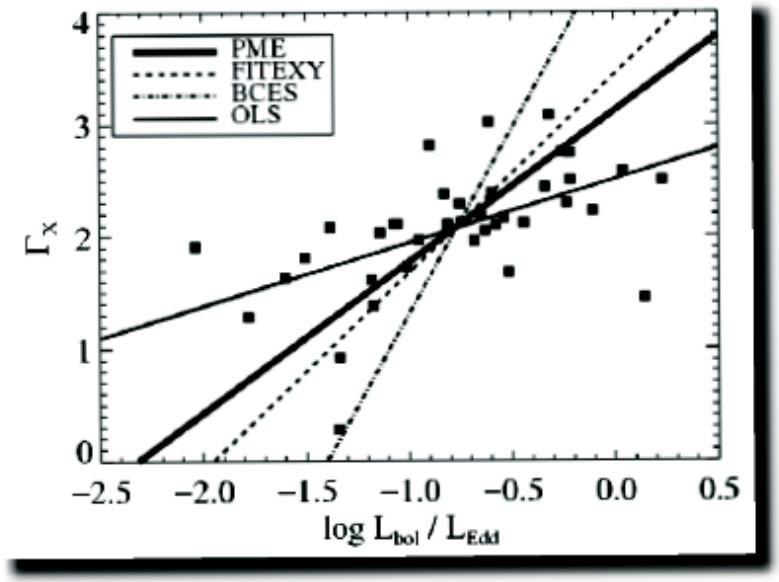


Slides courtesy of Brandon Kelly

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

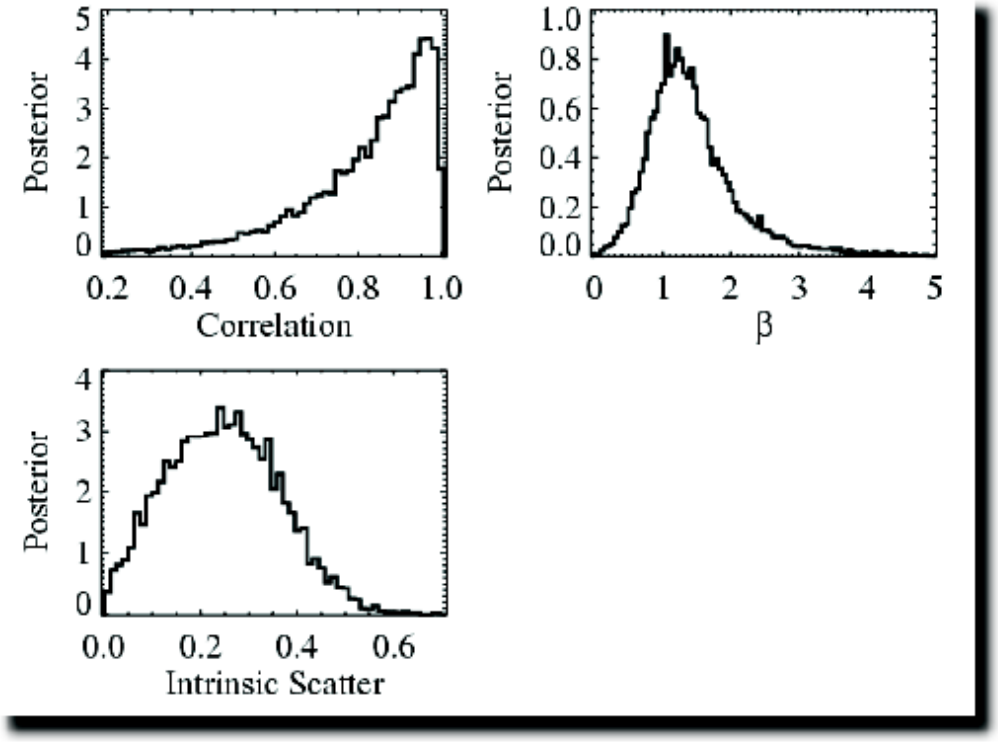


Solid line is posterior median,
Shaded region contains 95%
Of posterior probability.



Slides courtesy of Brandon Kelly

Posterior for Quasar Spectral Slope vs Eddington Ratio



For Comparison:

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

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

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

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

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

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

Slides courtesy of Brandon Kelly

Conclusions

Linear regression, even for unweighted bivariate data, is surprisingly complex. Pay attention to the precise scientific question and details of the situation:

- Functional vs. structural regression (measurement error vs. intrinsic scatter)
- Symmetrical vs. dependent regression
- Parameter confidence intervals (analytic, bootstrap, MCMC)
- Measurement error weightings
- Truncation & censoring due to flux limits (survival, Bayesian)
- Model selection (residual analysis, goodness-of-fit, BIC)
- Robust, multivariate, time series, spatial, ... regression