Measurement Error and Linear Regression of Astronomical Data

Brandon Kelly
Penn State Summer School in Astrostatistics, June 2007
Classical Regression Model

• Collect n data points, denote $i^{th}$ pair as $(\eta_i, \xi_i)$, where $\eta$ is the dependent variable (or ‘response’), and $\xi$ is the independent variable (or ‘predictor’, ‘covariate’)

• Assume usual additive error model:

$$\eta_i = \alpha + \beta \xi_i + \varepsilon_i$$

$$E(\varepsilon_i) = 0$$

$$Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$$
• Ordinary Least Squares (OLS) slope estimate is

\[ \hat{\beta} = \frac{\text{Cov}(\eta, \xi)}{\text{Var}(\xi)} \]

• Wait, that’s not in Bevington…

• The ‘error’ term, \( \varepsilon \), encompasses real physical variations in source properties, i.e., the intrinsic scatter.
Example: Photoionization Physics in Broad Line AGN

- Test if distance between BLR and continuum source set be photoionization physics
- From definition of ionization parameter, U:

\[
R^2 = \frac{L_{ion}}{4\pi c^2 U n_e \bar{E}}
\]

\[
\log r = \frac{1}{2} \log L_{ion} - \frac{1}{2} \log(4\pi c^2) - \frac{1}{2} (\log U + \log n_e + \log \bar{E})
\]

\[
\log r = \alpha + \beta \log L_{ion} + \epsilon
\]

\[
\beta = \frac{1}{2}
\]

\[
\alpha = -\frac{1}{2} [E(\log U + \log n_e + \log \bar{E}) + \log(4\pi c^2)]
\]

\[
\epsilon = -\frac{1}{2} [(\log U + \log n_e + \log \bar{E}) - E(\log U + \log n_e + \log \bar{E})]
\]
BLR Size vs Luminosity, uncorrected for host galaxy starlight (top) and corrected for starlight (bottom). Some scatter due to measurement errors, but some due to intrinsic variations.

Measurement Errors

• Don’t observe \((\eta, \xi)\), but measured values \((y, x)\) instead.

• Measurement errors add an additional level to the statistical model:

\[
\eta_i = \alpha + \beta \xi_i + \varepsilon_i, \quad E(\varepsilon_i) = 0, \quad E(\varepsilon_i^2) = \sigma^2
\]
\[
x_i = \xi_i + \varepsilon_{x,i}, \quad E(\varepsilon_{x,i}) = 0, \quad E(\varepsilon_{x,i}^2) = \sigma_{x,i}^2
\]
\[
y_i = \eta_i + \varepsilon_{y,i}, \quad E(\varepsilon_{y,i}) = 0, \quad E(\varepsilon_{y,i}^2) = \sigma_{y,i}^2
\]
\[
E(\varepsilon_{x,i} \varepsilon_{y,i}) = \sigma_{xy,i}
\]
Different Types of Measurement Error

• Produced by measuring instrument
  – CCD Read Noise, Dark Current

• Poisson, Counting Errors
  – Uncertainty in photon count rate creates measurement error on flux.

• Quantities inferred from fitting a parametric model
  – Using a spectral model to ‘measure’ flux density

• Using observable as proxies for unobservables
  – Using stellar velocity dispersion to ‘measure’ black hole mass, using galaxy flux to ‘measure’ star formation rate
  – Measurement error set by intrinsic source variations, won’t decrease with better instruments/bigger telescopes
Measurement errors alter the moments of the joint distribution of the response and covariate, bias the correlation coefficient and slope estimate

\[ \hat{b} = \frac{\text{Cov}(x,y)}{\text{Var}(x)} = \frac{\text{Cov}(\xi,\eta) + \sigma_{xy}}{\text{Var}(\xi) + \sigma_x^2} \]

\[ \hat{r} = \frac{\text{Cov}(x,y)}{\sqrt{\text{Var}(x)\text{Var}(y)}} = \frac{\text{Cov}(\xi,\eta) + \sigma_{xy}}{\sqrt{\left[\text{Var}(\xi) + \sigma_x^2\right]^{1/2}\left[\text{Var}(\eta) + \sigma_y^2\right]^{1/2}}} \]

- If measurement errors are uncorrelated, regression slope unaffected by measurement error in the response

- If errors uncorrelated, regression slope and correlation biased toward zero (attenuated).
Degree of Bias Depends on Magnitude of Measurement Error

- Define ratios of measurement error variance to observed variance:
  
  \[ R_X = \frac{\sigma_x^2}{\text{Var}(x)}, \quad R_Y = \frac{\sigma_y^2}{\text{Var}(y)}, \]
  
  \[ R_{XY} = \frac{\sigma_{xy}}{\text{Cov}(x, y)} \]

- Bias can then be expressed as:

  \[ \hat{b} = \frac{1 - R_X}{\hat{\beta} - 1 - R_{XY}}, \quad \hat{r} = \frac{(1 - R_X)^{1/2} (1 - R_Y)^{1/2}}{1 - R_{XY}} \]
BCES Estimator

• BCES Approach (Akritas & Bershady, ApJ, 1996, 470, 706; see also Fuller 1987, Measurement Error Models) is to ‘debias’ the moments:

\[ \hat{\beta}_{BCES} = \frac{\text{Cov}(x, y) - \bar{\sigma}_{xy}}{\text{Var}(x) - \bar{\sigma}_x^2}, \quad \hat{\alpha}_{BCES} = \bar{y} - \hat{\beta}_{BCES} \bar{x} \]

• Also give estimator for bisector and orthogonal regression slopes

• Asymptotically normal, variance in coefficients can be estimated from the data

• Variance of measurement error can depend on measured value
Advantages vs Disadvantages

- Asymptotically unbiased and normal, variance in coefficients can be estimated from the data
- Variance of measurement error can depend on measured value
- Easy and fast to compute
- Can be unstable, highly variable for small samples and/or large measurement errors
- Can be biased for small samples and/or large measurement errors
- Convergence to asymptotic limit may be slow, depends on size of measurement errors
FITEXY Estimator

• Press et al. (1992, *Numerical Recipes*) define an ‘effective $\chi^2$’ statistic:

$$
\chi^2_{EXY} = \sum_{i=1}^{n} \frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2_y, i + \beta^2 \sigma^2_{x, i}}
$$

• Choose values of $\alpha$ and $\beta$ that minimize $\chi^2_{EXY}$

• Modified by Tremaine et al. (2002, *ApJ*, 574, 740), to account for intrinsic scatter:

$$
\chi^2_{EXY} = \sum_{i=1}^{n} \frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2 + \sigma^2_{y, i} + \beta^2 \sigma^2_{x, i}}
$$
Advantages vs Disadvantages

- Simulations suggest that FITEXY is less variable than BCES under certain conditions.
- Fairly easy to calculate for a given value of $\sigma^2$.
- Can’t be minimized simultaneously with respect to $\alpha$, $\beta$, and $\sigma^2$, ad hoc procedure often used.
- Statistical properties poorly understood.
- Simulations suggest FITEXY can be biased, more variable than BCES.
- Not really a ‘true’ $\chi^2$, can’t use $\Delta \chi^2 = 1$ to find confidence regions.
- Only constructed for uncorrelated measurement errors.
Structural Approach

• Regard $\xi$ and $\eta$ as missing data, random variables with some assumed probability distribution

• Derive a complete data likelihood:

$$p(x, y, \xi, \eta \mid \theta, \psi) = p(x, y \mid \xi, \eta) p(\eta \mid \xi, \theta) p(\xi \mid \psi)$$

• Integrate over the missing data, $\xi$ and $\eta$, to obtain the observed (measured) data likelihood function

$$p(x, y \mid \theta, \psi) = \iint p(x, y, \xi, \eta \mid \theta, \psi) \, d\xi \, d\eta$$
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:

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

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

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

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

Integrate complete data likelihood to obtain observed data likelihood:

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

\[
= \prod_{i=1}^{n} \sum_{k=1}^{K} \frac{\pi_k}{2\pi \sqrt{\det V_{k,i}}} \exp\left\{ -\frac{1}{2} (z_i - \xi_k)^T V_{k,i}^{-1} (z_i - \xi_k) \right\}
\]

\[
z_i = (y_i \quad x_i)^T
\]

\[
\xi_k = (\alpha + \beta \mu_k \quad \mu_k)^T
\]

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

Can be used to calculate a maximum-likelihood estimate (MLE), perform Bayesian inference. See Kelly (2007) for generalization to multiple covariates.
What if we assume that $\xi$ has a uniform distribution?

- The likelihood for uniform $p(\xi)$ can be obtained in the limit $\tau \to \infty$:

$$ p(x, y | \theta) \propto \prod_{i=1}^{n} \left( \sigma^2 + \sigma_{y,i}^2 + \beta^2 \sigma_{x,i}^2 \right)^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \frac{(y_i - \alpha - \beta x_i)^2}{\sigma^2 + \sigma_{y,i}^2 + \beta^2 \sigma_{x,i}^2} \right\} $$

- Argument of exponential is $\chi^2_{EXY}$ statistic!
- But minimizing $\chi^2_{EXY}$ is not the same as maximizing the likelihood...
- For homoskedastic measurement errors, maximizing this likelihood leads to the ordinary least-squares estimate, so still biased.
Advantages vs Disadvantages

- Simulations suggest MLE for normal mixture approximately unbiased, even for small samples and large measurement error
- Lower variance than BCES and FITEXY
- MLE for $\sigma^2$ rarely, if ever, equal to zero
- Bayesian inference can be performed, valid for any sample size
- Can be extended to handle non-detections, truncation, multiple covariates
- Simultaneously estimates distribution of covariates, may be useful for some studies

- Computationally intensive, more complicated than BCES and FITEXY
- Assumes measurement error variance known, does not depend on measurement
- Needs a parametric form for the intrinsic distribution of covariates (but mixture model flexible and fairly robust).
Simulation Study: Slope

Dashed lines mark the median value of the estimator, solid lines mark the true value of the slope. Each simulated data set had 50 data points, and y-measurement errors of $\sigma_y \sim \sigma$. 
Simulation Study: Intrinsic Dispersion

Dashed lines mark the median value of the estimator, solid lines mark the true value of $\sigma$. Each simulated data set had 50 data points, and $y$-measurement errors of $\sigma_y \sim \sigma$. 
Effects of Sample Selection

- Suppose we select a sample of $n$ sources, out of a total of $N$ possible sources.
- Introduce indicator variable, $I$, denoting whether a source is included. $I_i = 1$ if the $i^{th}$ source is included, otherwise $I_i = 0$.
- Selection function for $i^{th}$ source is $p(I_i=1|y_i,x_i)$
- Selection function gives the probability that the $i^{th}$ source is included in the sample, given the values of $x_i$ and $y_i$.

Complete data likelihood is

\[
p(x, y, \xi, \eta, I | \theta, \psi, N) \propto \binom{N}{n} \prod_{i \in A_{\text{obs}}} p(x_i, y_i, \xi_i, \eta_i, I_i | \theta, \psi) \prod_{j \in A_{\text{mis}}} p(x_j, y_j, \xi_j, \eta_j, I_j | \theta, \psi)
\]

Here, \( A_{\text{obs}} \) denotes the set of \( n \) included sources, and \( A_{\text{mis}} \) denotes the set of \( N-n \) sources not included.

Binomial coefficient gives number of possible ways to select a subset of \( n \) sources from a sample of \( N \) sources.

Integrating over the missing data, the observed data likelihood now becomes

\[
p(x_{\text{obs}}, y_{\text{obs}}, I | \theta, \psi, N) \propto \binom{N}{n} \prod_{i \in A_{\text{obs}}} p(x_i, y_i | \theta, \psi)
\]

\[
\times \prod_{j \in A_{\text{mis}}} \iiint p(I_j = 0 | x_j, y_j) p(x_j, y_j | \theta, \psi) \, dx_j \, dy_j
\]
Selection only dependent on covariates

- If the sample selection only depends on the covariates, then $p(I|x,y) = p(I|x)$.
- Observed data likelihood is then

$$p(x_{obs}, y_{obs}, I | \theta, \psi, N) \propto \binom{N}{n} \prod_{i \in A_{obs}} p(x_i, y_i | \theta, \psi)$$

$$\times \prod_{j \in A_{mis}} \iint p(I_j = 0 | x_j) p(y_j | x_j, \theta, \psi) p(x_j | \psi) \, dx_j \, dy_j$$

$$\propto \binom{N}{n} \prod_{i \in A_{obs}} p(x_i, y_i | I_i = 1, \theta, \psi_{obs}) \prod_{j \in A_{mis}} \int p(x_j | I_j = 0, \psi_{mis}) p(I_j = 0 | \psi_{mis}) \, dx_j$$

- Therefore, if selection is only on the covariates, then inference on the regression parameters is unaffected by selection effects.
Selection depends on dependent variable: truncation

• Take Bayesian approach, posterior is

\[ p(\theta, \psi, N \mid x_{\text{obs}}, y_{\text{obs}}, I) \propto p(\theta, \psi, N) p(x_{\text{obs}}, y_{\text{obs}}, I \mid \theta, \psi, N) \]

• Assume uniform prior on \( \log N \), \( p(\theta, \psi, N) \propto N^{-1} p(\theta, \psi) \)

• Since we don’t care about \( N \), sum posterior over \( n < N < \infty \):

\[ p(\theta, \psi \mid x_{\text{obs}}, y_{\text{obs}}, I) \propto \left[ p(I = 1 \mid \theta, \psi) \right]^{-n} \prod_{i \in A_{\text{obs}}} p(x_i, y_i \mid \theta, \psi) \]

\[ p(I = 1 \mid \theta, \psi) = \iint p(I = 1 \mid x, y) p(x, y \mid \theta, \psi) \, dx \, dy \]
Covariate Selection vs Response Selection

Covariate Selection: No effect on distribution of $y$ at a given $x$

Response Selection: Changes distribution of $y$ at a given $x$
Non-detections: ‘Censored’ Data

• Introduce additional indicator variable, D, denoting whether a data point is detected or not: D=1 if detected.

• Assuming selection function independent of response, observed data likelihood becomes

\[
p(x, y, D \mid \theta, \psi) \propto \prod_{i \in A_{\text{det}}} p(x_i, y_i \mid \theta, \psi) \times \prod_{j \in A_{\text{cens}}} p(x_j \mid \psi) \int p(D_j = 0 \mid y_j, x_j) p(y_j \mid x_j, \theta, \psi) \, dy_j
\]

• \(A_{\text{det}}\) denotes set of detected sources, \(A_{\text{cens}}\) denotes set of censored sources.

• Equations for \(p(x \mid \psi)\) and \(p(y \mid x, \theta, \psi)\) under the mixture of normals models are given in Kelly (2007).
Bayesian Inference

• Calculate posterior probability density for mixture of normals structural model
• Posterior valid for any sample size, doesn’t rely on large sample approximations (e.g., asymptotic distribution of MLE).
• Assume prior advocated by Carroll et al. (1999)
• Use markov chain monte carlo (MCMC) to obtain draws from posterior
Gibbs Sampler

- Method for obtaining draws from posterior, easy to program
- Start with initial guesses for all unknowns
- Proceed in three steps:
  - Simulate new values of missing data, including non-detections, given the observed data and current values of the parameters
  - Simulate new values of the parameters, given the current values of the prior parameters and the missing data
  - Simulate new values of the prior parameters, given the current parameter values.
- Save random draws at each iterations, repeat until convergence.
- Treat values from latter part of Gibbs sampler as random draw form the posterior

Example: Simulated Data Set with Non-detections

Solid line is true regression line, Dashed-dotted is posterior Median, shaded region contains Approximately 95% of posterior probability

Filled Squares: Detected data points
Hollow Squares: Undetected data points
Posterior from MCMC for simulated data with non-detections

Solid vertical line marks true value. For comparison, a naïve MLE that ignores meas. error found

$$\beta_{MLE} = 0.229 \pm 0.077$$

This is biased toward zero at a level of 3.5σ
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:

\[ \hat{\beta}_{OLS} = 0.56 \pm 0.14 \]
\[ \hat{\beta}_{BCES} = 3.29 \pm 3.34 \]
\[ \hat{\beta}_{EXY} = 1.76 \pm 0.49 \]
\[ \hat{\sigma}_{OLS} = 0.41 \]
\[ \hat{\sigma}_{BCES} = 0.32 \]
\[ \hat{\sigma}_{EXY} = 0.0 \]
References

• Carroll, Roeder, & Wasserman, 1999, Biometrics, 55, 44
• Carroll, Ruppert, & Stefanski, 1995, Measurement Error in Nonlinear Models (London: Chapman & Hall)
• Fuller, 1987, Measurement Error Models, (New York: John Wiley & Sons)
• Gelman et al., 2004, Bayesian Data Analysis, (2nd Edition; Boca Raton: Chapman Hall/CRC)
• Schafer, 1987, Biometrica, 74, 385
• Schafer, 2001, Biometrics, 57, 53