Astrostatistics: Sat 03 Feb 2017

https://github.com/CambridgeAstroStat/PartIII-Astrostatistics

- Fitting Statistical Models to Astronomical Data
  - Maximum Likelihood vs. ~~minimising χ^2~~
  - Generative / Latent Variable Modeling / Bayes
  - Ivezic Ch 4 “Classical Statistical Inference” & Ch 5 “Bayesian Statistical Inference”
  - F&B Ch 3 “Statistical Inference”
  - Hogg, Bovy & Lang. “Data analysis recipes: Fitting a model to data”. https://arxiv.org/abs/1008.4686

Statistical Modelling Wisdom

- Have an objective function [e.g. Likelihood or posterior] that you optimise or sample to fit the data - not just a procedure/recipe
- Objective function helps you evaluate relative fits of data with under different parameter values / models
- Derive your objective function from your modelling assumptions (physical or statistical)
- Write down your assumptions!
- First question: what is the likelihood L(θ) ? Derive it from the assumptions underlying your sampling distribution P(D | θ)!
- Second question: what is your prior P(θ) ? (if Bayesian)
- Third question: How do I optimise/sample objective function to fit the data?

Fitting Models to Astro Data

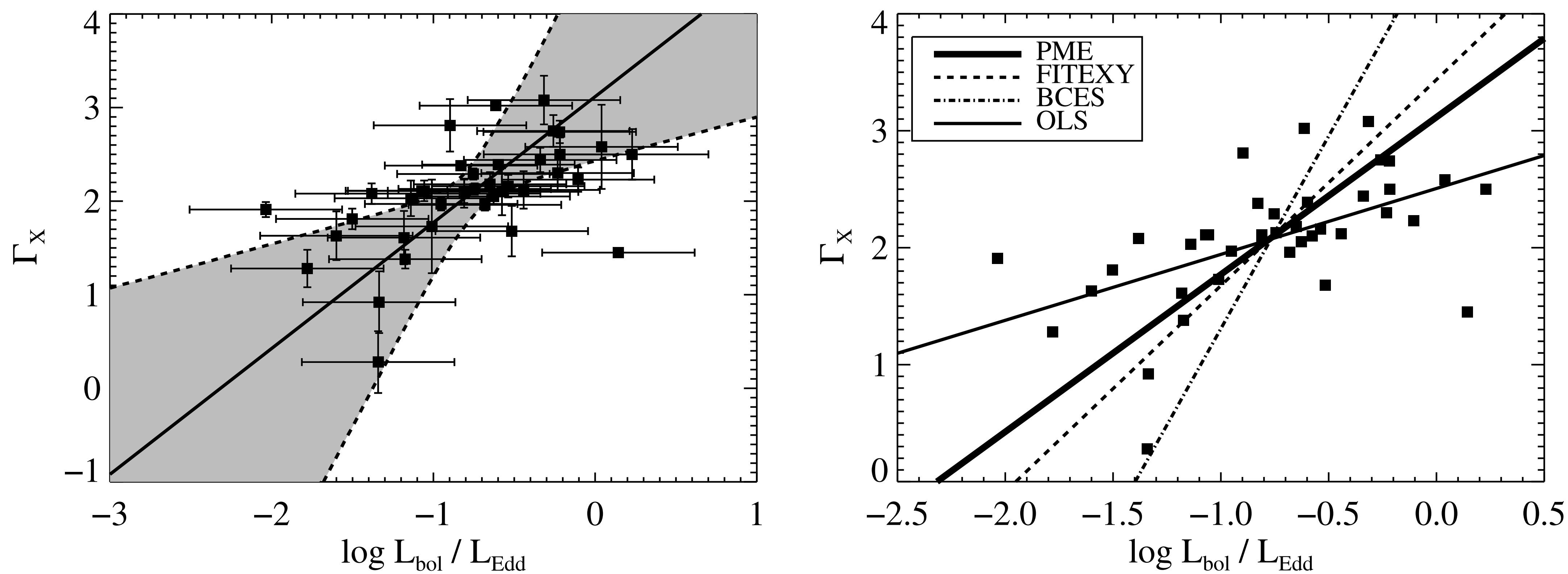


FIG. 10.—X-ray photon index Γ_X as a function of $\log L_{bol}/L_{Edd}$ for 39 $z \lesssim 0.8$ radio-quiet quasars. In both plots, the thick solid line shows the posterior median estimate (PME) of the regression line. In the left panel, the shaded region denotes the 95% ($2\,\sigma$) pointwise confidence intervals on the regression line. In the right panel, the thin solid line shows the OLS estimate, the dashed line shows the FITEXY estimate, and the dot-dashed line shows the BCES($Y|X$) estimate; the error bars have been omitted for clarity. A significant positive trend is implied by the data.

Modelling heteroskedastic, correlated measurement errors in both y and x, intrinsic scatter, nondetections, selection effects

B. Kelly et al. 2007, “Some Aspects of Measurement Error in Linear Regression of Astronomical Data.” ApJ, 665, 1489

Ad-hoc “χ^2” approaches vs. Likelihood formulation

FITEXY Estimator

- Press et al.(1992, *Numerical Recipes*) define an ‘effective χ^2’ statistic:

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

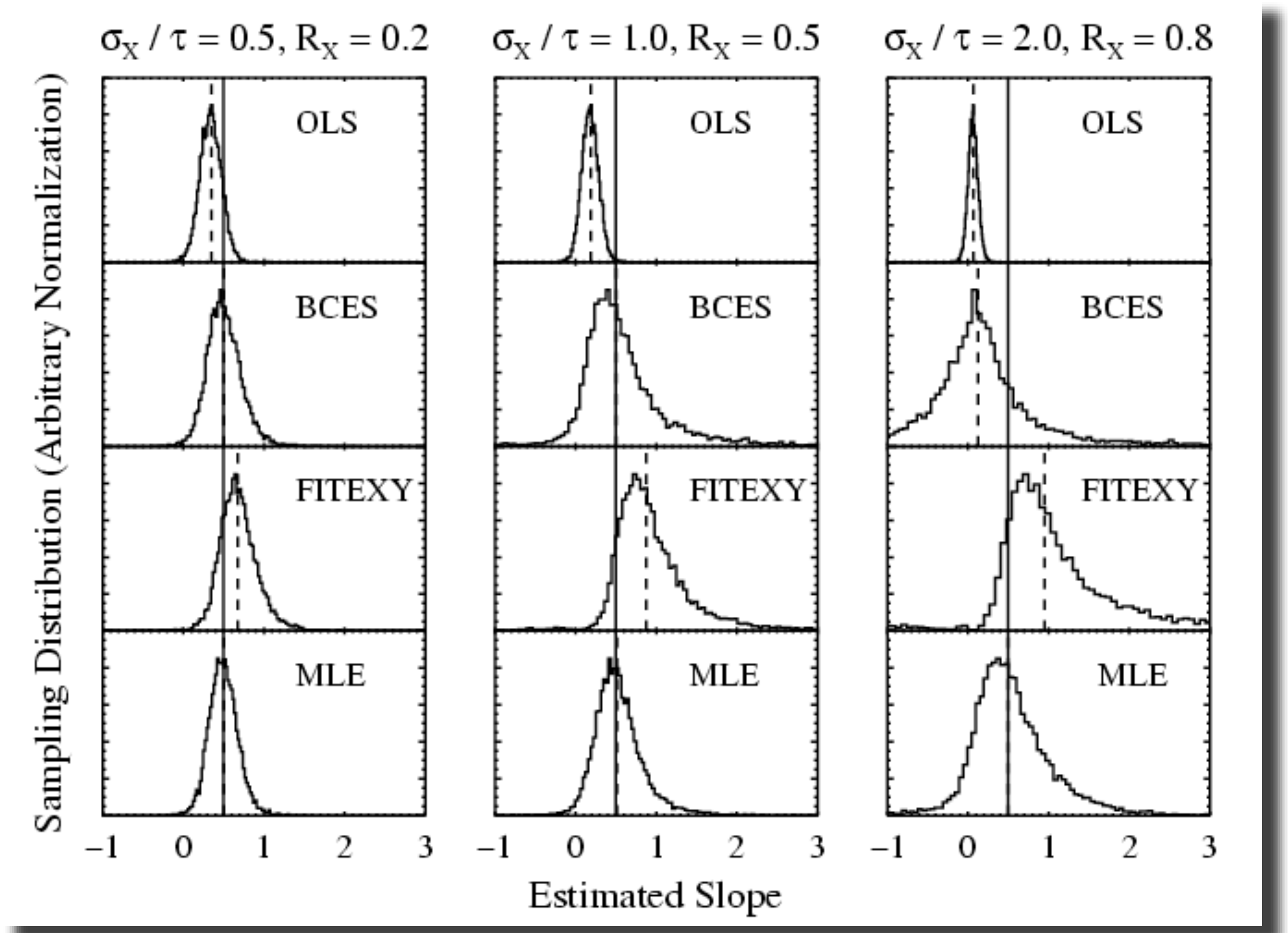
- Choose values of α and β that minimize χ^2_{EXY}
- Modified by Tremaine et al.(2002, ApJ, 574, 740), to account for intrinsic scatter:

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

http://astrostatistics.psu.edu/su07/kelley_measerr07.pdf

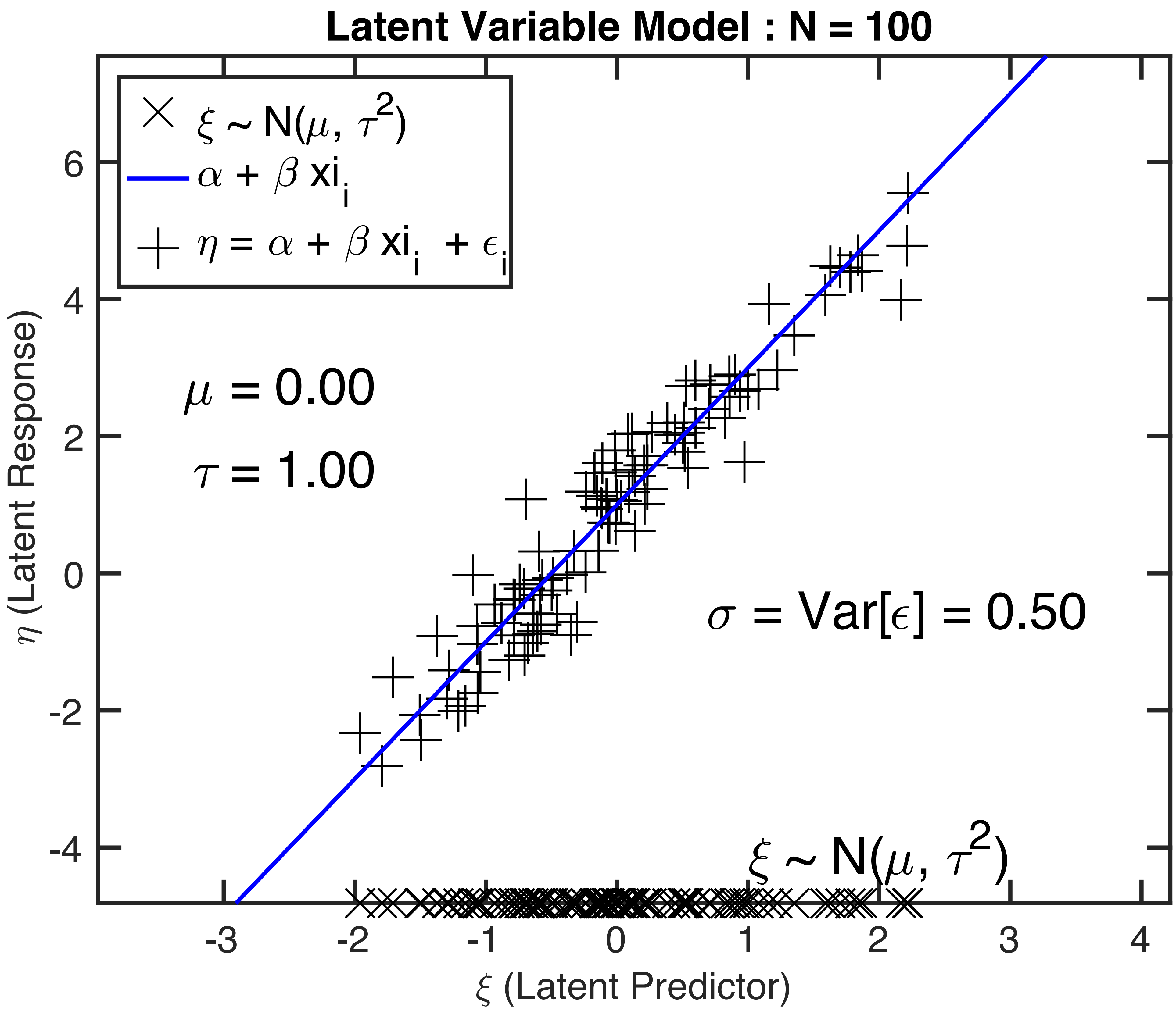
Kelly et al. 2017, Latent Variable Likelihood approach vs. Bad

Simulation Study: Slope

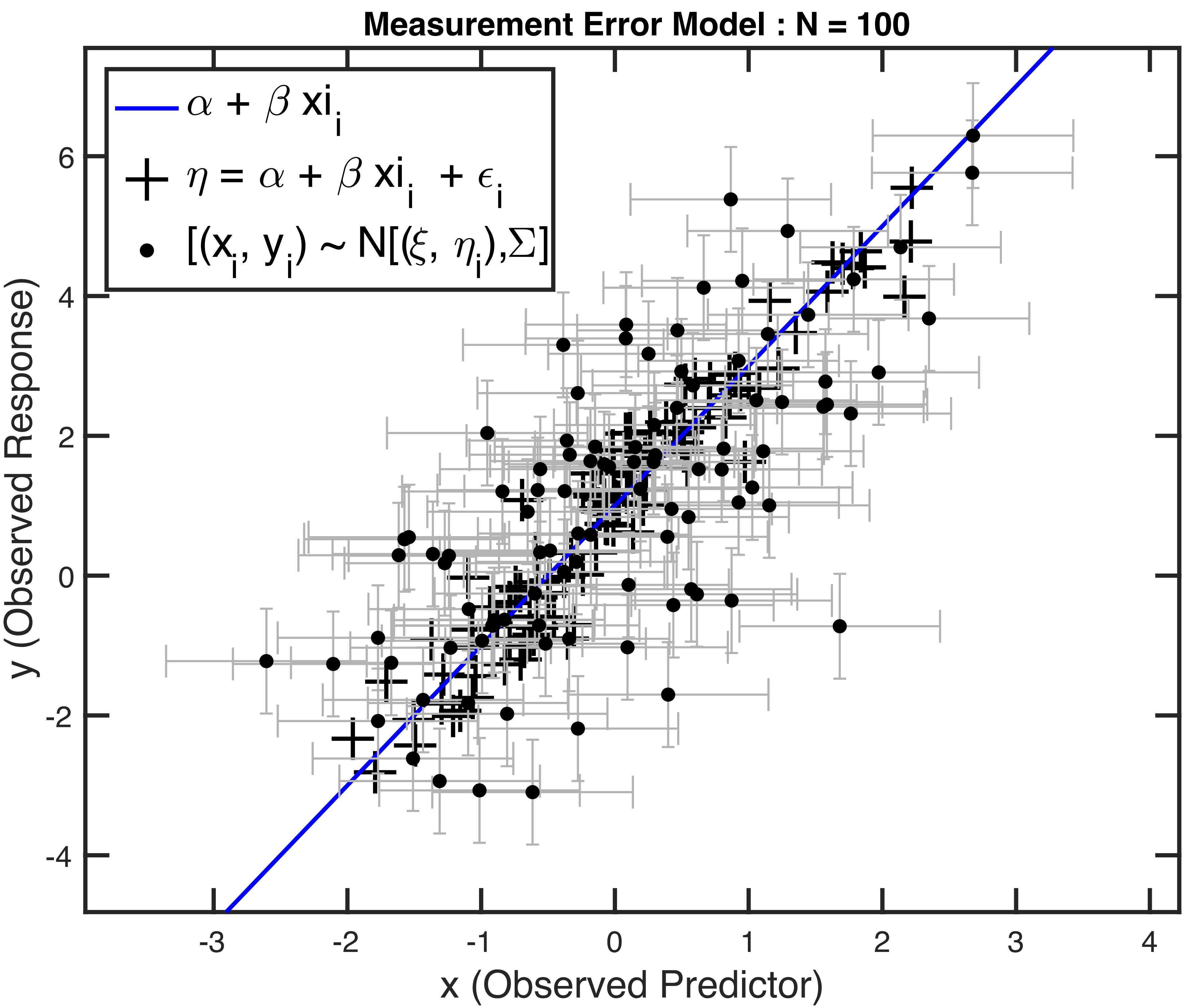


Dashed lines mark the median value of the estimator, solid lines mark the true value of the slope. Each simulated data set had 50 data points, and y-measurement errors of $\sigma_y \sim \sigma$.

http://astrostatistics.psu.edu/su07/kelley_measerr07.pdf



Latent Variable Model : N = 100
$\xi \sim N(\mu, \tau^2)$
$\alpha + \beta$ xi$_i$
$\eta = \alpha + \beta$ xi$_i$ $+ \epsilon_i$
$\mu = 0.00$
$\tau = 1.00$
$\sigma = Var[\epsilon] = 0.50$
$\xi \sim N(\mu, \tau^2)$
η (Latent Response)
ξ (Latent Predictor)
6
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
4



Measurement Error Model : N = 100
$\alpha + \beta$ xi$_i$
$\eta = \alpha + \beta$ xi$_i$ + ϵ_i
[(x$_i$, y$_i$) ~ N[(ξ, η_i),Σ]
y (Observed Response)
x (Observed Predictor)
6
4
2
0
-2
-4
-3
-2
-1
0
1
2
3
4