Regression in Astronomy: Errors–In–Variables, Censoring, Heteroskedasticity, and Intrinsic Scatter

October 5, 2016

Outline

Regression

- y is approximately some function of x

$$y = f(x) + \epsilon$$

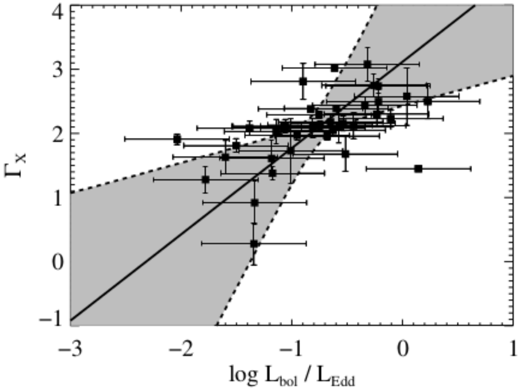
- regression is used to:
  1. estimate, quantify uncertainty in f
  2. predict y values for new x

- common to assume linear relation:

$$f(x) = \beta_0 + \beta_1 x$$

- linear regression is often complicated in astronomy due to:
  - *heteroskedastic measurement error*
  - *intrinsic scatter*
  - *errors–in–variables* (measurement error in x)
  - *censoring*

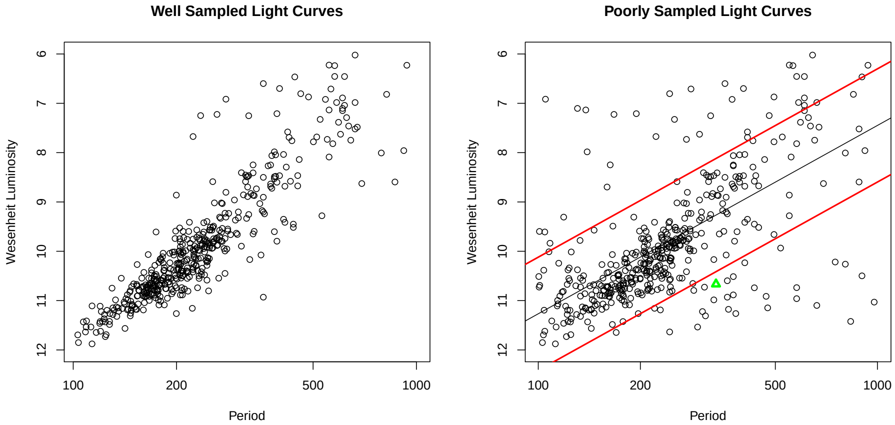
Example: Eddington Ratio



- Γ_X and $\log L_{bol}/L_{edd}$ are both measured with error (crosses)
- `intrinsic scatter`: even if no measurement error in x and y, still not a perfect linear relation.

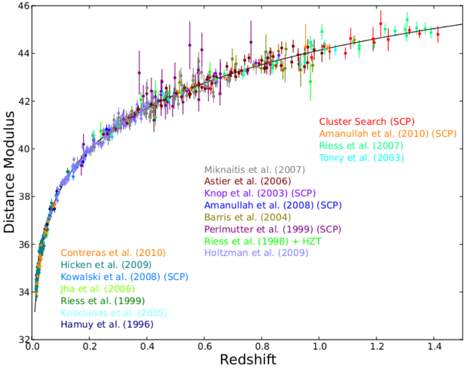
Source: "Some aspects of measurement error in linear regression of astronomical data" Kelly ApJ 2007
`http://iopscience.iop.org/article/10.1086/519947/meta`

Example: Period Luminosity Relation



- roughly a linear relationship between luminosity and log(*period*)
- with poorly sampled light curves, errors in period estimates

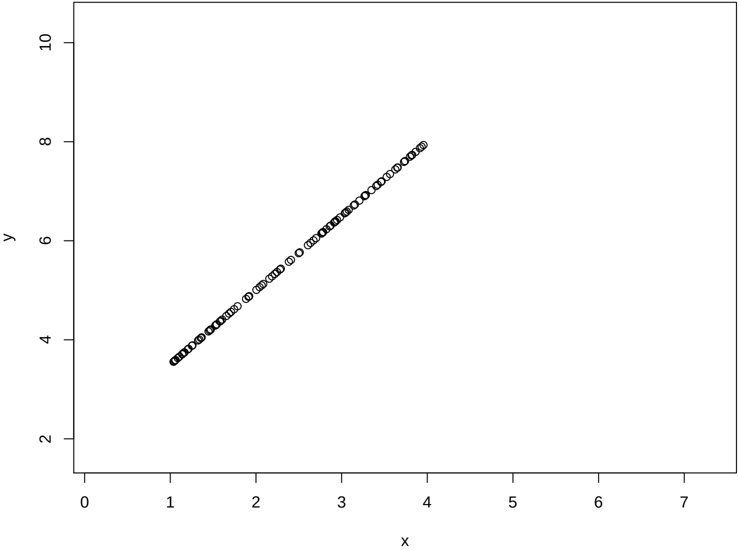
Supernovae Cosmology



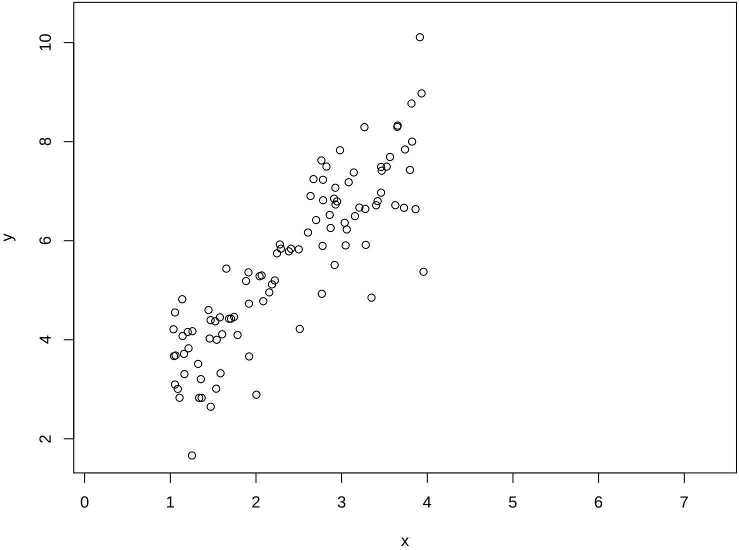
- non–linear relationship between distance modulus and redshift
- equations from cosmology determine model form

Source: "THE HUBBLE SPACE TELESCOPE CLUSTER SUPERNOVA SURVEY. V. IMPROVING THE DARK-ENERGY CONSTRAINTS ABOVE $z > 1$ AND BUILDING AN EARLY-TYPE-HOSTED SUPERNOVA SAMPLE" Suzuki 2012
`http://iopscience.iop.org/article/10.1088/0004-637X/746/1/85/meta`

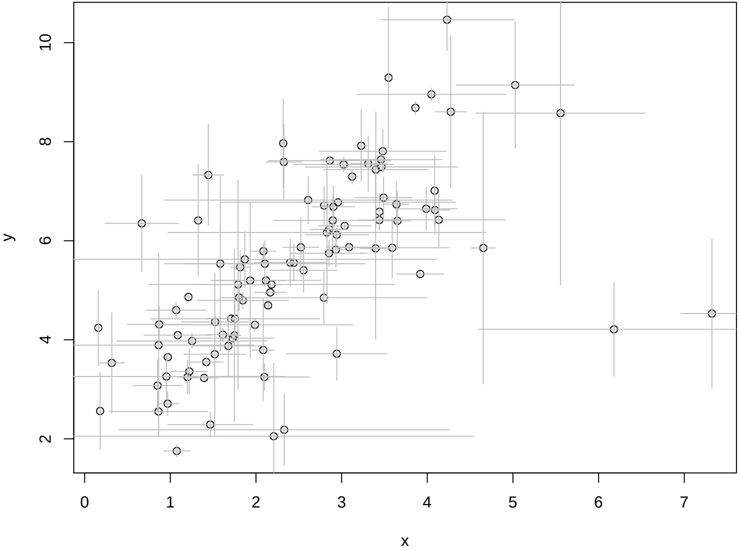
Perfect Linear Relationship



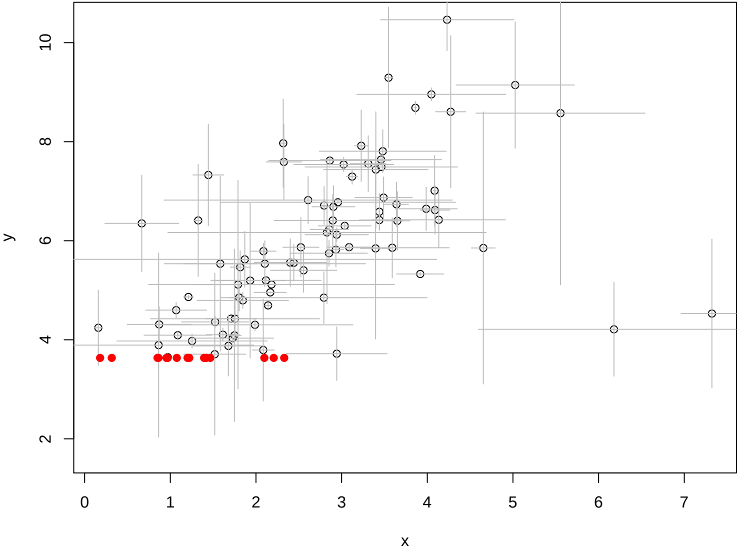
Intrinsic Scatter



Heteroskedastic Error on x and y



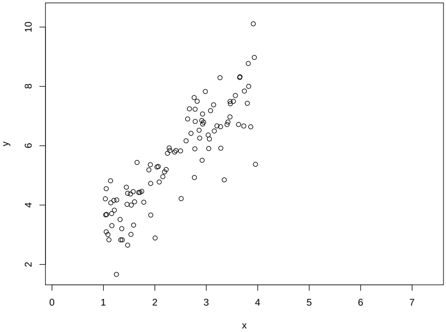
Censoring



Outline

Ordinary Least Squares Model

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$
- parameters: $(\sigma^2, \beta_0, \beta_1)$.
- only intrinsic scatter present



Estimate $(\sigma^2, \beta_0, \beta_1)$ with Maximum Likelihood

$$\begin{aligned}
\widehat{\sigma}^2, \widehat{\beta_0}, \widehat{\beta_1} &= \underset{(\sigma^2, \beta_0, \beta_1)}{\operatorname{argmax}} \; L((\sigma^2, \beta_0, \beta_1)|D) \\
&= \underset{(\sigma^2, \beta_0, \beta_1)}{\operatorname{argmax}} \; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y_i - \beta_0 - \beta_1 x_i)^2/(2\sigma^2)}
\end{aligned}$$

After some calculus

$$\begin{aligned}
\widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x} \\
\widehat{\beta}_1 &= \frac{n^{-1} \sum x_i y_i - \bar{x}\bar{y}}{n^{-1} \sum x_i^2 - \bar{x}^2} \\
\widehat{\sigma}^2 &= \frac{1}{n} \sum (y_i - \widehat{\beta}_0 - \widehat{\beta}_1)^2
\end{aligned}$$

Can replace $1/n$ with $1/(n-2)$ in $\widehat{\sigma}^2$ formula.

Use Matrices

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^{n\times 1} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathbb{R}^{n\times 2} \qquad \epsilon \sim N(0, \sigma^2 I) \in \mathbb{R}^{n\times 1}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

linear regression is now

$$Y = X\beta + \epsilon$$

maximum Likelihood in matrix form

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

$$\widehat{\sigma}^2 = n^{-1}(Y - X\widehat{\beta})^T (Y - X\widehat{\beta})$$

$$\widehat{\mathrm{Var}}\,(\widehat{\beta}) = \widehat{\sigma}^2 (X^T X)^{-1}$$

Outline

Weighted Least Squares

- intrinsic scatter is 0.
- known $\sigma_{yi} \neq 0$ (y_i measured with error)

In statistics this is called heteroskedastic error.

Statistical Model:

$$Y = X\beta + \epsilon$$

where

$$\epsilon \sim N(0, \Sigma)$$

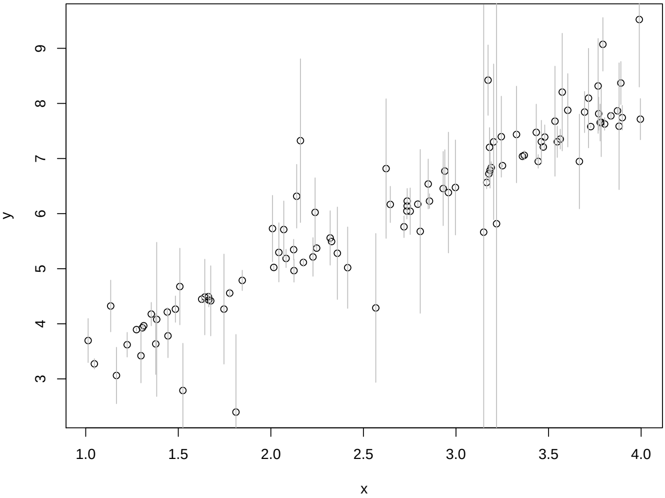
where Σ is a diagonal matrix with $\Sigma_{ii} = \sigma_{yi}^2$.

(non–matrix form)

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_{yi}^2)$ independent across i.

Example



Only accounts for measurement error in y, not intrinsic scatter.

Maximum Likelihood for Heteroskedastic Error

- **Trick:** $\epsilon \sim N(0, \Sigma)$ and

  $$Y = X\beta + \epsilon$$

  is the same as

  $$\Sigma^{-1/2}Y = \Sigma^{-1/2}X\beta + \Sigma^{-1/2}\epsilon$$

  where $\Sigma^{-1/2}\epsilon \sim N(0, I)$.

- **Maximum Likelihood** from the homoskedastic case tells us

  $$\widehat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

Or write out likelihood, take derivatives, set equal to 0, solve.

Uncertainty on $\widehat{\beta}$

Recall from OLS model

$$\text{Var}\,(\widehat{\beta}) = \sigma^2(X^T X)^{-1}.$$

With heteroskedastic error $X \to \Sigma^{-1/2}X$ and $\sigma \to 1$, so

$$\text{Var}\,(\widehat{\beta}) = (X^T \Sigma^{-1} X)^{-1}.$$

Intrinsic Scatter + Measurement Error

- OLS covers intrinsic scatter, but no measurement error in y
- WLS covers measurement error in y, but no intrinsic scatter

Intrinsic Scatter and y (Normal) Measurement Error

$$Y = X\beta + \epsilon$$

where

$$\epsilon \sim N(0, \Sigma)$$

where Σ is a diagonal matrix with $\Sigma_{ii} = \sigma^2 + \sigma_{yi}^2$.

β and σ are unknown parameters.

General Weighted Least Squares Estimators

- let W be a diagonal weight matrix
- consider estimators of the form

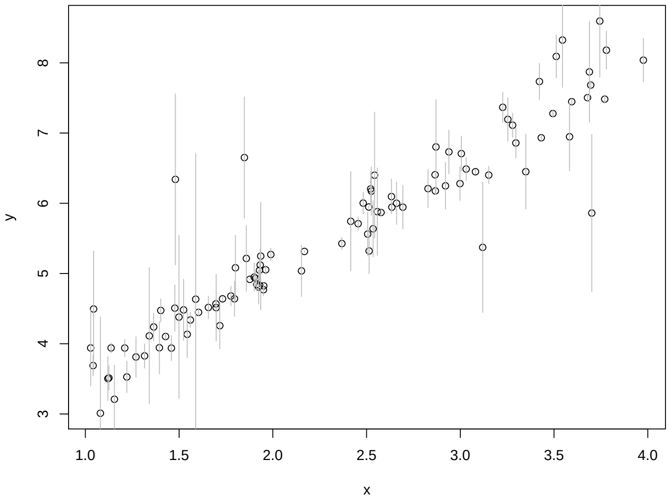
$$\widehat{\beta}(W) = (X^T W X)^{-1} X^T W Y.$$

Possible Weight Matrices:

- $W_{1,ii} = 1$
- $W_{2,ii} = \sigma_{yi}^{-2}$
- $W_{3,ii} = (\sigma_{yi}^2 + \sigma^2)^{-1}$

Recall W_3 is not known because σ^2 is unknown.

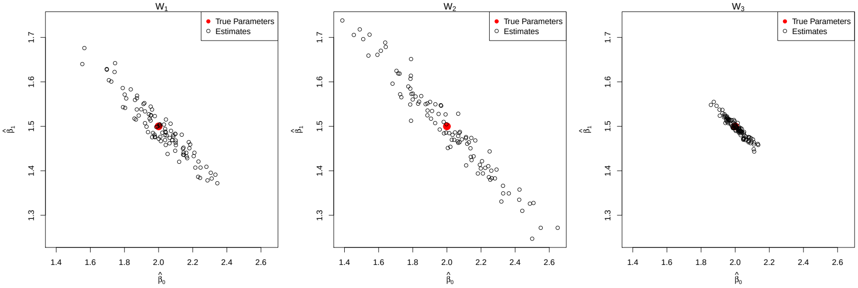
$\beta = (2, 1.5)^T, \sigma = 0.1$ with Heteroskedastic Error



What is sampling distribution using W_1, W_2, and W_3?

Sampling Distributions

Regenerate data 100 times from model, compute estimators with each weight matrix.



- W_3 is best, but it depends on σ which is unknown.
- W_1 overweights observations with <u>large</u> error (σ^2_{yi} large)
- W_2 overweights observations with <u>small</u> error (σ^2_{yi} small)

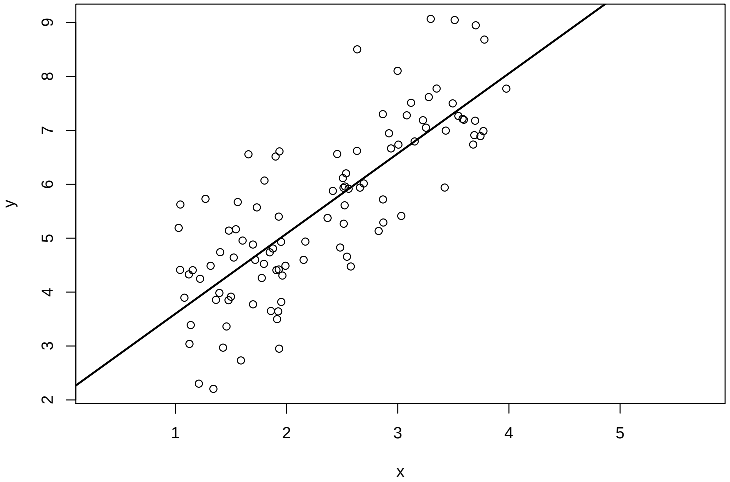
Maximum Likelihood with Intrinsic Scatter

$$\begin{aligned}
\widehat{\sigma}^2, \widehat{\beta}_0, \widehat{\beta}_1 &= \underset{(\sigma^2,\beta_0,\beta_1)}{\text{argmax}} \; L((\sigma^2, \beta_0, \beta_1)|D) \\
&= \underset{(\sigma^2,\beta_0,\beta_1)}{\text{argmax}} \; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi(\sigma^2 + \sigma_i^2)}} e^{-(y_i-\beta_0-\beta_1 x_i)^2/(2(\sigma^2+\sigma_i^2))}
\end{aligned}$$

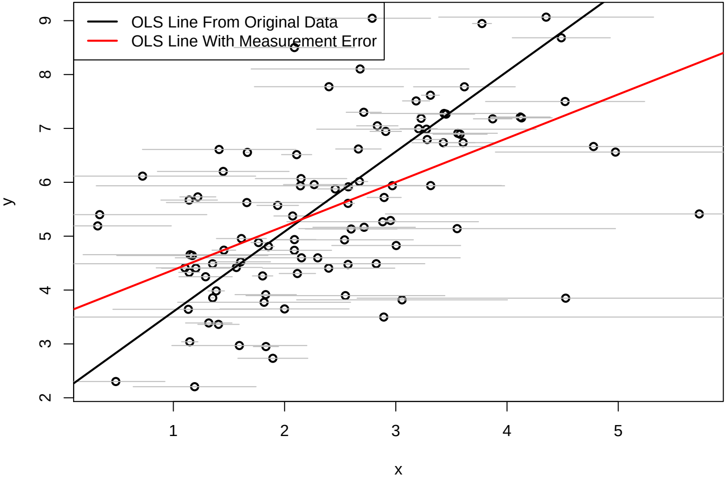
- computation
  - no closed form solution
  - at fixed σ, closed form solution
  - evaluate likelihood at each σ in grid
- uncertainty quantification
  - compute standard errors using information matrix

Outline

Simulation



Simulation



OLS Line From Original Data
OLS Line With Measurement Error
y
9
8
7
6
5
4
3
2
x
1
2
3
4
5

Observations

- ignoring error in x produces biased estimators
- worse than ignoring intrinsic scatter / heteroskedasticity in y
  - increased variance, not biased
- large sample sizes do not alleviate bias

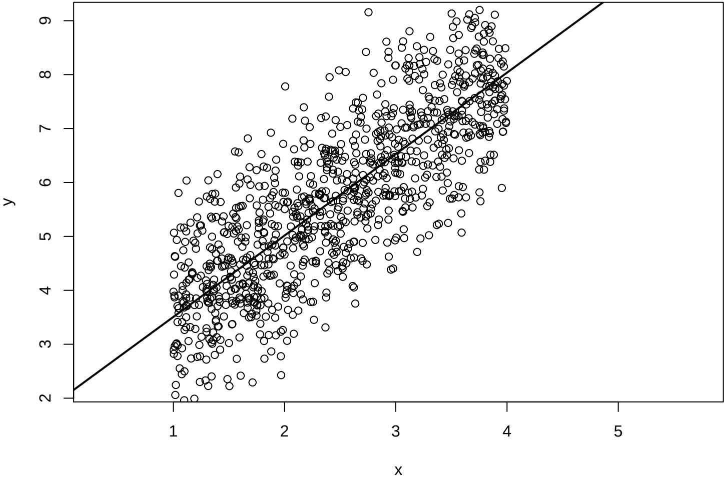
Essentially if $\widehat{\beta}_{zj}$ is parameter fit with noisy x, then

$$\lim_{n\to\infty} \widehat{\beta}_{z0} \not\to \beta_0$$
$$\lim_{n\to\infty} \widehat{\beta}_{z1} \not\to \beta_1$$

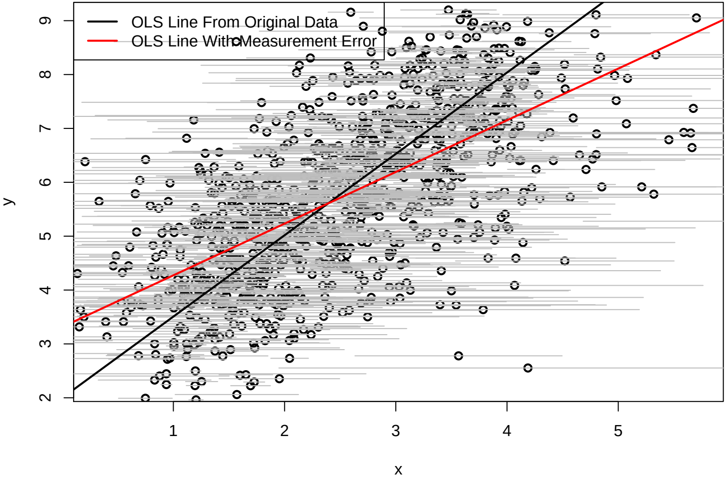
In particular

$$|\lim_{n\to\infty} \widehat{\beta}_{z1}| < |\beta_1|$$

Simulation with Larger Sample Size



Simulation with Larger Sample Size



More Rigorously

$$
\begin{aligned}
x_i &\sim N(\mu_x, \sigma_x^2) \\
y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\
\epsilon_i &\sim N(0, \sigma^2)
\end{aligned}
$$

But we do not observe x_i, instead

$$
\begin{aligned}
z_i &= x_i + \delta_i \\
\delta_i &\sim N(0, \sigma_\delta^2)
\end{aligned}
$$

We observe $(z_i, y_i)_{i=1}^n$ iid. We matrix notation

$$
\mathbf{Z} = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{pmatrix} \in \mathbb{R}^{n \times 2}
$$

More Rigorously

The estimator ignoring the measurement error is

$$\widehat{\beta}_Z = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T Y$$

But some algebra shows that

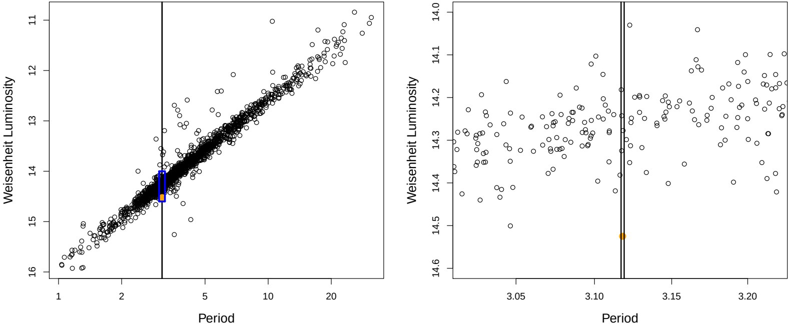
$$\mathbb{E}[\widehat{\beta}_Z] = \begin{pmatrix} \beta_0 + \frac{\sigma_\delta^2}{\sigma_\delta^2+\sigma_x^2}\mu_x\beta_1 \\ \frac{\sigma_x^2}{\sigma_\delta^2+\sigma_x^2}\beta_1 \end{pmatrix}$$

- unbiased only if $\sigma_\delta^2 = 0$
- level of bias in slope proportional to

$$\frac{\sigma_x^2}{\sigma_\delta^2 + \sigma_x^2}$$

Normality not necessary, only second moments.

Recall from Time Domain Lecture



Here

$$\sigma_x^2 \approx 25$$

$$\sigma_\delta^2 < 0.001^2$$

$$\mathbb{E}[\widehat{\beta}_{Z1}] = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\delta^2}\beta_1 \approx \beta_1$$

The Double Whammy

Whammy Number 1: Parameter estimates are biased.

Whammy Number 2: Power diminished in hypothesis tests.

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Result: We draw incorrect conclusions (Whammy 1) and believe sample sizes need to be larger than actually necessary (Whammy 2).

Source: "Double Whammy" defined in "Measurement error in nonlinear models: a modern perspective" Carroll et al
`http://www.stat.tamu.edu/~carroll/eiv.SecondEdition/Table_of_Contents.pdf`

Functional versus Structural Models

Models for addressing measurement error in predictors:

- functional models: do not assume distribution for predictors. sampling distribution of estimators calculated conditional on $x_1, \ldots, x_n$.
- structural models: assume distribution for x. can marginalize out x in inference.

I will only discuss structural solutions. This means we need to assume distributions for predictors x.

Everything Normal Model

Recall $(z_i, y_i)_{i=1}^n$ i.i.d. where

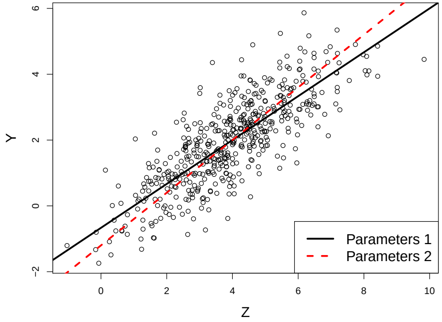
$$
\begin{aligned}
x_i &\sim N(\mu_x, \sigma_x^2) \\
y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\
z_i &= x_i + \delta_i \\
\epsilon_i &\sim N(0, \sigma^2) \\
\delta_i &\sim N(0, \sigma_\delta^2)
\end{aligned}
$$

The marginal distribution of observation (z, y) is

$$
\begin{pmatrix} z \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \beta_0 + \beta_1 \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_x^2 + \sigma_\delta^2 & \beta_1 \sigma_x^2 \\ \beta_1 \sigma_x^2 & \beta_1^2 \sigma_x^2 + \sigma^2 \end{pmatrix} \right)
$$

Identifiability Issue

	Param. 1	Param. 2
β_0	-2/3	-6/5
β_1	2/3	4/5
μ_x	4	4
σ_x^2	3	5/2
σ^2	2/3	2/5
σ_δ^2	0	1/2



Problem: (z, y) has same distribution under Parameters 1 and 2.
One Solution: Assume σ_δ^2 is known. (Fairly realistic in many astronomy applications.)

MLE Solution

Compute MLEs for means and covariances of (z, y). Then solve for parameters of interest:

$$\begin{aligned}
\bar{z} &= \widehat{\mu}_x \\
\bar{y} &= \widehat{\beta}_0 + \widehat{\beta}_1 \widehat{\mu}_x \\
n^{-1} S_{zz} &= \sigma_\delta^2 + \widehat{\sigma}_x^2 \\
n^{-1} S_{yy} &= \widehat{\sigma}^2 + \widehat{\beta}^2 \widehat{\sigma}_x^2 \\
n^{-1} S_{zy} &= \widehat{\beta}_1 \widehat{\sigma}_x^2
\end{aligned}$$

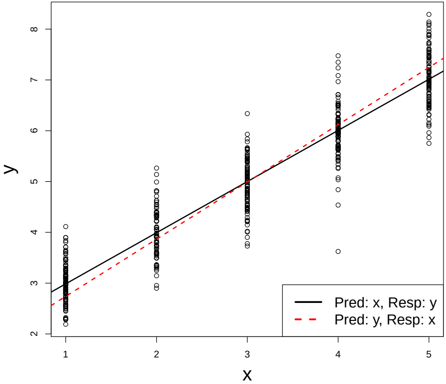
See "Statistical Inference" Second Edition by Casella and Berger 12.2.4 for discussion of constructing confidence sets for parameters.

Notes

- distribution of covariates (x) is important
  - without measurement error, inference is done conditional on x
  - here used normality assumption for x, very restrictive
  - mixture of gaussians would add more flexibility
- identifiability issue shows measurement error in predictors not necessarily detectable from (z, y) data alone
- if goal is prediction of y using z, ignoring measurement error may be okay*
- can we reverse the roles of x and y?
  - y may be linear function of x, but x not linear function of y
  - reversing roles of x and y results in different lines
  - causality / prediction considerations may influence role of x, y

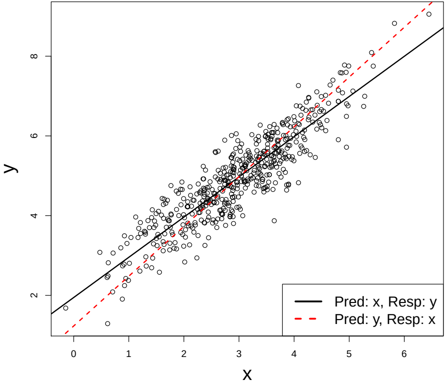
* See `https://www.stat.tamu.edu/~carroll/talks/Seio_Carroll_04-19-2012.pdf` for discussion on classification and `https://www.stat.tamu.edu/~carroll/ftp/2009.papers.directory/prediction_MEM.pdf` for nonparametric prediction

OLS Valid Only y on x



OLS model valid for regression y on x but not for x on y. Red dotted line does not have clear model interpretation.

OLS Valid y on x and x on y



OLS model valid for both y on x and x on y. But estimators (and parameters themselves) are different.

See "Linear Regression in Astronomy I" ApJ 1990 Isobe, Feigelson, Akritas, Babu
`http://adsabs.harvard.edu/full/1990ApJ...364..104I`

Outline

Censoring in Astronomy

- we do not observe y, but know that y is below or above a certain value
- sometimes called non–detections
- censoring often due to limiting magnitude in astronomy
- notation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

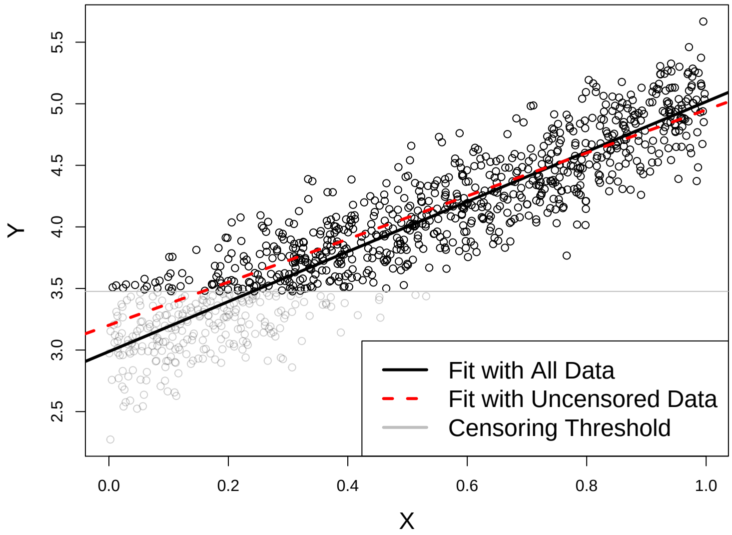
$$c_i = \begin{cases} 1 & : y_i < t_i \\ 0 & : y_i \geq t_i \end{cases}$$

$$y_i^* = \begin{cases} y_i & : c_i = 0 \\ t_i & : c_i = 1 \end{cases}$$

where we observe (x_i, y_i^*, c_i)

Censoring of data is studied extensively in the field of `Survival Analysis`

Simulation of Censoring



Conclusion: Censoring biases slope estimate towards 0.

Outline

Good Reference

THE ASTROPHYSICAL JOURNAL, 665:1489–1506, 2007 August 20

SOME ASPECTS OF MEASUREMENT ERROR IN LINEAR REGRESSION OF ASTRONOMICAL DATA

BRANDON C. KELLY
Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721; bkelly@as.arizona.edu
Received 2006 December 7; accepted 2007 May 8

ABSTRACT

I describe a Bayesian method to account for measurement errors in linear regression of astronomical data. The method allows for heteroscedastic and possibly correlated measurement errors and intrinsic scatter in the regression relationship. The method is based on deriving a likelihood function for the measured data, and I focus on the case when the intrinsic distribution of the independent variables can be approximated using a mixture of Gaussian functions. I generalize the method to incorporate multiple independent variables, nondetections, and selection effects (e.g., Malmquist bias). A Gibbs sampler is described for simulating random draws from the probability distribution of the parameters, given the observed data

- **next few slides**: overview (simplified version) of Kelly
  - Bayesian hierarchical model
  - accounts for heteroskedastic measurement error in x,y, intrinsic scatter, censoring, multivariate regression, truncation
  - code available in `IDL` (astronomy programming language)
- **project idea**: code model in `Stan` with an `R` or `Python` wrapper

Gaussian Mixture Model for X

- $\Psi = (\vec{\pi}, \vec{\mu}, \vec{\sigma})$ models distribution of x

$$f(x|\Psi) = \sum_{k=1}^{K} \pi_j N(x|\mu_k, \sigma_k^2)$$

  - Ψ are nuisance parameters, without much physical meaning
  - with K large, very flexible
  - Gaussian mixture is computationally advantageous
- $z_i = x_i + \delta_i$ where $\delta_i \sim N(0, \sigma_{\delta i}^2)$
- $f(z_i|\Psi) = \sum \pi_k N(z_i|\mu_k, \sigma_k^2 + \sigma_{\delta i}^2)$

Likelihood for (z, y)

- $\theta = (\sigma, \beta_0, \beta_1)$ are regression parameters
- the response is

  $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

  where $\epsilon_i \sim N(0, \sigma^2 + \sigma_{yi}^2)$ (intrinsic scatter + error)
- the likelihood $f((z_i, y_i)|\theta, \Psi) =$

$$\sum_{k=1}^{K} \pi_k N\left(\begin{pmatrix} \mu_k \\ \beta_0 + \beta_1 \mu_k \end{pmatrix}, \begin{pmatrix} \sigma_k^2 + \sigma_{\delta i}^2 & \beta_1 \sigma_k^2 \\ \beta_1 \sigma_k^2 & \beta_1^2 \sigma_k^2 + \sigma^2 + \sigma_{yi}^2 \end{pmatrix}\right)$$

Censoring

- observation i is censored if $y_i < t_i$

$$c_i = \begin{cases} 1 & : y_i < t_i \\ 0 & : y_i \geq t_i \end{cases}$$

$$y_i^* = \begin{cases} y_i & : c_i = 0 \\ t_i & : c_i = 1 \end{cases}$$

- the likelihood for data with censoring is

$$p(z_i, y_i^*, c_i) = p(z_i, y_i|\theta, \Psi)^{1-c_i} \left(\int_{-\infty}^{t_i} p(z_i, y|\theta, \Psi) dy \right)^{c_i}$$

- special cases
  - $t_i = t$ for all i, censoring at same level
  - $t_i = -\infty$ for all i, then $c_i = 0 \; \forall i$ and $p(z_i, y_i^*, c_i|\theta, \Psi) = p(z_i, y_i|\theta, \Psi)$

Overview of Priors, Computation, Results

- bayesian: priors on (θ, Ψ)
- need proper priors on Ψ to avoid improper posteriors
- gibbs sampler for generating posterior samples

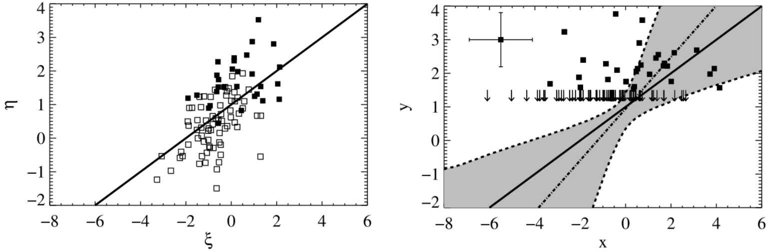


FIG. 8.—Distribution of η and ξ (*left*) and the measured values of y and x (*right*), from a simulated censored data set of $n = 50$ data points, $\sigma_x \sim \tau$, and $\sigma_y \sim \sigma$ (see § 7.2). In the plot of η and ξ, the filled squares denote the values of ξ and η for the detected data points, and the open squares denote the values of ξ and η for the undetected data points. The solid line in both plots is the true regression line. In the plot of y and x, the squares denote the measured values of x and y for the detected data points, and the arrows denote the "upper limits" on y for the undetected data points. The fictitious data point with error bars illustrates the median values of the error bars. The dash-dotted line shows the best-fit regression line, as calculated from the posterior median of α and β, and the shaded region defines the approximate 95% (2 σ) pointwise confidence intervals on the regression line. The true values of the regression line are contained within the 95% confidence intervals.

Source: "Some aspects of measurement error in linear regression of astronomical data" Kelly ApJ 2007
`http://iopscience.iop.org/article/10.1086/519947/meta`

Related Possible Projects

Density Estimation and Classification with Measurement Error

The Annals of Applied Statistics
2011, Vol. 5, No. 2B, 1657–1677
DOI: 10.1214/10-AOAS439

EXTREME DECONVOLUTION: INFERRING COMPLETE DISTRIBUTION FUNCTIONS FROM NOISY, HETEROGENEOUS AND INCOMPLETE OBSERVATIONS

BY JO BOVY[1], DAVID W. HOGG[1,2] AND SAM T. ROWEIS[3]

New York University

We generalize the well-known mixtures of Gaussians approach to density estimation and the accompanying Expectation–Maximization technique for finding the maximum likelihood parameters of the mixture to the case where each data point carries an individual d-dimensional uncertainty covariance

THE ASTROPHYSICAL JOURNAL, 729:141 (20pp), 2011 March 10 doi:10.1088/0004-637X/729/2/141

THINK OUTSIDE THE COLOR BOX: PROBABILISTIC TARGET SELECTION AND THE *SDSS-XDQSO* QUASAR TARGETING CATALOG

JO BOVY[1], JOSEPH F. HENNAWI[2], DAVID W. HOGG[1,2], ADAM D. MYERS[2,3], JESSICA A. KIRKPATRICK[4], DAVID J. SCHLEGEL[4], NICHOLAS P. ROSS[4], ERIN S. SHELDON[5], IAN D. MCGREER[6], DONALD P. SCHNEIDER[7], AND BENJAMIN A. WEAVER[1]
[1] Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place, New York, NY 10003, USA; jo.bovy@nyu.edu
[2] Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany
[3] Department of Astronomy, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[4] Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 92420, USA
[5] Brookhaven National Laboratory, Upton, NY 11973, USA
[6] Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, AZ 85721, USA
[7] Department of Astronomy and Astrophysics, The Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802, USA
Received 2010 November 24; accepted 2011 January 12; published 2011 February 22

ABSTRACT

We present the *SDSS-XDQSO* quasar targeting catalog for efficient flux-based quasar target selection down to the faint limit of the Sloan Digital Sky Survey (*SDSS*) catalog, even at medium redshifts ($2.5 \lesssim z \lesssim 3$) where the stellar contamination is significant. We build models of the distributions of stars and quasars in flux space down to the flux limit by applying the extreme-deconvolution method to estimate the underlying density. We convolve this density with the flux uncertainties when evaluating the probability that an object is a quasar. This approach results

“Extreme deconvolution . . .”

- density estimation with measurement error
- uses Gaussian mixture model
- emphasis on fast computation

“Think outside . . .”

- uses “Extreme deconvolution . . “ to estimate class densities
- class density used to construct classifier