STATISTICS
TEXAS A&M UNIVERSITY

Regression in Astronomy II

October 26, 2015

Intrinsic Scatter Continued

Cramer–Rao Bound and Fisher Information

Measurement Error in x

Outline

Intrinsic Scatter Continued

Cramer–Rao Bound and Fisher Information

Measurement Error in x

Intrinsic Scatter + Measurement Error

Intrinsic Scatter and y (Normal) Measurement Error

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

where

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

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

$\beta = (\beta_0, \beta_1)$ and σ^2 are unknown parameters.

Maximum Likelihood with Intrinsic Scatter

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

- No closed form solution.
- But at fixed σ, closed form solution.

Minimization Procedure

Define $W(\sigma^2)$ to be diagonal matrix with $W(\sigma^2)_{ii} = (\sigma_i^2 + \sigma^2)^{-1}$.

$$\widehat{\sigma}^2, \widehat{\beta}_0, \widehat{\beta}_1 = \underset{(\sigma^2,\beta_0,\beta_1)}{\operatorname{argmin}} \sum_{i=1}^{n} \log(\sigma^2 + \sigma_i^2) + (Y - X\beta)^T W(\sigma^2)(Y - X\beta)$$

So

$$\begin{aligned}\widehat{\sigma}^2 &= \underset{\sigma^2}{\operatorname{argmin}} \min_{\beta_0,\beta_1} \sum_{i=1}^{n} \log(\sigma^2 + \sigma_i^2) + (Y - X\beta)^T W(\sigma^2)(Y - X\beta) \\ &= \underset{\sigma^2}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} \log(\sigma^2 + \sigma_i^2) + (Y - X\widehat{\beta}(\sigma^2))^T W(\sigma^2)(Y - X\widehat{\beta}(\sigma^2))}_{\equiv SSML(\sigma^2)}\end{aligned}$$

where

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

- Grid search on σ to find $\widehat{\sigma}$.
- $\widehat{\beta} = \widehat{\beta}(\widehat{\sigma})$.

"χ^2 Minimization" for Estimating Parameters

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{(\sigma^2 + \sigma_i^2)}$$

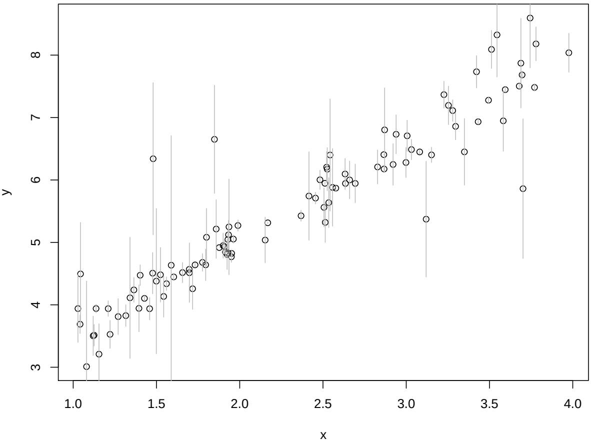
- One could minimize chi–squared:

$$\widehat{\sigma}^2, \widehat{\beta_0}, \widehat{\beta_1} = \operatorname*{argmin}_{\sigma^2, \beta_0, \beta_1} \chi^2$$

- Computational issue is same as with ML, but at fixed σ^2 easy. So compute:

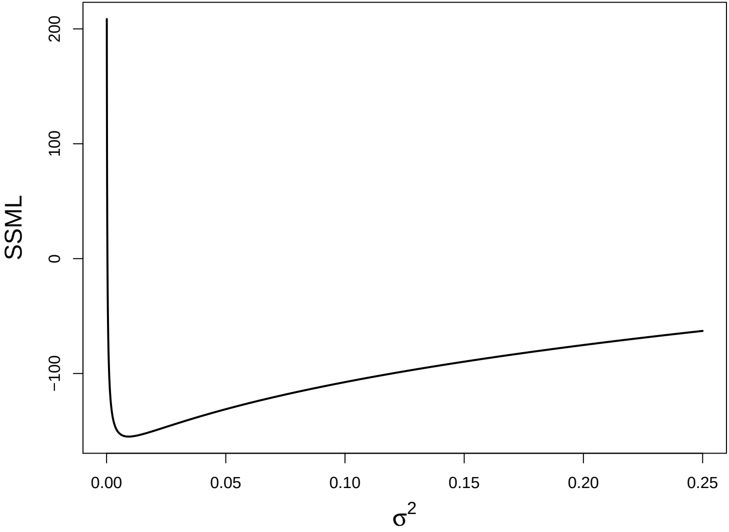
$$\widehat{\sigma}^2 = \operatorname*{argmin}_{\sigma^2} \underbrace{\min_{\beta_0, \beta_1} \chi^2}_{\equiv SS\chi^2(\sigma^2)}$$

Simulation



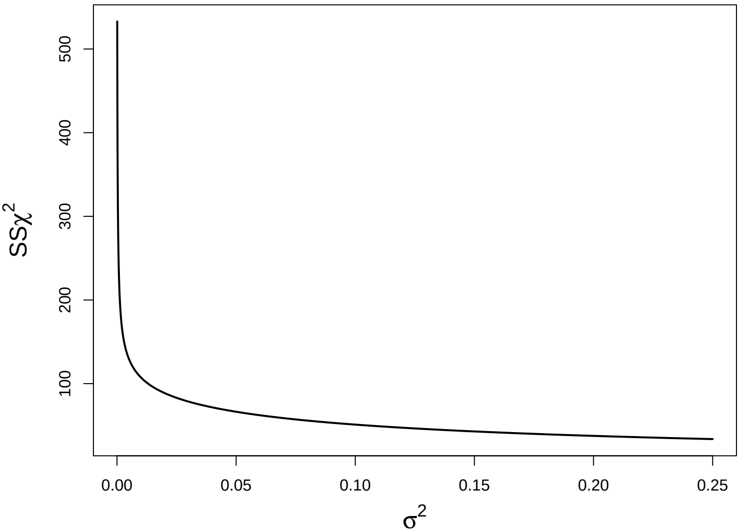
Parameters: $\beta_0 = 2$, $\beta_1 = 1.5$, $\sigma^2 = 0.1^2$
Data: $\{(y_i, x_i, \sigma_{yi})\}_{i=1}^{n}$

Maximum Likelihood



Looks reasonable.

Chi–Squared



$\chi^2 \to 0$ as $\sigma \to \infty$. Not good.

Quantify Uncertainty on ML Estimates

The maximum likelihood estimate for the parameters is

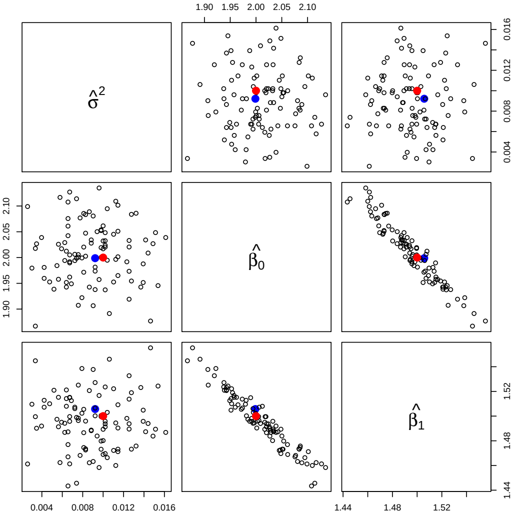
$$(\widehat{\sigma}^2, \widehat{\beta}_0, \widehat{\beta}_1) = (0.0092, 1.9988, 1.5057)$$

- Since this is simulation we know the truth $(0.01, 2, 1.5)$.
- In practice, need to report uncertainty on our estimates.

Sampling Distribution

- Generate the data many times.
- Calculate $(\widehat{\sigma}^2, \widehat{\beta}_0, \widehat{\beta}_1)$ each time.
- Calculate variance of resulting data.

Empirical Sampling Distribution of ML Estimator



Red point is truth. Blue point is our 1 actual sample ML estimates.

Variance of $(\widehat{\sigma}^2, \widehat{\beta})$

Variance (based on simulation) is:

$$\text{Var}\left((\widehat{\sigma}^2, \widehat{\beta})\right) = \begin{pmatrix} 9.46 \times 10^{-6} & -1.76 \times 10^{-6} & 1.27 \times 10^{-6} \\ -1.76 \times 10^{-6} & 3.31 \times 10^{-3} & -1.23 \times 10^{-3} \\ 1.27 \times 10^{-6} & -1.23 \times 10^{-3} & 4.97 \times 10^{-4} \end{pmatrix}$$

So

$$sd(\widehat{\sigma}^2) = \sqrt{\text{Var}\left(\widehat{\sigma}^2\right)} \approx \sqrt{9.46 \times 10^{-6}} \approx 0.0031$$
$$sd(\widehat{\beta}_0) = \sqrt{\text{Var}\left(\widehat{\beta}_0\right)} \approx \sqrt{3.31 \times 10^{-3}} \approx 0.0576$$
$$sd(\widehat{\beta}_1) = \sqrt{\text{Var}\left(\widehat{\beta}_1\right)} \approx \sqrt{4.97 \times 10^{-4}} \approx 0.0223$$

Simulation Has Major Weaknesses:

- What about $\beta \neq (2, 1.5)^T$ or $\sigma^2 \neq 0.1^2$?
- Since I don't know β or σ, how can this be used?

Analytic Method

$$
\begin{aligned}
\text{Var}\ (\widehat{\beta}) &= \text{Var}\ \left((\widehat{\sigma}, \widehat{\beta}_0, \widehat{\beta}_1)\right) \\
&= \text{Var}\ \left(\underset{(\sigma^2, \beta_0, \beta_1)}{\text{argmin}} \sum_{i=1}^{n} \left(\log(\sigma^2 + \sigma_i^2) + \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{(\sigma^2 + \sigma_i^2)} \right) \right) \\
&= \text{ummm}\ .\ .\ .
\end{aligned}
$$

Need more powerful statistical tools.

Outline

Intrinsic Scatter Continued

Cramer–Rao Bound and Fisher Information

Measurement Error in x

Selecting an Estimator

There are an infinite number of estimators for any problem:

$$\widehat{\theta}_1 = \text{maximum likelihood estimator}$$
$$\widehat{\theta}_2 = \text{“chi–squared minimization”}$$
$$\widehat{\theta}_3 = (\widehat{\theta}_1 + \widehat{\theta}_2)/2$$
$$\widehat{\theta}_4 = \text{“first estimate } \theta_1 \text{ using chi–squared, then } \ldots \text{”}$$
$$\vdots$$

Which is best?

Cramer–Rao Bound

Under regularity conditions on the model, for any (approximately) unbiased estimator $\widehat{\theta}$

$$\mathrm{Var}\,(\widehat{\theta}) \succeq I(\theta)^{-1}$$

where

$$I(\theta) = \mathbb{E}\left[\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right]$$

is called the Fisher information matrix.

Significance: Among all possible (approximately) unbiased estimators, there is a best possible (ie lowest variance) performance.

Variance of Maximum Likelihood Estimator

Theorem: Under regularity conditions, the maximum likelihood estimator acheives this bound ie

$$\text{Var}_{\theta}(\widehat{\theta}_{ML}) \approx I(\theta)^{-1}$$

Significance: You can't do better than maximum likelihood (when n is large and model satisfies "regularity" conditions).

Estimating $I(\theta)^{-1}$

$I(\theta)^{-1}$ is unknown, but we can estimate it:

$$\begin{aligned} I(\theta) &= \mathbb{E}\left[\left(\frac{d}{d\theta}\log f(X|\theta)\right)^2\right] \\ &= -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log f(X|\theta)\right] \\ &\approx -\frac{d^2}{d\theta^2}\log f(X|\theta)\Big|_{\theta=\widehat{\theta}_{ML}} \\ &\equiv J(\widehat{\theta}_{ML}) \end{aligned}$$

Significance: Not only is maximum likelihood the best, we can quantify its performance even when there is no closed form solution to maximizing the likelihood (by computing $J(\widehat{\theta}_{ML})$).

Some Caveats

- $n \approx p$: Maximum likelihood theory is asymptotic so not informative at small sample sizes or where the number of parameters is similar to number of samples.
- Nonparametric and semi–parametric models: Nadaraya–Watson, Kernel Density Estimators. Maximum likelihood does not work here.
- Bayesian Arguments: The Bayesian says: The sampling distribution is not what's important.
- Prior Information: What if I have pre–existing notions about the value of θ?
- Model Misspecification: What if I have an approximate model?

Application to Intrinsic Scatter

- $\widehat{\theta}_{ML} = (\widehat{\sigma}^2, \widehat{\beta}_0, \widehat{\beta}_0)$
- Var $(\widehat{\theta}_{ML}) \approx J(\widehat{\theta}_{ML})^{-1}$.

$$J(\widehat{\theta}_{ML}) = -\begin{pmatrix} \frac{d^2 \log(f(X|\theta))}{(d\sigma^2)^2} & \frac{d^2 \log(f(X|\theta))}{d\sigma^2 d\beta} \\ \frac{d^2 \log(f(X|\theta))}{d\sigma^2 d\beta}^T & \frac{d^2 \log(f(X|\theta))}{d\beta^2} \end{pmatrix} \Bigg|_{\theta=\widehat{\theta}_{ML}}$$

$J(\widehat{\theta}_{ML})$ is the negative Hessian evaluated at $\widehat{\theta}_{ML}$. Also known as the observed information.

Computing $J(\widehat{\theta}_{ML})$

$$\log(f(X|\theta)) \propto -\frac{1}{2}\sum \log(\sigma_{yi}^2 + \sigma^2) - \frac{1}{2}(Y - X\beta)^T W(\sigma^2)(Y - X\beta)$$

So

$$\frac{d^2 \log(f(X|\theta))}{d\beta^2} = -X^T W(\sigma^2) X$$

$$\frac{d^2 \log(f(X|\theta))}{(d\sigma^2)^2} = \frac{1}{2}(\sigma_{yi}^2 + \sigma^2)^{-2} - (Y - X\beta)^T W(\sigma^2)^3 (Y - X\beta)$$

$$\frac{d^2 \log(f(X|\theta))}{d\sigma^2 d\beta} = -Y^T W(\sigma^2)^2 X + \beta^T X^T W(\sigma^2)^2 X$$

Solution

For the intrinsic scatter problem:

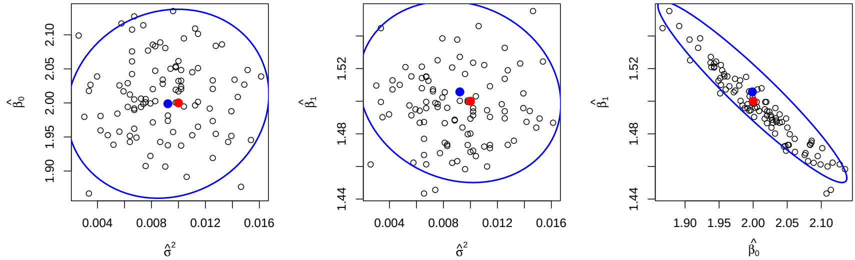
$$(\widehat{\sigma}^2, \widehat{\beta}_0, \widehat{\beta}_1) = (0.0092, 1.9988, 1.5057)$$

and the estimate of the variance is

$$\text{Var}\left((\widehat{\sigma}^2, \widehat{\beta})\right) = \begin{pmatrix} 9.36 \times 10^{-6} & 1.75 \times 10^{-5} & -9.19 \times 10^{-6} \\ 1.75 \times 10^{-5} & 3.21 \times 10^{-3} & -1.22 \times 10^{-3} \\ -9.19 \times 10^{-6} & -1.22 \times 10^{-3} & 5.16 \times 10^{-4} \end{pmatrix}$$

This is done using a single sample.

Estimate, Truth, Sampling Distribution, 95% CI



95% Confidence regions. Elliptical regions computed only from 1 sample (blue dot).

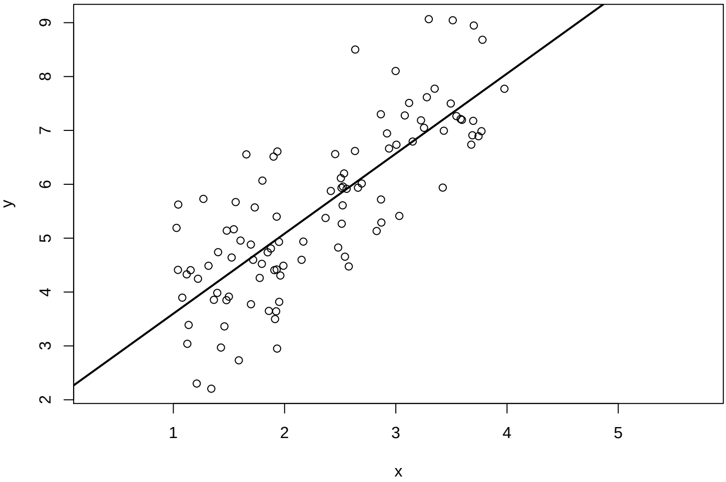
Outline

Intrinsic Scatter Continued

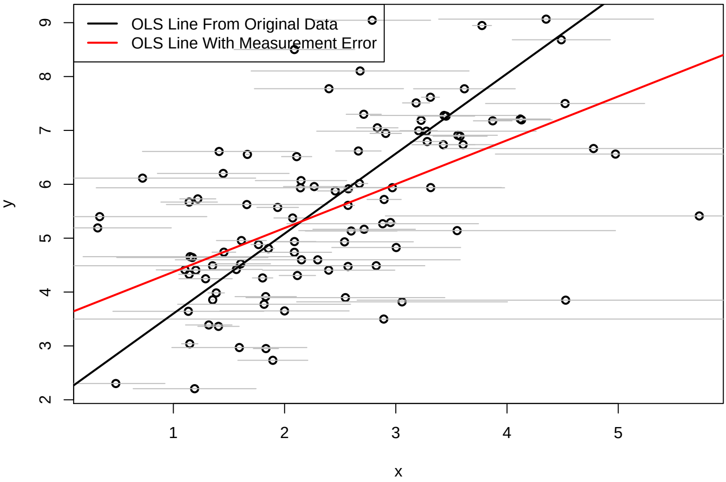
Cramer–Rao Bound and Fisher Information

Measurement Error in x

Simulation



Simulation



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

Observations

- Ignoring error in x creates bias in estimators
- This is (in many ways) worse than ignoring intrinsic scatter or photometric errors in y
  - Only increase the variance, not biased.
- Having a large sample size <u>does not</u> help with errors in x.

Essentially

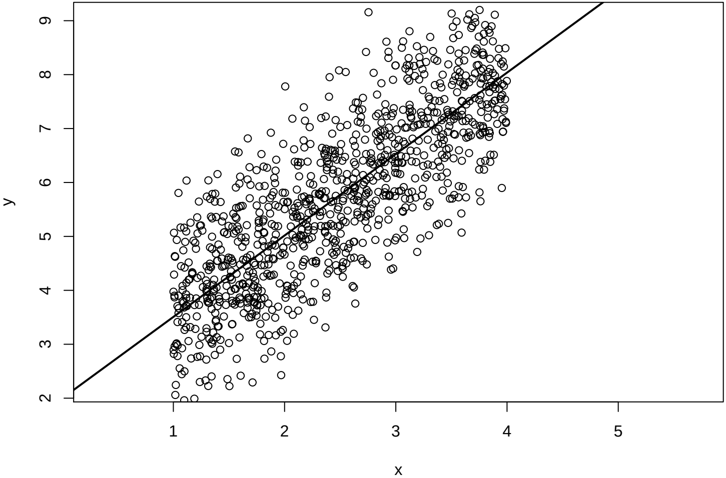
$$\lim_{n\to\infty} \widehat{\beta}_0 \not\to \beta_0$$

$$\lim_{n\to\infty} \widehat{\beta}_1 \not\to \beta_1$$

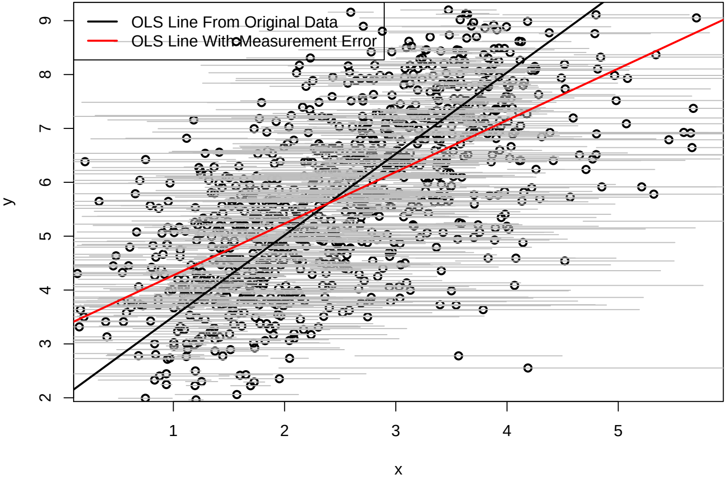
In particular

$$|\lim_{n\to\infty} \widehat{\beta}_1| < |\beta_1|$$

Simulation with Larger Sample Size



Simulation with Larger Sample Size



Overview of Solutions

Notation: Observe data $\{(y_i, w_i, \sigma_{xi})\}_{i=1}^{n}$.

$$w_i = x_i + \delta_i$$

where $\delta_i \sim N(0, \sigma_{xi}^2)$. Linear relationship between y and x ie

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

where $\epsilon_i \sim N(0, \sigma^2)$.

The x_i are unobserved, latent variables.

Some References

- Section 7.5 in textbook
- "Linear Regression for Astronomical Data with Measurement Errors and Intrinsic Scatter" Akritas [1]
- "Some aspects of measurement error in linear regression of astronomical data" Kelly [2]

Bibliography I

[1] Michael G Akritas and Matthew A Bershady.
Linear regression for astronomical data with measurement errors and intrinsic scatter.
The Astrophysical Journal, 470:706, 1996.

[2] Brandon C Kelly.
Some aspects of measurement error in linear regression of astronomical data.
The Astrophysical Journal, 665(2):1489, 2007.