STATISTICS
TEXAS A&M UNIVERSITY

Regression in Astronomy

October 20, 2015

Outline

Regression

- y is approximately some function of x

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

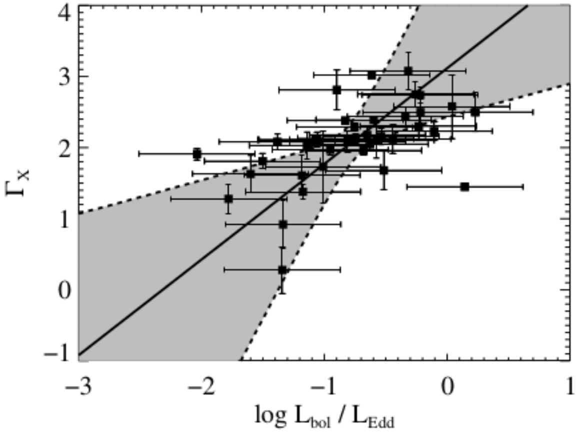
- Regression is used to:
  1. Estimate f.
  2. Quantify uncertainty in estimate of f.
  3. Predict y values for new x.

- Common to assume linear relation:

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

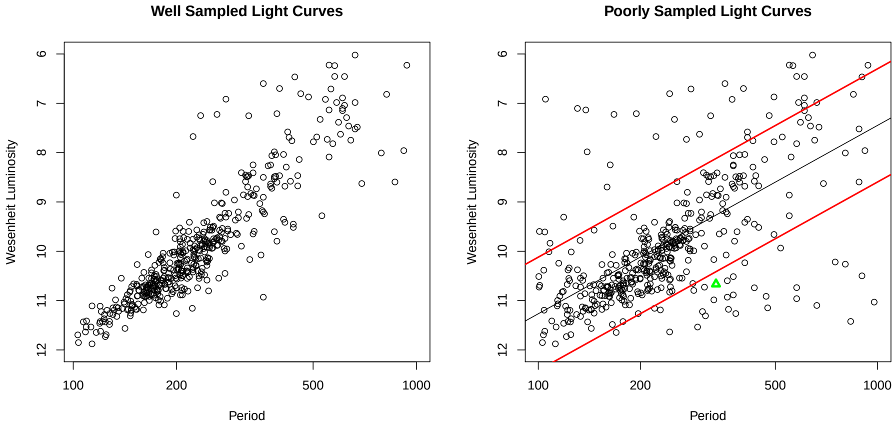
- Linear regression is often complicated in astronomy due to *measurement error* and *censoring*.

Example: Eddington Ratio (see [1])



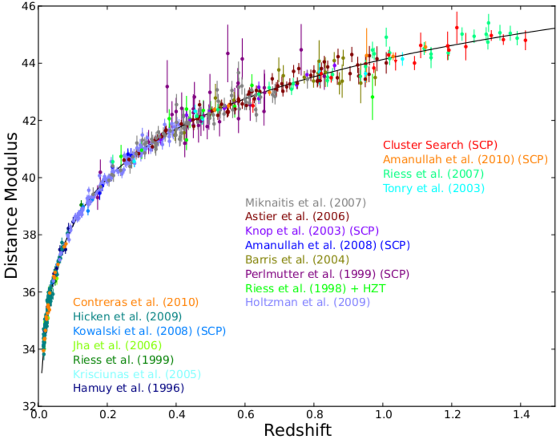
- Γ_X and $\log L_{bol}/L_{edd}$ are both measured with error (cross).
- There is `intrinsic scatter` ie even if no measurement error in x and y, still not a perfect linear relation.

Example: Period Luminosity Relation



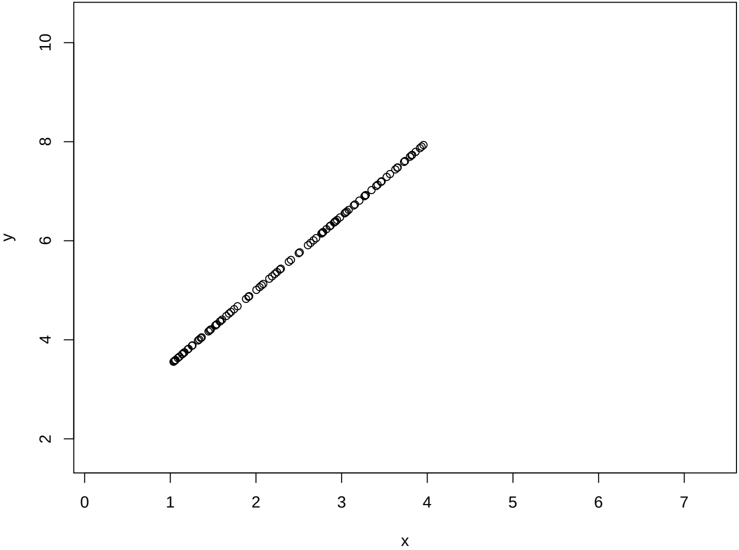
- Roughly a linear relationship between luminosity and log(*period*).
- With poorly sampled light curves, measurement error in period.

Supernovae Cosmology (from [2])

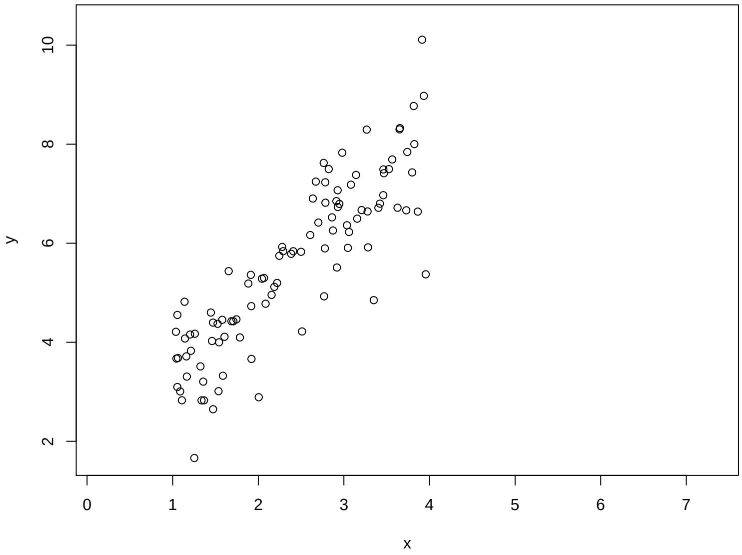


- Non–linear relationship between distance modulus and redshift.
- Equations from cosmology determine model form.

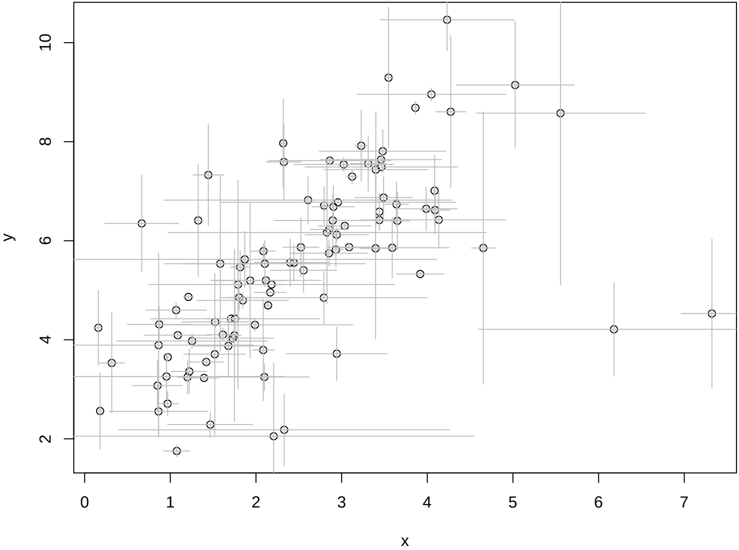
Perfect Linear Relationship



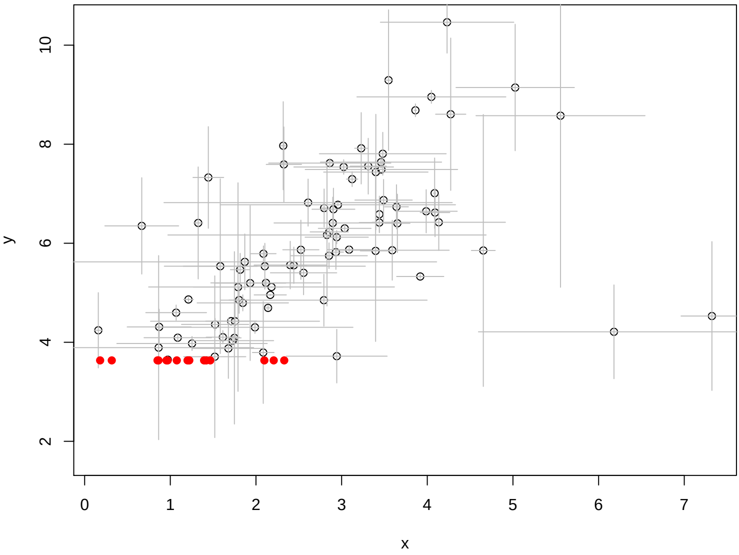
Intrinsic Scatter



Heteroskedastic Measurement Error on x and y



Censoring of x



Outline of Next Four Lectures

- Oct. 20: Background, Heteroskedasticity, Intrinsic Scatter
- Oct. 22: Errors–in–variables (measurement error in x)
- Oct. 27: Bayesian Methods for Linear Regression I
- Oct. 29: Bayesian Methods for Linear Regression II

Outline

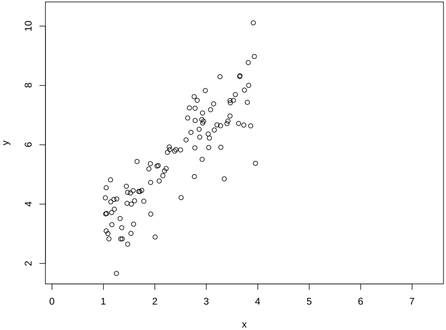
Introduction

Linear Regression Basics

Intrinsic Scatter and Heteroskedastic y Error

Ordinary Least Squares Model

- $\sigma_{xi} = \sigma_{yi} = 0$ for all i
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ 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

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

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

Linear regression is now

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

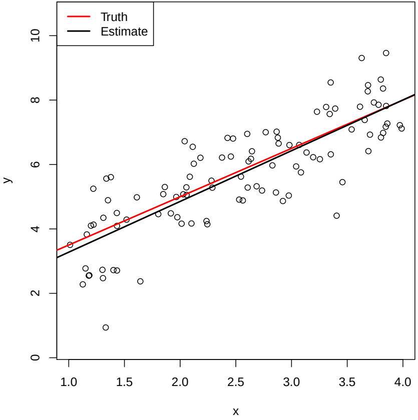
Maximum Likelihood in Matrix Form

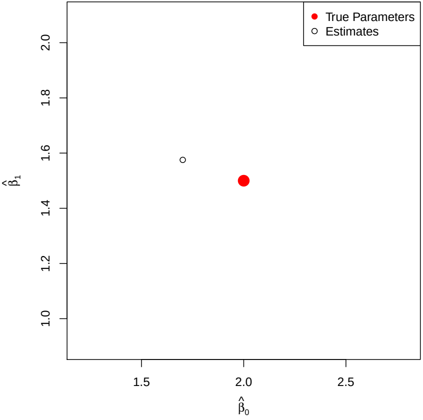
$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$
$$\widehat{\sigma}^2 = n^{-1}(Y - X\widehat{\beta})^T (Y - X\widehat{\beta})$$

Uncertainty on β

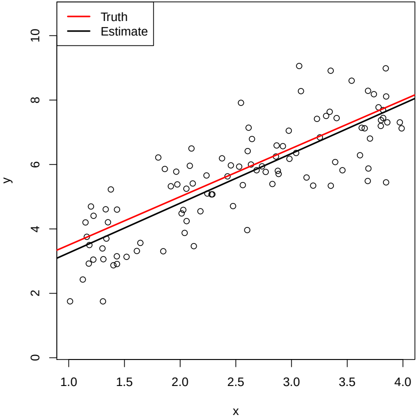
- We are in **frequentist** mode (no priors).
- Assess uncertainty with **sampling distribution**:
  1. Repeat data collection process over and over.
  2. Compute $\widehat{\beta}$ each time.
  3. Uncertainty on $\widehat{\beta}$ is some function (usually variance) of sampling distribution.

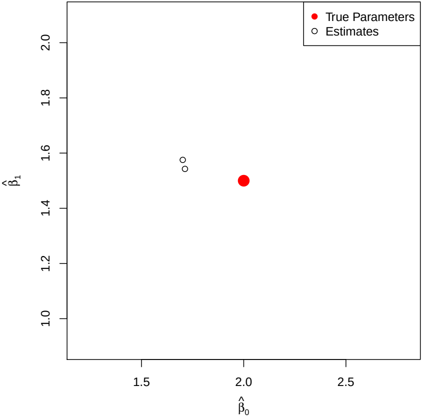
Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$



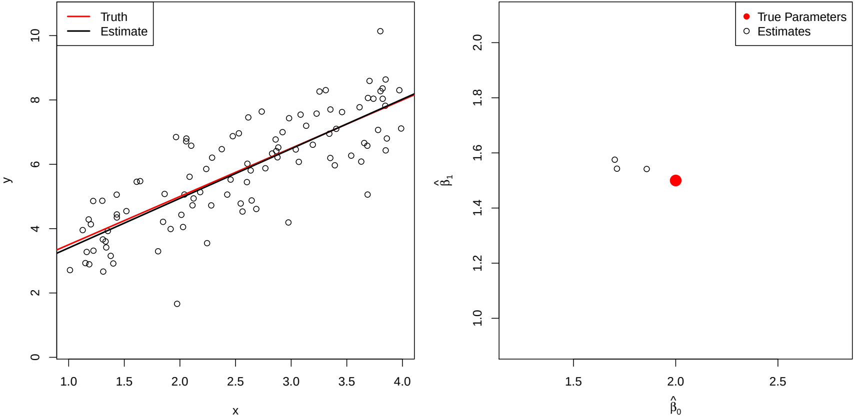


Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

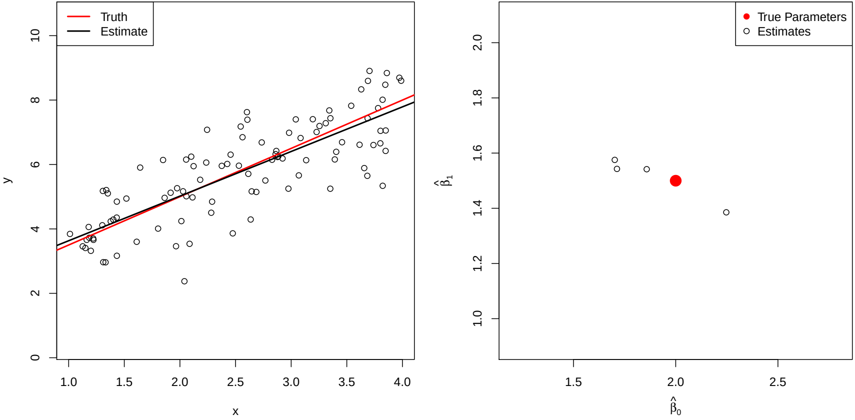




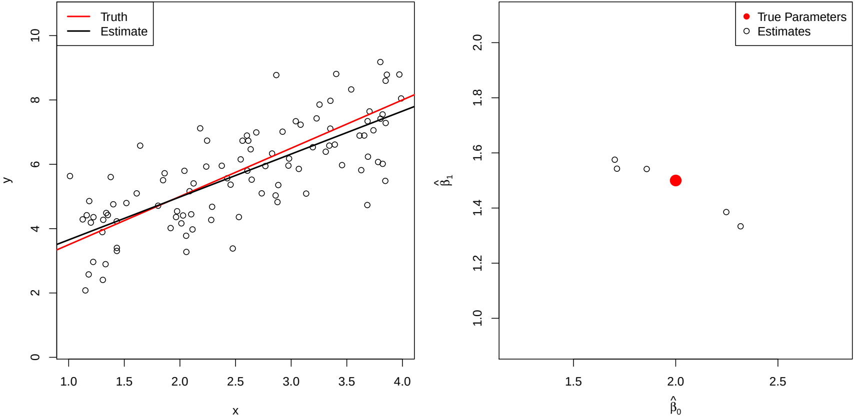
Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$



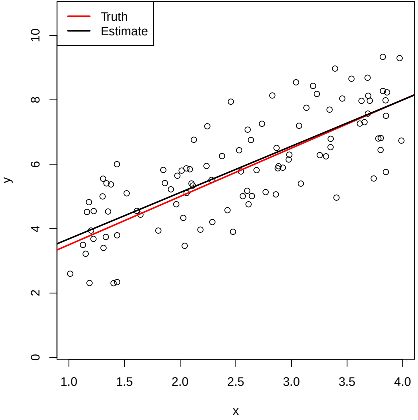
Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

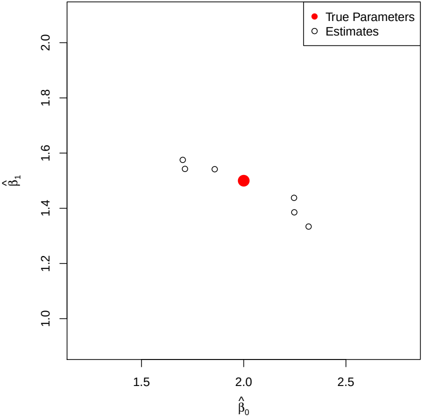


Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

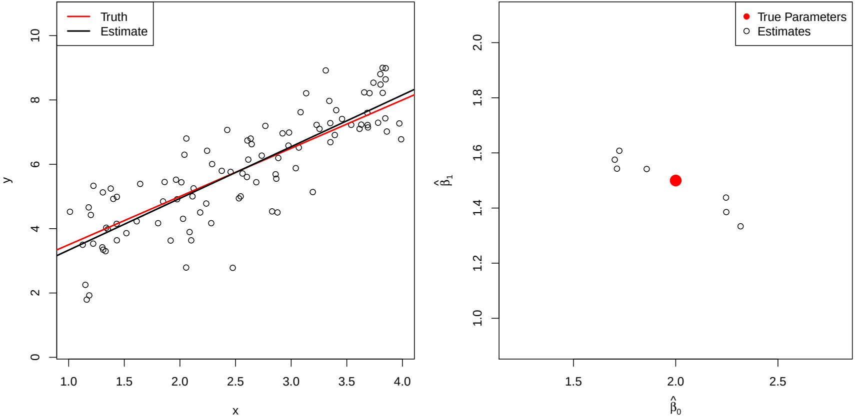


Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

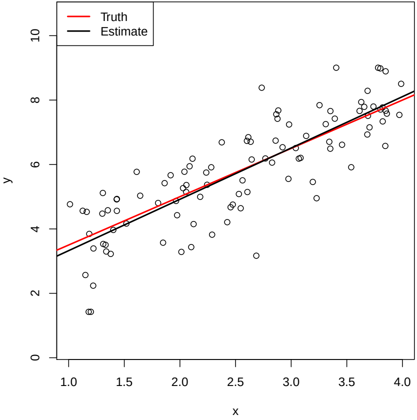


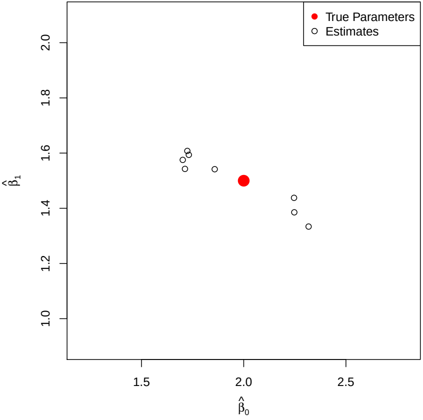


Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

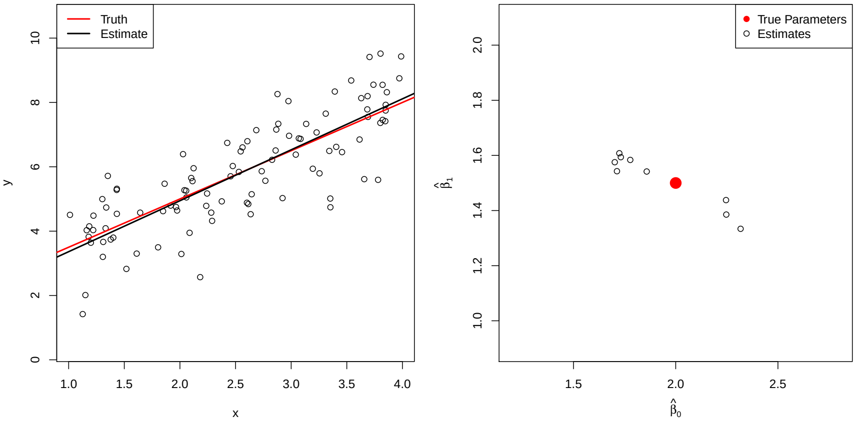


Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

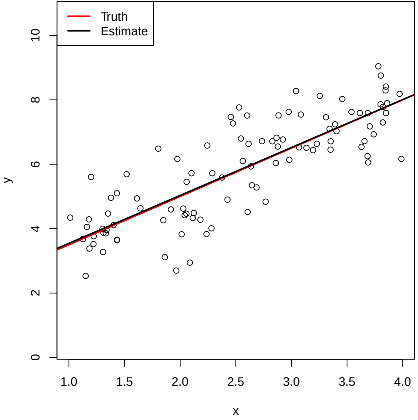


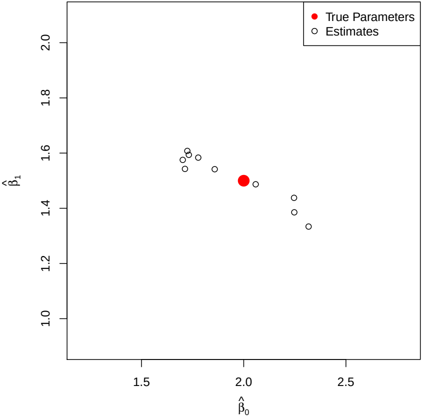


Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$



Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

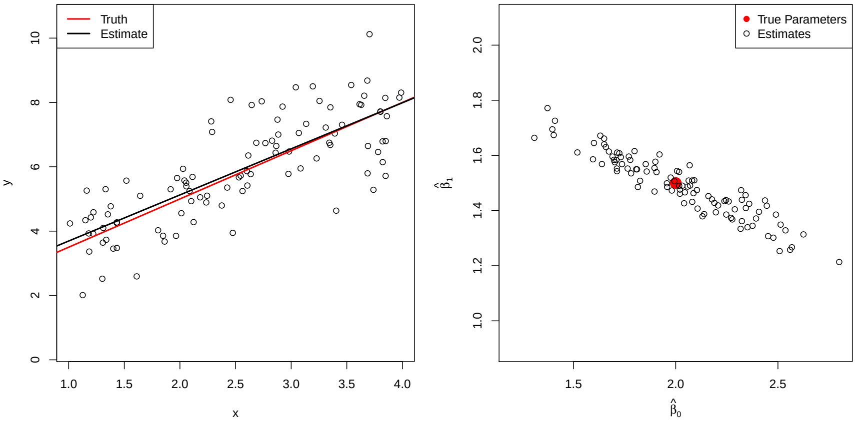




Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

Repeat 89 more times.

Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$



Covariance of β

Covariance (based on simulation) is:

$$\text{Cov}\ (\widehat{\beta}) = \begin{pmatrix} 0.080 & -0.029 \\ -0.029 & 0.012 \end{pmatrix}$$

So

$$sd(\widehat{\beta}_0) = \sqrt{\text{Var}\ (\widehat{\beta}_0)} \approx \sqrt{0.08} \approx 0.28$$
$$sd(\widehat{\beta}_1) = \sqrt{\text{Var}\ (\widehat{\beta}_1)} \approx \sqrt{0.012} \approx 0.11$$

Simulation Has Major Weaknesses:

- What about $\beta \neq (2, 1.5)^T$ or $\sigma^2 \neq 1$?
- Since I don't know β or σ^2, how can this be used?

Better Solution: Statistical Theory

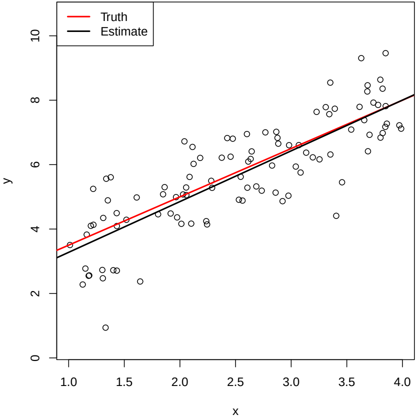
$$
\begin{aligned}
\text{Var}\ (\widehat{\beta}) &= \text{Var}\ ((X^TX)^{-1}X^TY) \\
&= \text{Var}\ ((X^TX)^{-1}X^T(X\beta+\epsilon)) \\
&= \text{Var}\ (\beta+(X^TX)^{-1}X^T\epsilon) \\
&= (X^TX)^{-1}X^T\text{Var}\ (\epsilon)X(X^TX)^{-1}) \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}
$$

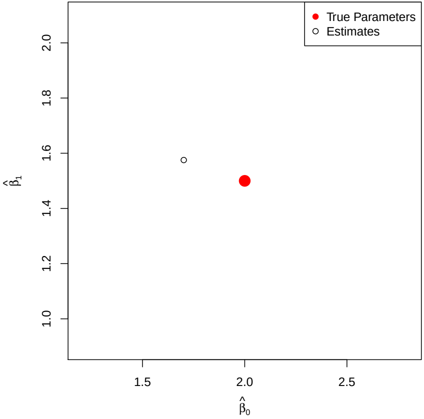
So

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

Variances for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are derived from this. n is “built–into” X^TX.

For First Simulation Run





$$\widehat{\beta} = \begin{pmatrix} 1.70 \\ 1.58 \end{pmatrix} \qquad \widehat{\text{Var}}\left(\widehat{\beta}\right) = \begin{pmatrix} 0.087 & -0.030 \\ -0.030 & 0.012 \end{pmatrix}$$

Weighted Least Squares

- Intrinsic scatter is 0.
- $\sigma_{xi} = 0$ for all i.
- $\sigma_{yi} \neq 0$

In statistics this is called heteroskedastic measurement error.

Statistical Model:

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

where

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

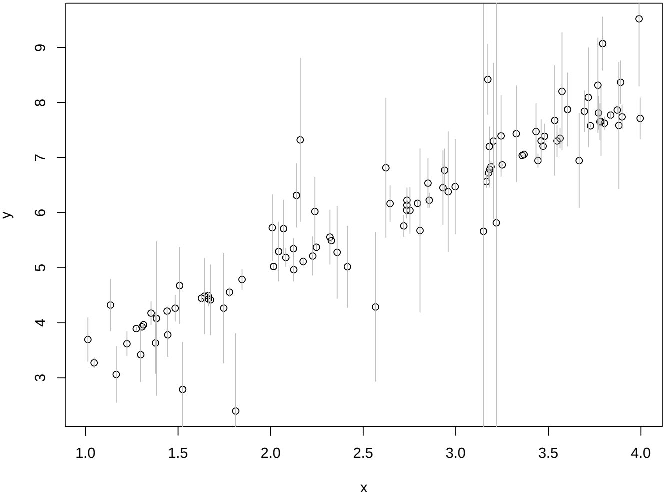
where Σ is a diagonal matrix with $\Sigma_{ii} = \sigma_{yi}^2$.

(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



This model 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

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

Outline

Intrinsic Scatter + Measurement Error

- First model (OLS) covered intrinsic scatter, but no measurement error in y.
- Second model (WLS) covered 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

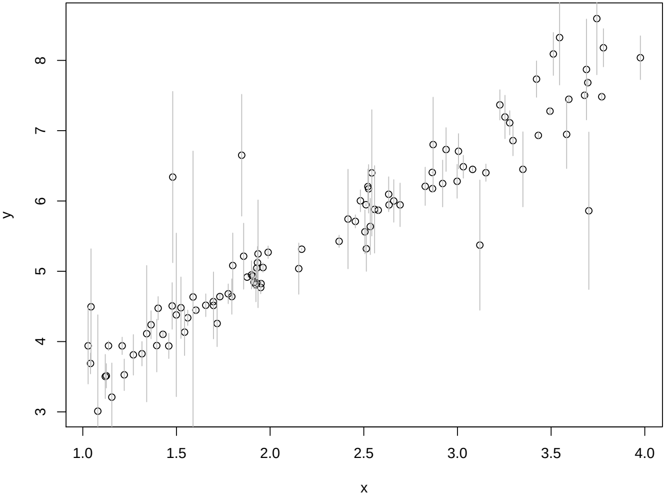
$$\widehat{\beta}(W) = (X^T W X)^{-1} X^T W Y.$$

Possible Weight Matrices:

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

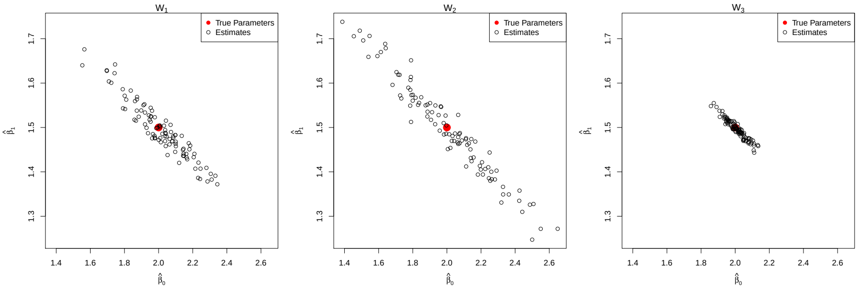
Recall W_3 is not known because σ^2 is unknown.

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



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

Sampling Distributions



W_3 is best, but it depends on σ which is unknown.

Maximum Likelihood with Intrinsic Scatter

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

- No closed form solution.
- But at fixed σ, closed form solution.
- Evaluate likelihood at each σ in grid.
- Choose value of σ which maximizes likelihood.

Bibliography I

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

[2] N Suzuki, D Rubin, C Lidman, G Aldering, R Amanullah, K Barbary, LF Barrientos, J Botyanszki, M Brodwin, N Connolly, et al.
The hubble space telescope cluster supernova survey. v. improving the dark-energy constraints above z¿ 1 and building an early-type-hosted supernova sample.
The Astrophysical Journal, 746(1):85, 2012.