STATISTICS
TEXAS A&M UNIVERSITY

# Sampling Distributions, Fisher Information, and MLEs

February 26, 2018

Standard Regression Model

Intrinsic Scatter and Heteroskedastic $y$ Error

MLEs and Fisher Information

# Outline

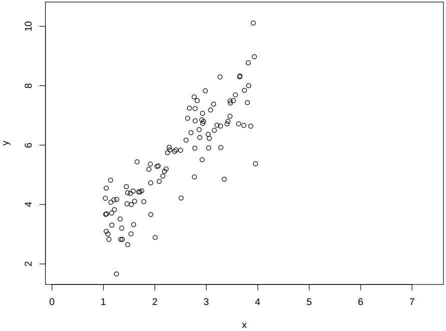
Standard Regression Model

Intrinsic Scatter and Heteroskedastic $y$ Error

MLEs and Fisher Information

# Ordinary Least Squares Model

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$
- Parameters: $(\sigma^2, \beta_0, \beta_1)$.



# 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

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

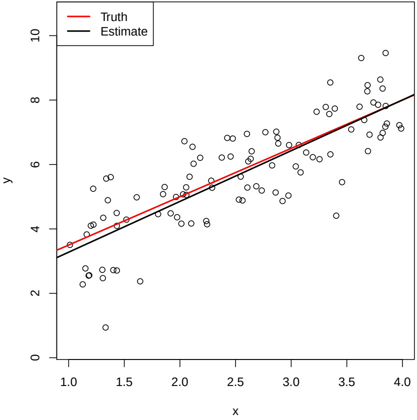
Maximum Likelihood in Matrix Form

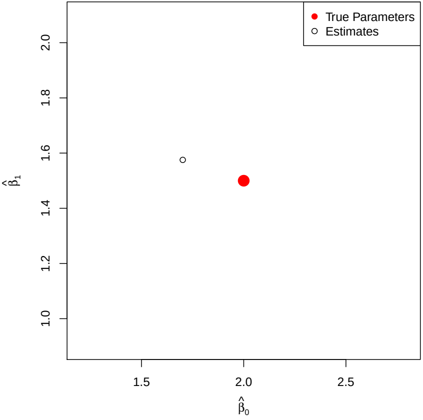
$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$
$$\widehat{\sigma}^2 = n^{-1}(Y - X\widehat{\beta})^T (Y - X\widehat{\beta})$$

# Uncertainty on $\beta$

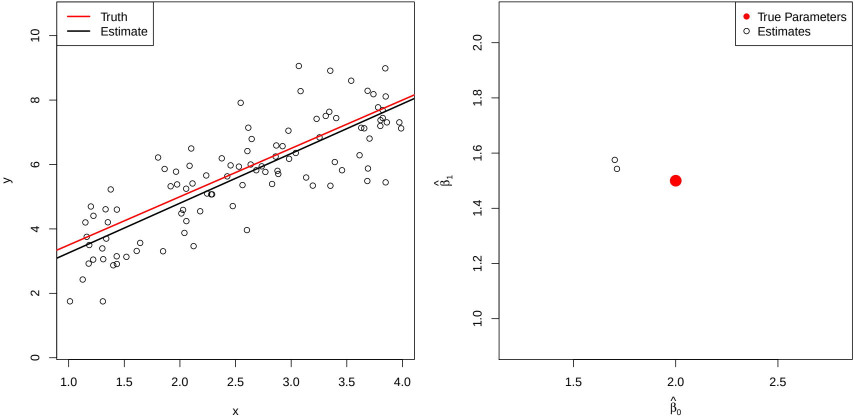
- We are in **frequentist** mode (no priors).
- Assess uncertainty with **sampling distribution**:
    1. Repeat data collection process over and over.
    2. Compute $\widehat{\beta}$ each time.
    3. Uncertainty on $\widehat{\beta}$ is some function (usually variance) of sampling distribution.

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

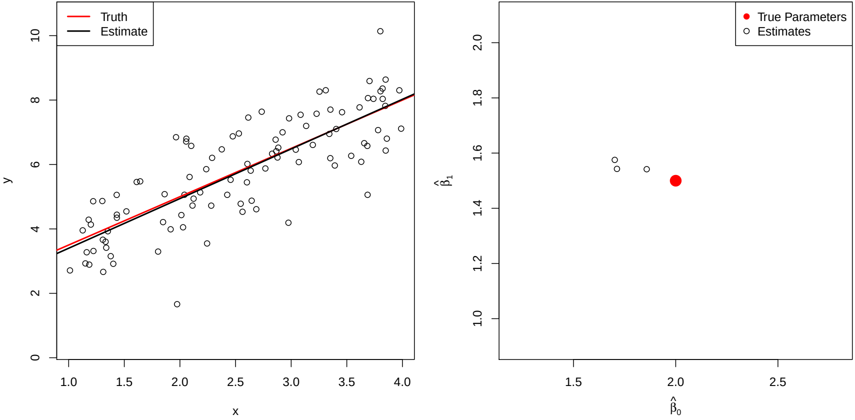


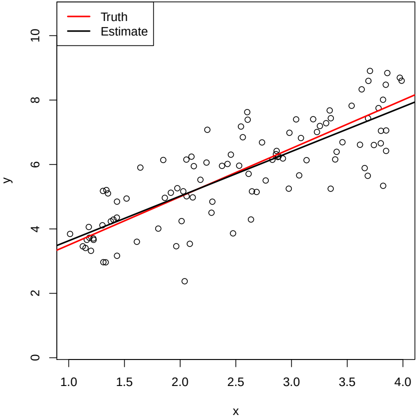
# Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

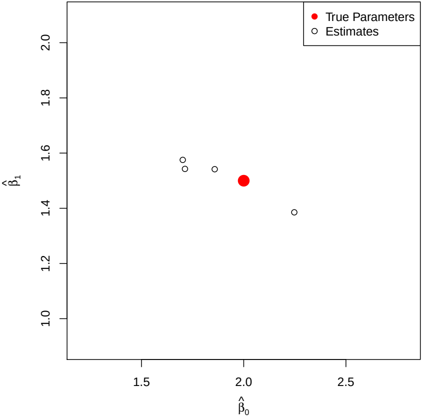


# Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

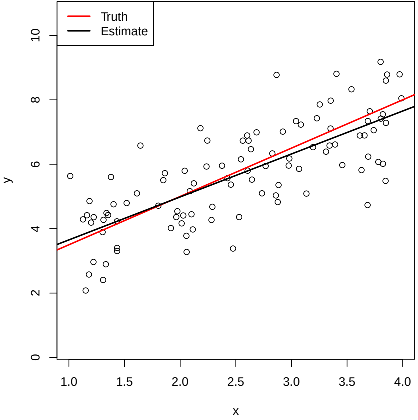


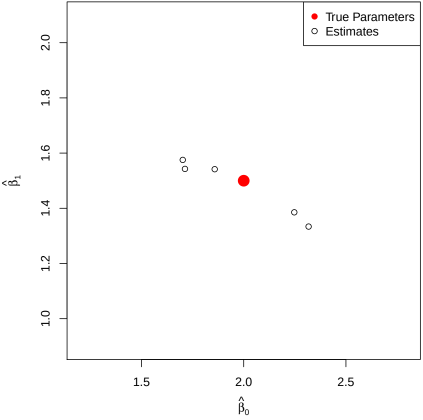
# Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$



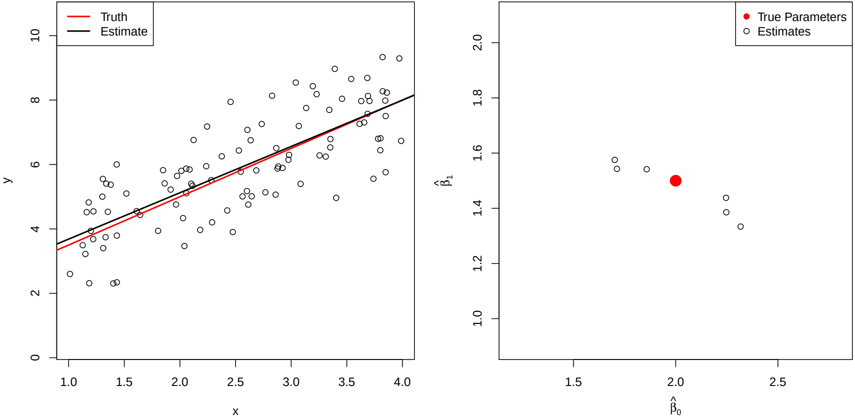



# Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

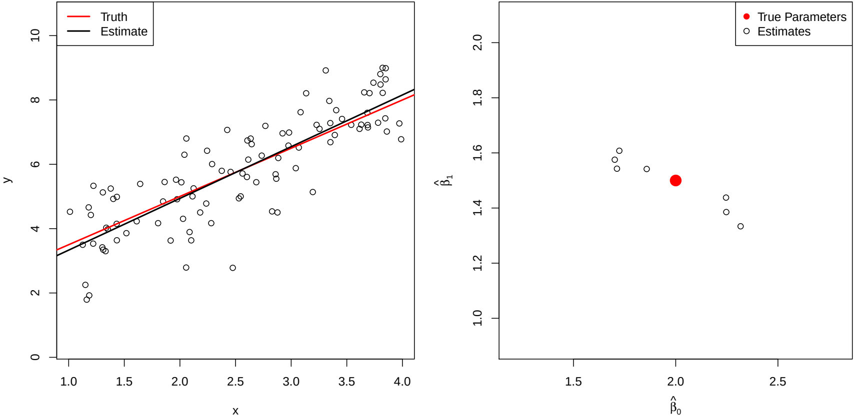


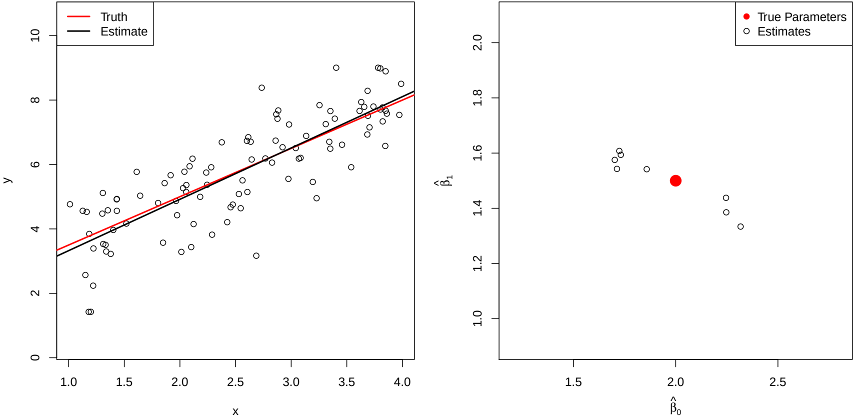
# Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$



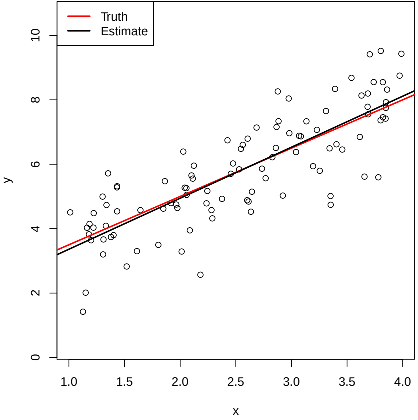
# Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

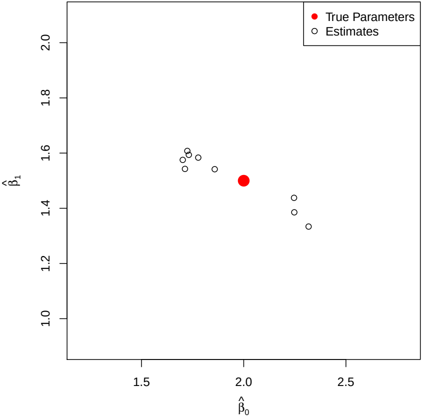


# Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$

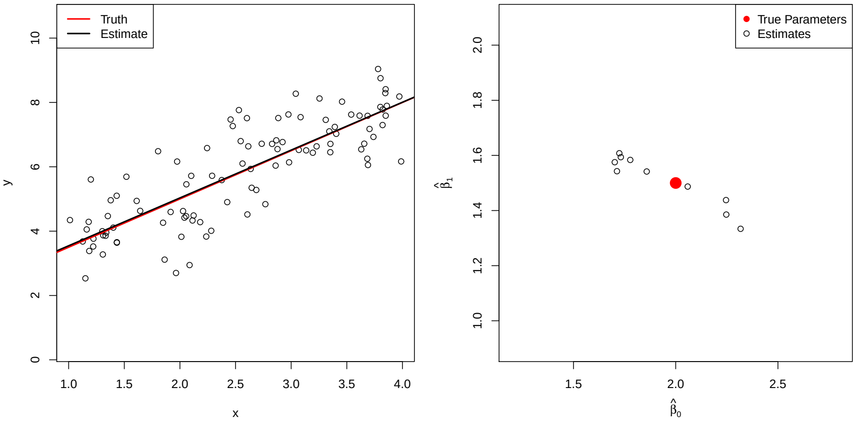


# Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$



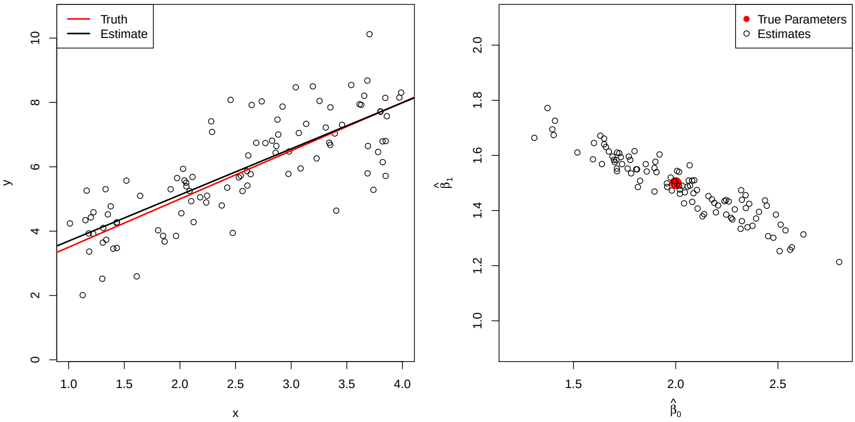
# Example: $\beta = (2, 1.5)^T$, $\sigma^2 = 1$



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

Repeat 89 more times.

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



# Covariance of $\beta$

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 $\beta$ or $\sigma^2$, how can this be used?

# Better Solution: Statistical Theory

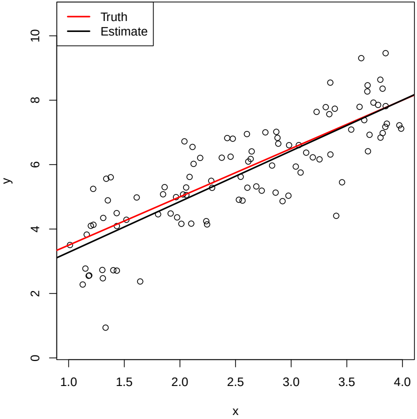
$$\begin{aligned}
\text{Var}\ (\widehat{\beta}) &= \text{Var}\ ((X^T X)^{-1} X^T Y) \\
&= \text{Var}\ ((X^T X)^{-1} X^T (X\beta + \epsilon)) \\
&= \text{Var}\ (\beta + (X^T X)^{-1} X^T \epsilon) \\
&= (X^T X)^{-1} X^T \text{Var}\ (\epsilon) X (X^T X)^{-1}) \\
&= \sigma^2 (X^T X)^{-1}
\end{aligned}$$

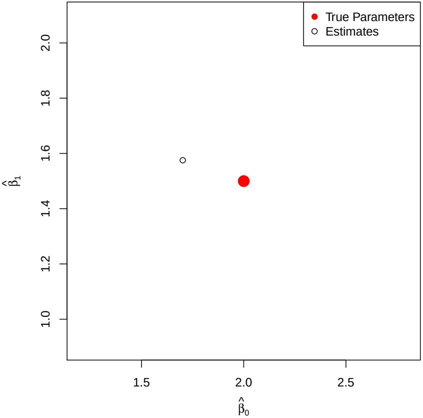
So

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

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

# For First Simulation Run




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

# Outline

# Intrinsic Scatter + Measurement Error

- Each observation may come with its own $y$ measurement error.
- Assume that error in $y$ is now due to intrinsice scatter around the line ($\sigma$) and observation specific uncertainty ($\sigma_{yi}$)
- We assume $\sigma_{yi}$ known, could loosen this assumption.

**Intrinsic Scatter and $y$ (Normal) Measurement Error**

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

where

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

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

$\beta$ and $\sigma$ are unknown parameters.

# General Weighted Least Squares Estimators

- Let $W$ be a diagonal weight matrix.
- Consider estimators of the form

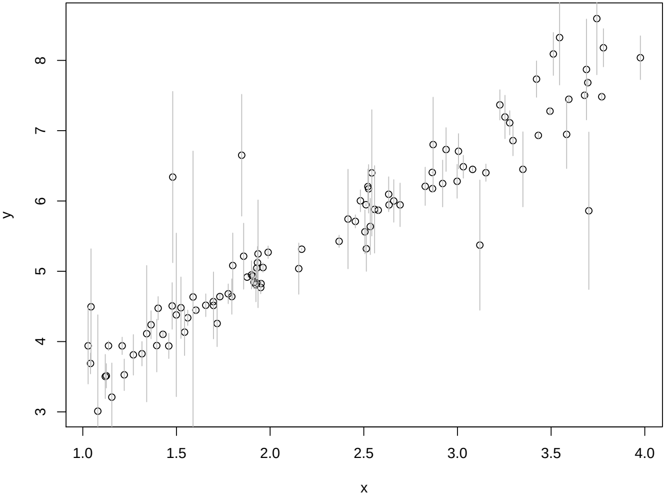
$$\widehat{\beta}(W) = (X^T W X)^{-1} X^T W Y.$$

**Possible Weight Matrices:**

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

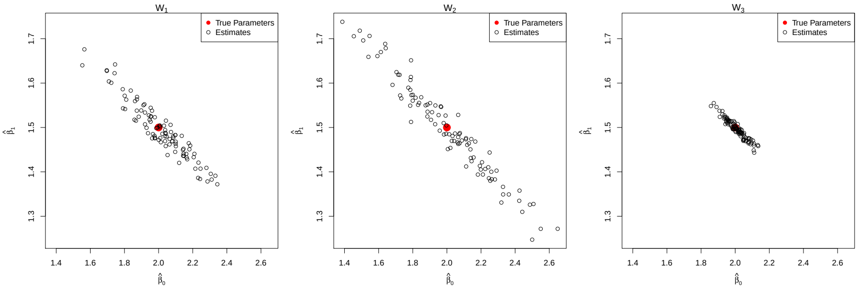
Recall $W_3$ is not known because $\sigma^2$ is unknown.

# $\beta = (2, 1.5)^T, \sigma = 0.1$ 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 $\sigma$ 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 $\sigma$, closed form solution.
- Evaluate likelihood at each $\sigma$ in grid.
- Choose value of $\sigma$ which maximizes likelihood.

# Minimize Negative Log Likelihood

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

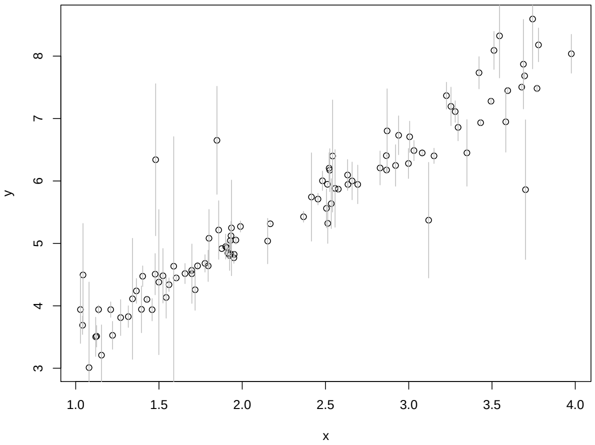
$$\begin{aligned}\widehat{\sigma}^2 &= \underset{\sigma^2}{\operatorname{argmin}} \min_{\beta_0, \beta_1} \sum_{i=1}^{n} \log(\sigma^2 + \sigma_i^2) + (Y - X\beta)^T W(\sigma^2)(Y - X\beta) \\ &= \underset{\sigma^2}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} \log(\sigma^2 + \sigma_i^2) + (Y -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

where

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

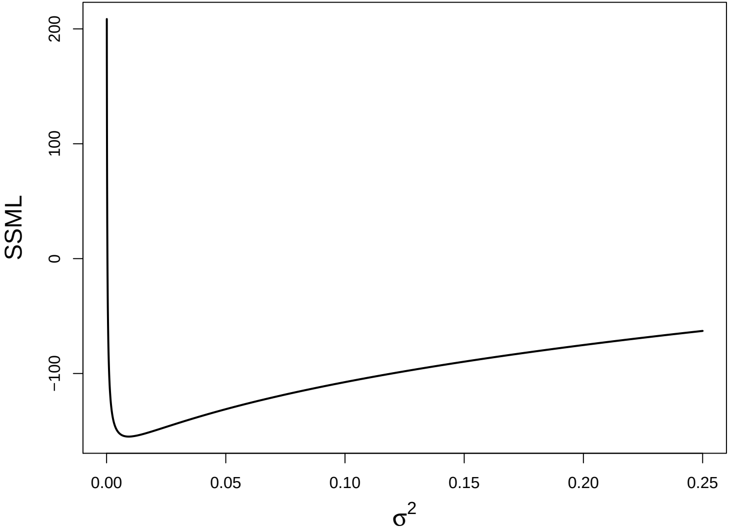
- Grid search on $\sigma$ to find $\widehat{\sigma}$.
- $\widehat{\beta} = \widehat{\beta}(\widehat{\sigma})$.

# 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



# Quantify Uncertainty on ML Estimates

The maximum likelihood estimate for the parameters is

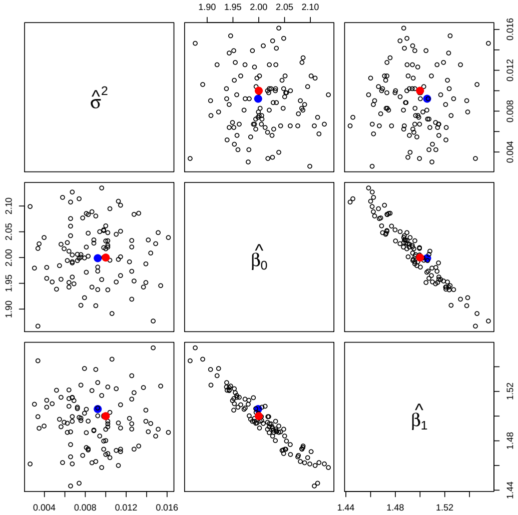
$$(\widehat{\sigma}^2, \widehat{\beta}_0, \widehat{\beta}_1) = (0.0092, 1.9988, 1.5057)$$

- 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

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

**Simulation Has Major Weaknesses:**

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

# Analytic Method

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

**Need more powerful statistical tools.**

# Outline

# MLE Asymptotics

Asymptotics: The study of how estimators behave as the sample sizes gets larger.

Consistency of MLEs:

$$\widehat{\theta}_{MLE} \to_P \theta \ \ (as \ n \to \infty)$$

Asymptotic Normality of MLE:

$$\sqrt{n}(\widehat{\theta}_{MLE} - \theta) \to_d N(0, I(\theta)^{-1})$$

where

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

is called the Fisher information matrix.

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

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

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

**Significance:** We can quantify the MLE uncertainty by computing the negative Hessian of the log likelihood at the MLE.

# Application to Intrinsic Scatter Model

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

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

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

# Computing $\widehat{I}(\widehat{\theta}_{ML})$: Calculus Exercise

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

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:

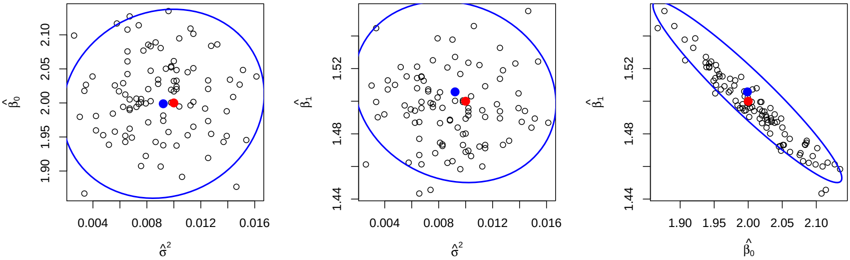
$$(\widehat{\sigma}^2, \widehat{\beta}_0, \widehat{\beta}_1) = (0.0092, 1.9988, 1.5057)$$

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).

# Important Points

- Statistical theory shows that uncertainty in MLE is approximately the negative Hessian of the log likelihood.
- In some models (such as intrinsic scatter model), analytically computing Hessian is not too bad. If so, estimating uncertainty is straightforward.
- In other models, we must numerically approximate Hessian.
- `optim` in R and `scipy.optimize` in python (with BFGS) approximate Hessians, simultaneously providing parameter estimates and uncertainties for MLEs in complex models.