Why Most Weighted Regressions Are Wrong

David Fletcher

Mathematics and Statistics

University of Otago

Summary

- Example
- Hierarchical models and MCMC
- Weighted regression

    Why it is often carried out wrongly

    Simple algorithm for doing it correctly

- When can we safely use standard (unweighted) regression?

Summary

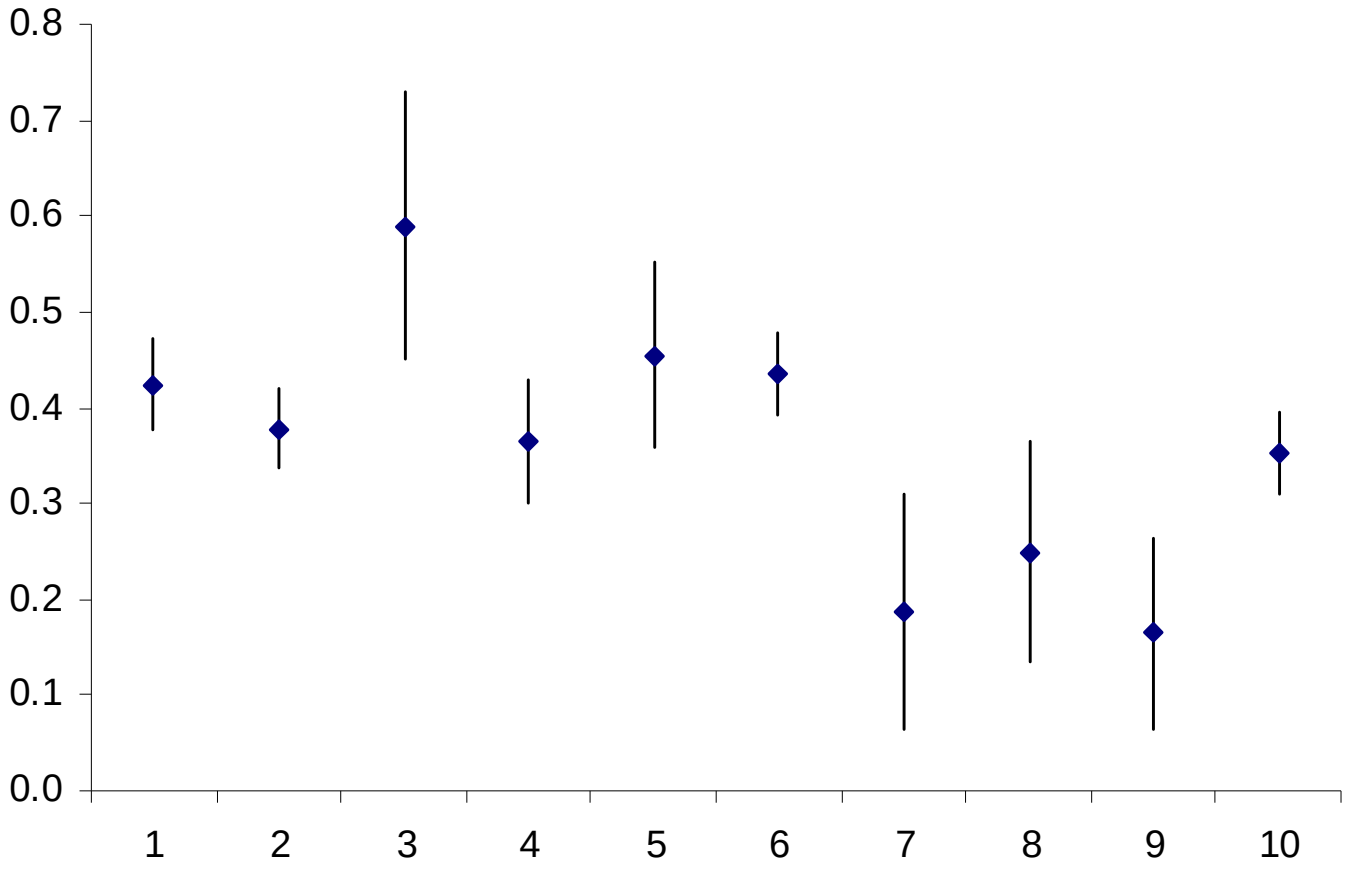
- Example
- Hierarchical models and MCMC
- Weighted regression

  Why it is often carried out wrongly

  Simple algorithm for doing it correctly
- When can we safely use standard (unweighted) regression?
- Simplicity and complexity...
- Statistical education...

Example

- Annual breeding rate (young/adult) over a number of years
- Relate to explanatory variable (e.g. year, SOI, SST,...)
- Or just calculate an overall mean rate for that period...



Modern Approach

- Hierarchical model fitted to all the data (using MCMC)
- Mixed model / random effects / variance components

r_i	Number of young
n_i	Number of adults
π_i	Probability of an adult producing a young individual

$$r_i \sim Binomial\left(n_i, \pi_i\right)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i$$

Modern Approach

- Hierarchical model fitted to all the data (using MCMC)
- Mixed model / random effects / variance components

r_i	Number of young
n_i	Number of adults
π_i	Probability of an adult producing a young individual

$$r_i \sim Binomial\left(n_i, \pi_i\right)$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta x_i + \varepsilon_i \qquad \varepsilon_i \sim N\left(0, \sigma^2\right)$$

Modern Approach

- Hierarchical model fitted to all the data (using MCMC)
- Mixed model / random effects / variance components

Modern Approach

- Hierarchical model fitted to all the data (using MCMC)
- Mixed model / random effects / variance components

But...

Modern Approach

- Hierarchical model fitted to all the data (using MCMC)
- Mixed model / random effects / variance components

But...

- Data unavailable?

  Published results

  Separate procedure used to obtain estimates
- Analysis slower?
- Analysis harder for client to implement and/or understand?

Modern Approach

- Hierarchical model fitted to all the data (using MCMC)
- Mixed model / random effects / variance components

But...

- Data unavailable?

  Published results

  Separate procedure used to obtain estimates
- Analysis slower?
- Analysis harder for client to implement and/or understand?

  Murtagh (Ecology 2007)

  Bayesian methods: more familiar?

  uses explicit modelling

Modern Approach

- Hierarchical model fitted to all the data (using MCMC)
- Mixed model / random effects / variance components

But...

- Data unavailable?

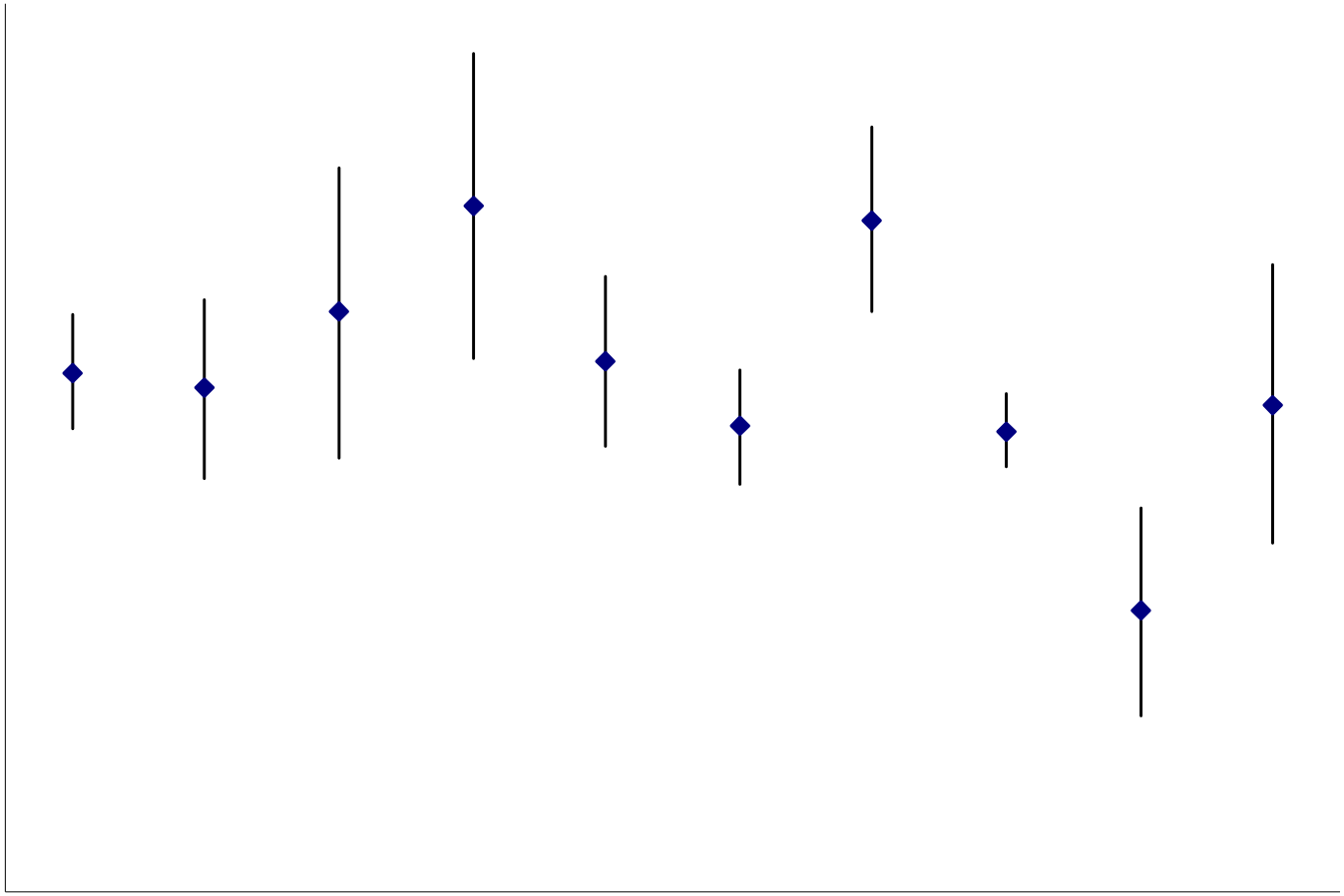
  Published results

  Separate procedure used to obtain estimates
- Analysis slower?
- Analysis harder for client to implement and/or understand?

  Use weighted regression? Calculate a weighted mean?

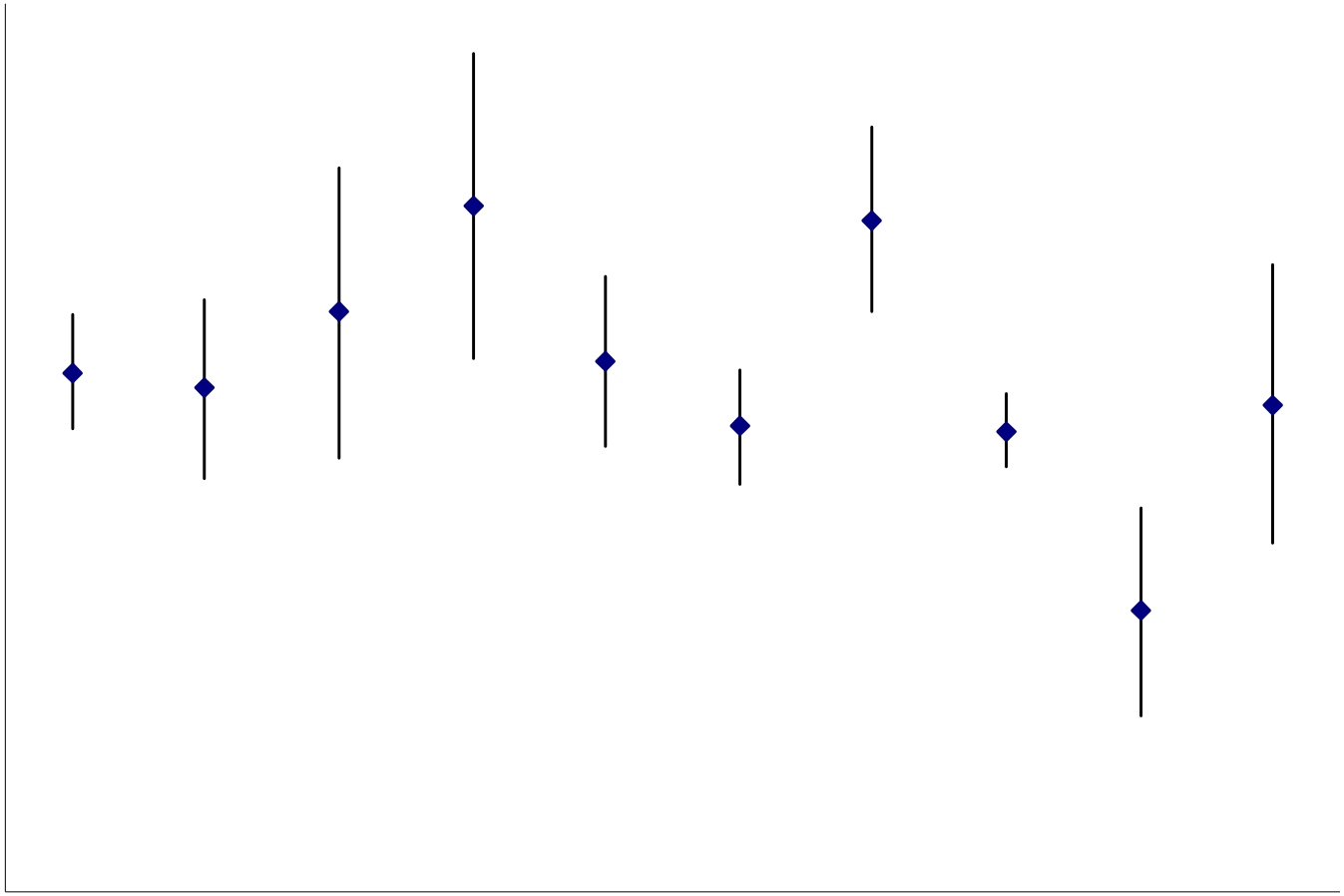
Generalised Example

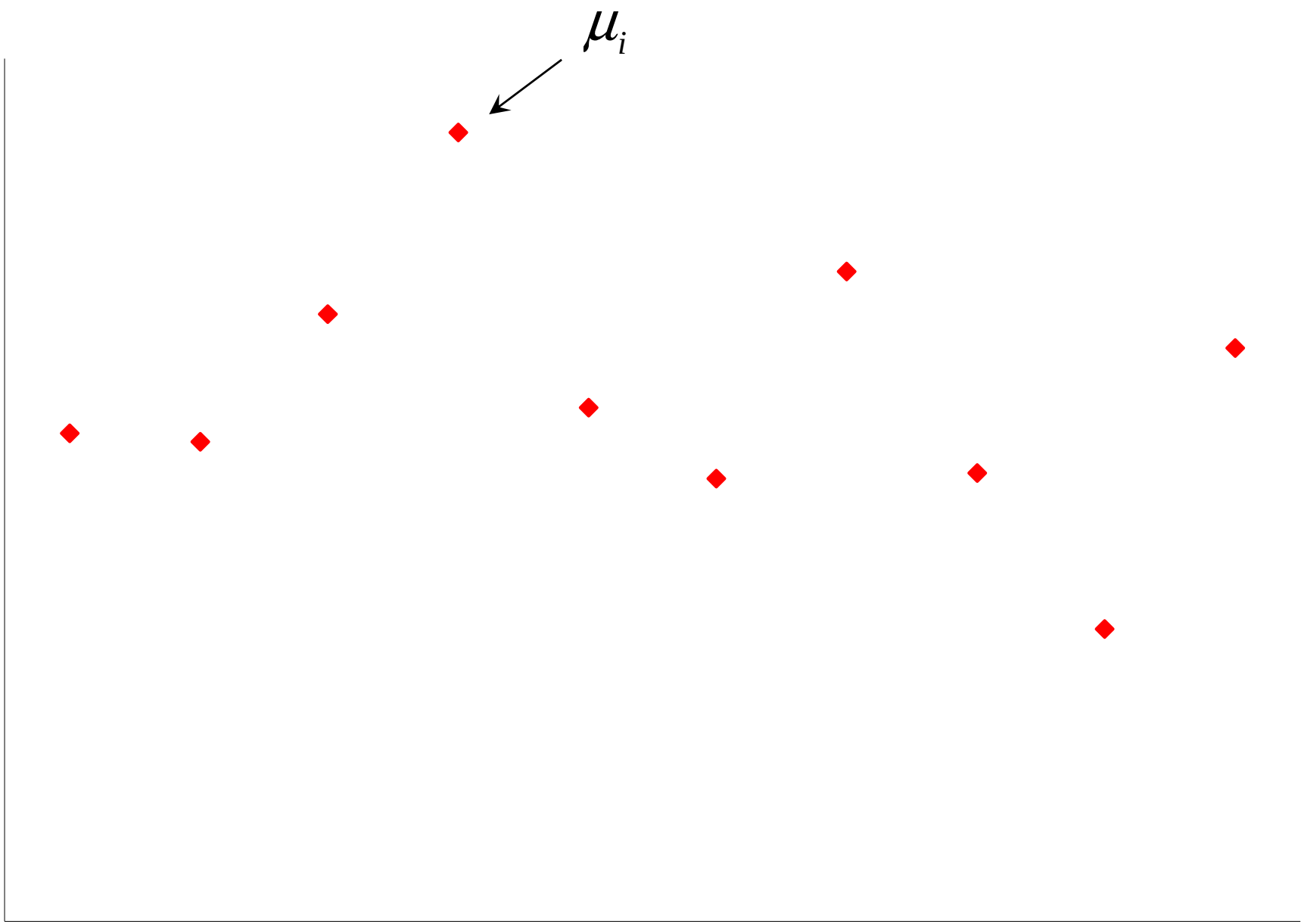
- Several estimates of the same quantity available
- Want to relate to one or more explanatory variables
- Or just calculate an overall mean

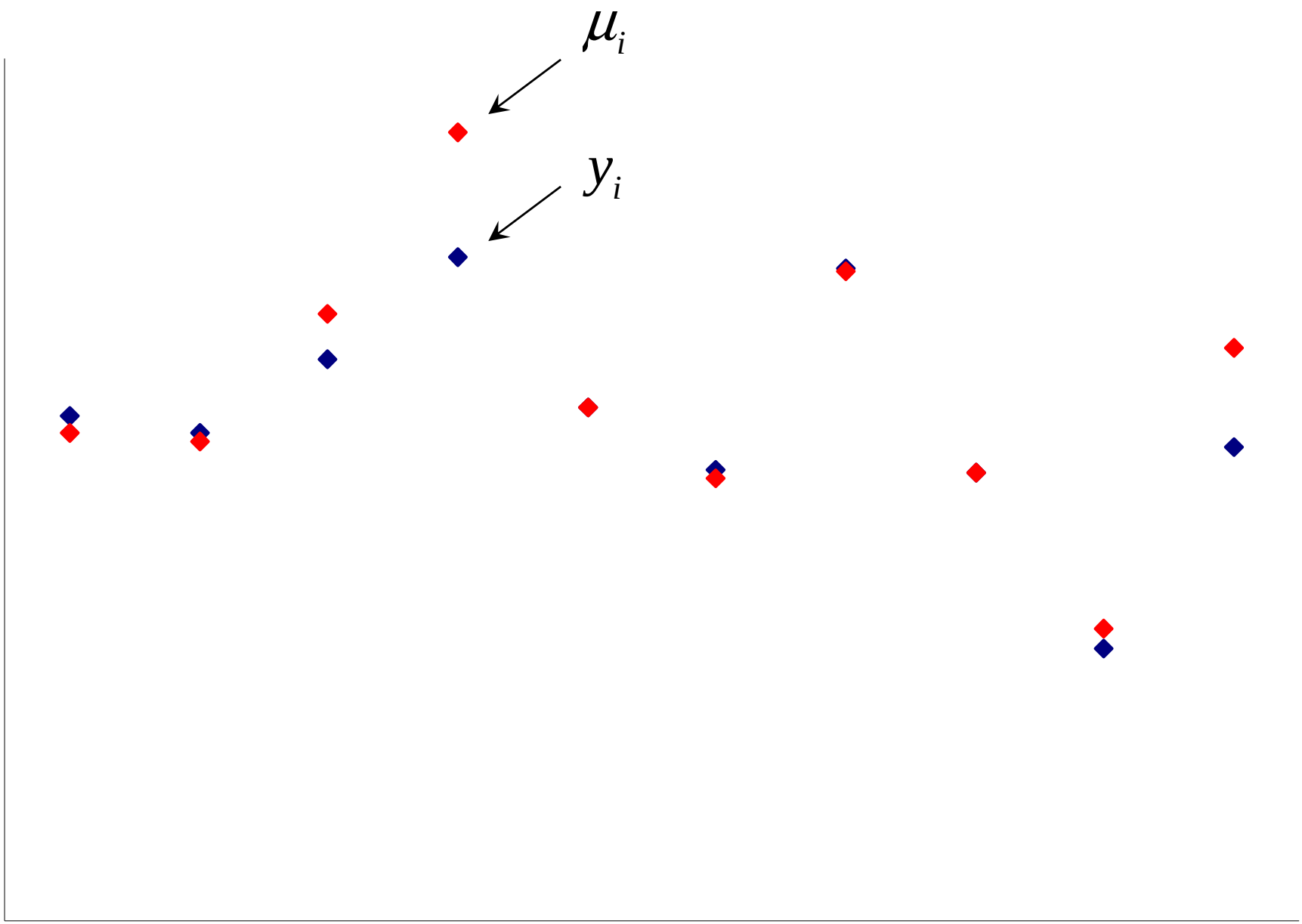


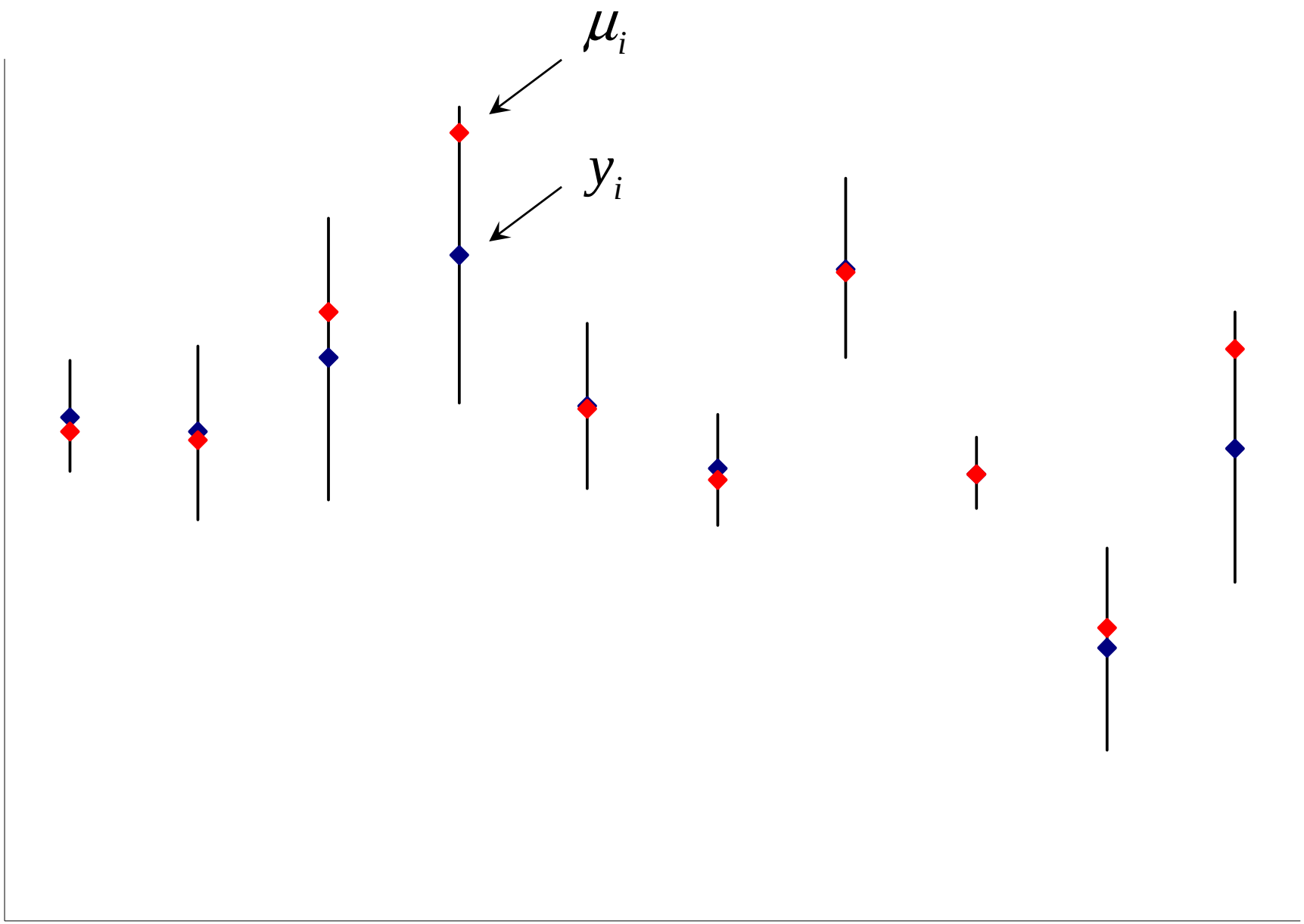
Generalised Example

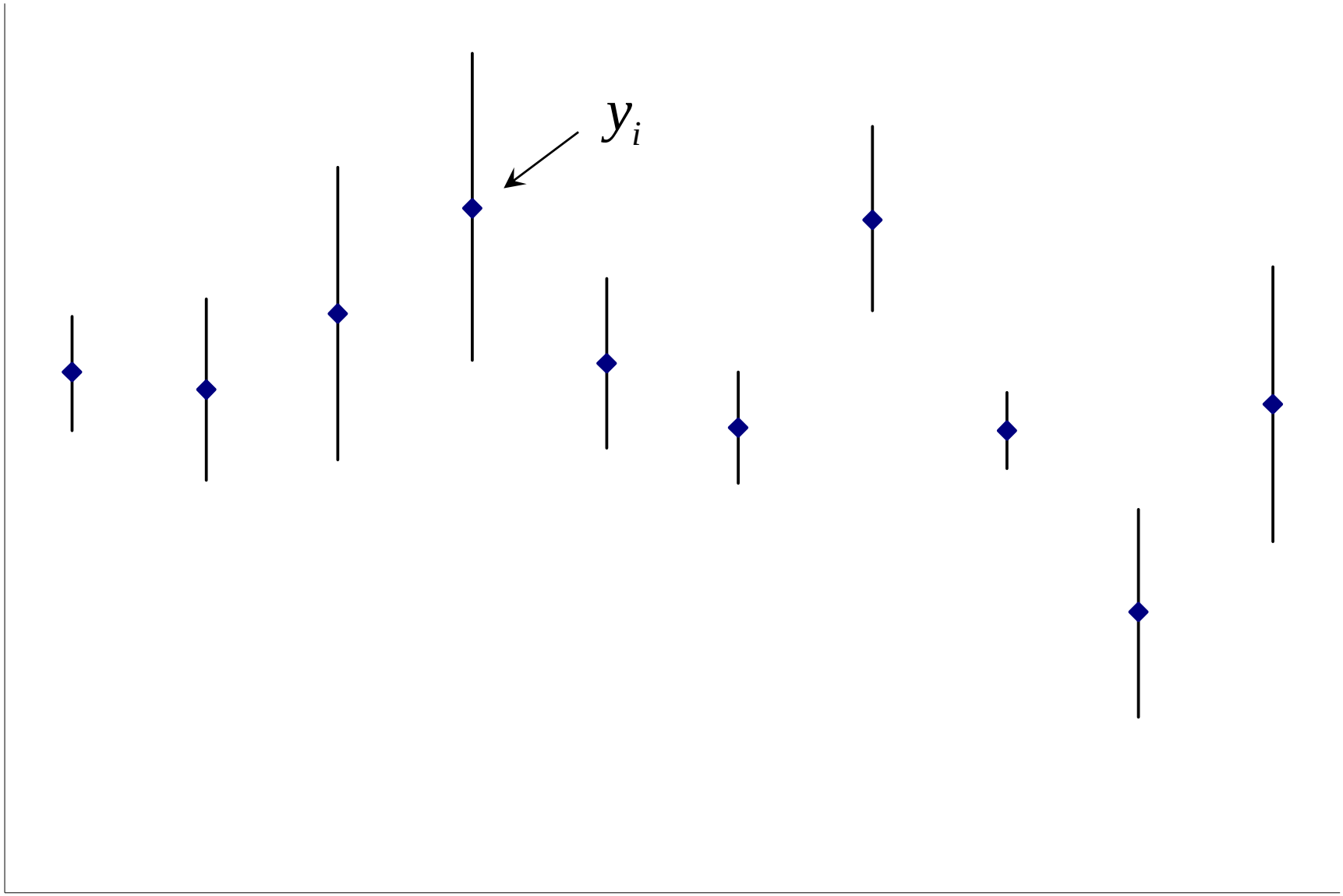
- Several estimates of the same quantity available
- Want to relate to one or more explanatory variables
- Or just calculate an overall mean (c.f. meta analysis)




μ_i


μ_i
y_i


μ_i
y_i


y_i

Notation

$$E\left(y_i \middle| \mu_i\right) = \mu_i \qquad V\left(y_i \middle| \mu_i\right) = \sigma_i^2$$

"Measurement" error

Notation

$$E\left(y_i \middle| \mu_i\right) = \mu_i \qquad V\left(y_i \middle| \mu_i\right) = \sigma_i^2$$

"Measurement" error

$$\mu_i = \alpha + \beta x_i + \varepsilon_i$$

$$E\left(\mu_i\right) = \alpha + \beta x_i$$

$$V\left(\mu_i\right) = V\left(\varepsilon_i\right) = \sigma^2$$

"Process" error

Notation

$$E\left(y_i \middle| \mu_i\right) = \mu_i \qquad V\left(y_i \middle| \mu_i\right) = \sigma_i^2$$

"Measurement" error

$$\mu_i = \alpha + \beta x_i + \varepsilon_i$$

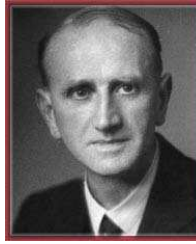
$$E\left(\mu_i\right) = \alpha + \beta x_i$$

$$V\left(\mu_i\right) = V\left(\varepsilon_i\right) = \sigma^2$$

"Process" error

$$V\left(y_i\right) = \sigma^2 + \sigma_i^2$$

$$\left(\sigma^2 \neq 0\right)$$



Weighted Regression (Aitken; GLS; Burnham and White 2002)

1. Select initial value for σ^2

   e.g. $\sigma^2 = RMS - \bar{\sigma}^2$

   RMS Residual MS from unweighted regression

   $\bar{\sigma}^2$ Mean of the σ_i^2

2. Run weighted regression with weights

$$w_i = \frac{1}{\sigma^2 + \sigma_i^2} \qquad \left(\text{NOT} \quad w_i = \frac{1}{\sigma_i^2} \right)$$

   N.B. Analysis assumes σ_i^2 known (unlike "full" analysis)

3. Modify value of σ^2 until RMS from weighted regression = 1

   May need to try $\sigma^2 < 0$

Example (using R)

σ^2	Weighted RMS	Slope Est	Slope SE
0.00450	0.989	-0.00966	0.00818
0.00400	1.035	-0.00960	0.00816
0.00410	1.025	-0.00961	0.00816
0.00430	1.007	-0.00964	0.00817
0.00435	1.002	-0.00964	0.00817
0.00436	1.002	-0.00965	0.00817
0.00437	1.001	-0.00965	0.00817
0.00438	1.000	-0.00965	0.00817

Unweighted analysis	-0.01038	0.00831
Setting $\sigma^2 = 0$	-0.00737	0.00605

Example (using R)

σ^2	Weighted RMS	Mean Est	Mean SE
0.00560	0.928	0.34722	0.02403
0.00600	0.902	0.34737	0.02407
0.00500	0.970	0.34697	0.02397
0.00450	1.011	0.34670	0.02390
0.00460	1.002	0.34677	0.02392
0.00465	0.998	0.34679	0.02392
0.00462	1.001	0.34678	0.02392
0.00463	1.000	0.34678	0.02392

Unweighted analysis	0.35020	0.02461
Setting $\sigma^2 = 0$	0.33247	0.01732

Simple guidelines

- Unweighted analysis OK if σ_i^2 roughly constant

Simple guidelines

- Unweighted analysis OK if σ_i^2 roughly constant

  How close to constant?

Simple guidelines

- Unweighted analysis OK if σ_i^2 roughly constant

  How close to constant?

- Ignoring process error OK if $\sigma^2 \approx 0$

Simple guidelines

- Unweighted analysis OK if σ_i^2 roughly constant

  How close to constant?

- Ignoring process error OK if $\sigma^2 \approx 0$

  How close to zero?

Summary

- Example
- Hierarchical models and MCMC
- Weighted regression

  Why it is often carried out wrongly

  Simple algorithm for doing it correctly
- When can we safely use standard (unweighted) regression?

Summary

- Example
- Hierarchical models and MCMC (WinBugs helps...)
- Weighted regression

  Why it is often carried out wrongly

  Simple algorithm for doing it correctly
- When can we safely use standard (unweighted) regression?
- Simplicity and complexity...
- Statistical education...