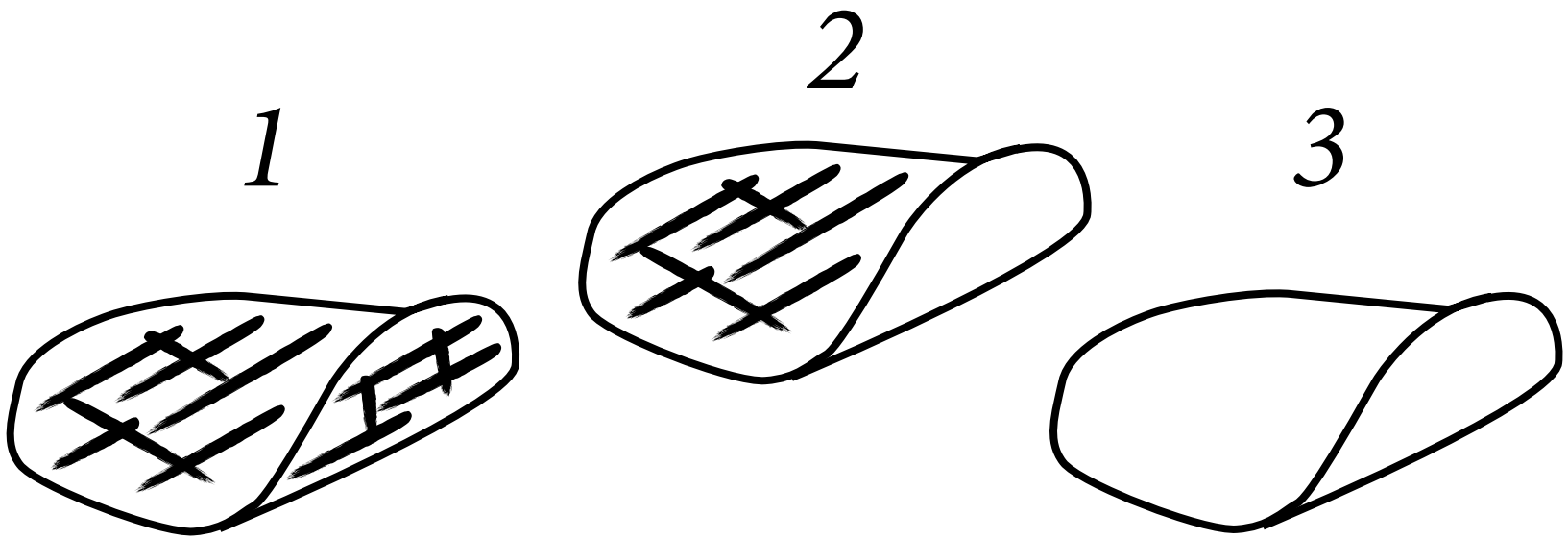
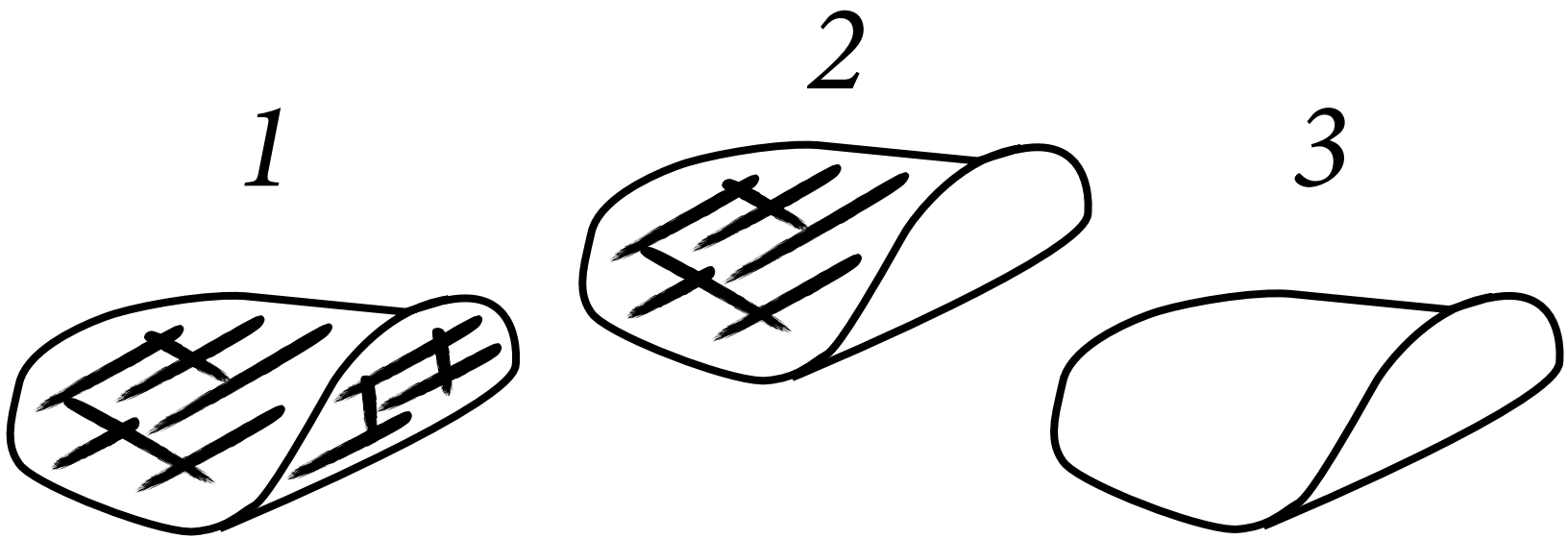
Statistical Rethinking

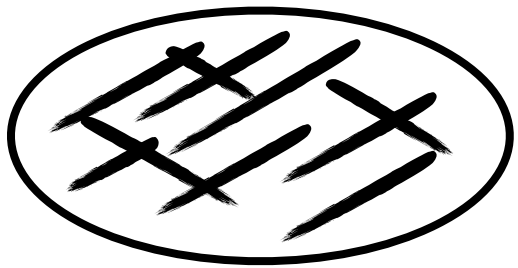
Week 10:
Missing Data &
Other Opportunities

Richard McElreath


1
2
3

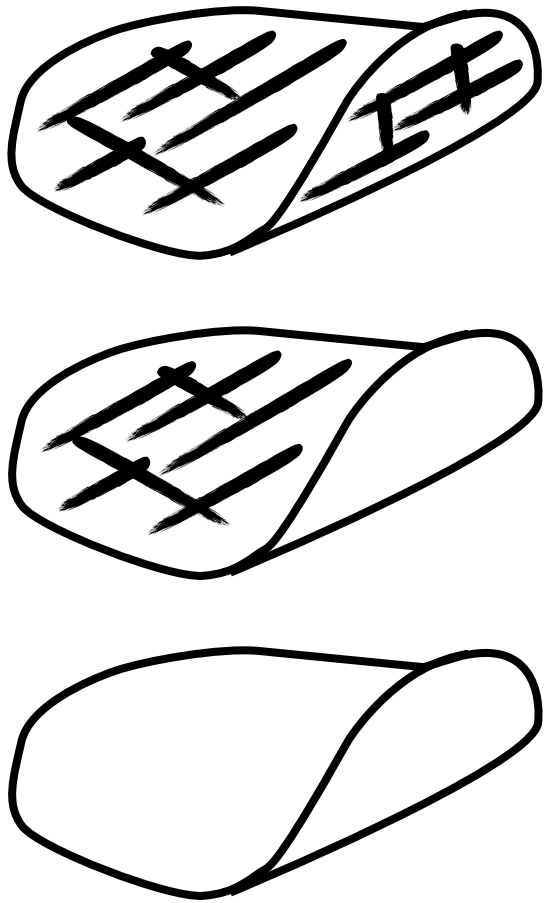


You are served:

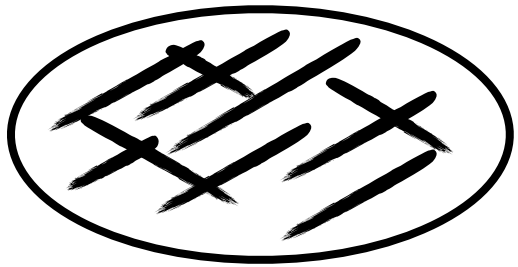


Probability other side is burnt?

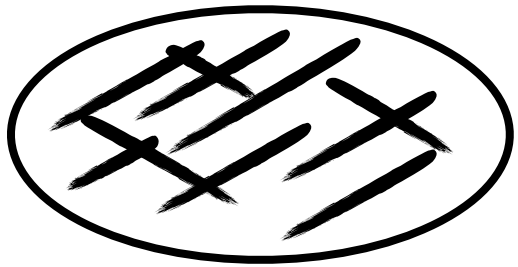
Avoid being clever



- Intuition terrible guide to probability
- No need to be clever; just ruthlessly apply conditional probability
  - Pr(want to know|already know)

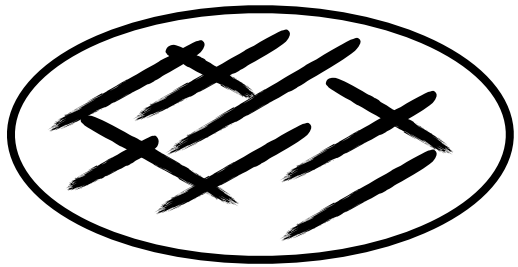


$$\Pr(\text{burnt down}|\text{burnt up}) = \frac{\Pr(\text{burnt up, burnt down})}{\Pr(\text{burnt up})}$$



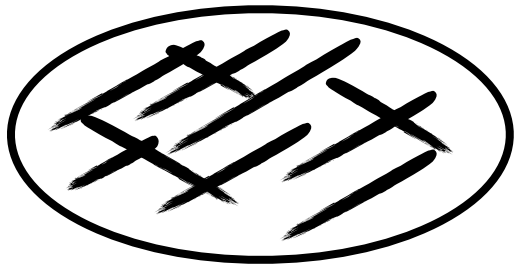
$$\Pr(\text{burnt down}|\text{burnt up}) = \frac{\Pr(\text{burnt up, burnt down})}{\Pr(\text{burnt up})}$$

$$\Pr(\text{burnt up}) = \Pr(\text{BB})(1) + \Pr(\text{BU})(0.5) + \Pr(\text{UU})(0)$$



$$\Pr(\text{burnt down}|\text{burnt up}) = \frac{\Pr(\text{burnt up, burnt down})}{\Pr(\text{burnt up})}$$

$$\begin{aligned}\Pr(\text{burnt up}) &= \Pr(\text{BB})(1) + \Pr(\text{BU})(0.5) + \Pr(\text{UU})(0) \\ &= (1/3) + (1/3)(1/2) = 0.5\end{aligned}$$



$$\Pr(\text{burnt down}|\text{burnt up}) = \frac{\Pr(\text{burnt up, burnt down})}{\Pr(\text{burnt up})}$$

$$\begin{aligned}\Pr(\text{burnt up}) &= \Pr(\text{BB})(1) + \Pr(\text{BU})(0.5) + \Pr(\text{UU})(0) \\ &= (1/3) + (1/3)(1/2) = 0.5\end{aligned}$$

$$\Pr(\text{burnt down}|\text{burnt up}) = \frac{1/3}{1/2} = \frac{2}{3}$$

Getting Ruthless

- Express information as constraints and distributions => let logic discover implications
- No need to be clever
- Examples:
  - Measurement error
  - Missing data

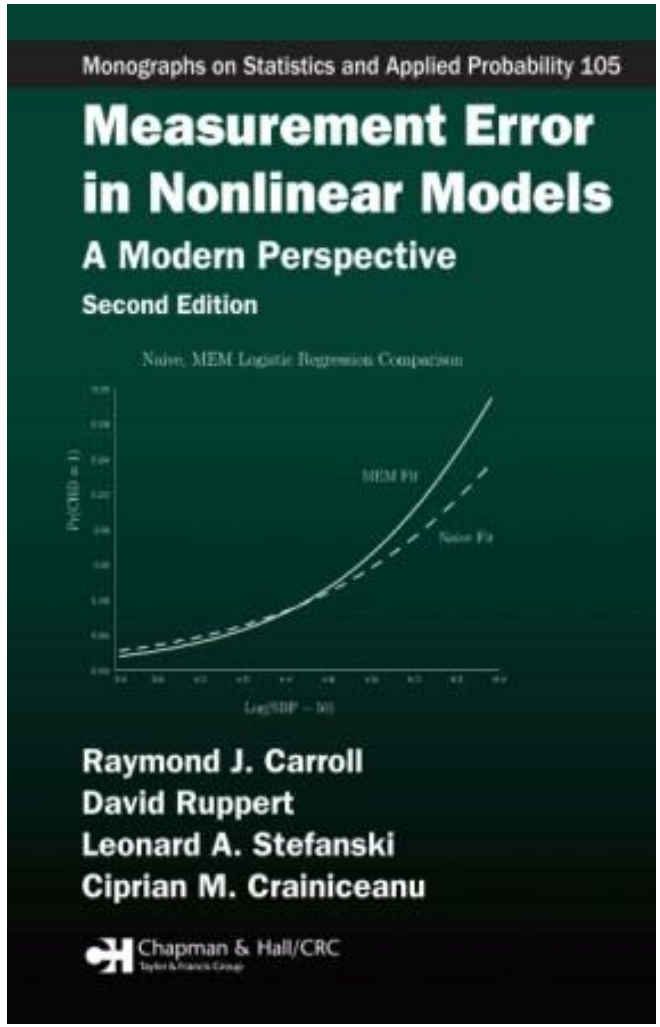
Decolonizing Bayes

- Bayes taught using non-Bayesian vocabulary
  - "data": An observed variable
  - "parameter": An unobserved variable
  - "likelihood": Probability assignment for observed var
  - "prior": Probability assignment for unobserved var
- Even term "Bayesian" not Bayesian!
- Distinction btw data and parameter relevant *after* observation
- Can exploit this fact to address common modeling issues

Sir Ronald Fisher (1890–1962) named it "Bayesian"

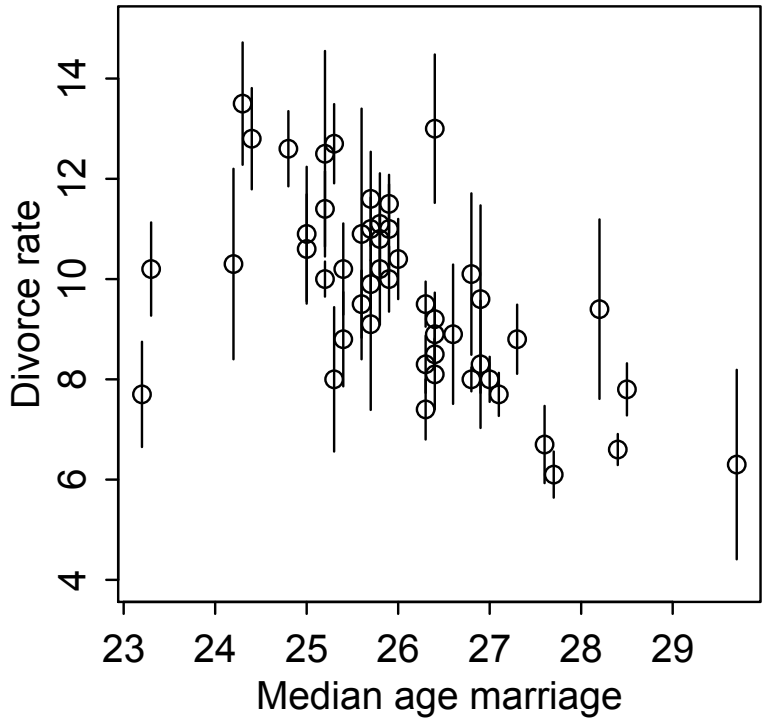
Measurement error

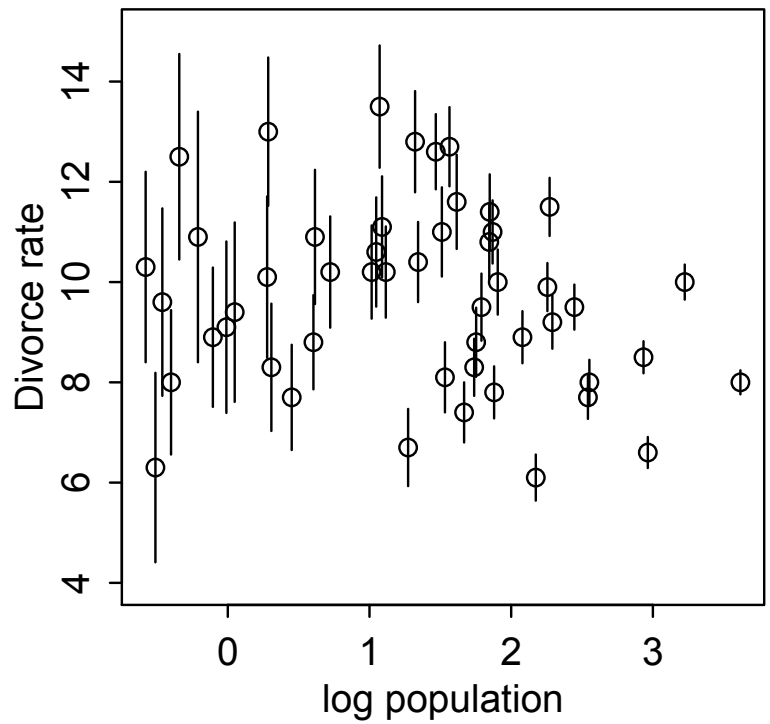
- Measurement always entails error
- Typical linear regression: interpret sigma as “error” on outcome
- What if error isn’t constant?
- What if error is on predictors?



Error on outcome

- `data(WaffleDivorce)`
- Consider error on outcome, divorce rate
- Heterogeneity in error
- Small State => large error





Error on outcome

- Approach:
  - Treat true divorce rate as unknown parameter
  - Observed rate is sample from Gaussian distribution:

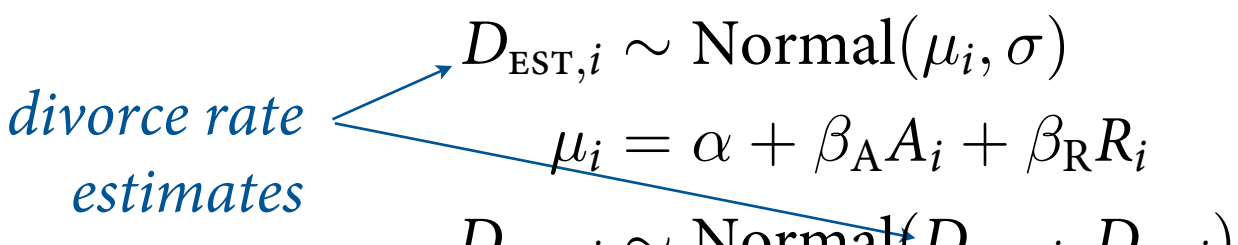
$$D_{\text{OBS},i} \sim \text{Normal}(D_{\text{EST},i}, D_{\text{SE},i})$$

observed (data) — *true (parameter)* — *std error (data)*

Error on outcome: model

$$
\begin{aligned}
D_{\text{EST},i} &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_{\text{A}} A_i + \beta_{\text{R}} R_i \\
D_{\text{OBS},i} &\sim \text{Normal}(D_{\text{EST},i}, D_{\text{SE},i}) \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta_{\text{A}} &\sim \text{Normal}(0, 10) \\
\beta_{\text{R}} &\sim \text{Normal}(0, 10) \\
\sigma &\sim \text{Cauchy}(0, 2.5)
\end{aligned}
$$

Error on outcome: model



$$D_{\text{EST},i} \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_{\text{A}} A_i + \beta_{\text{R}} R_i$$

$$D_{\text{OBS},i} \sim \text{Normal}(D_{\text{EST},i}, D_{\text{SE},i})$$

$$\alpha \sim \text{Normal}(0, 10)$$

$$\beta_{\text{A}} \sim \text{Normal}(0, 10)$$

$$\beta_{\text{R}} \sim \text{Normal}(0, 10)$$

$$\sigma \sim \text{Cauchy}(0, 2.5)$$

Error on outcome: model

likelihood for each estimate

$$D_{\text{EST},i} \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_{\text{A}} A_i + \beta_{\text{R}} R_i$$

likelihood for each observation

$$D_{\text{OBS},i} \sim \text{Normal}(D_{\text{EST},i}, D_{\text{SE},i})$$

estimate

standard error of observation

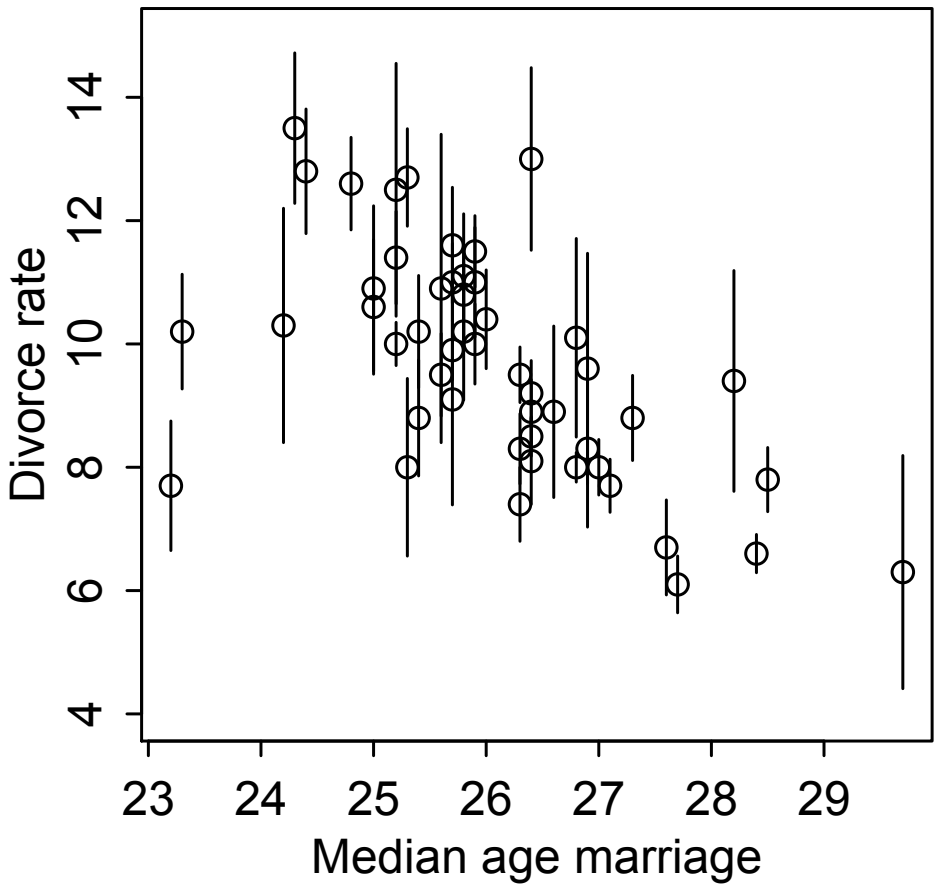
Error on outcome: fitting

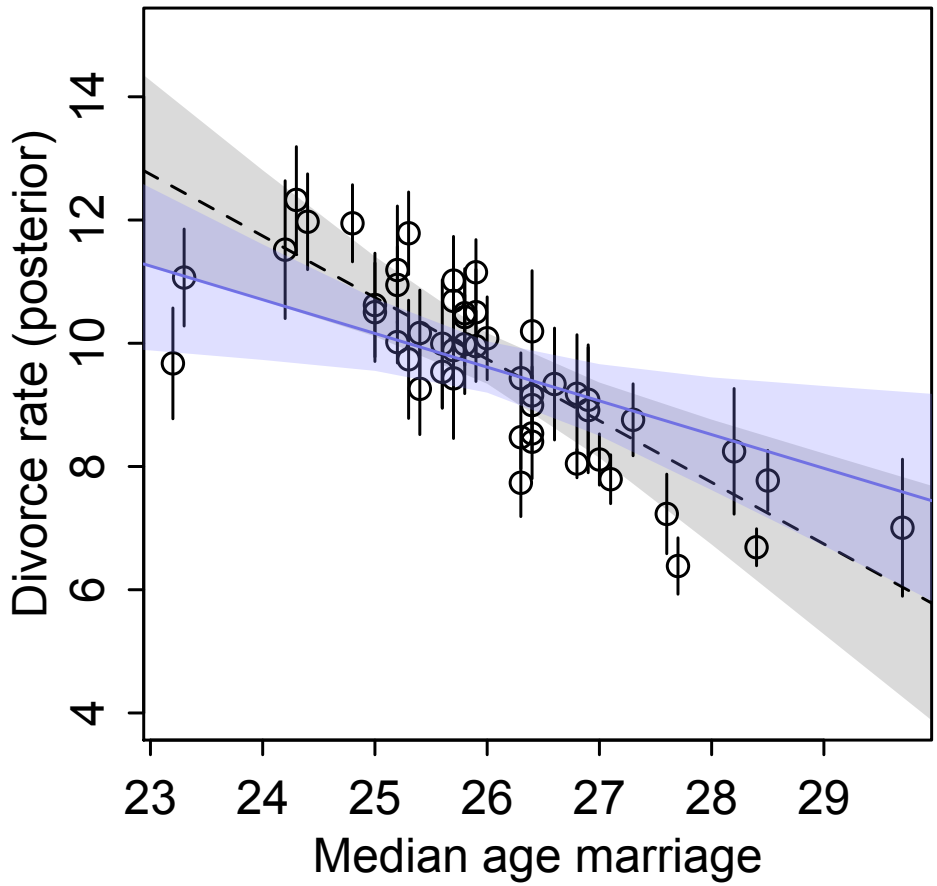
R code 14.3

```
dlist <- list(
    div_obs=d$Divorce,
    div_sd=d$Divorce.SE,
    R=d$Marriage,
    A=d$MedianAgeMarriage
)

m14.1 <- map2stan(
    alist(
        div_est ~ dnorm(mu,sigma),
        mu <- a + bA*A + bR*R,
        div_obs ~ dnorm(div_est,div_sd),
        a ~ dnorm(0,10),
        bA ~ dnorm(0,10),
        bR ~ dnorm(0,10),
        sigma ~ dcauchy(0,2.5)
    ) ,
    data=dlist ,
    start=list(div_est=dlist$div_obs) ,
    WAIC=FALSE , iter=5000 , chains=2 )
```

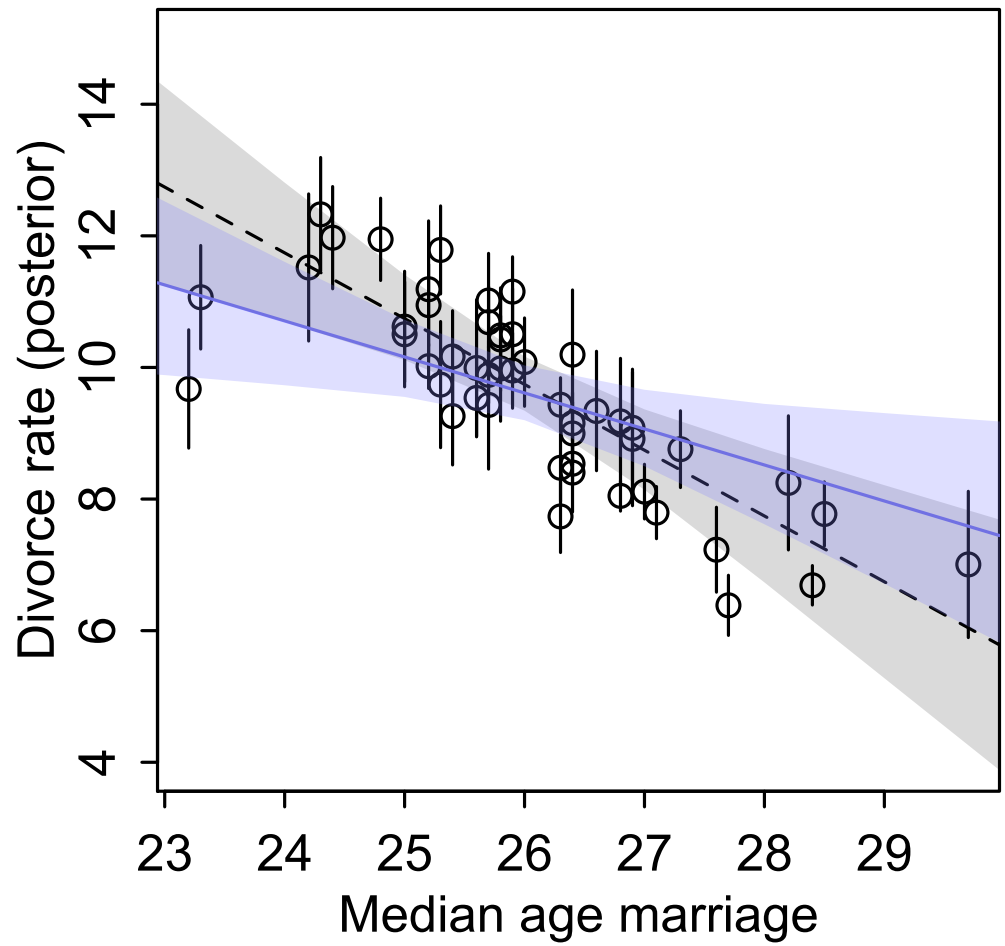
$$
\begin{aligned}
D_{\text{EST},i} &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_{\text{A}} A_i + \beta_{\text{R}} R_i \\
D_{\text{OBS},i} &\sim \text{Normal}(D_{\text{EST},i}, D_{\text{SE},i}) \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta_{\text{A}} &\sim \text{Normal}(0, 10) \\
\beta_{\text{R}} &\sim \text{Normal}(0, 10) \\
\sigma &\sim \text{Cauchy}(0, 2.5)
\end{aligned}
$$


Divorce rate
4
6
8
10
12
14
23
24
25
26
27
28
29
Median age marriage


Divorce rate (posterior)
4
6
8
10
12
14
23
24
25
26
27
28
29
Median age marriage

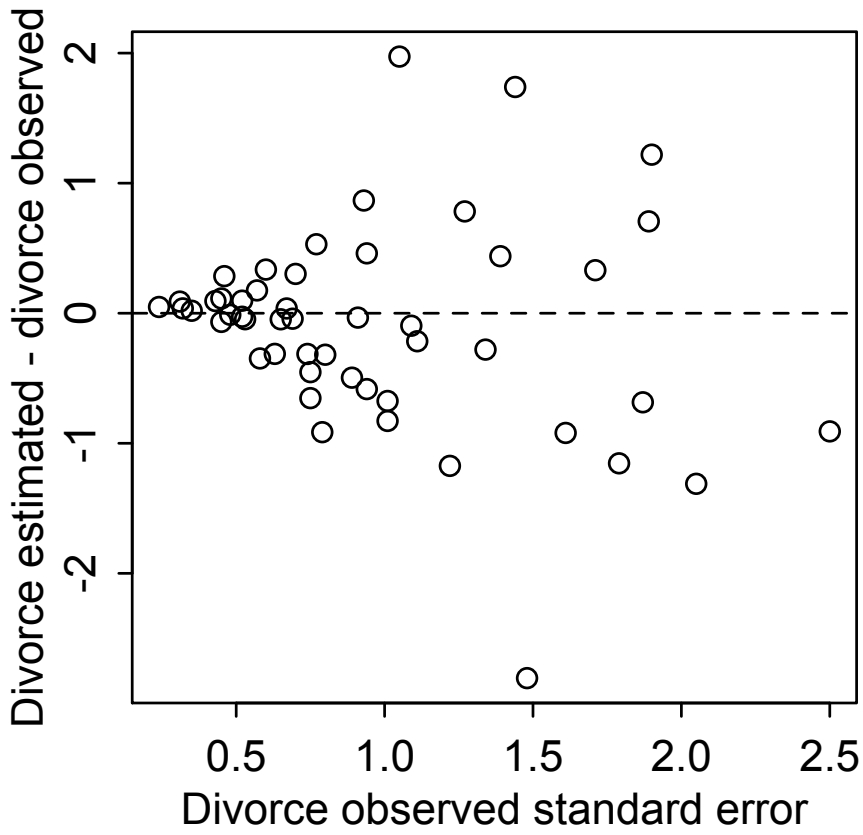
Error on outcome: results

- Divorce rate estimates move from observed values.
- Why?



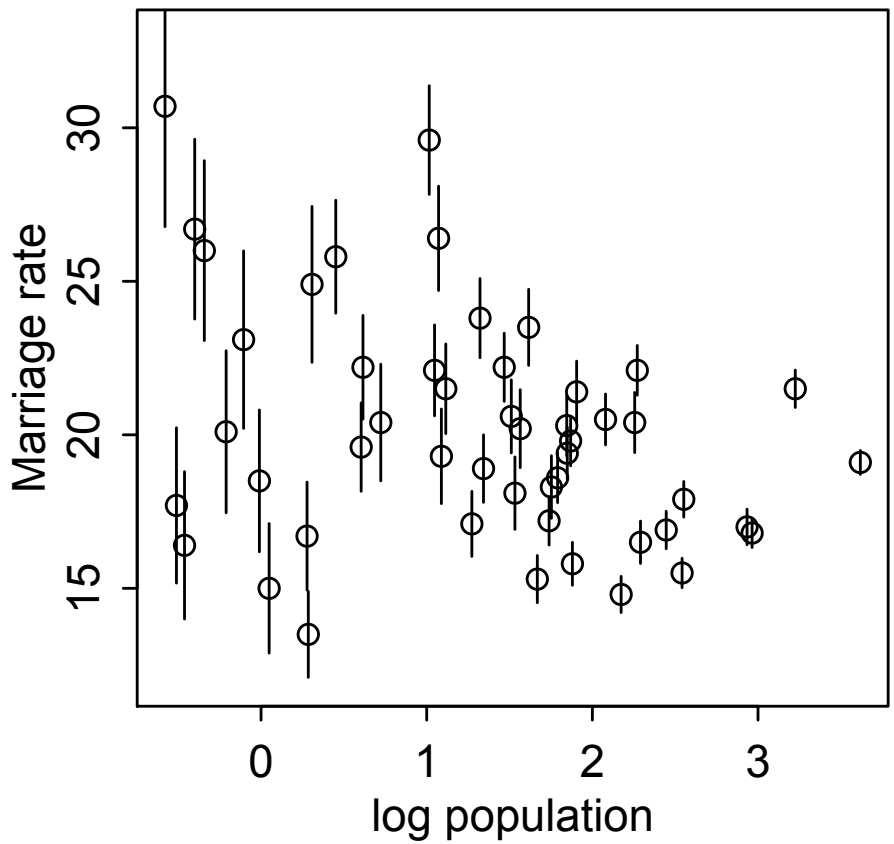
Error on outcome: results

- Q: Why do divorce rate estimates move?
- A: Pooling!
  - Small States have highly uncertain rates => low influence on regression
  - Large States have more certain rates => high influence on regression
  - Divorce estimates should be consistent with regression => update estimates of each State's divorce rate
  - Noisier estimates shrink more



Error on predictor

- What about error on predictor?
- Many procedures invented
  - errors-in-variables
  - reduced major axis
  - total least squares
- Our approach will be logical
  - State information
  - Deduce implications
  - Garbage in? You know what comes out.



Error on predictor: model

$$
\begin{aligned}
D_{\text{EST},i} &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_{\text{A}} A_i + \beta_{\text{R}} R_{\text{EST},i} \\
D_{\text{OBS},i} &\sim \text{Normal}(D_{\text{EST},i}, D_{\text{SE},i}) \\
R_{\text{OBS},i} &\sim \text{Normal}(R_{\text{EST},i}, R_{\text{SE},i}) \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta_{\text{A}} &\sim \text{Normal}(0, 10) \\
\beta_{\text{R}} &\sim \text{Normal}(0, 10) \\
\sigma &\sim \text{Cauchy}(0, 2.5)
\end{aligned}
$$

Error on predictor: model

$$D_{\text{EST},i} \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_{\text{A}} A_i + \beta_{\text{R}} R_{\text{EST},i}$$

$$D_{\text{OBS},i} \sim \text{Normal}(D_{\text{EST},i}, D_{\text{SE},i})$$

$$R_{\text{OBS},i} \sim \text{Normal}(R_{\text{EST},i}, R_{\text{SE},i})$$

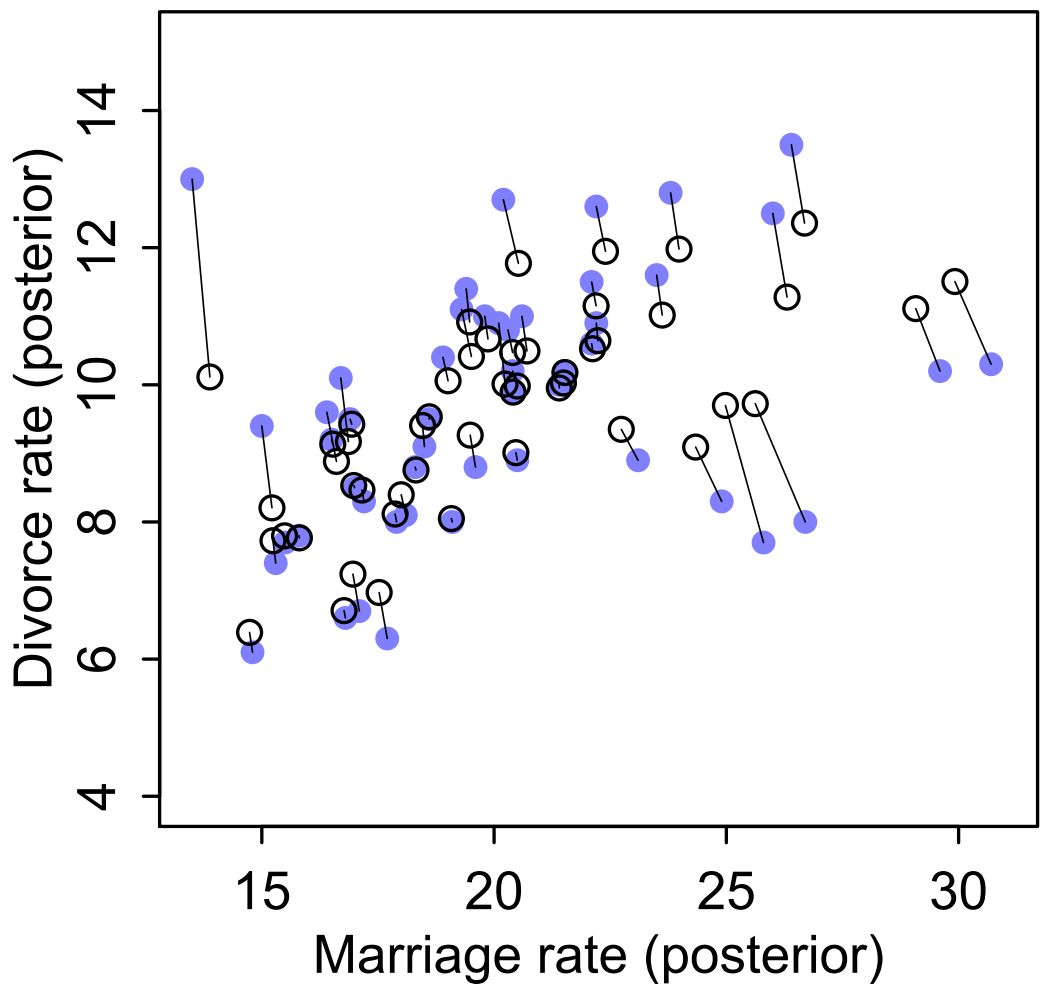
use estimates in regression

likelihood for each observed rate

estimated marriage rate

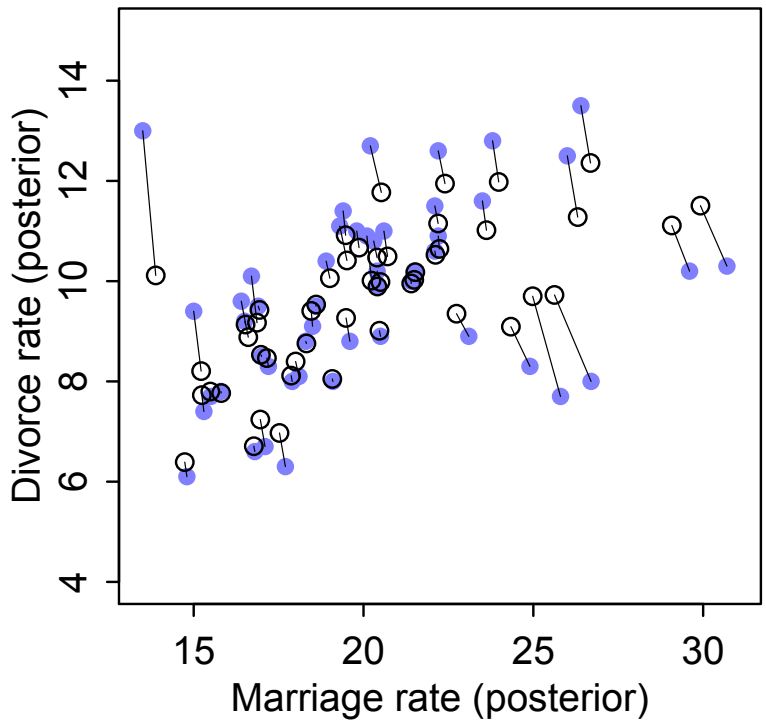
standard error of marriage rate

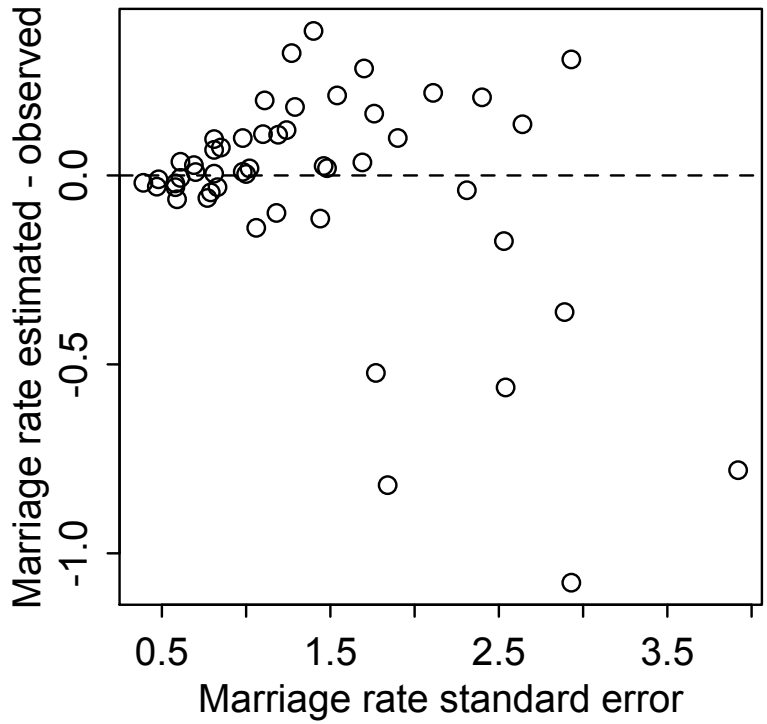
filled circles: observed
open circles: estimated
lines connect points for same State



Error on predictor

- Both divorce rate and marriage rate shrink
- Divorce shrinks much more. Why?
- Marriage rate not strongly associated with outcome => not much pooling through regression => not much shrinkage





Measurement error

- Common malady: “data” come from uncertain procedure, but uncertainty discarded at analysis
- Examples:
  - Predicting with averages; use posterior of average
  - DNA sequence data: respect error rate
  - Parentage analysis: probability distribution over possible parents
  - Phylogenetics: distribution of trees
  - Archaeology/paleontology/forensics: identification, sexing, aging, dating
- Propagate uncertainty

Missing data

- Missing values commonplace
- Usual approach: ***complete-case*** analysis
  - drop all cases with any missing values
- Discards a lot of information
- Alternatives
  - replace missing with mean of column: NEVER DO THIS
  - Multiple imputation
  - Bayesian imputation
  - others

im•pute | imˈpyo͞ot |

verb [with obj.]

represent (something, esp. something undesirable) as being done, caused, or possessed by someone; attribute: *the crimes* ***imputed to*** *Richard.*

- Finance assign (a value) to something by inference from the value of the products or processes to which it contributes: (as adj. **imputed**) *: recovering the initial outlay plus imputed interest.*
- Theology ascribe (righteousness, guilt, etc.) to someone by virtue of a similar quality in another: *Christ's righteousness has been imputed to us.*