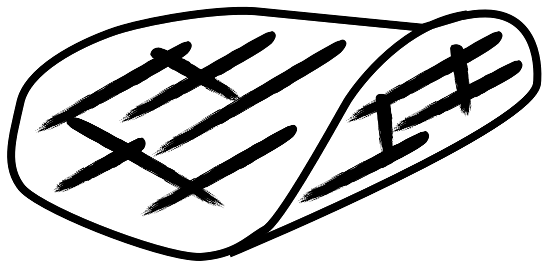
Statistical Rethinking
Winter 2019

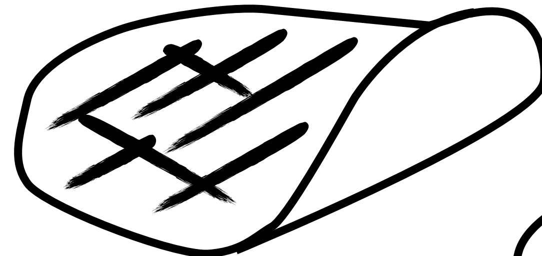
Lecture 20 / Week 10

Missing Data &
Other Opportunities

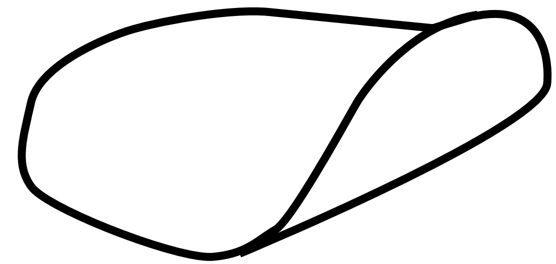
1

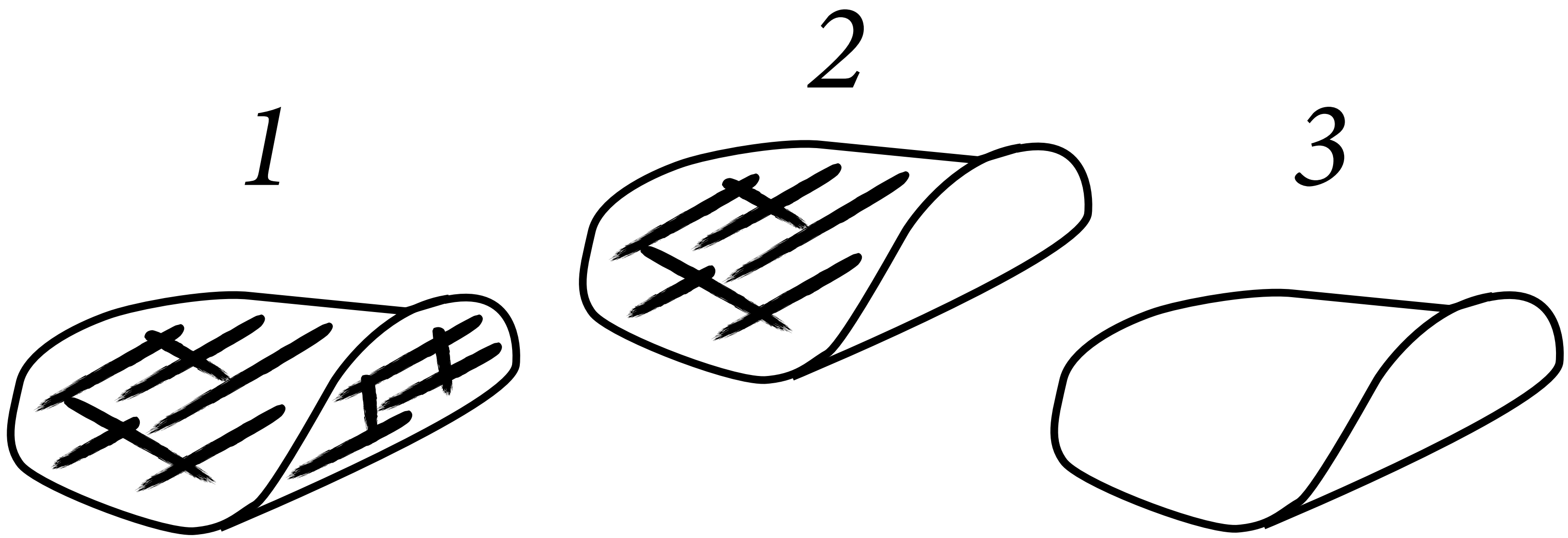


2



3



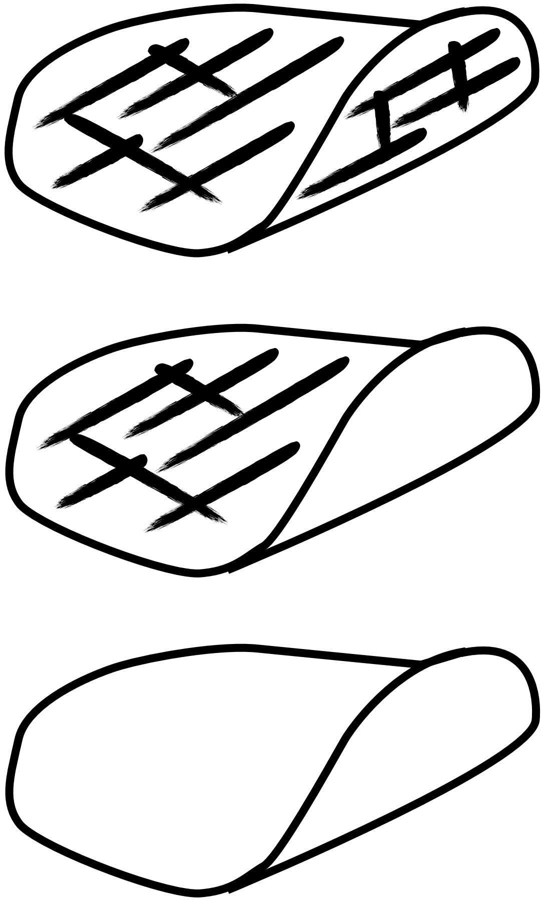


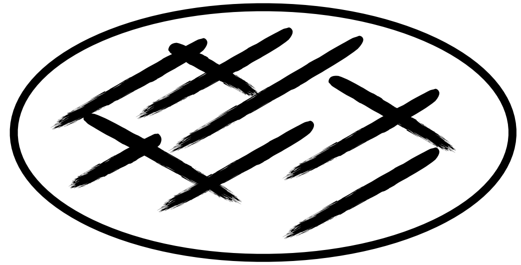
You are served:

Probability other side is burnt?

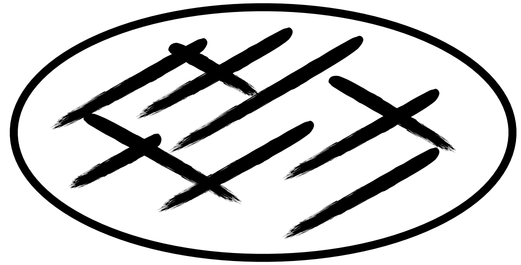
Avoid being clever

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



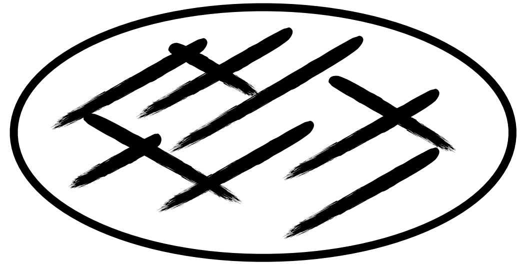


$$\Pr(\text{burnt down}|\text{burnt up}) = \frac{\Pr(\text{burnt up}, \text{burnt down})}{\Pr(\text{burnt up})}$$



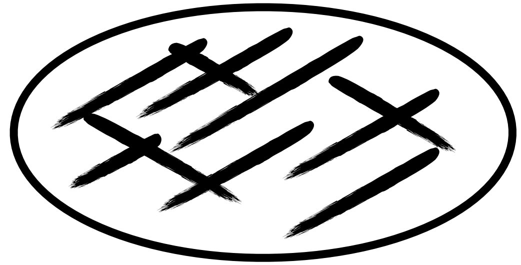
$$\Pr(\text{burnt down}|\text{burnt up}) = \frac{\Pr(\text{burnt up, burnt down})}{\Pr(\text{burnt up})}$$

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



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

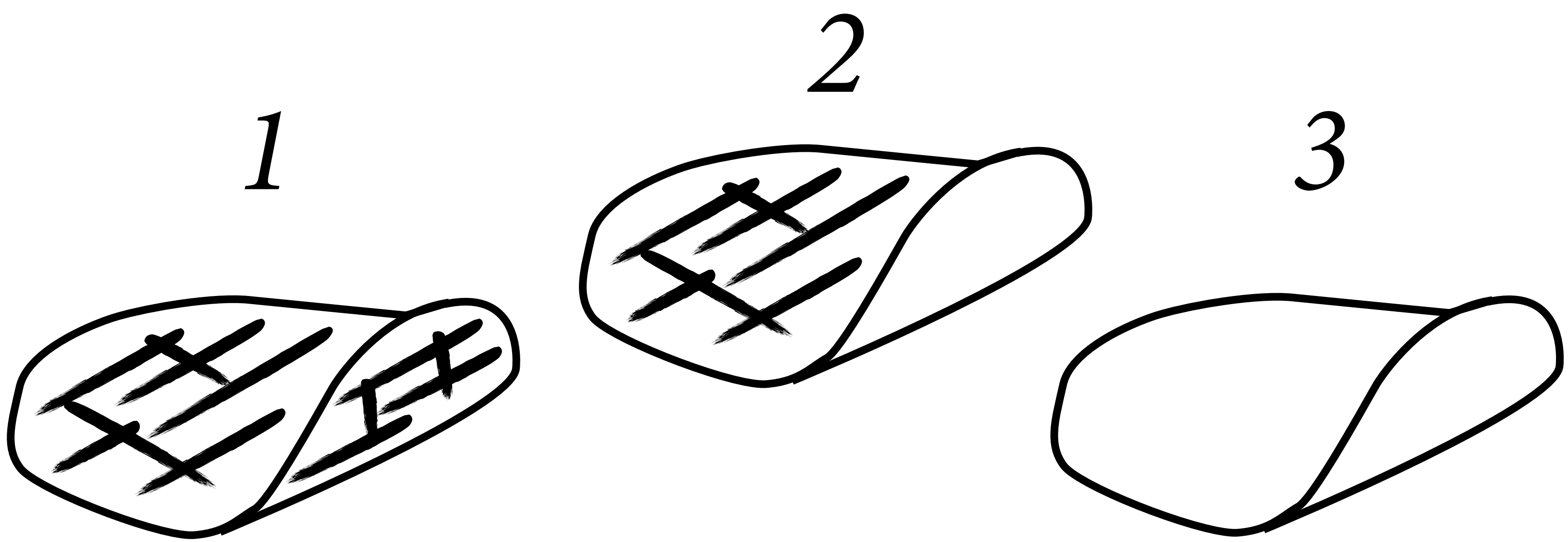
$$\begin{aligned}\Pr(\text{burnt up}) &= \Pr(\text{BB})(1) + \Pr(\text{BU})(0.5) + \Pr(\text{UU})(0) \\ &= (1/3) + (1/3)(1/2) = 0.5\end{aligned}$$



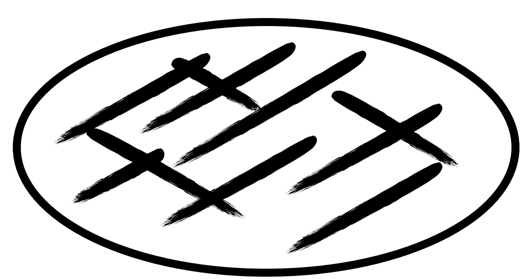
$$\Pr(\text{burnt down}|\text{burnt up}) = \frac{\Pr(\text{burnt up, burnt down})}{\Pr(\text{burnt up})}$$

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

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



You are served:



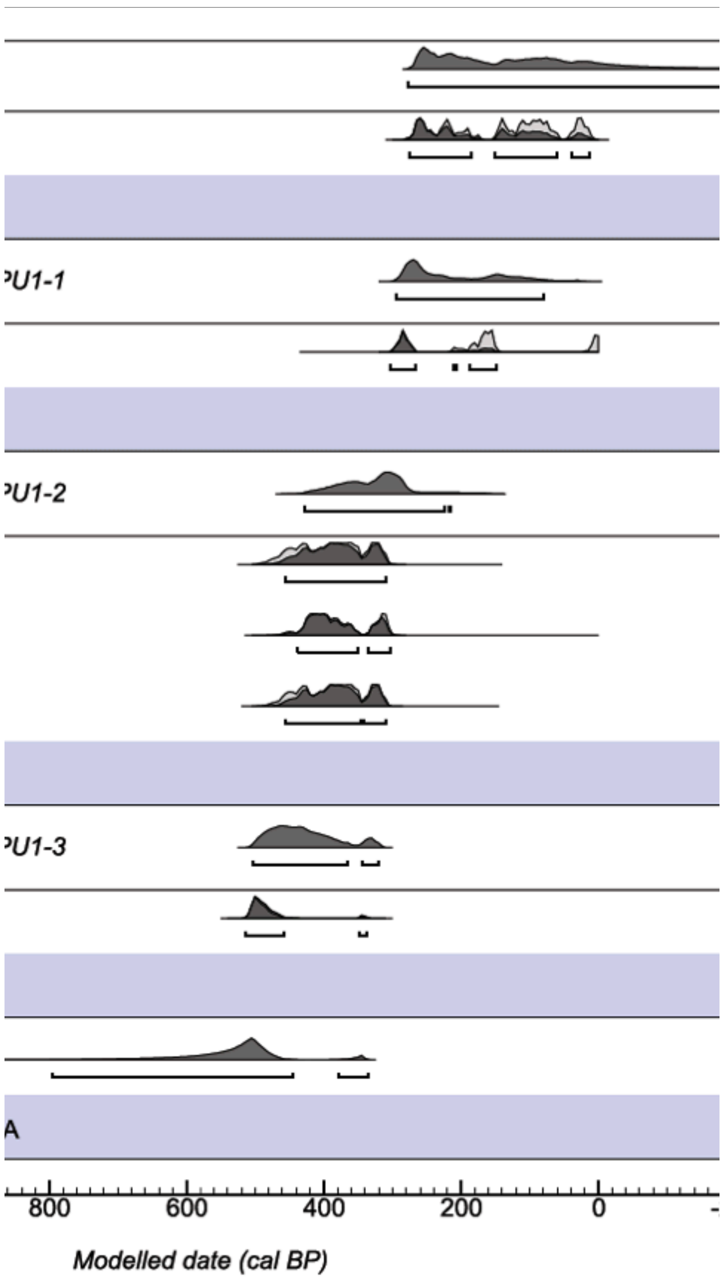
Probability other side is burnt?

Getting Ruthless

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

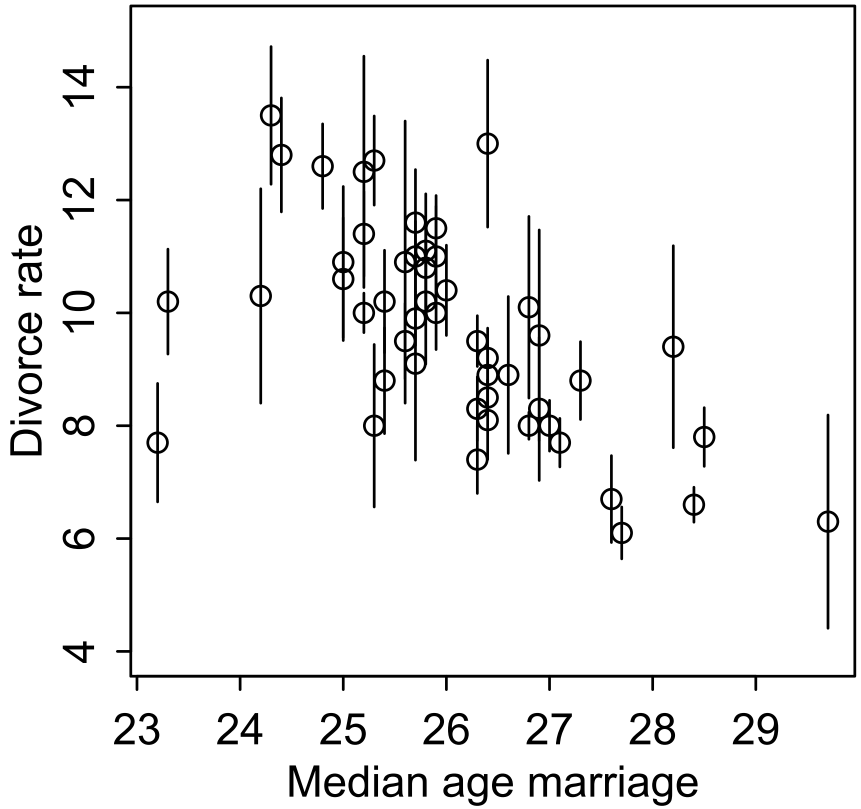
Measurement error

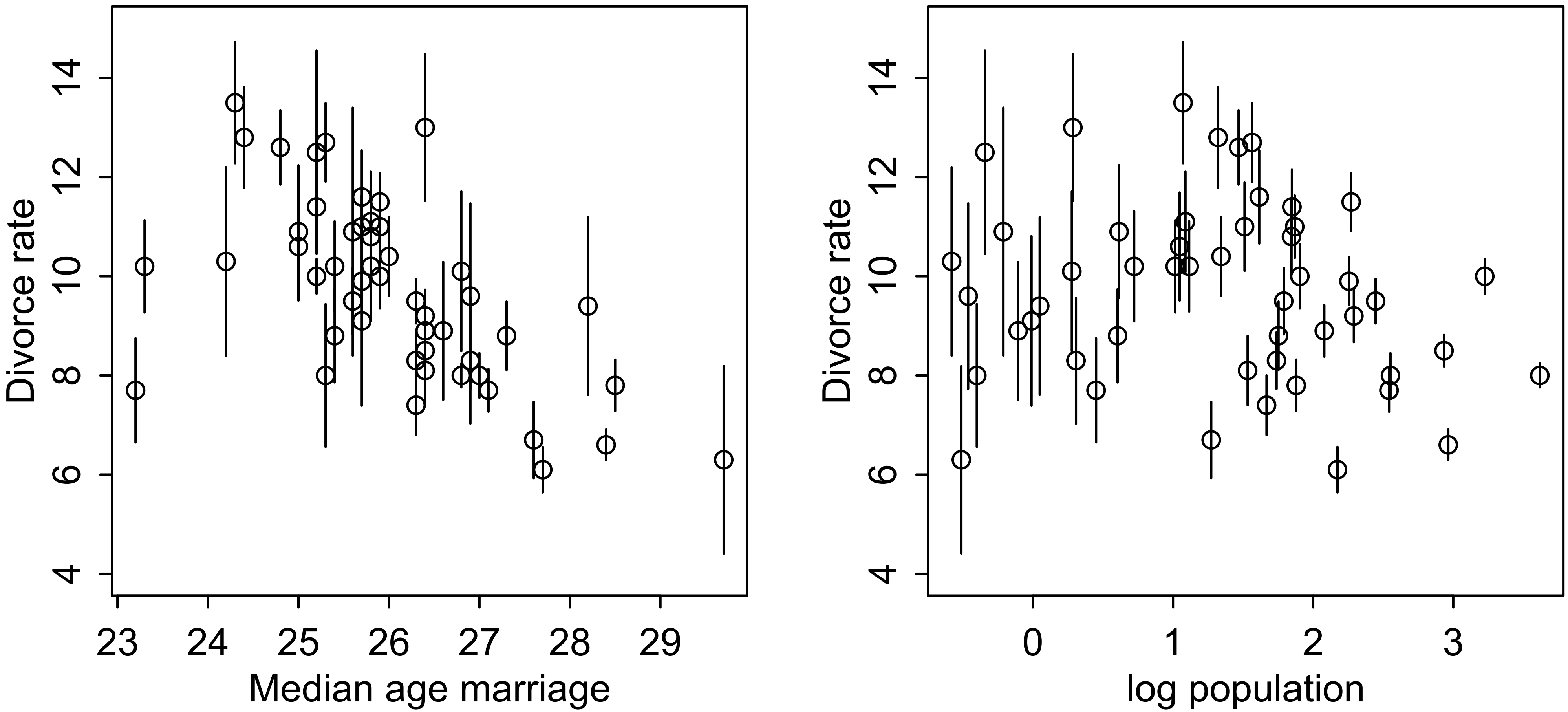
- Measurement always entails error
- Typical linear regression: interpret sigma as "error" on outcome
- What if error isn't constant?
- What if error is on predictors?



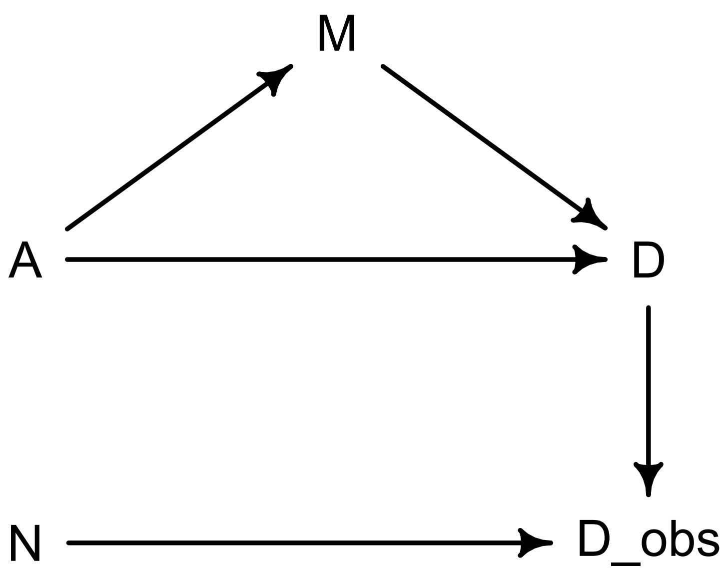
Error on outcome

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

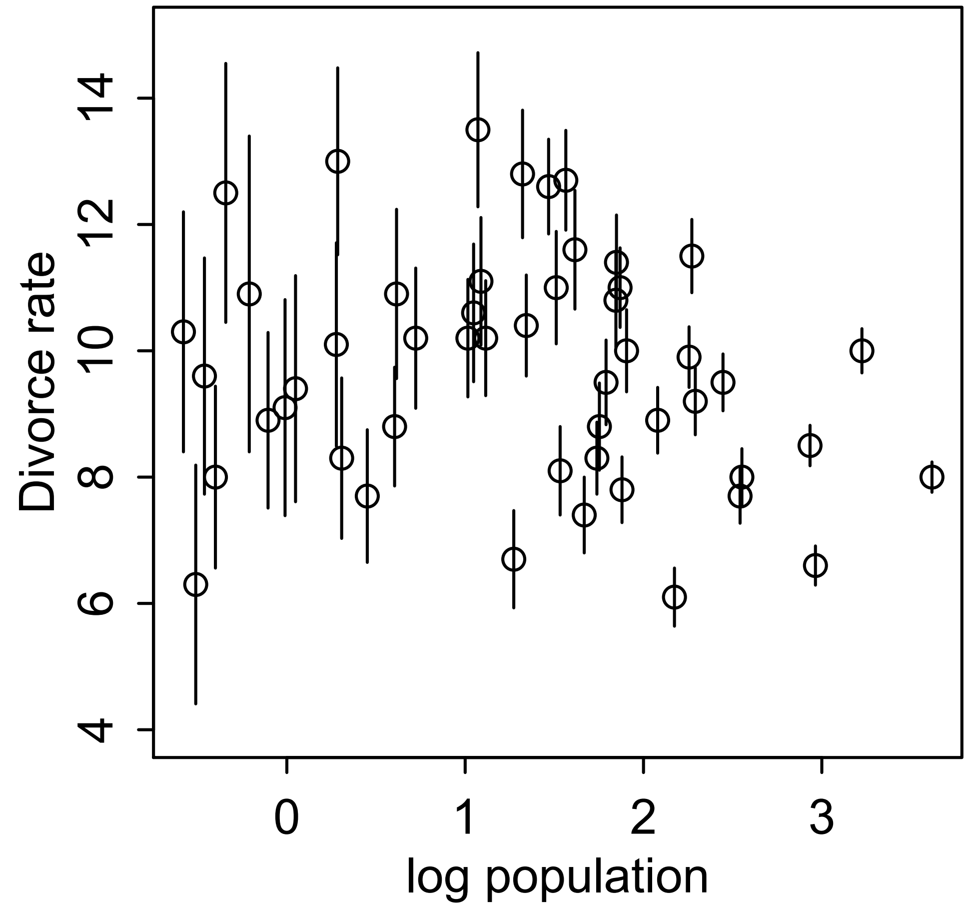




Divorce rate
4
6
8
10
12
14
23
24
25
26
27
28
29
Median age marriage
Divorce rate
4
6
8
10
12
14
0
1
2
3
log population



M
A
D
N
D_obs



Divorce rate
log population
4
6
8
10
12
14
0
1
2
3

Error on outcome

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

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

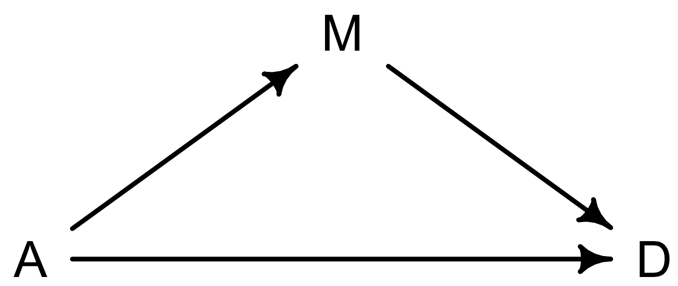
observed (data)

true (parameter)

std error (data)

Error on outcome: model

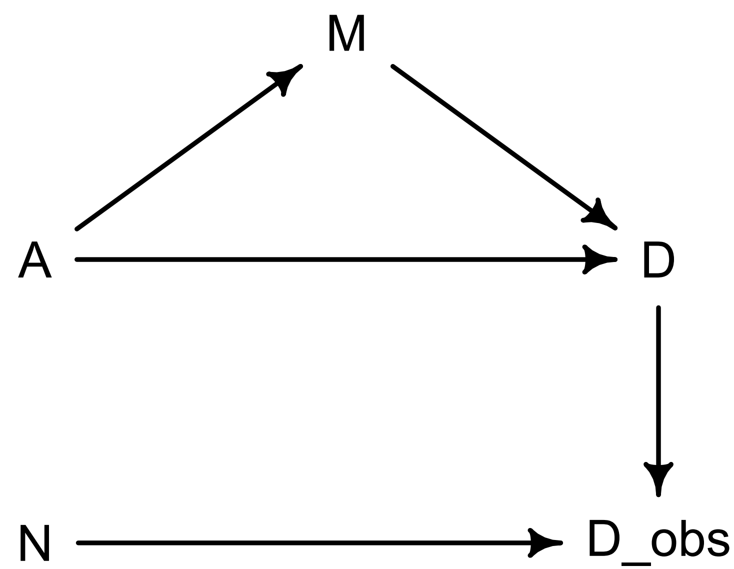
$$
\begin{aligned}
\color{blue}{D_{\text{TRUE},i}} &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_{\text{A}} A_i + \beta_{\text{M}} M_i \\
\alpha &\sim \text{Normal}(0, 0.2) \\
\beta_{\text{A}} &\sim \text{Normal}(0, 0.5) \\
\beta_{\text{M}} &\sim \text{Normal}(0, 0.5) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$



estimate

standard error of observation

$$
\begin{aligned}
D_{\text{OBS},i} &\sim \text{Normal}(D_{\text{TRUE},i}, D_{\text{SE},i}) \\
D_{\text{TRUE},i} &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_{\text{A}} A_i + \beta_{\text{M}} M_i \\
\alpha &\sim \text{Normal}(0, 0.2) \\
\beta_{\text{A}} &\sim \text{Normal}(0, 0.5) \\
\beta_{\text{M}} &\sim \text{Normal}(0, 0.5) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

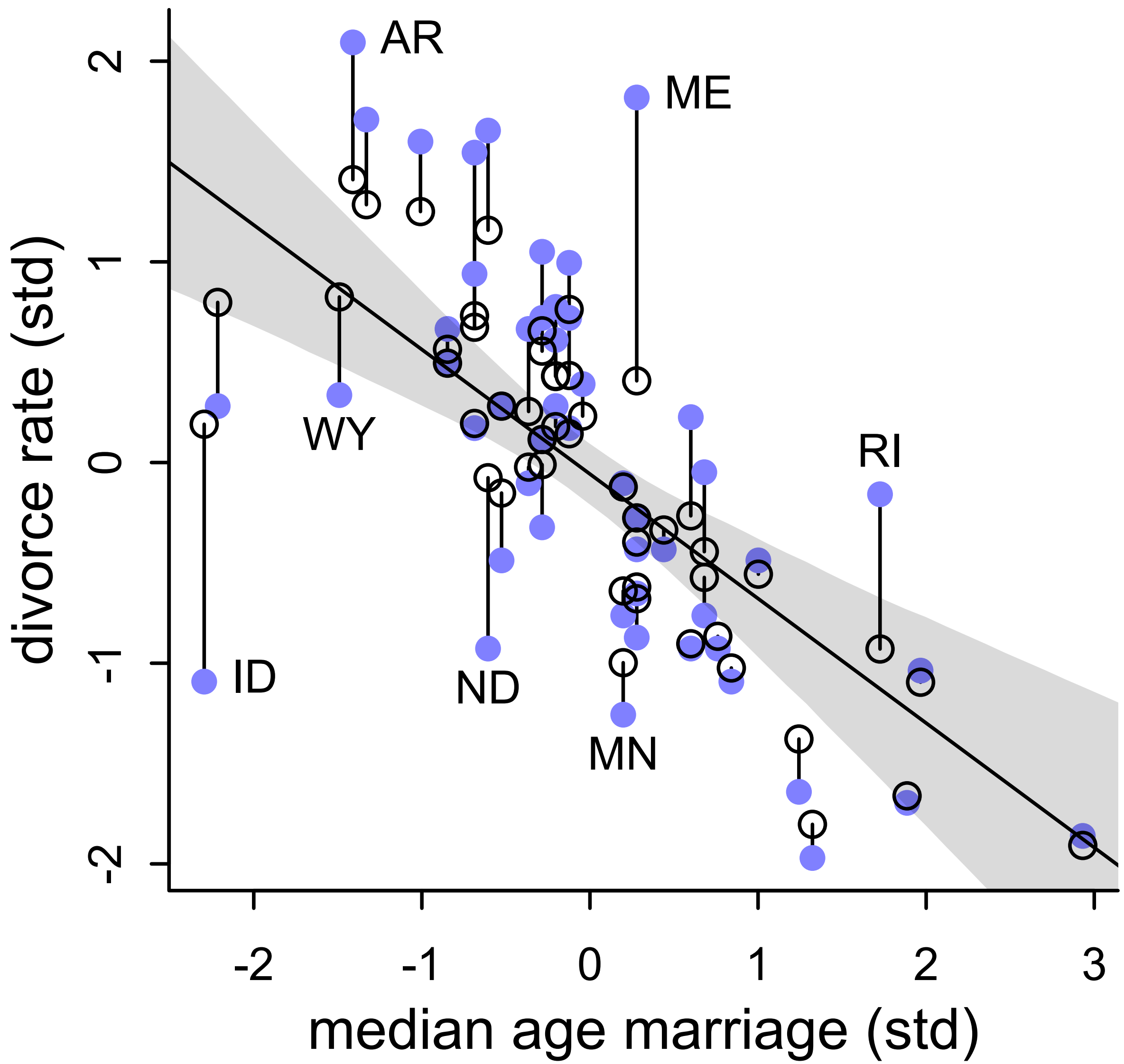


Error on outcome: fitting

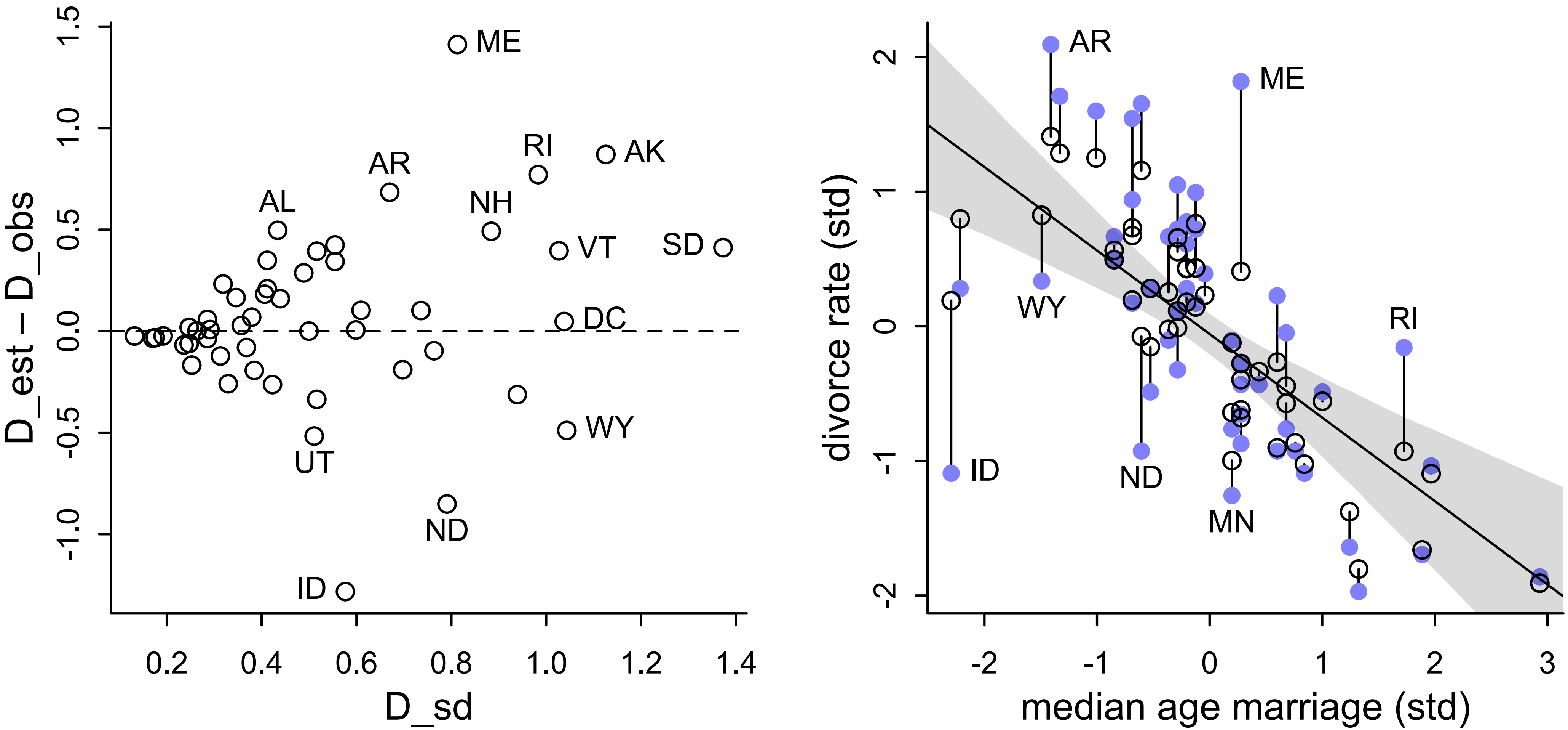
```
m15.1 <- ulam(
    alist(
        D_obs ~ dnorm( D_true , D_sd ),
        vector[N]:D_true ~ dnorm( mu , sigma ),
        mu <- a + bA*A + bM*M,
        a ~ dnorm(0,0.2),
        bA ~ dnorm(0,0.5),
        bM ~ dnorm(0,0.5),
        sigma ~ dexp(1)
    ) , data=dlist , chains=4 , cores=4 )
```

Error on outcome: fitting

```
m15.1 <- ulam(
    alist(
        D_obs ~ dnorm( D_true , D_sd ),
        vector[N]:D_true ~ dnorm( mu , sigma ),
        mu <- a + bA*A + bM*M,
        a ~ dnorm(0,0.2),
        bA ~ dnorm(0,0.5),
        bM ~ dnorm(0,0.5),
        sigma ~ dexp(1)
    ) , data=dlist , chains=4 , cores=4 )
```



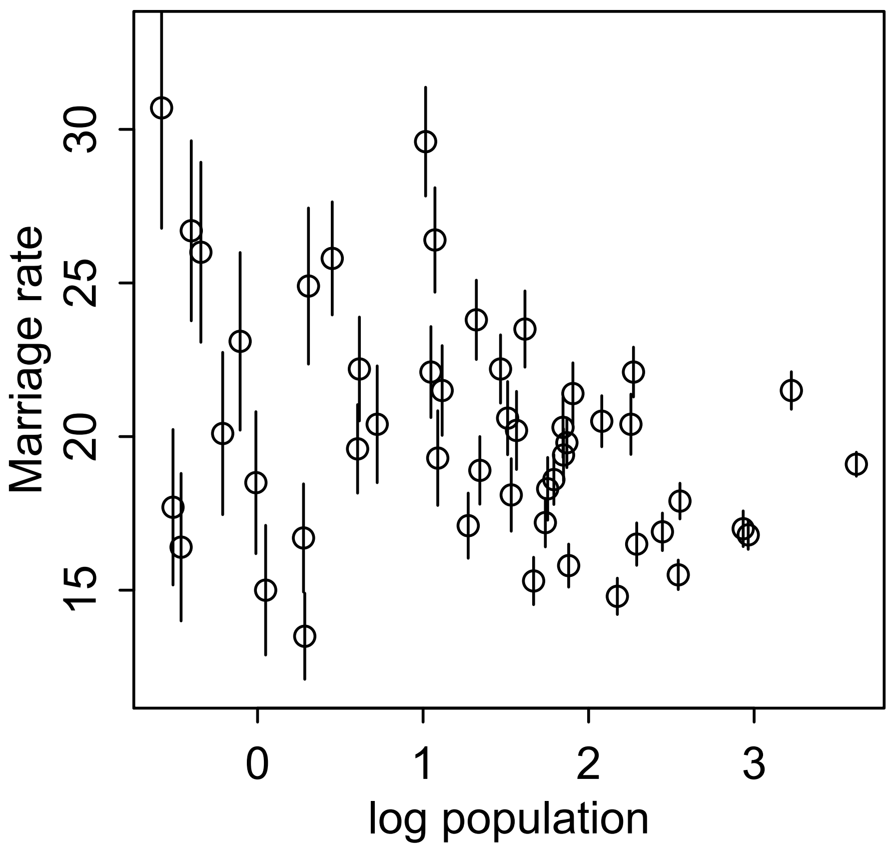
- Divorce rate estimates move from observed values. Why?



- Shrinkage! Uncertain or extreme states shrink to regression line.

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

$$
\begin{aligned}
D_{\text{OBS},i} &\sim \text{Normal}(D_{\text{TRUE},i}, D_{\text{SE},i}) \\
D_{\text{TRUE},i} &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_{\text{A}} A_i + \beta_{\text{M}} M_{\text{TRUE},i} \\
M_{\text{OBS},i} &\sim \text{Normal}(M_{\text{TRUE},i}, M_{\text{SE},i}) \\
M_{\text{TRUE},i} &\sim \text{Normal}(0, 1) \\
\alpha &\sim \text{Normal}(0, 0.2) \\
\beta_{\text{A}} &\sim \text{Normal}(0, 0.5) \\
\beta_{\text{M}} &\sim \text{Normal}(0, 0.5) \\
\sigma &\sim \text{Exponential}(1)
\end{aligned}
$$

Error on predictor: model

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

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

$$\mu_i = \alpha + \beta_{\text{A}} A_i + \beta_{\text{M}} M_{\text{TRUE},i}$$

$$M_{\text{OBS},i} \sim \text{Normal}(M_{\text{TRUE},i}, M_{\text{SE},i})$$

use estimates in regression

likelihood for observed rate

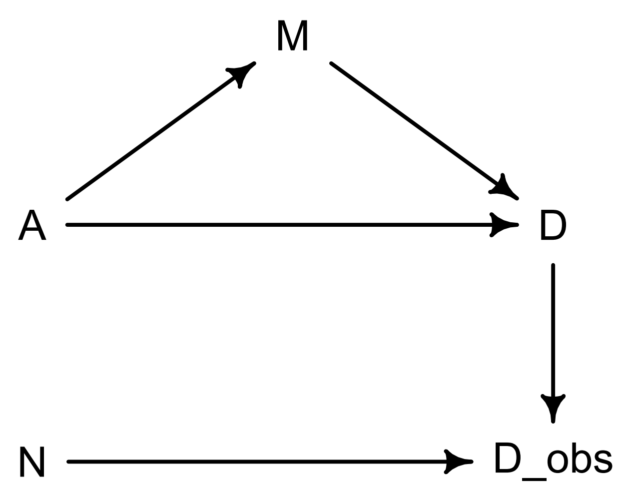
estimated marriage rate

standard error of marriage rate

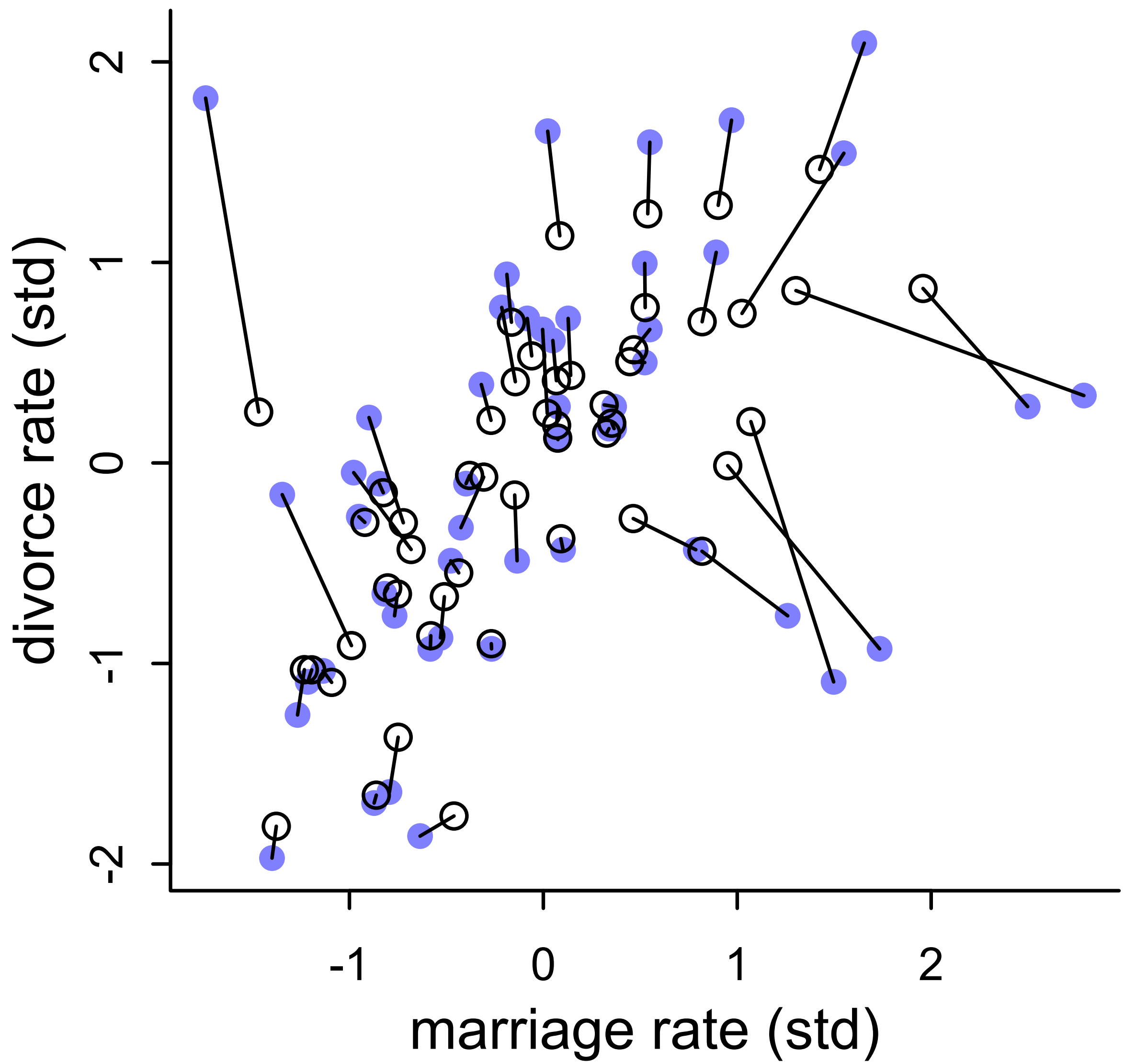
Error on predictor: model

$$
\begin{aligned}
D_{\text{OBS},i} &\sim \text{Normal}(D_{\text{TRUE},i}, D_{\text{SE},i}) \\
D_{\text{TRUE},i} &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_{\text{A}} A_i + \beta_{\text{M}} M_{\text{TRUE},i} \\
M_{\text{OBS},i} &\sim \text{Normal}(M_{\text{TRUE},i}, M_{\text{SE},i}) \\
\textit{prior rates} \longrightarrow \quad M_{\text{TRUE},i} &\sim \text{Normal}(0, 1)
\end{aligned}
$$

Not the best approach:
M and A are associated!
Will do better later on.

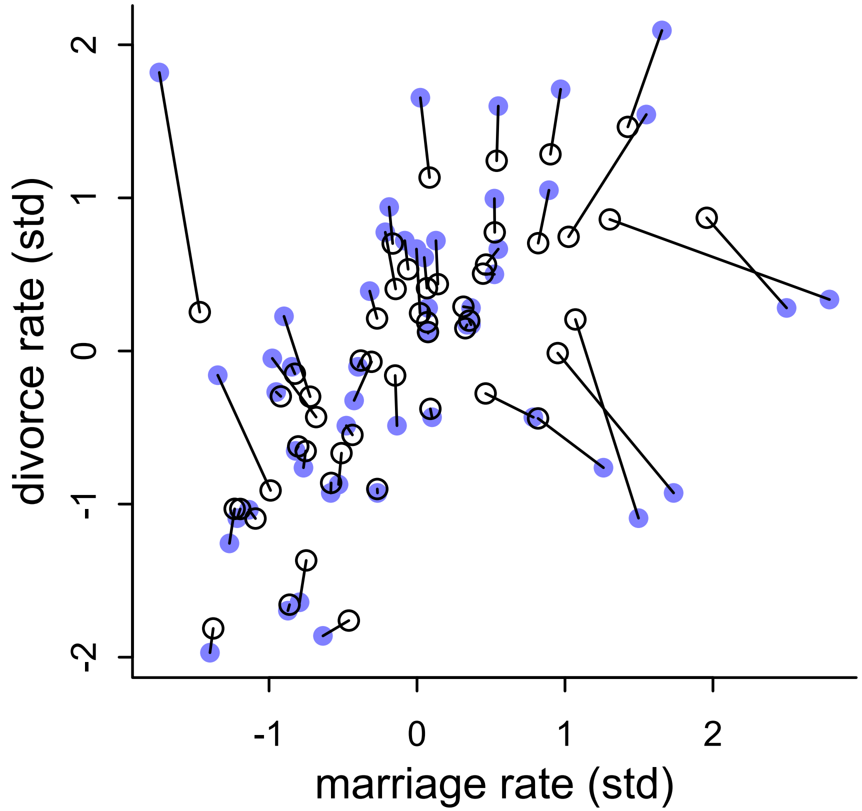


filled circles: observed
open circles: estimated
lines connect points for same State



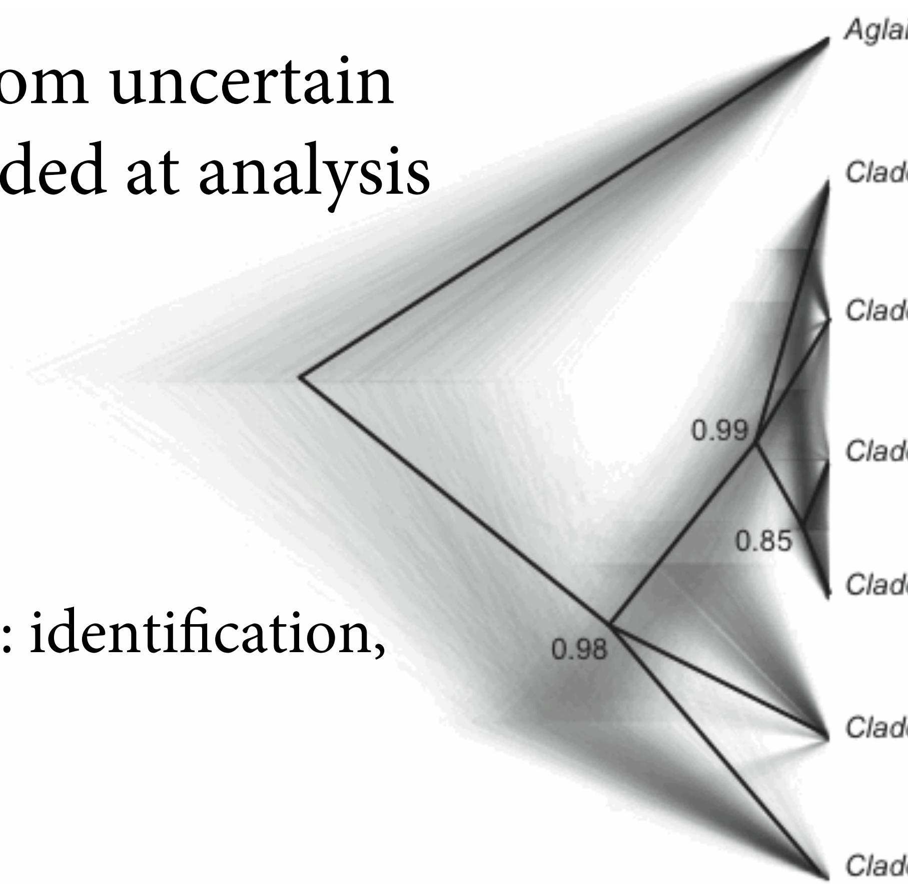
Error on predictor

- Both divorce rate and marriage rate shrink
- Divorce shrinks 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
  - Parentage analysis
  - Phylogenetics: distribution of trees
  - Archaeology/paleontology/forensics: identification, sexing, aging, dating
- Propagate uncertainty



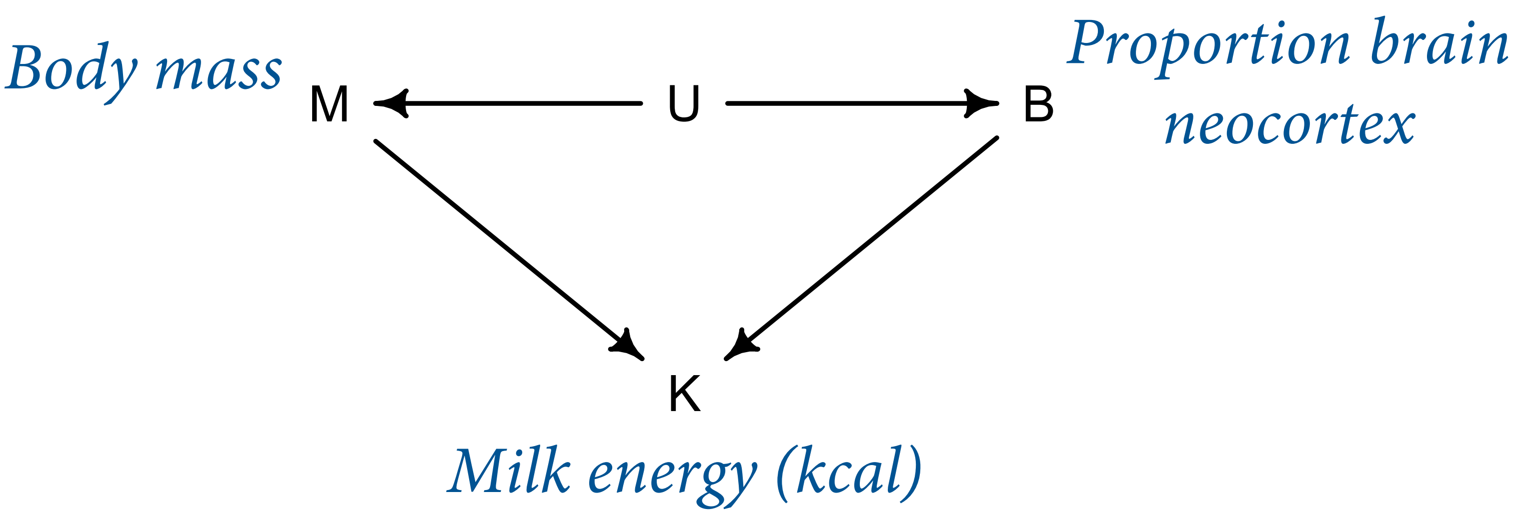
Missing data

- 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.*

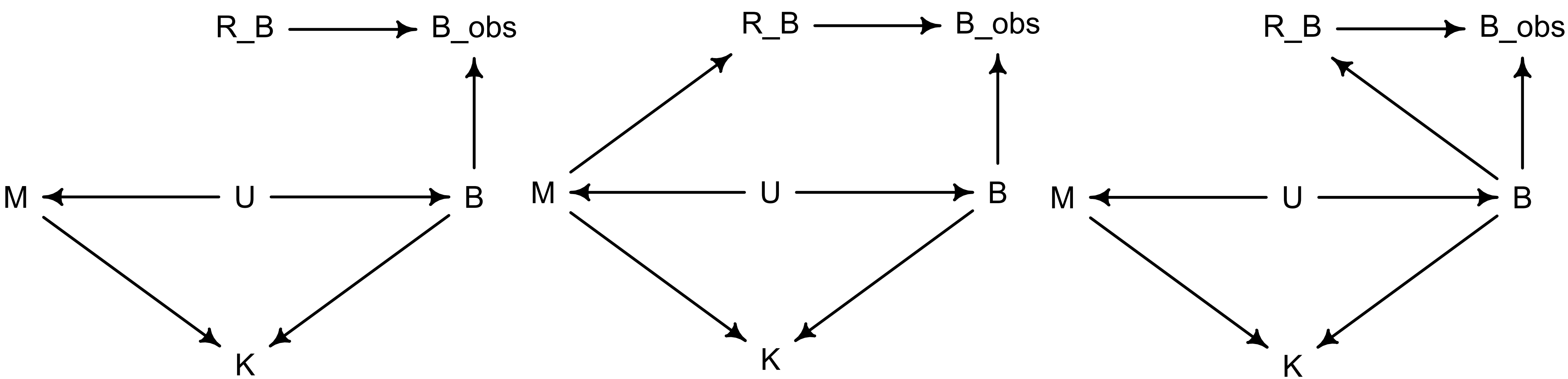
Why impute?

- Missingness can be a confound



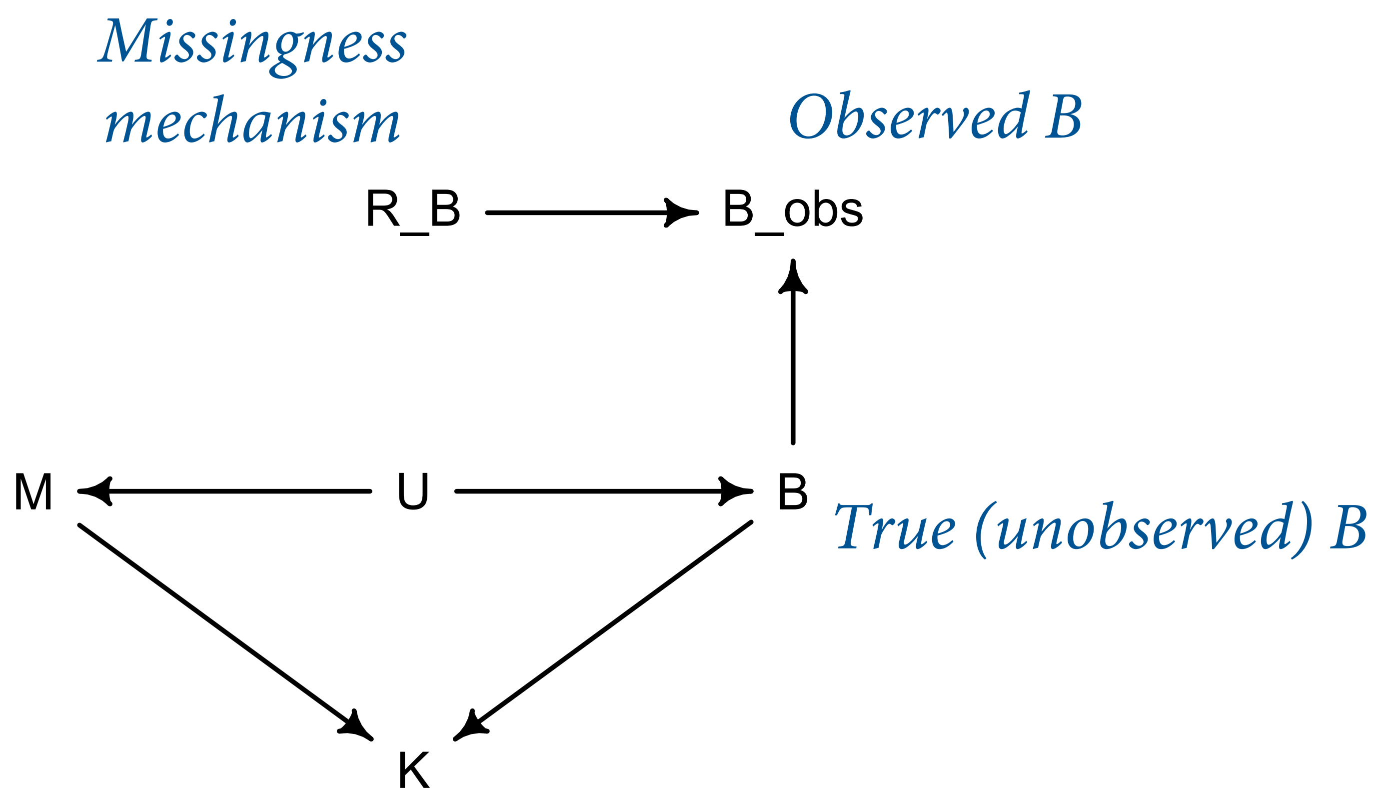
Three Types of Missingness

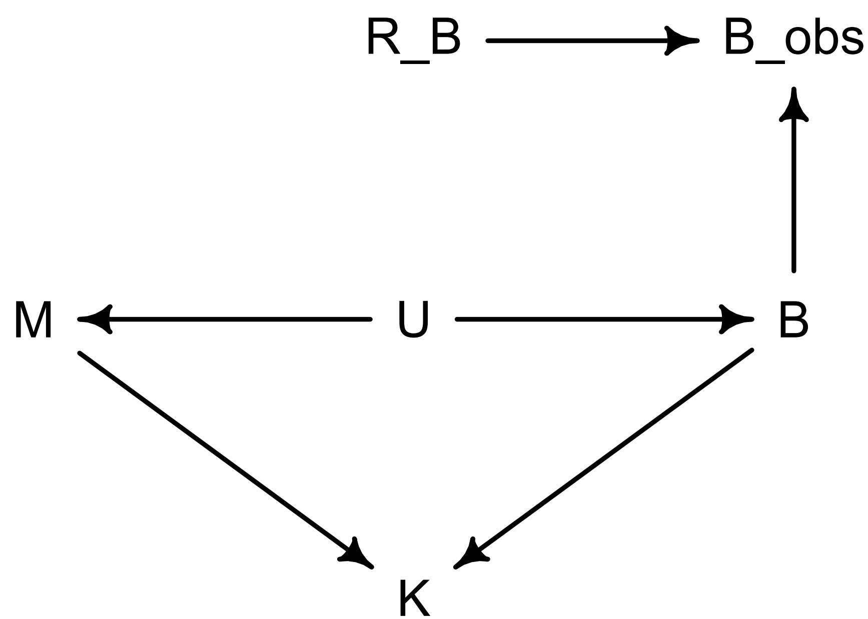
MCAR | MAR | MNAR



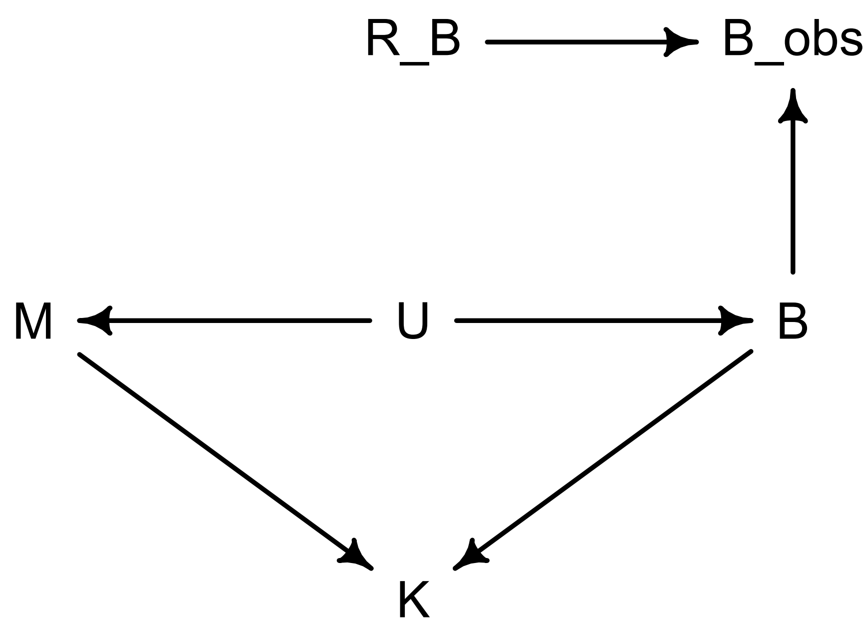
MISSING COMPLETELY AT RANDOM | MISSING AT RANDOM | MISSING NOT AT RANDOM

Possibly most confusing statistical terms ever invented.


Missingness mechanism
Observed B
R_B
B_obs
M
U
B
True (unobserved) B
K



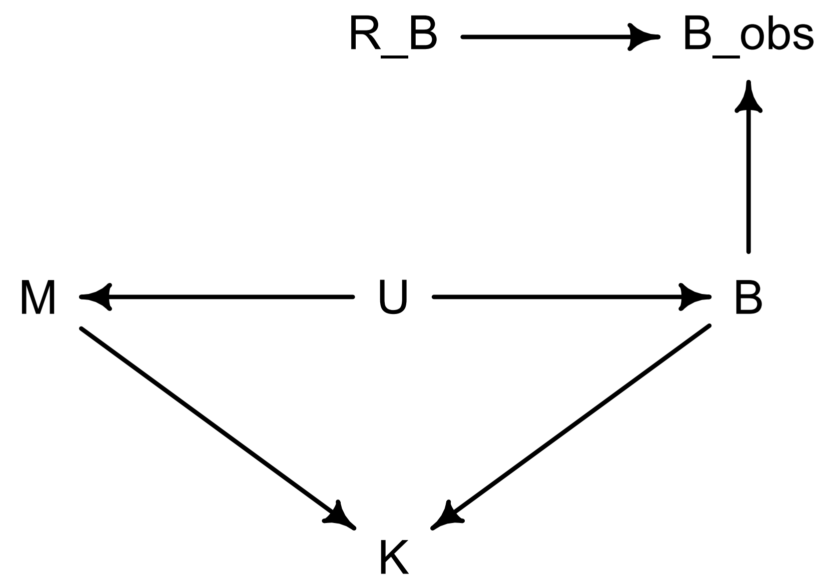
Are there any backdoors from B_obs to K?

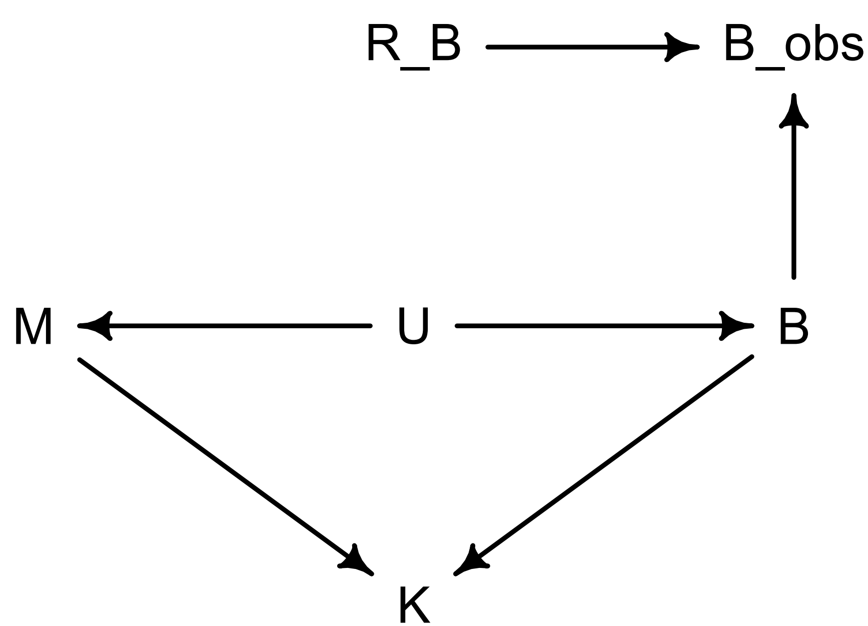


Can condition on M for direct effect.
Either way, R_B is ignorable.

Missing Completely At Random

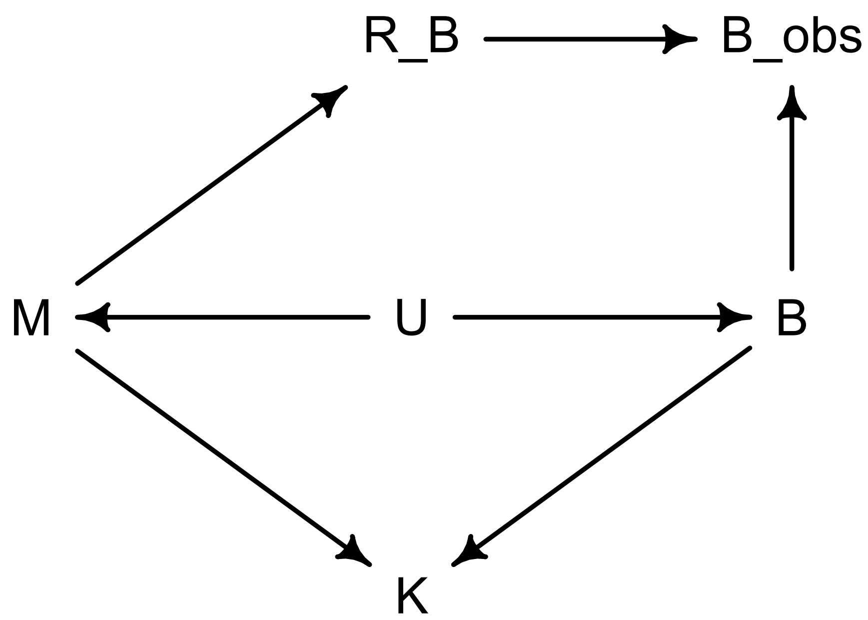
- MCAR: K is unconditionally independent of R_B
- Do not need to condition on anything for R_B not to be a confound
- On right, no path through R_B, conditioning on B_obs
- Do not NEED to impute
- But imputation adds precision





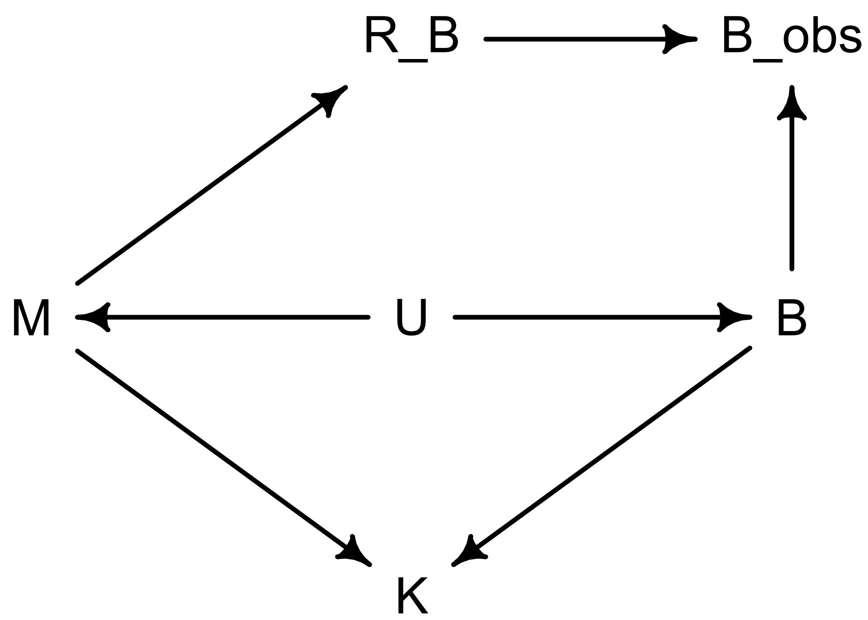
Does MCAR ever happen in real data?
Research assistant randomly deletes values?

Missing At Random



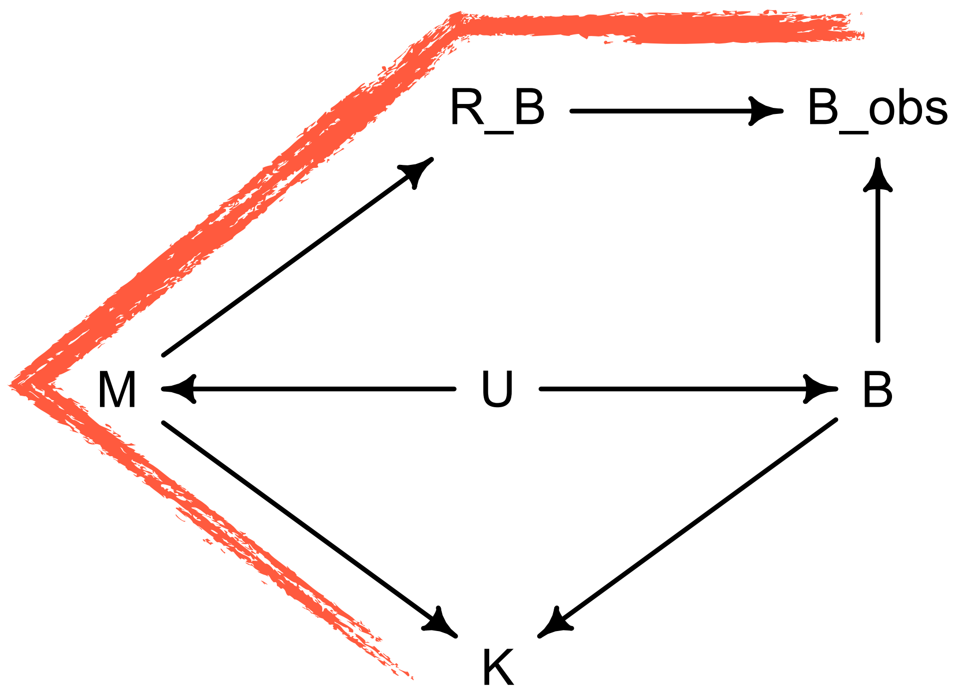
Missingness more likely for specific values of M.
How can this happen?

Missing At Random



Backdoor path from B_obs to K?

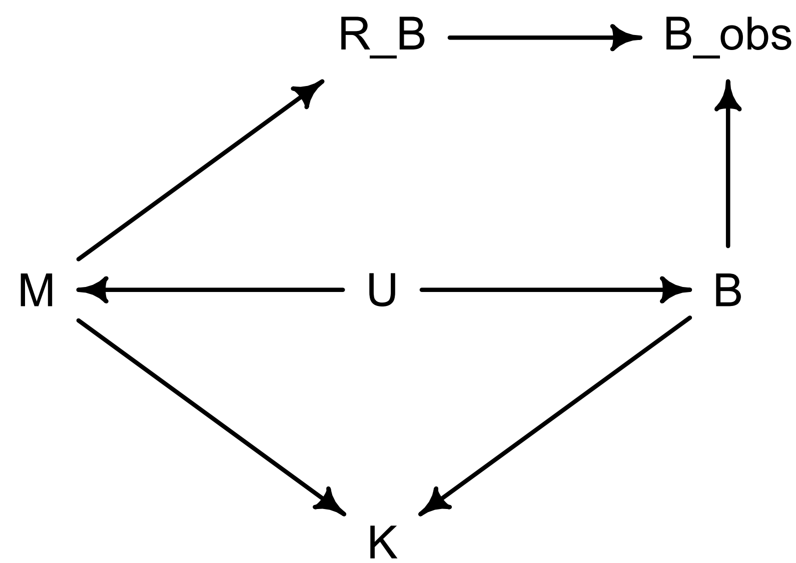
Missing At Random



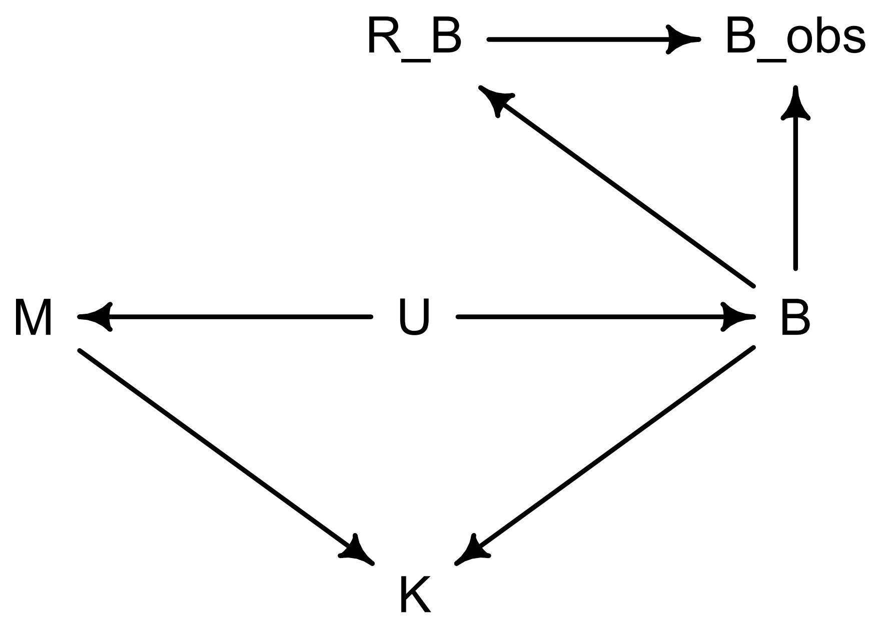
Backdoor path from B_obs to K?
Can condition on M to close.

Missing (Simply) At Random

- MAR: K is conditionally independent of R_B
- Must to condition on M for R_B not to be a confound
- Still must impute to de-bias estimates
- Why? If you delete cases of M/K where B is missing, missingness obscures causation.

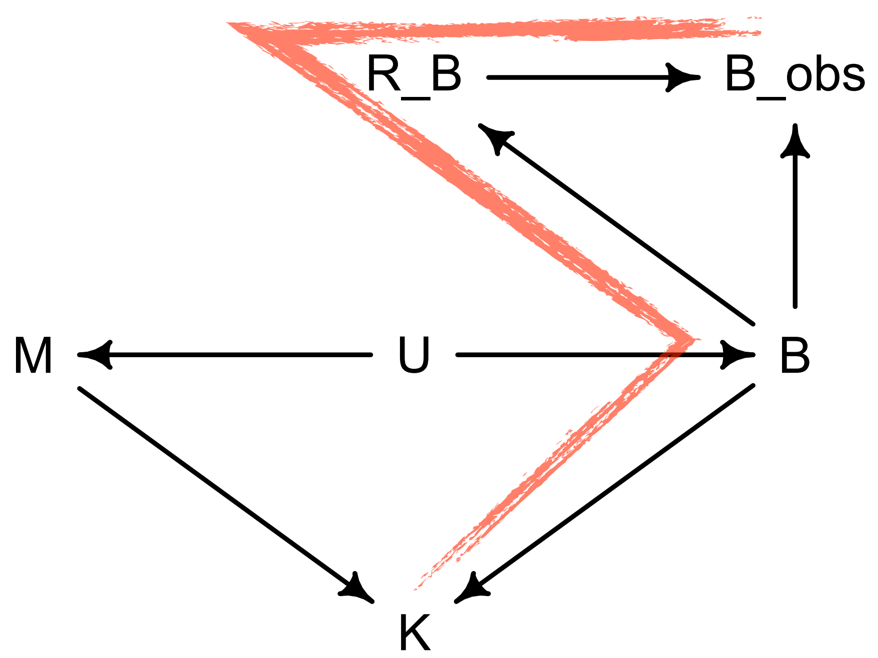


Missing Not At Random



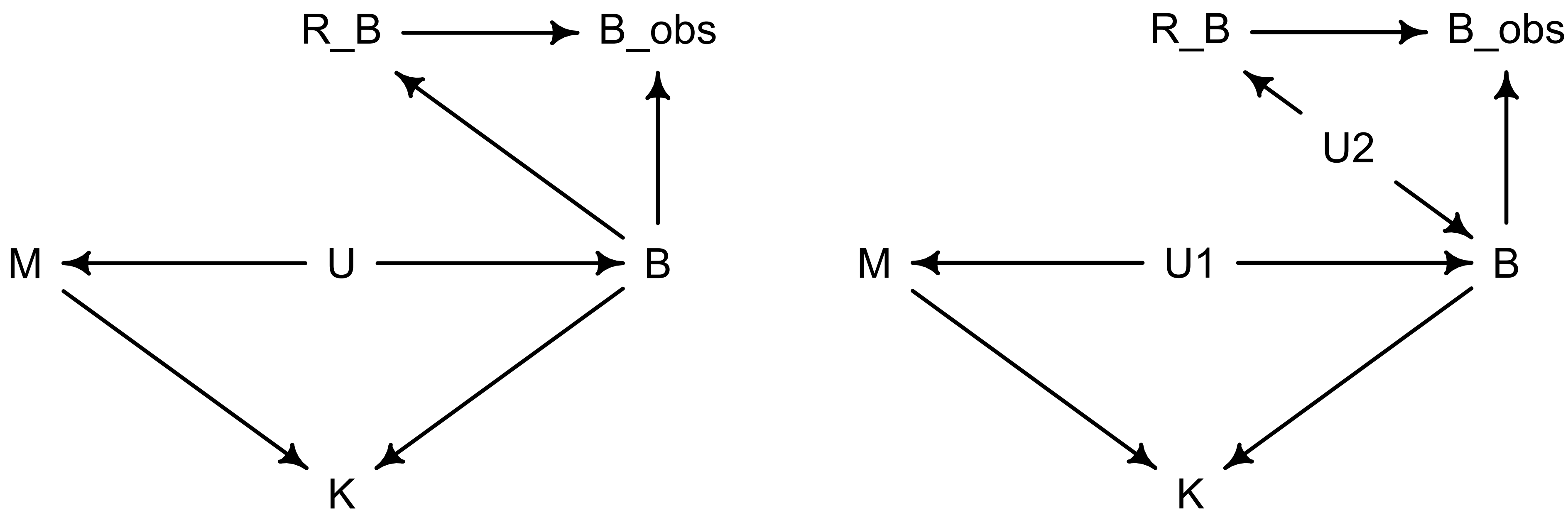
Missingness more likely for specific values of B.
How can this happen?

Missing Not At Random



No way to shut the backdoor!

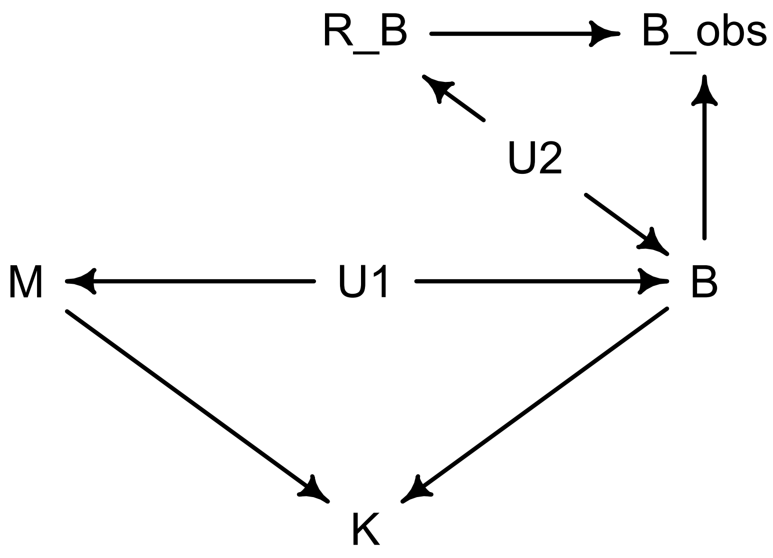
Missing Not At Random



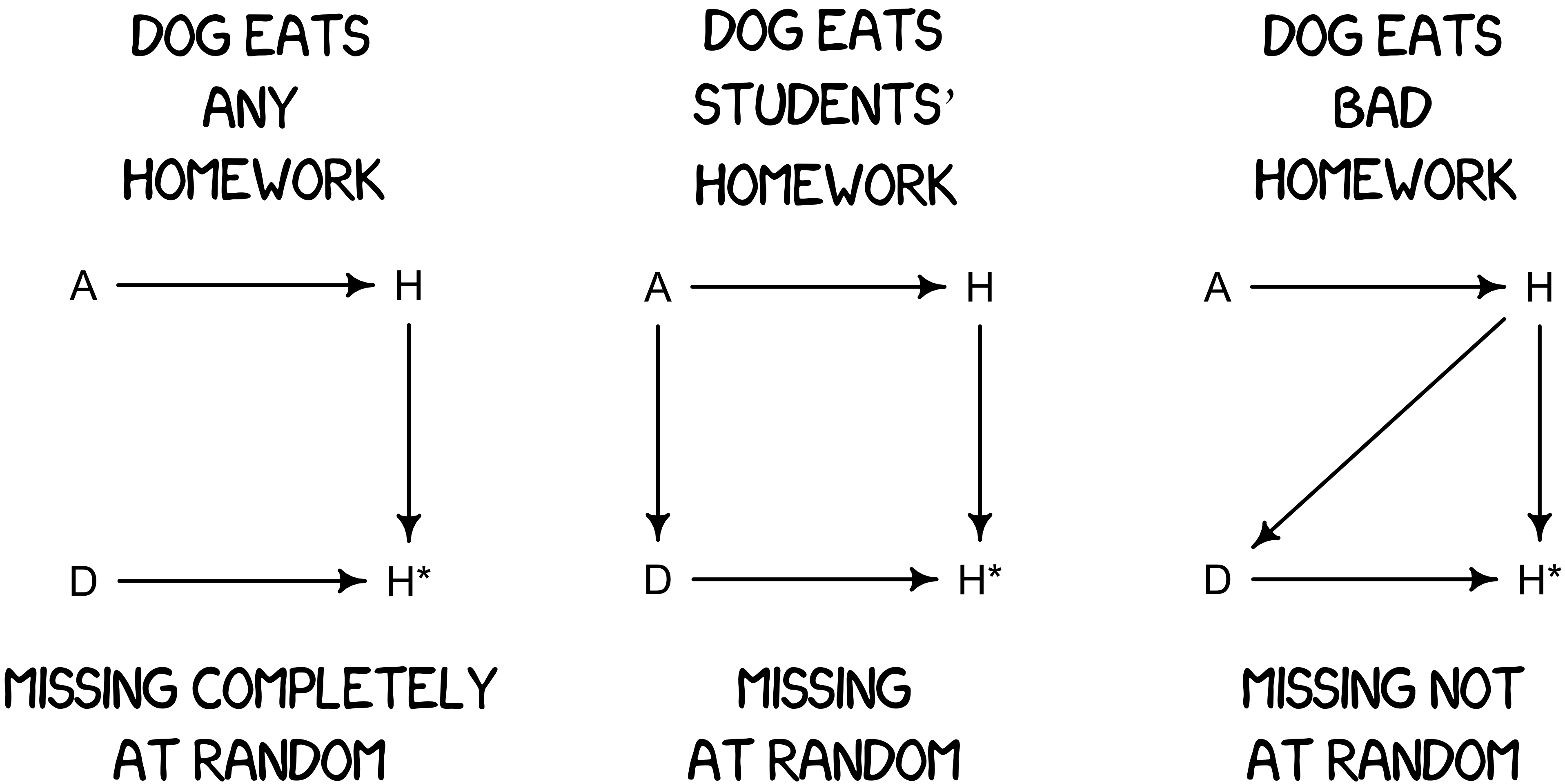
Can also arise through unobserved variables (right).

Missing Not At Random

- MNAR: K is unconditionally **dependent** on R_B
- If you can model R_B, might be okay
- No guarantees



H: Homework
H*: Homework with missing values
A: Attribute of student
D: Dog (missingness mechanism)



Milk imputation

- `data(milk)`
- 12 missing values for neocortex
- Suppose values are ***Missing At Random*** (**MAR**)
  - Distribution of observed values provides information
  - Can use to impute missing values
  - Same procedure for MCAR

	kcal.per.g	mass	neocortex.perc
1	0.49	1.95	55.16
2	0.51	2.09	NA
3	0.46	2.51	NA
4	0.48	1.62	NA
5	0.60	2.19	NA
6	0.47	5.25	64.54
7	0.56	5.37	64.54
8	0.89	2.51	67.64
9	0.91	0.71	NA
10	0.92	0.68	68.85
11	0.80	0.12	58.85
12	0.46	0.47	61.69
13	0.71	0.32	60.32
14	0.71	0.60	NA
15	0.73	3.47	NA
16	0.68	1.55	69.97
17	0.72	7.08	NA
18	0.97	3.24	70.41
19	0.79	7.94	NA
20	0.84	12.30	73.40
21	0.48	7.59	NA
22	0.62	5.37	67.53
23	0.51	10.72	NA
24	0.54	35.48	71.26
25	0.49	79.43	72.60
26	0.53	97.72	NA
27	0.48	40.74	70.24
28	0.55	33.11	76.30
29	0.71	54.95	75.49

Milk energy MAR

- Consider just neocortex variable:
  - Q: What is your best guess of each missing value?
  - A: Posterior distribution derived from remaining data

	neocortex.perc
1	55.16
2	NA
3	NA
4	NA
5	NA
6	64.54
7	64.54
8	67.64
9	NA
10	68.85
11	58.85
12	61.69
13	60.32
14	NA
15	NA
16	69.97
17	NA
18	70.41
19	NA
20	73.40
21	NA
22	67.53
23	NA
24	71.26
25	72.60
26	NA
27	70.24
28	76.30
29	75.49

Milk energy MCAR

- Place a unique parameter for each missing value
  - NC1 ... NC12
  - These are values to be imputed

	neocortex.perc
1	55.16
2	NC1
3	NC2
4	NC3
5	NC4
6	64.54
7	64.54
8	67.64
9	NC5
10	68.85
11	58.85
12	61.69
13	60.32
14	NC6
15	NC7
16	69.97
17	NC8
18	70.41
19	NC9
20	73.40
21	NC10
22	67.53
23	NC11
24	71.26
25	72.60
26	NC12
27	70.24
28	76.30
29	75.49

Milk energy MAR: model

$$B = [0.55, B_2, B_3, B_4, 0.65, 0.65, ..., 0.76, 0.75]$$

$$\begin{aligned}
K_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_B B_i + \beta_M \log M_i \\
B_i &\sim \text{Normal}(\nu, \sigma_B) \\
\alpha &\sim \text{Normal}(0, 0.5) \\
\beta_B &\sim \text{Normal}(0, 0.5) \\
\beta_M &\sim \text{Normal}(0, 0.5) \\
\sigma &\sim \text{Exponential}(1) \\
\nu &\sim \text{Normal}(0.5, 1) \\
\sigma_B &\sim \text{Exponential}(1)
\end{aligned}$$

Milk energy MAR: model

$$B = [0.55, B_2, B_3, B_4, 0.65, 0.65, ..., 0.76, 0.75]$$

linear model using mix of observed and imputed values

$$
\begin{aligned}
K_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_B \boxed{B_i} + \beta_M \log M_i \\
B_i &\sim \text{Normal}(\nu, \sigma_B) \\
\alpha &\sim \text{Normal}(0, 0.5) \\
\beta_B &\sim \text{Normal}(0, 0.5) \\
\beta_M &\sim \text{Normal}(0, 0.5) \\
\sigma &\sim \text{Exponential}(1) \\
\nu &\sim \text{Normal}(0.5, 1) \\
\sigma_B &\sim \text{Exponential}(1)
\end{aligned}
$$

Milk energy MAR: model

$$B = [0.55, B_2, B_3, B_4, 0.65, 0.65, ..., 0.76, 0.75]$$

$$K_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_B B_i + \beta_M \log M_i$$
$$B_i \sim \text{Normal}(\nu, \sigma_B)$$

when obs, likelihood; when imputed, prior

mean neocortex (to be estimated)

std dev of neocortex (to be estimated)

Fitting

```
m15.3 <- ulam(
    alist(
        K ~ dnorm( mu , sigma ),
        mu <- a + bB*B + bM*M,
        B ~ dnorm( nu , sigma_B ),
        c(a,nu) ~ dnorm( 0 , 0.5 ),
        c(bB,bM) ~ dnorm( 0, 0.5 ),
        sigma_B ~ dexp( 1 ),
        sigma ~ dexp( 1 )
    ) , data=dat_list , chains=4 , cores=4 )
```

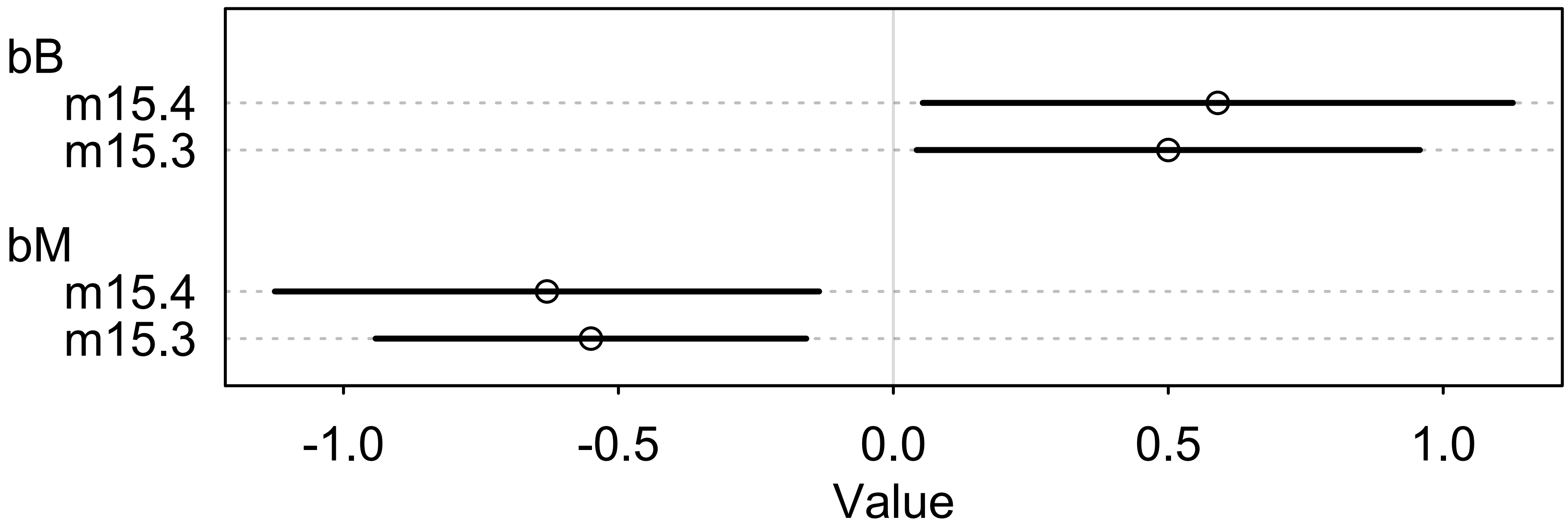
ulam detects NA values and tries to cope.
More explicit example in text.

R code 15.9

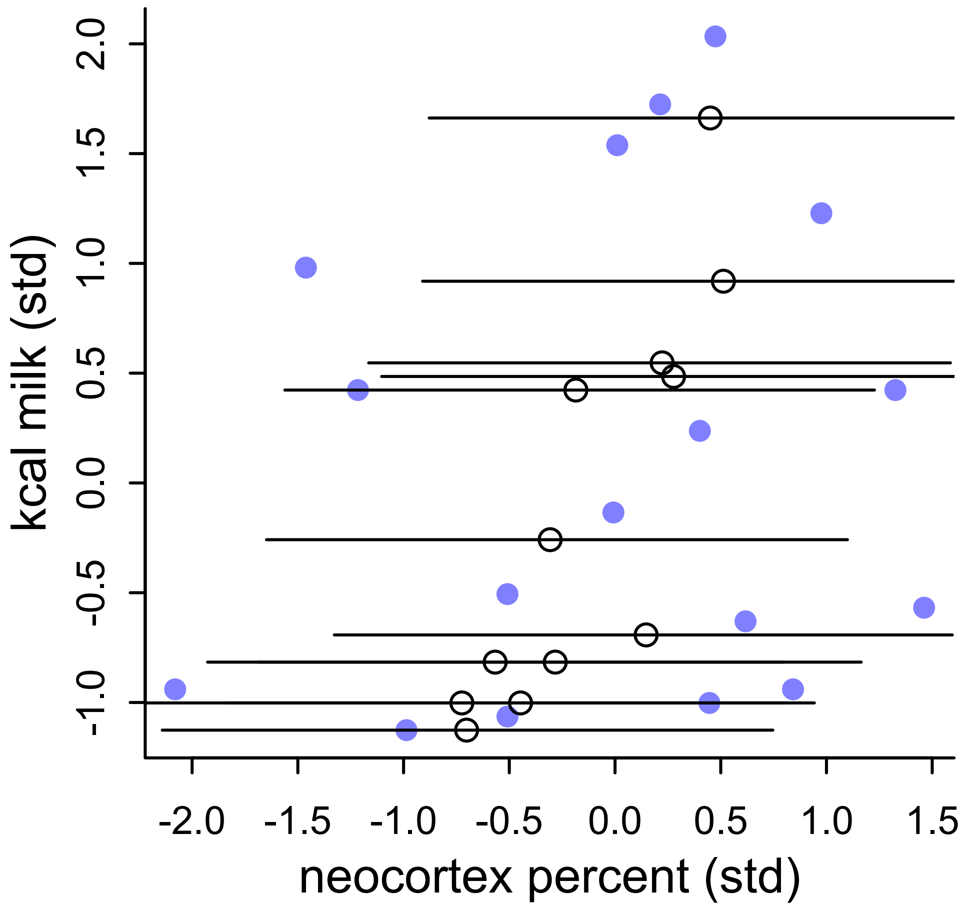
```
precis( m15.3 , depth=2 )
```

```
               mean   sd  5.5% 94.5% n_eff Rhat
nu            -0.04 0.20 -0.36  0.28  1806    1
a              0.02 0.17 -0.25  0.28  2080    1
bM            -0.55 0.20 -0.87 -0.22  1084    1
bB             0.50 0.23  0.12  0.86   818    1
sigma_B        1.01 0.17  0.77  1.31  1245    1
sigma          0.84 0.14  0.64  1.07  1010    1
B_impute[1]   -0.57 0.87 -1.93  0.80  2408    1
B_impute[2]   -0.70 0.92 -2.14  0.75  1940    1
B_impute[3]   -0.72 0.96 -2.24  0.81  1832    1
B_impute[4]   -0.31 0.87 -1.65  1.10  2345    1
B_impute[5]    0.45 0.88 -0.88  1.83  2416    1
B_impute[6]   -0.19 0.89 -1.56  1.23  2187    1
B_impute[7]    0.22 0.87 -1.16  1.59  2428    1
B_impute[8]    0.28 0.85 -1.10  1.62  2271    1
B_impute[9]    0.51 0.87 -0.91  1.88  2877    1
B_impute[10]  -0.45 0.89 -1.84  0.94  2668    1
B_impute[11]  -0.28 0.88 -1.69  1.16  2404    1
B_impute[12]   0.15 0.91 -1.33  1.60  2483    1
```

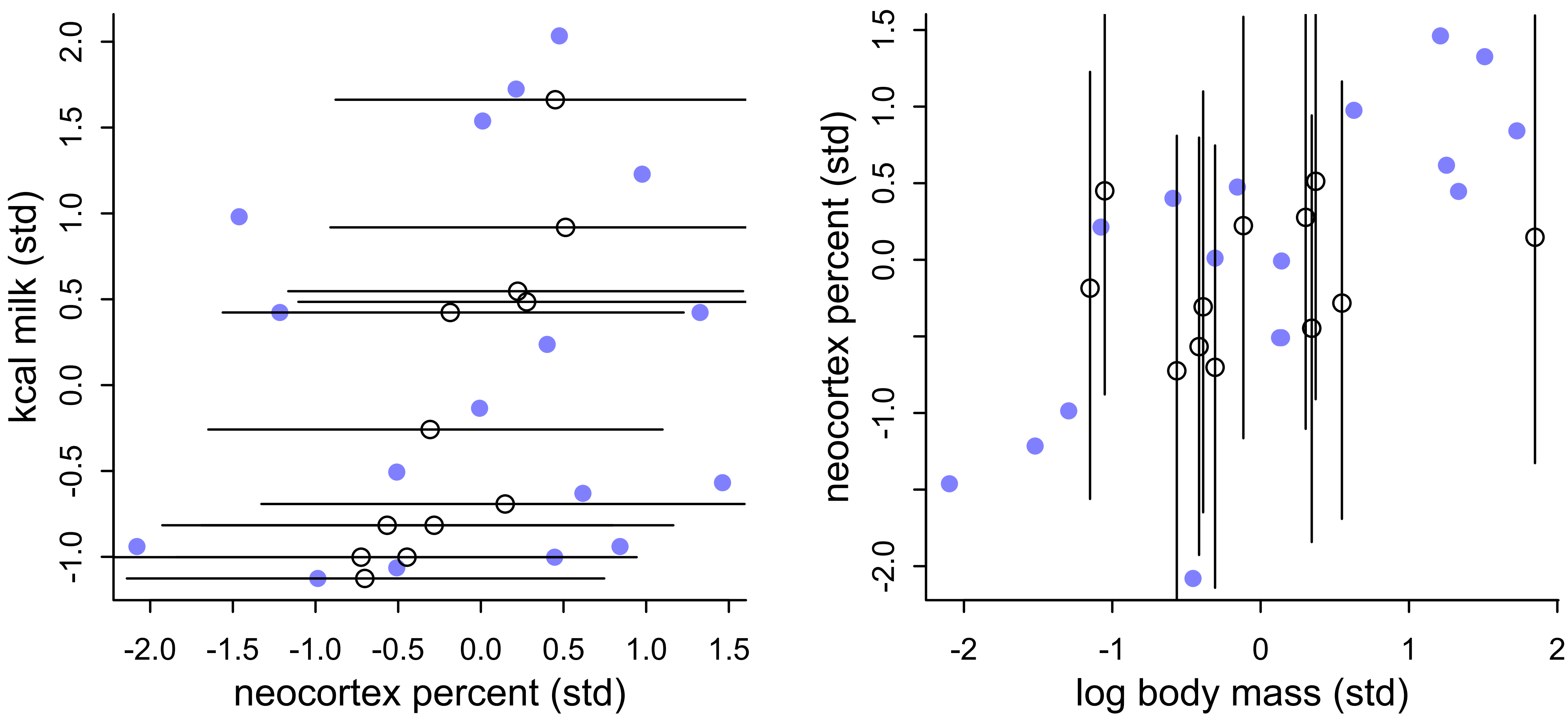
Compared to complete-cases



m15.3: full sample (with imputation)
m15.4: complete-cases only



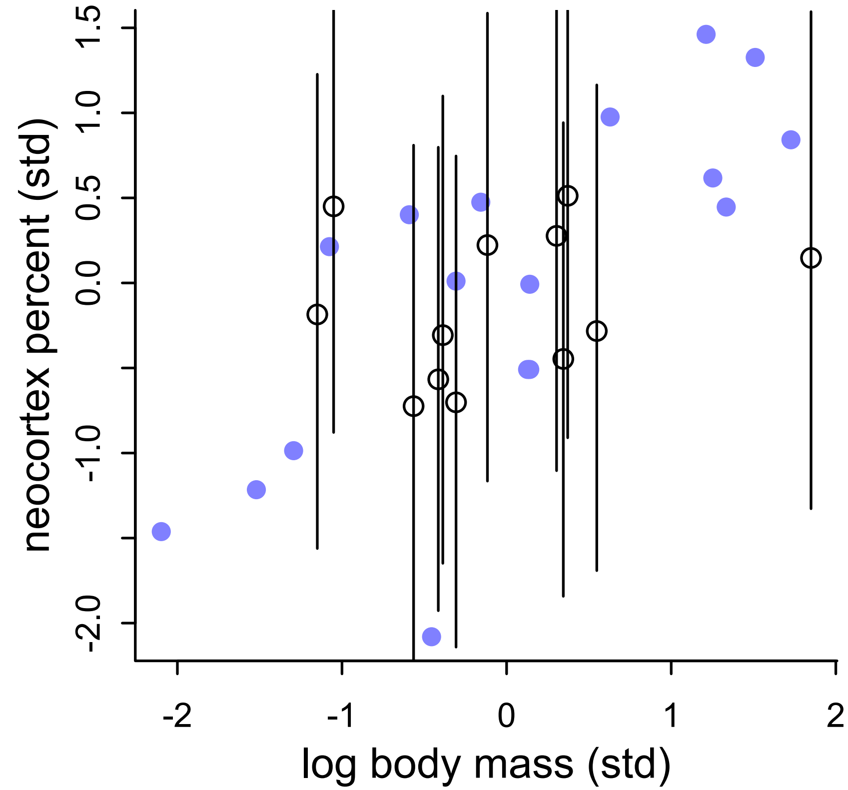
Imputed values track regression trend

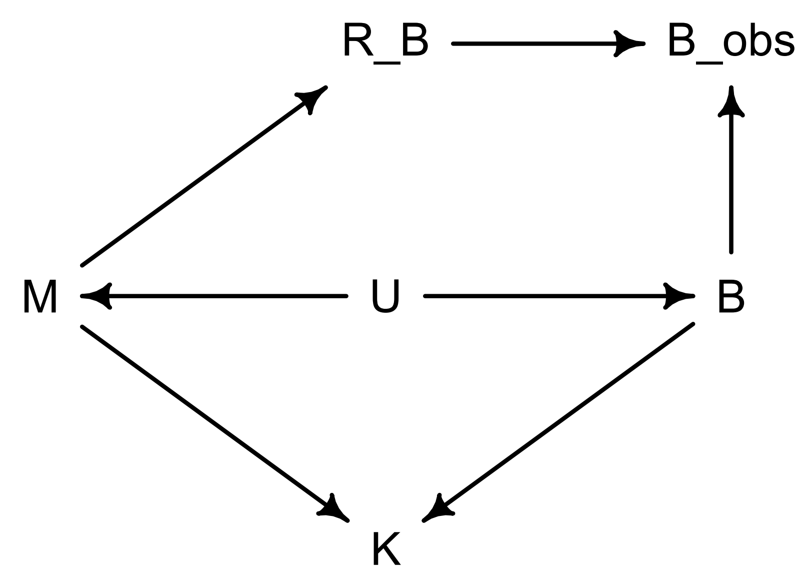


Imputed values do not track other predictor!

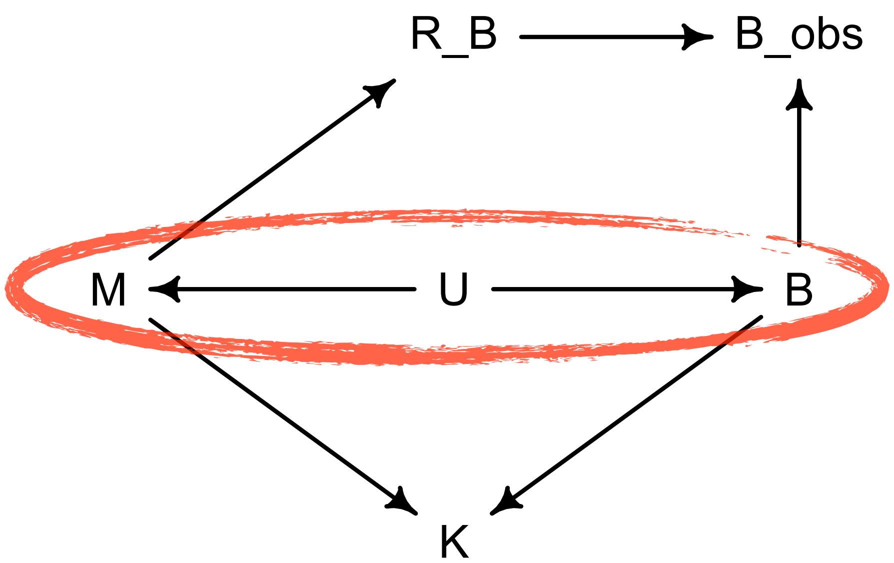
Results

- Observed neocortex positively associated with observed body mass
- Imputed neocortex NOT associated with observed body mass
- Can do better
    - Imputation model can use full causal model





Full Flavor Imputation

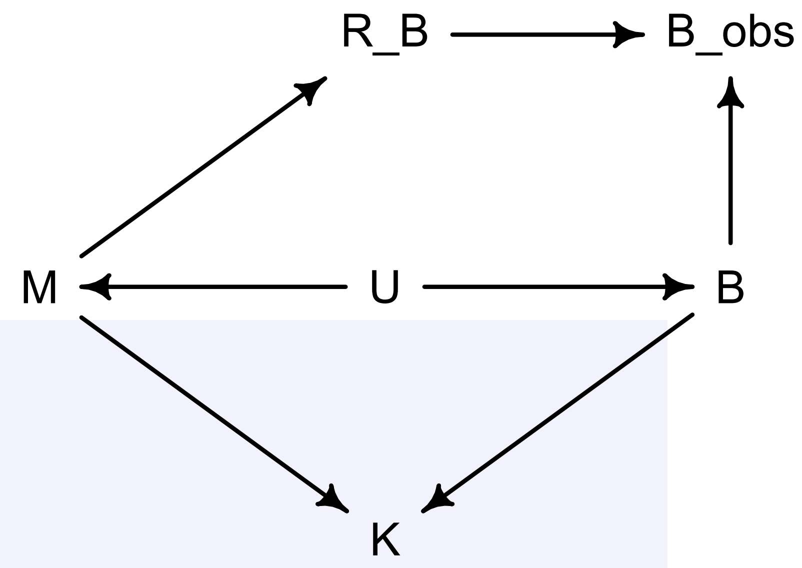


$$K_i \sim \text{Normal}(\mu_i, \sigma)$$

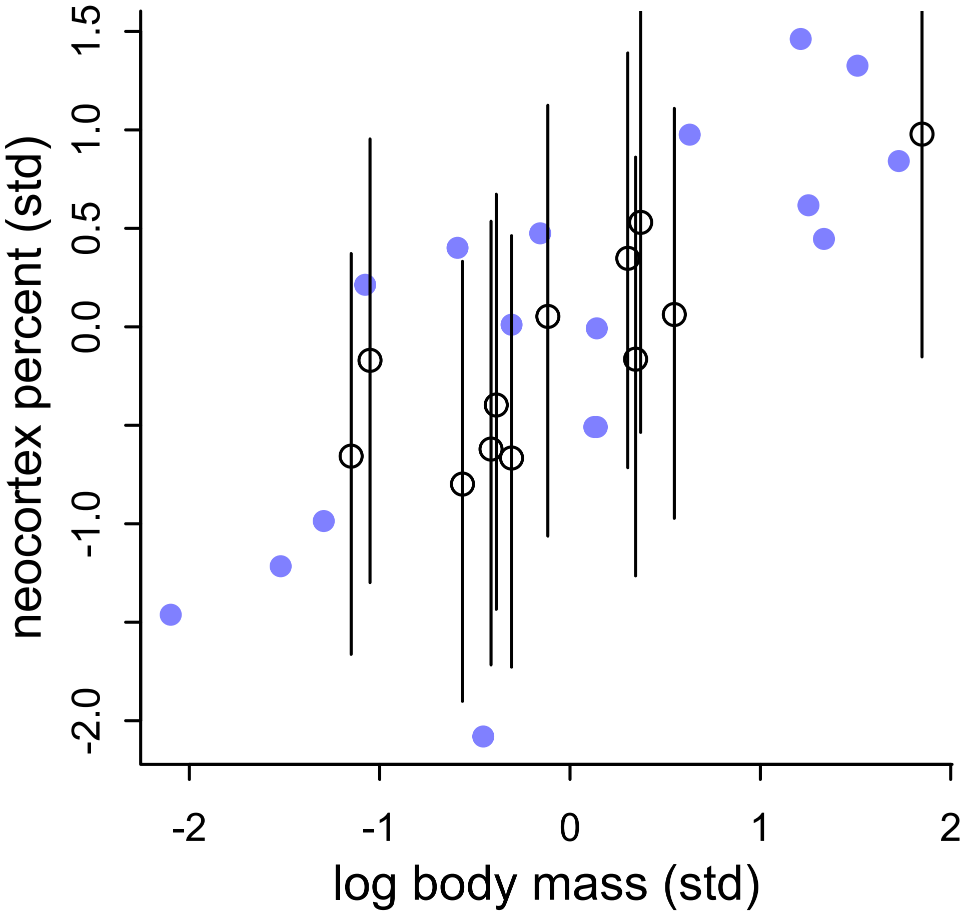
$$\mu_i = \alpha + \beta_B B_i + \beta_M \log M_i$$

$$(M_i, B_i) \sim \text{MVNormal}((\mu_M, \mu_B), \mathbf{S})$$

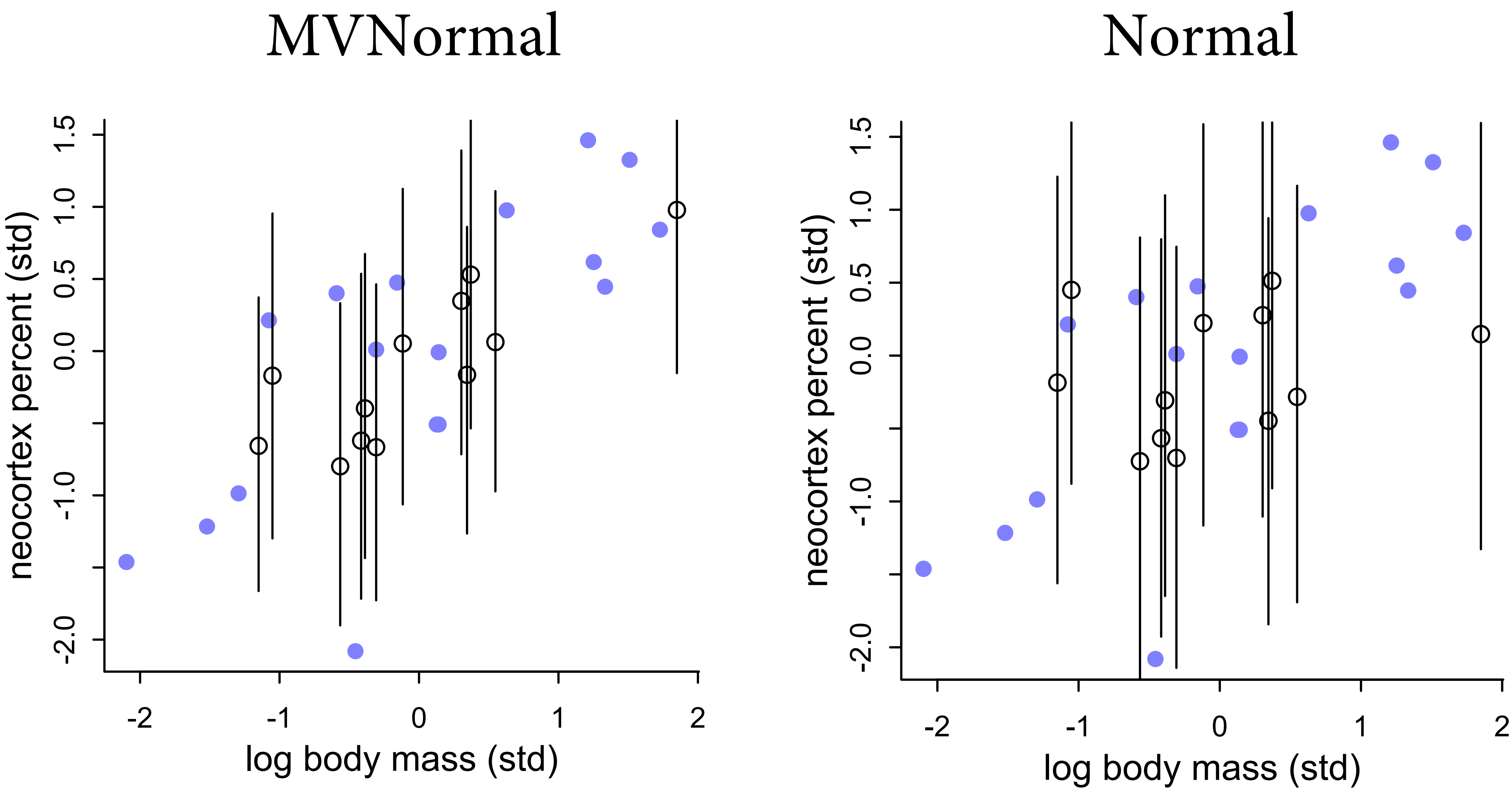
Full Flavor Imputation




```
m15.5 <- ulam(
    alist(
       # K as function of B and M
        K ~ dnorm( mu , sigma ),
        mu <- a + bB*B_merge + bM*M,

       # M and B correlation
        MB ~ multi_normal( c(muM,muB) , Rho_BM , Sigma_BM ),
        matrix[29,2]:MB <<- append_col( M , B_merge ),

       # define B_merge as mix of observed and imputed values
        vector[29]:B_merge <- merge_missing( B , B_impute ),

       # priors
        c(a,muB,muM) ~ dnorm( 0 , 0.5 ),
        c(bB,bM) ~ dnorm( 0, 0.5 ),
        sigma ~ dexp( 1 ),
        Rho_BM ~ lkj_corr(2),
        Sigma_BM ~ exponential(1)
    ) , data=dat_list , chains=4 , cores=4 )
```



	mean	sd	5.5%	94.5%
bM	-0.65	0.22	-1.00	-0.30
bB	0.58	0.26	0.16	0.99
Rho_BM[1,1]	1.00	0.00	1.00	1.00
Rho_BM[1,2]	0.60	0.13	0.37	0.78
Rho_BM[2,1]	0.60	0.13	0.37	0.78
Rho_BM[2,2]	1.00	0.00	1.00	1.00

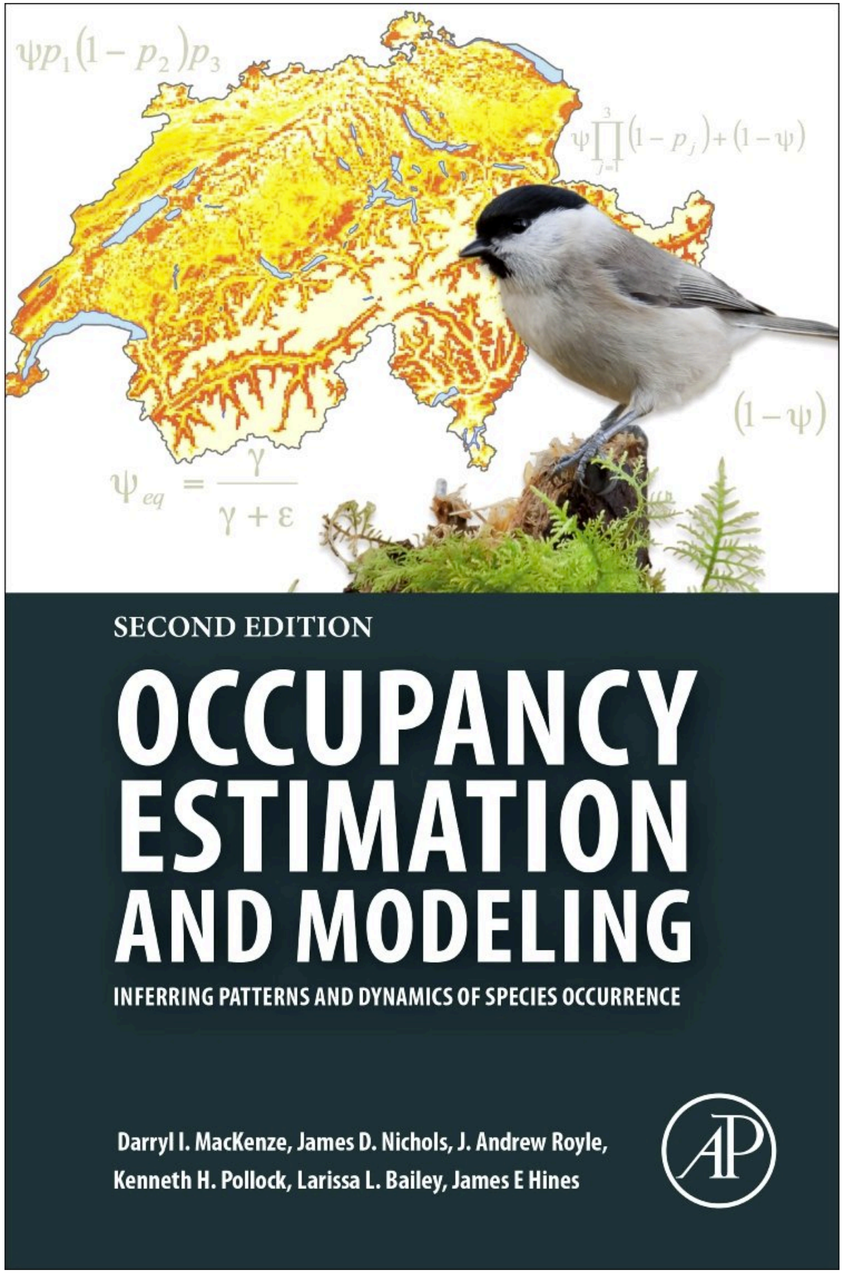


MVNormal
Normal
neocortex percent (std)
log body mass (std)
1.5
1.0
0.5
0.0
-1.0
-2.0
-2
-1
0
1
2

Missing data

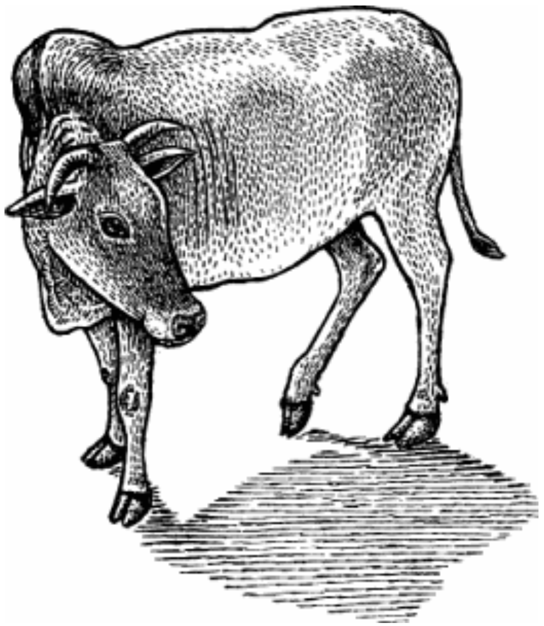
- Can also impute discrete values, but need another technique (see text)

- Extends to many model types:
    - Mark-recapture, occupancy (presence/absence)
    - Latent-state models (hidden Markov models)



Final Homework

- A little imputation practice
- Finish for complete sense of accomplishment



Bos primigenius

The Golem of Prague

"Even the most perfect of Golem, risen to life to protect us, can easily change into a destructive force. Therefore let us treat carefully that which is strong, just as we bow kindly and patiently to that which is weak."

Rabbi Judah Loew ben Bezalel (1512–1609)

From *Breath of Bones: A Tale of the Golem*