Statistical Rethinking
Winter 2019

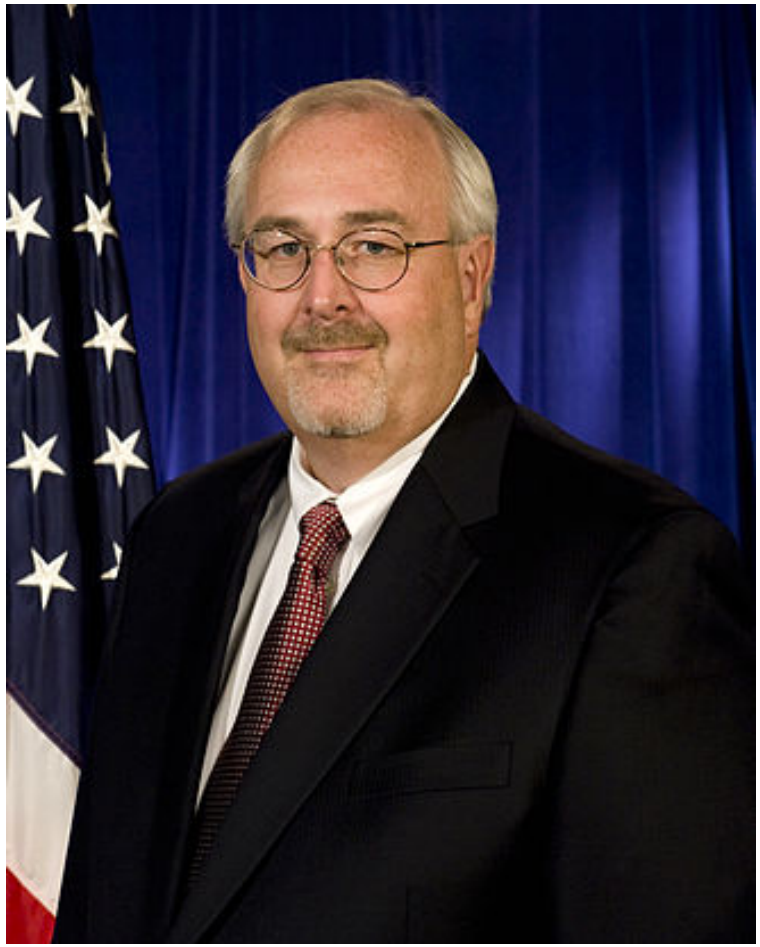
Lecture 04 / Week 3

The Many Variables & The Spurious Waffles


WAFFLE HOUSE
WAFFLE HOUSE
WAFFLE HOUSE

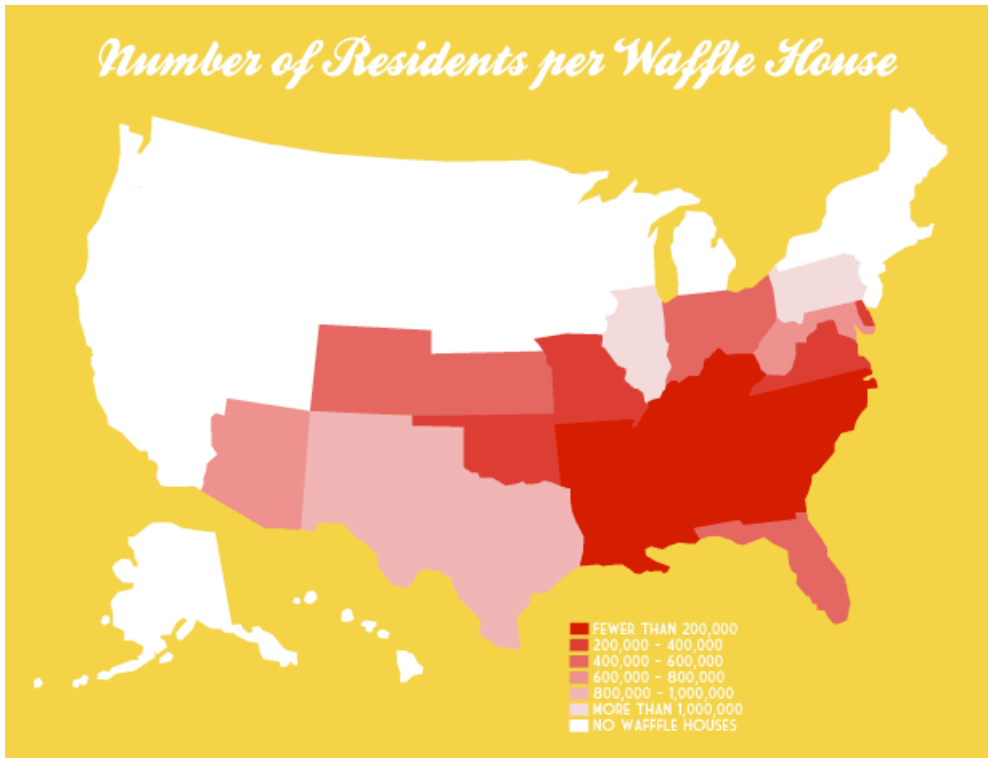

WAFFLE
HOUSE

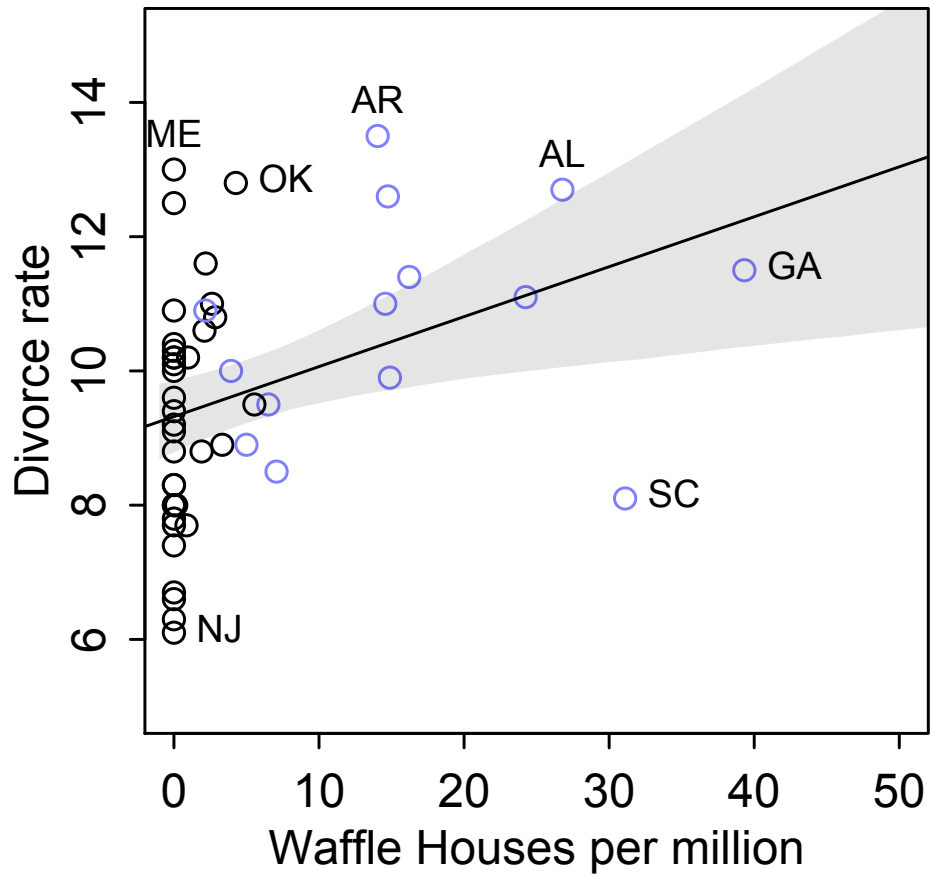
"If you get there and the Waffle House is closed? That's really bad. That's when you go to work."



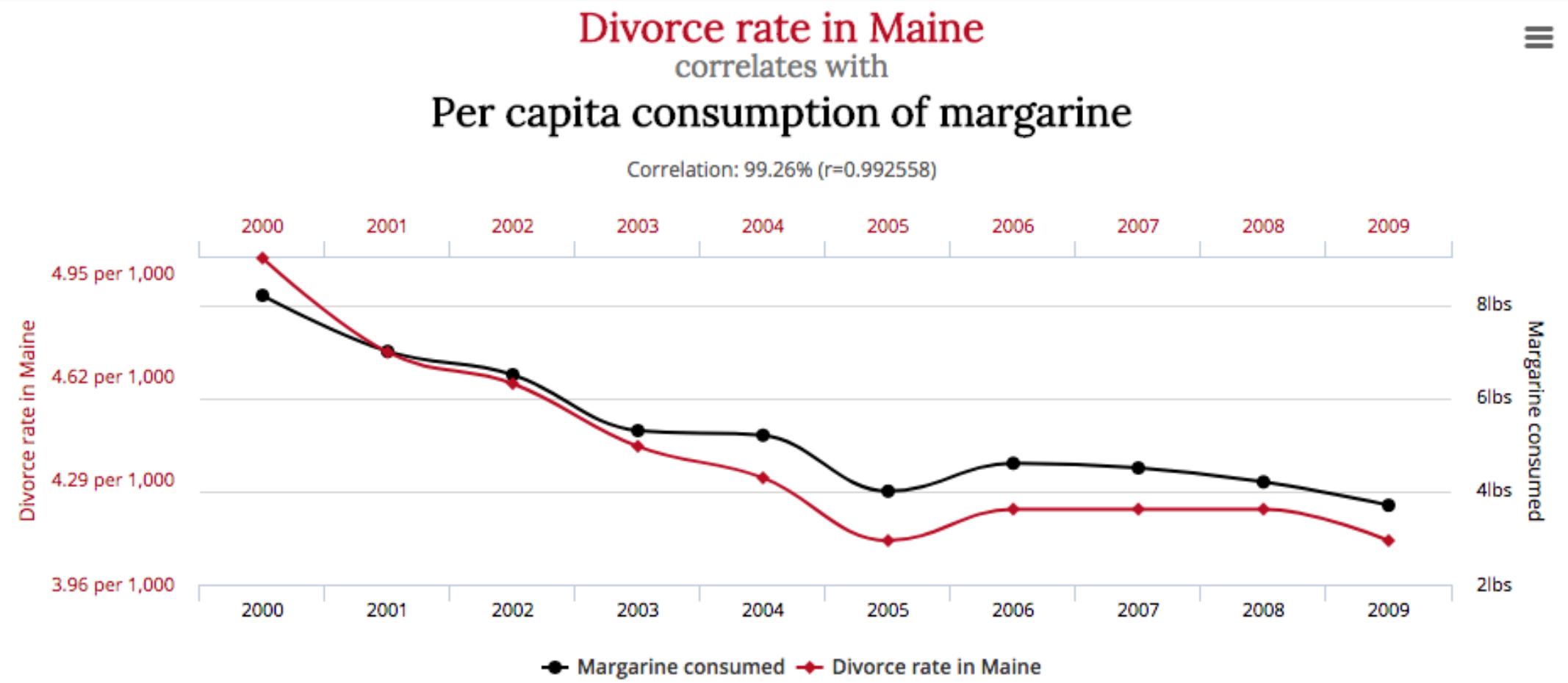
Craig Fugate, director (2009–2017)
USA Federal Emergency Management Agency (FEMA)

Does Waffle House cause divorce?





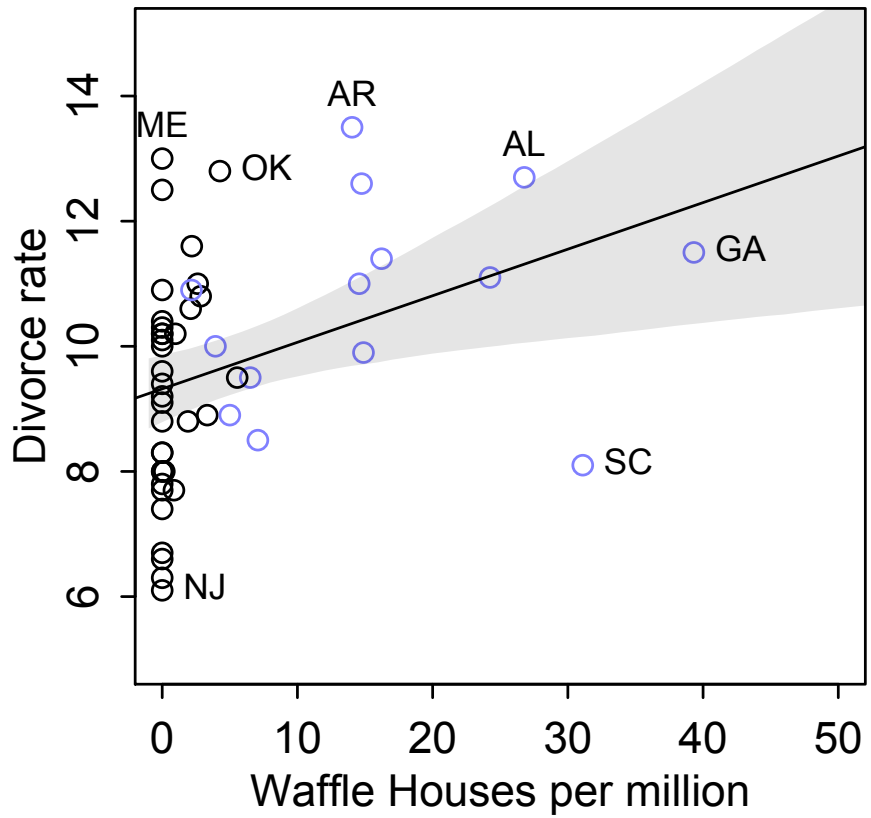
Correlation is commonplace



http://www.tylervigen.com/spurious-correlations

Goals this week

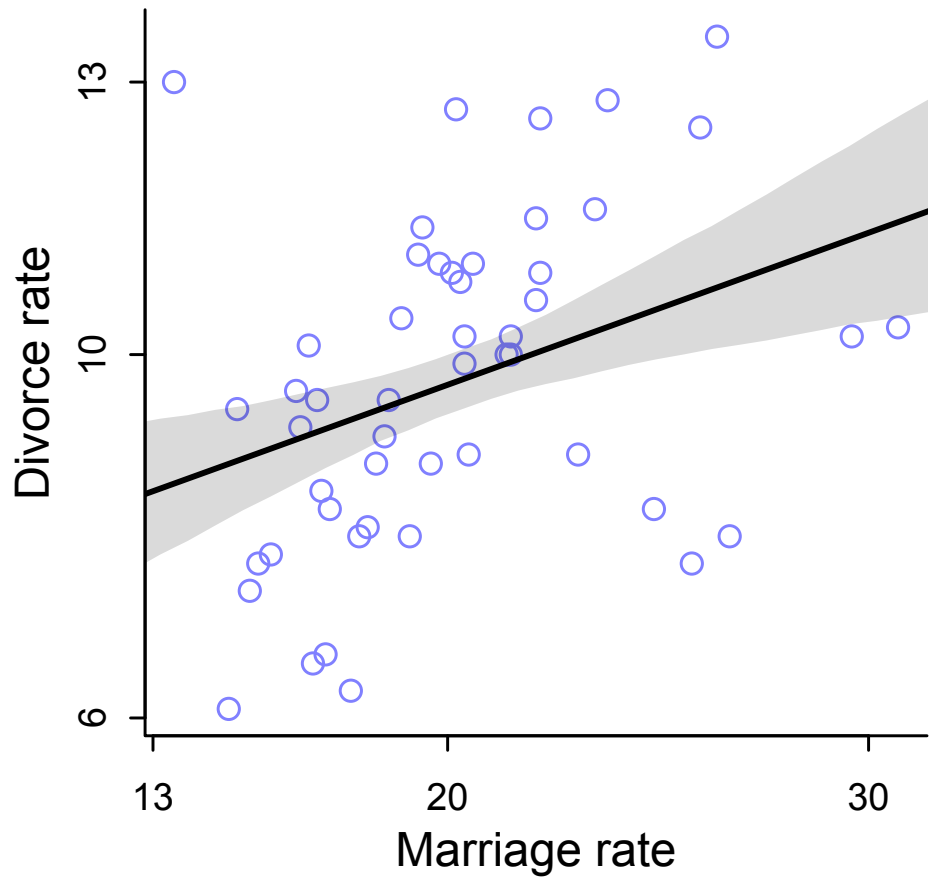
- Multiple regression models
- The good:
  - Reveal spurious correlation
  - Uncover masked association
- The bad:
  - *Cause* spurious correlation
  - Hide real associations
- Learn basics of causal inference
  - Directed acyclic graphs
  - forks, pipes, colliders, oh my!
  - Backdoor criterion



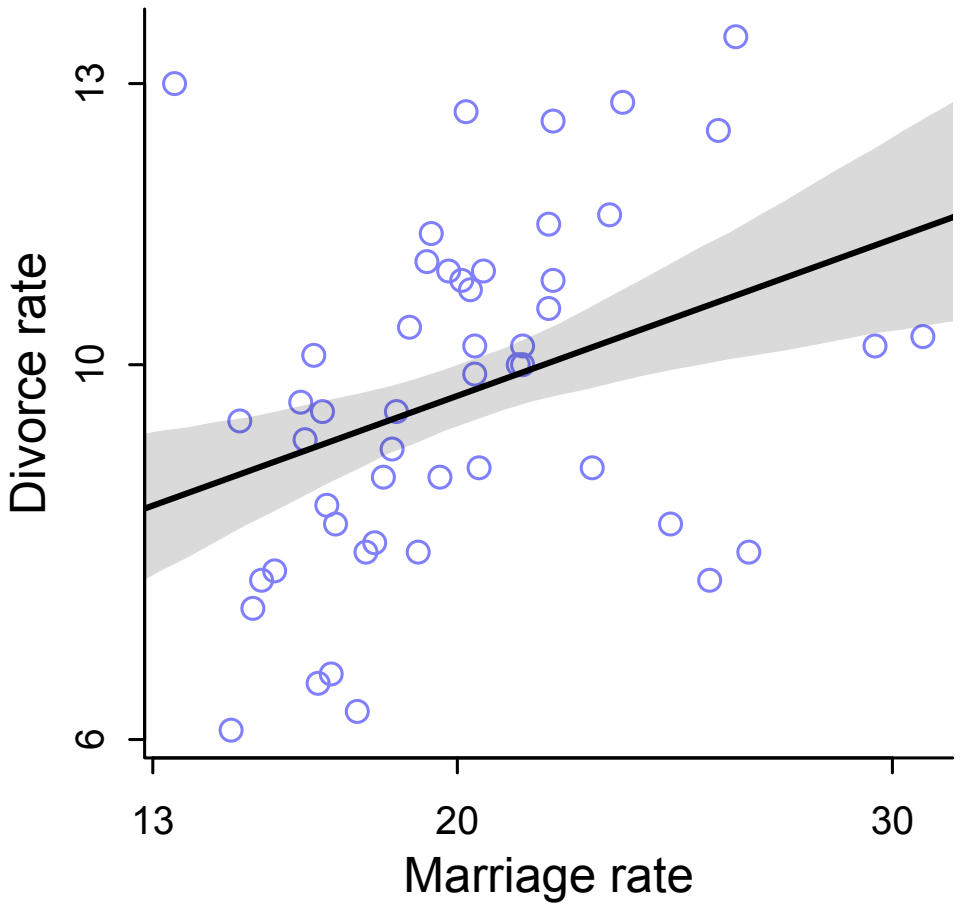
Spurious association

- Correlation does not imply causation
- Causation does not imply correlation
- Causation implies conditional correlation
- Need more than just models

- Q: Does marriage cause divorce?



Spurious association



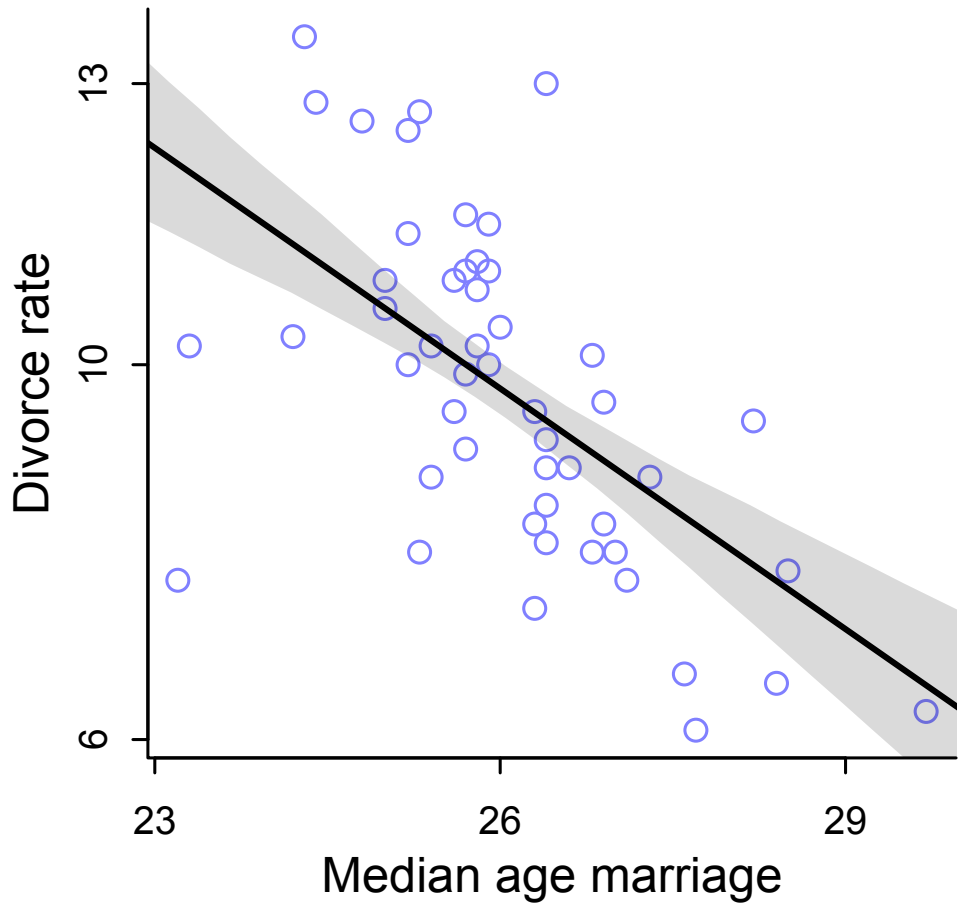


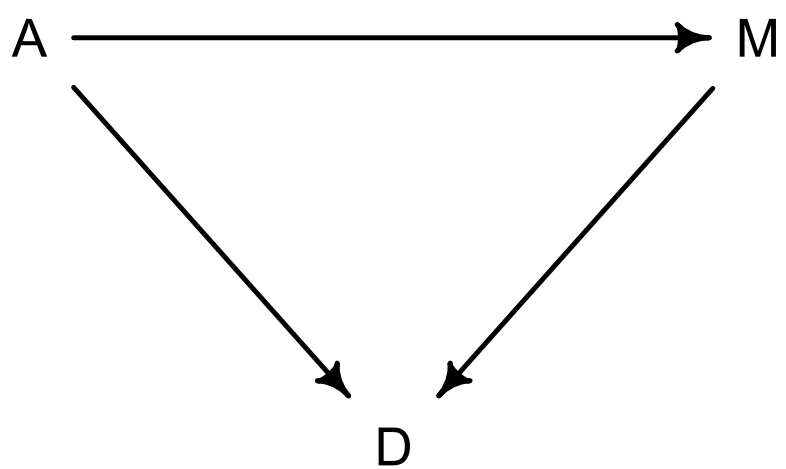
Figure 5.2

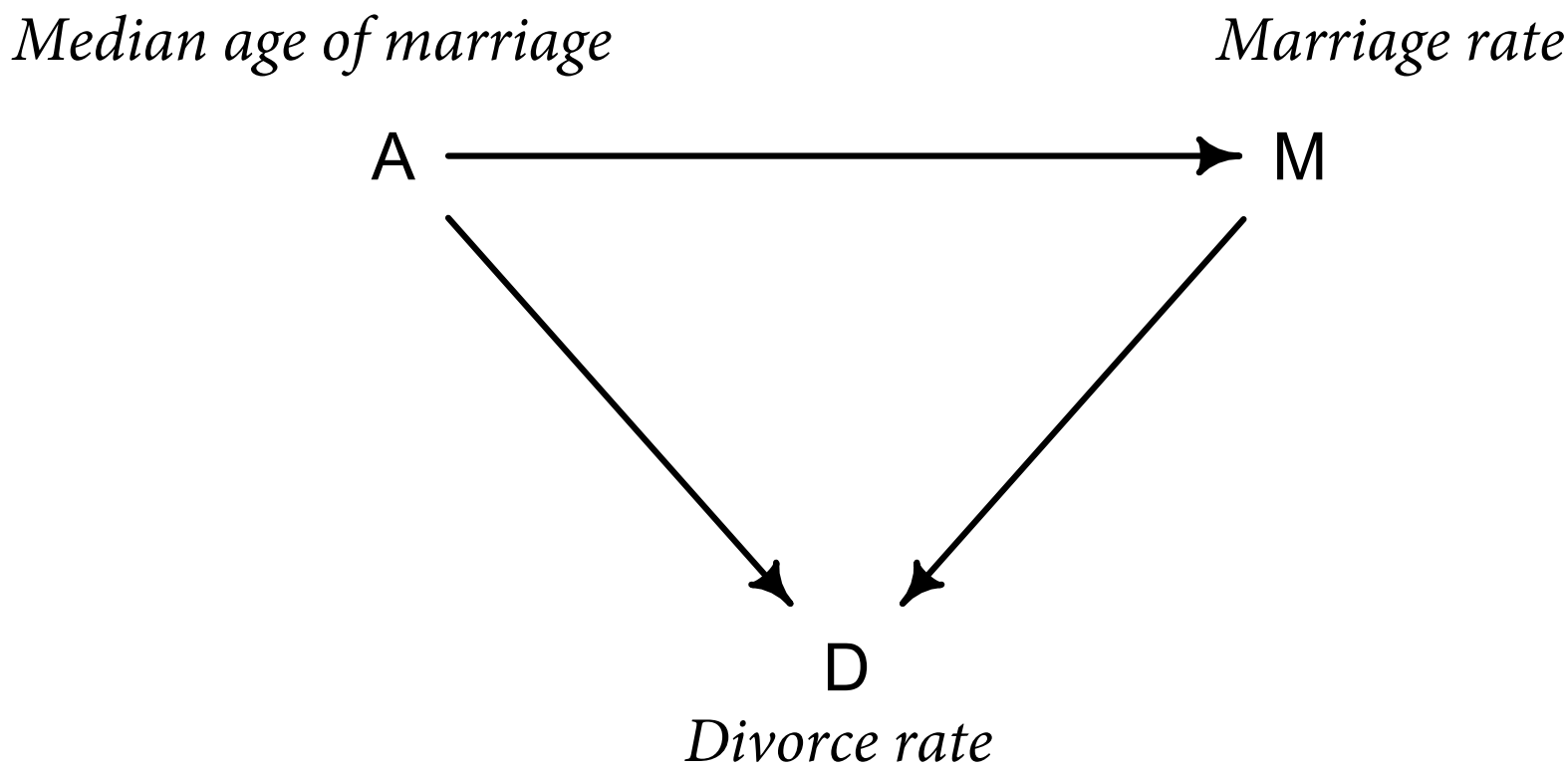
Multiple causes of divorce

- Want to know: *what is value of a predictor, once we know the other predictors?*
  - What is value of knowing marriage rate, once we already know median age at marriage?
  - What is value of knowing median age marriage, once we know marriage rate?

They're good DAGs, Brent

- Directed Acyclic Graphs — tools for causal models
  - Directed: Arrows
  - Acyclic: Arrows don't make loops
  - Graphs: Nodes and edges
- Unlike statistical model, has causal implications





Implications:

(1) M is a function of A

(2) D is a function of A and M

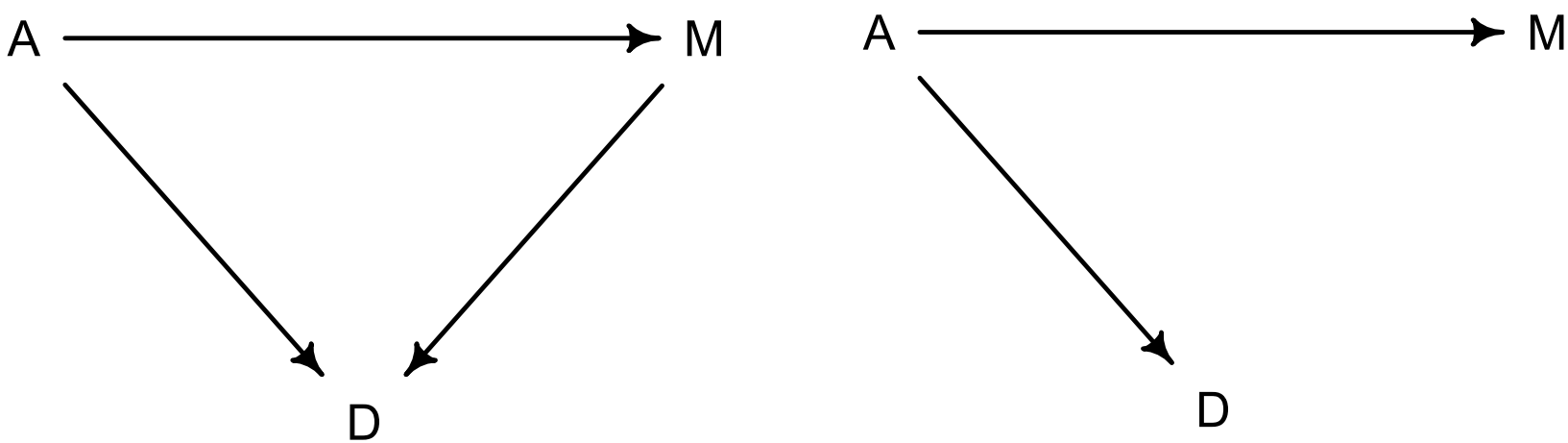
(3) The total causal effect of A has two *paths*:

(a) A –> M –> D

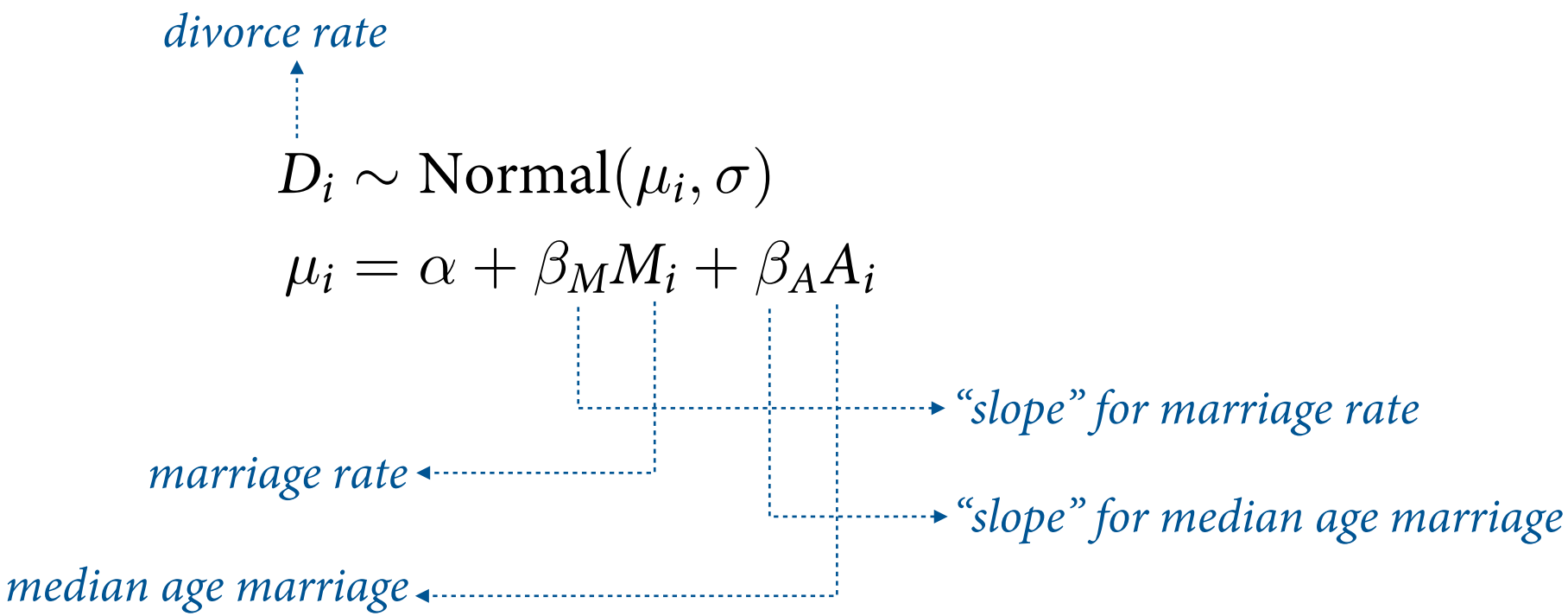
(b) A –> D

Good DAGs

- Given association M <–> D, cannot tell difference between:



- Need conditional association: M <–> D | A


divorce rate
$D_i \sim \text{Normal}(\mu_i, \sigma)$
$\mu_i = \alpha + \beta_M M_i + \beta_A A_i$
"slope" for marriage rate
marriage rate
"slope" for median age marriage
median age marriage

Priors

$$D_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_M M_i + \beta_A A_i$$

- Standardize divorce rate D, marriage rate M, median age at marriage A
- We expect alpha to be near zero

$$\alpha \sim \text{Normal}(0, 0.2)$$

- Slopes should not produce impossibly strong relationships

$$\beta_M \sim \text{Normal}(0, 0.5)$$
$$\beta_A \sim \text{Normal}(0, 0.5)$$

Prior predictive simulation

R code 5.3

```
m5.1 <- quap(
    alist(
        D ~ dnorm( mu , sigma ) ,
        mu <- a + bA * A ,
        a ~ dnorm( 0 , 0.2 ) ,
        bA ~ dnorm( 0 , 0.5 ) ,
        sigma ~ dexp( 1 )
    ) , data = d )
```

R code 5.4

```
set.seed(10)
prior <- extract.prior( m5.1 )
mu <- link( m5.1 , post=prior , data=list( A=c(-2,2) ) )
plot( NULL , xlim=c(-2,2) , ylim=c(-2,2) )
for ( i in 1:50 ) lines( c(-2,2) , mu[i,] , col=col.alpha("black",0.4) )
```

Prior predictive simulation

R code 5.4

```
set.seed(10)
prior <- extract.prior( m5.1 )
mu <- link( m5.1 , post=prior , data=list( A=c(-2,2) ) )
plot( NULL , xlim=c(-2,2) , ylim=c(-2,2) )
for ( i in 1:50 ) lines( c(-2,2) , mu[i,] , col=col.alpha("black",0.4) )
```

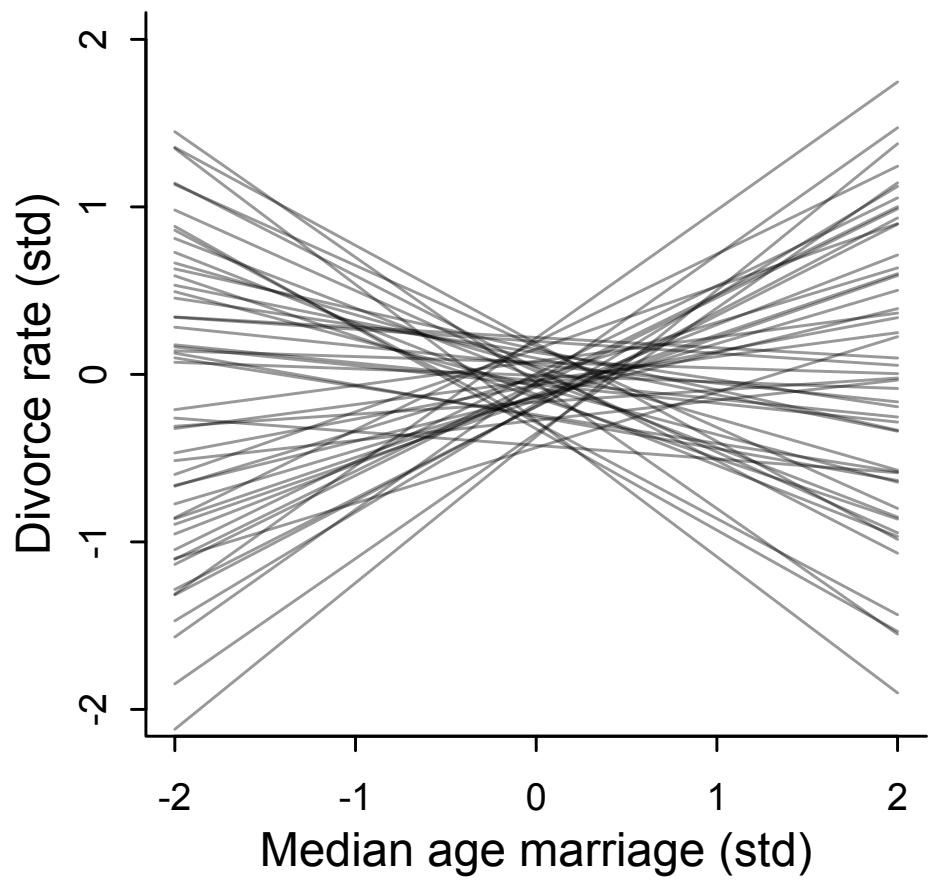


Figure 5.3

$$D_i \sim \text{Normal}(\mu_i, \sigma) \quad \text{[probability of data]}$$

$$\mu_i = \alpha + \beta_M M_i + \beta_A A_i \quad \text{[linear model]}$$

$$\alpha \sim \text{Normal}(0, 0.2) \quad \text{[prior for } \alpha\text{]}$$

$$\beta_M \sim \text{Normal}(0, 0.5) \quad \text{[prior for } \beta_M\text{]}$$

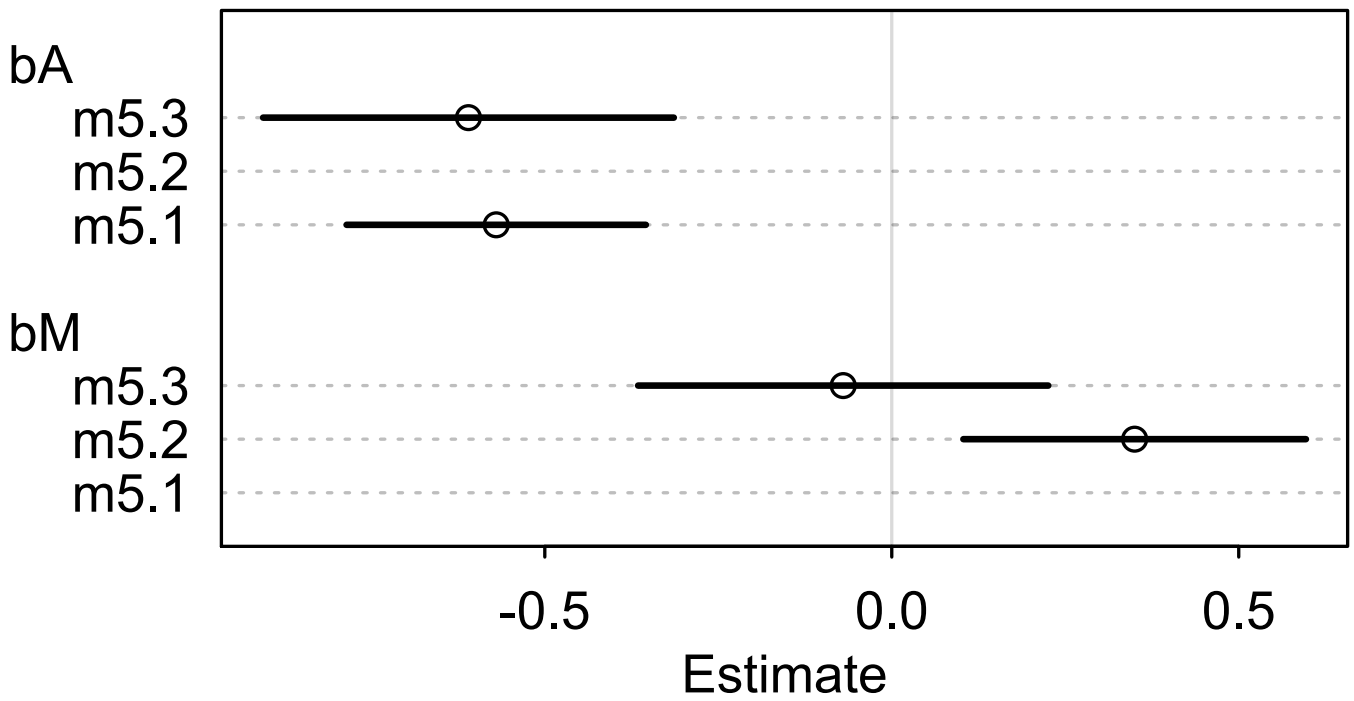
$$\beta_A \sim \text{Normal}(0, 0.5) \quad \text{[prior for } \beta_A\text{]}$$

$$\sigma \sim \text{Exponential}(1) \quad \text{[prior for } \sigma\text{]}$$

R code 5.8

```
m5.3 <- quap(
    alist(
        D ~ dnorm( mu , sigma ) ,
        mu <- a + bM*M + bA*A ,
        a ~ dnorm( 0 , 0.2 ) ,
        bM ~ dnorm( 0 , 0.5 ) ,
        bA ~ dnorm( 0 , 0.5 ) ,
        sigma ~ dexp( 1 )
    ) , data = d )
precis( m5.3 )
```

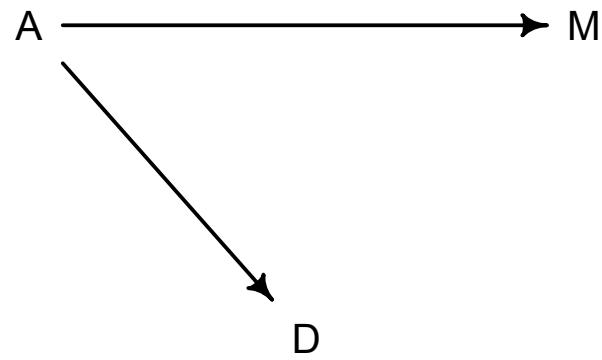
```
       mean   sd  5.5% 94.5%
a      0.00 0.10 -0.16  0.16
bM    -0.07 0.15 -0.31  0.18
bA    -0.61 0.15 -0.85 -0.37
sigma  0.79 0.08  0.66  0.91
```



m5.1: age of marriage only D ~ A
m5.2: marriage rate only D ~ M
m5.3: multiple regression D ~ A + M

Multiple regression

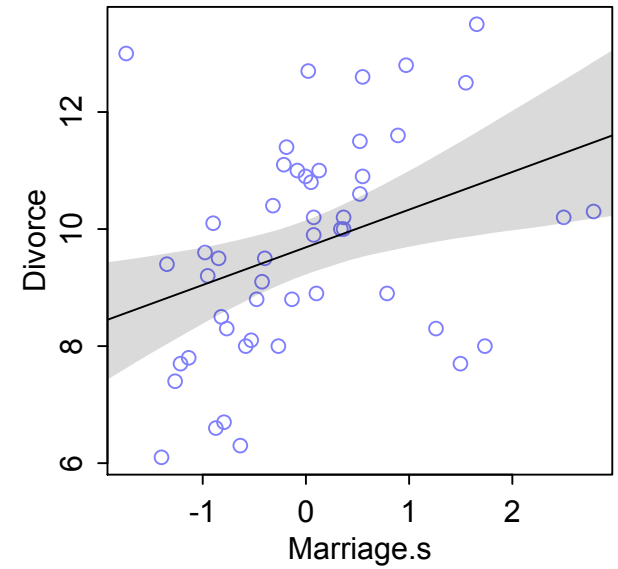
	mean	sd	5.5%	94.5%
a	0.00	0.10	-0.16	0.16
bM	-0.07	0.15	-0.31	0.18
bA	-0.61	0.15	-0.85	-0.37
sigma	0.79	0.08	0.66	0.91

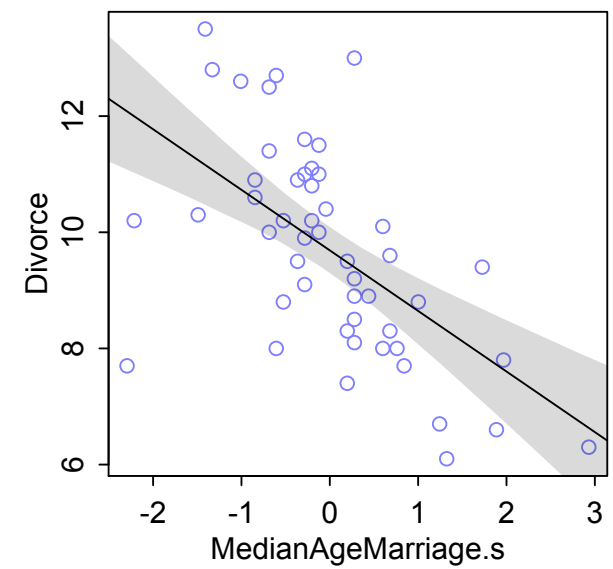


- Once we know median age marriage, little additional value in knowing marriage rate.
- Once we know marriage rate, still value in knowing median age marriage.
- If we *don't know* median age marriage, still useful to know marriage rate.

Posterior predictions

- Lots of plotting options now
  1. Predictor residual plots
  2. Counterfactual plots
  3. Posterior prediction plots





Predictor residual plots

- Goal: Show association of each predictor with outcome, “controlling” for other predictors
- Useful intuition
- Never analyze residuals!
- Recipe:
  1. Regress predictor on other predictors
  2. Compute predictor *residuals*
  3. Regress outcome on residuals

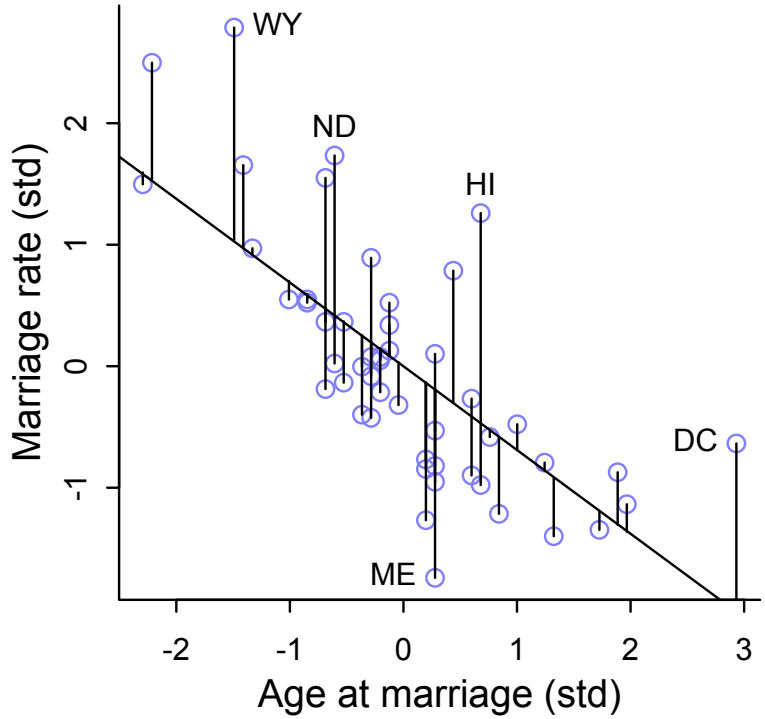
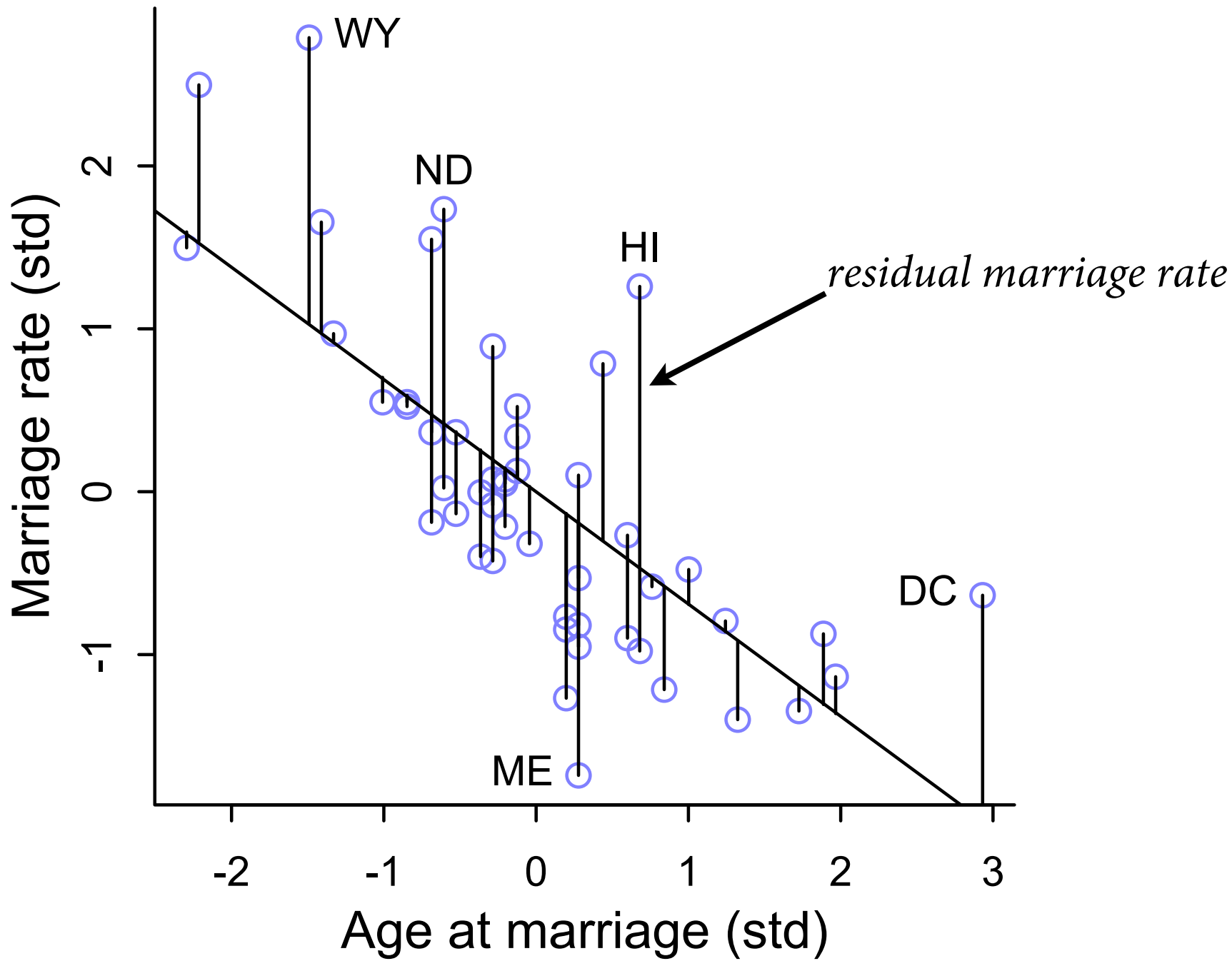
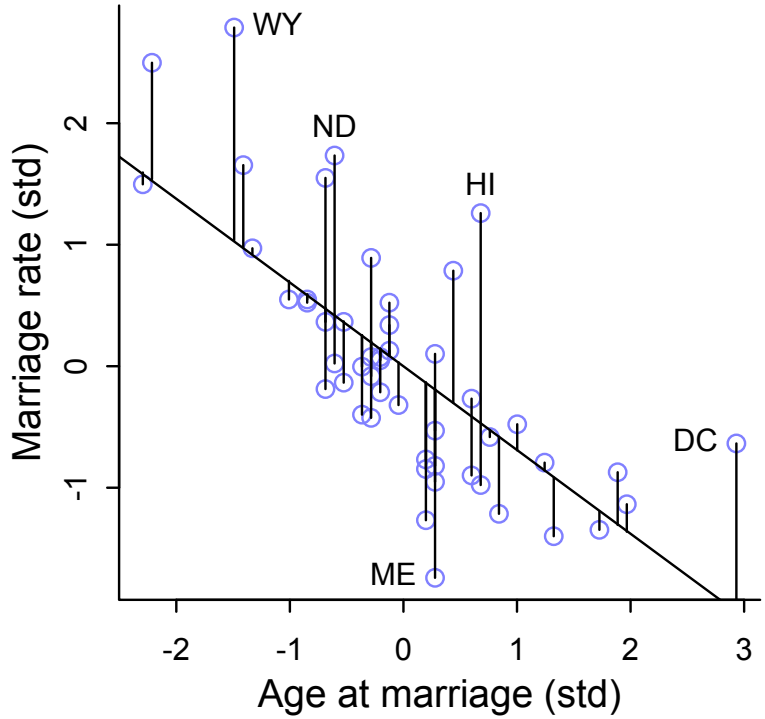

WY
ND
HI
DC
ME
Marriage rate (std)
Age at marriage (std)
-2
-1
0
1
2
3
-1
0
1
2

Figure 5.4


WY
ND
HI
DC
ME
residual marriage rate
Marriage rate (std)
Age at marriage (std)
-2
-1
0
1
2
3
-1
0
1
2


WY
ND
HI
DC
ME
Marriage rate (std)
Age at marriage (std)
-2
-1
0
1
2
3

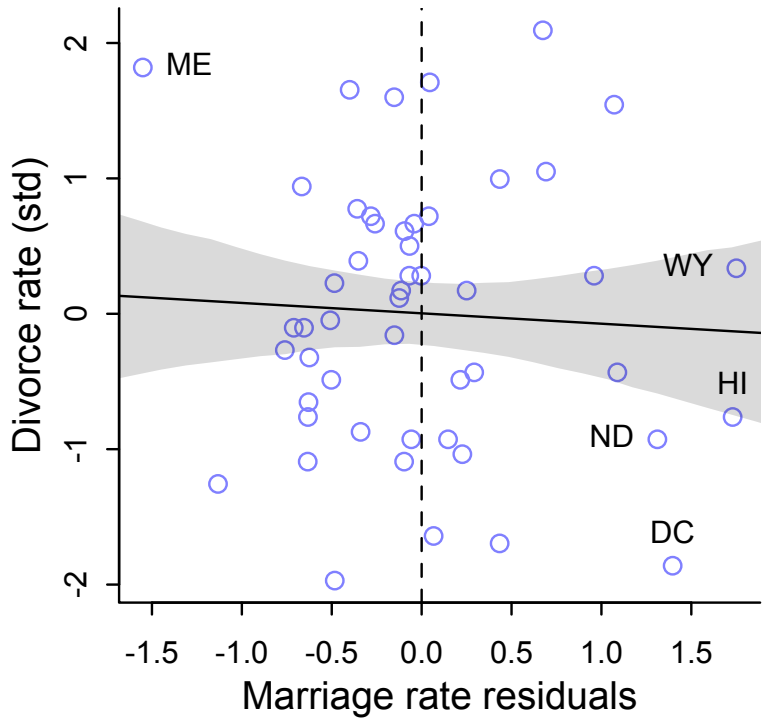
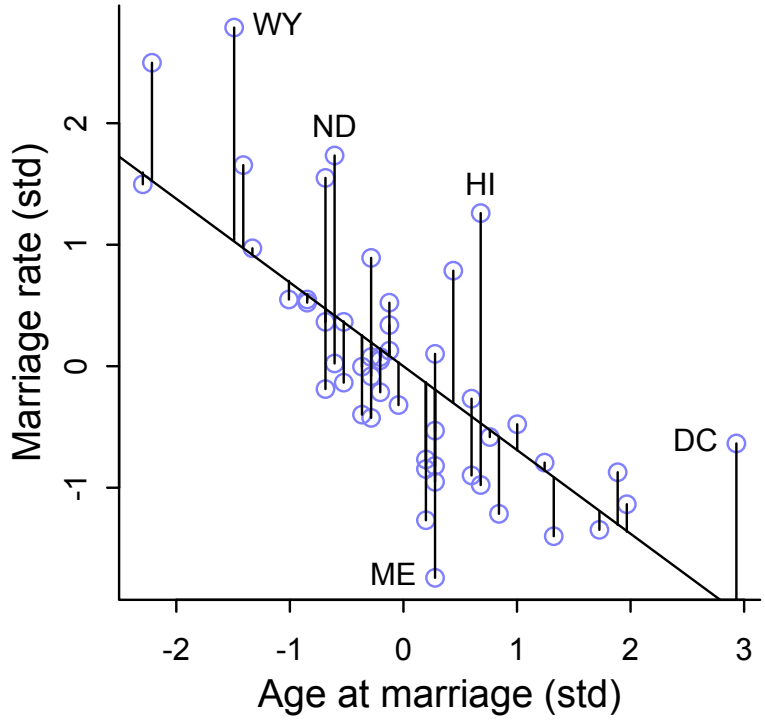
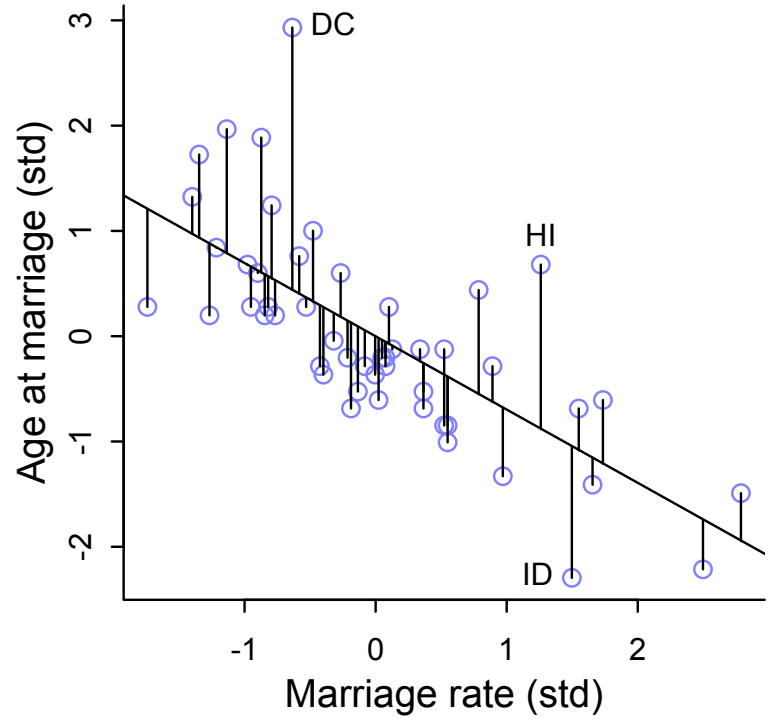

ME
WY
HI
ND
DC
Divorce rate (std)
Marriage rate residuals
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Figure 5.4


WY
ND
HI
DC
ME
Marriage rate (std)
Age at marriage (std)
-2
-1
0
1
2
3


DC
HI
ID
Age at marriage (std)
Marriage rate (std)
-2
-1
0
1
2
3

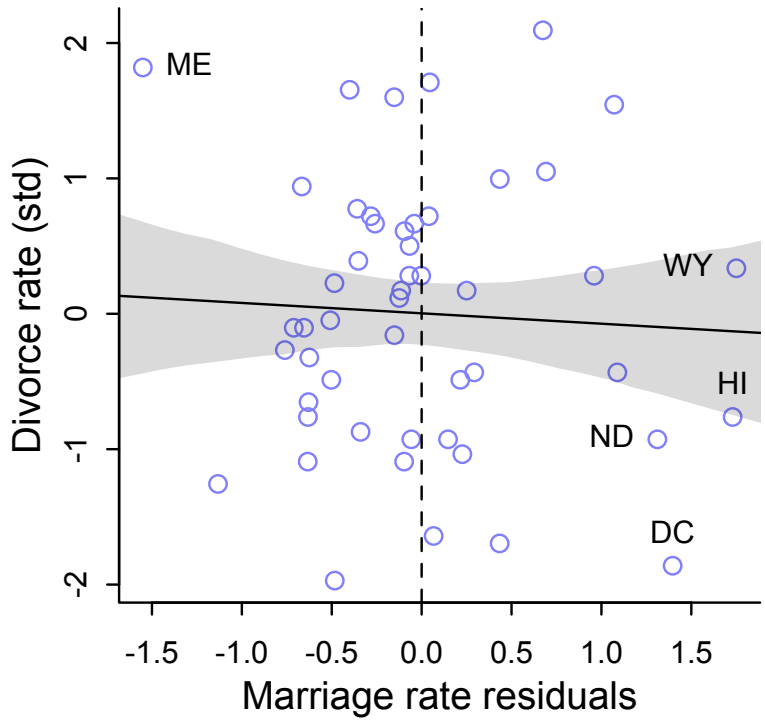

ME
WY
HI
ND
DC
Divorce rate (std)
Marriage rate residuals
-2
-1
0
1
2
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Figure 5.4

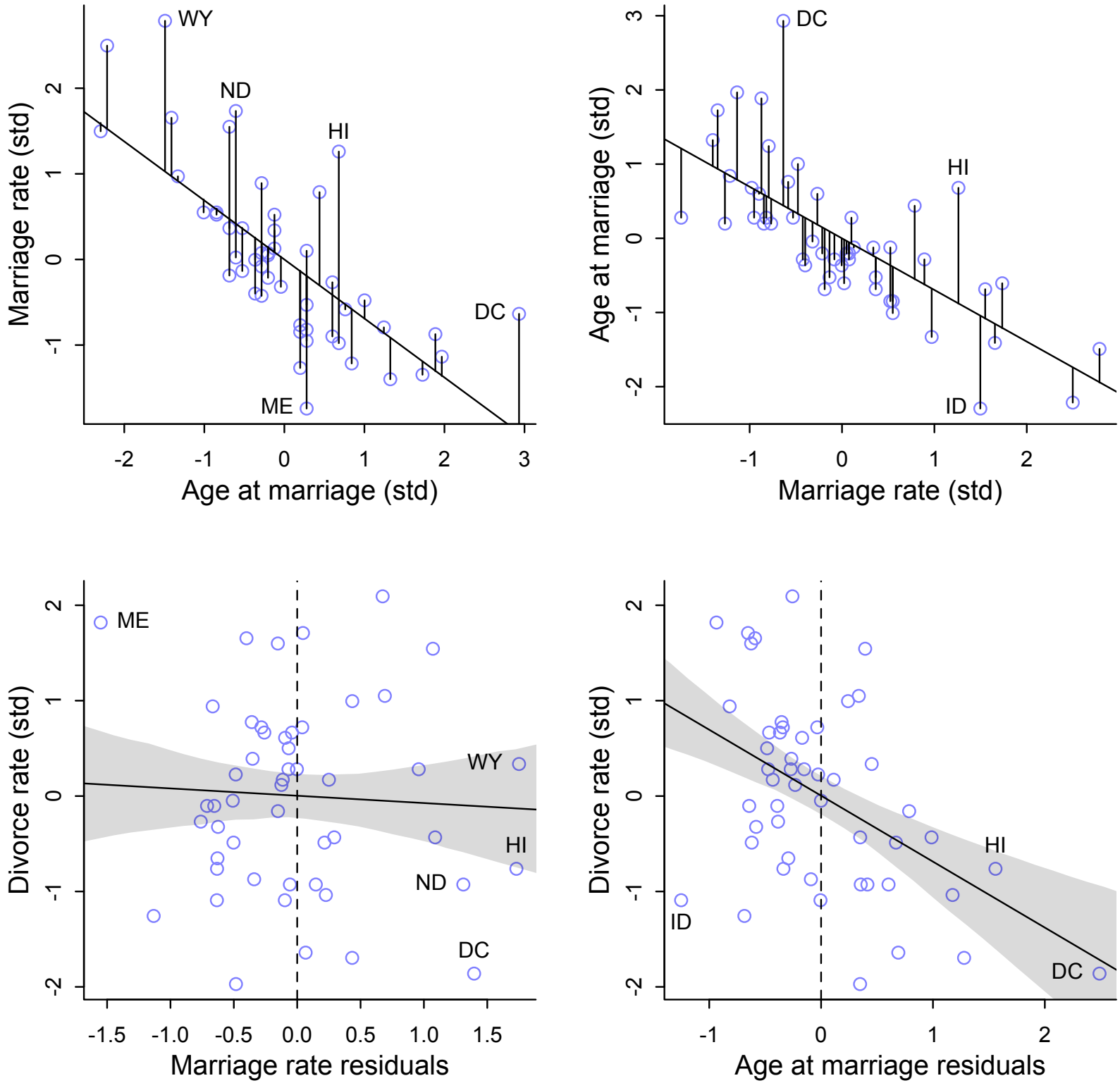
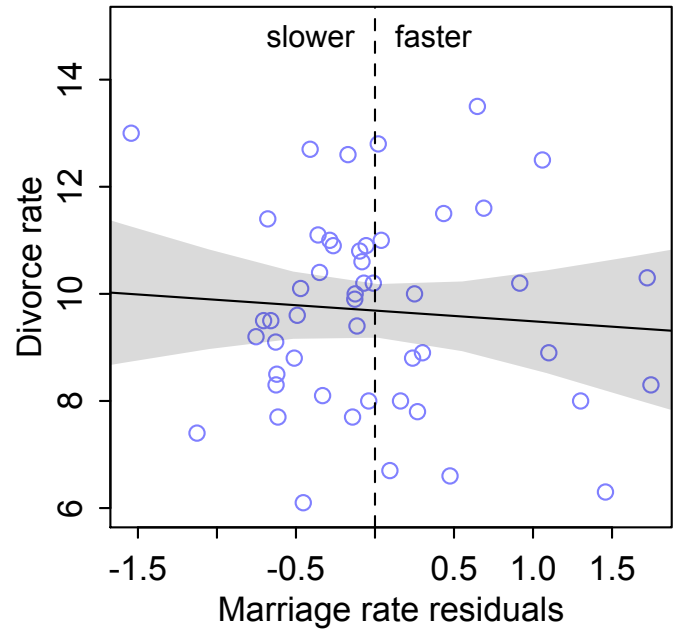

WY
ND
HI
DC
ME
Marriage rate (std)
Age at marriage (std)
DC
HI
ID
Age at marriage (std)
Marriage rate (std)
ME
WY
HI
ND
DC
Divorce rate (std)
Marriage rate residuals
ID
HI
DC
Divorce rate (std)
Age at marriage residuals

Figure 5.4

Statistical "control"

- Multiple linear regression answers question: *How is each predictor associated with outcome, once we know all the other predictors?*
  - Uses model to build expected outcomes — not magic!
  - Don't get cocky: Marriage rate may still be associated with divorce, for some *subset* of States
  - Can't make strong causal inferences from averages; need data on individuals



Counterfactual plots

- Goal: Explore model implications for outcomes
  - Fix other predictor(s)
  - Compute predictions across values of predictor
- Compute for unobserved (impossible?) cases, hence "counterfactual"

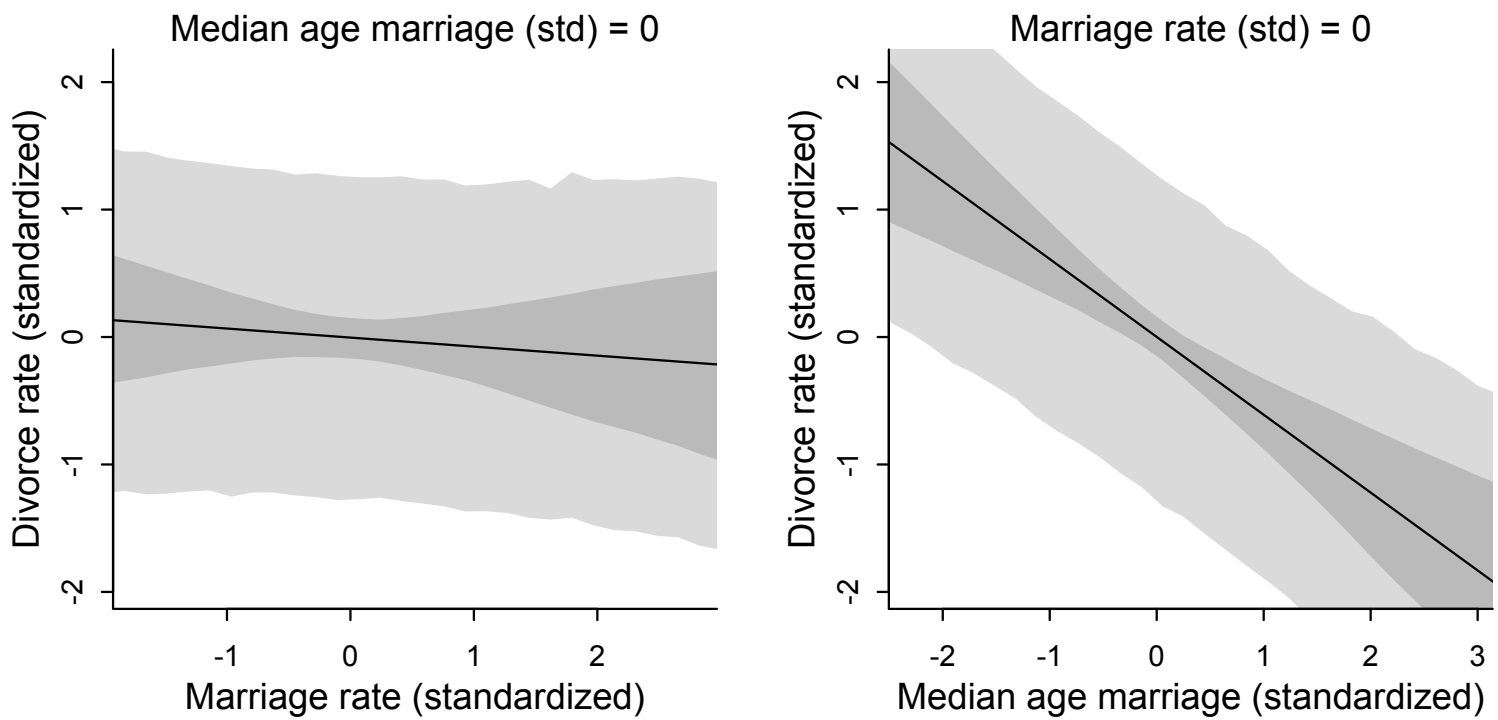


Figure 5.5

Posterior prediction checks

- Goal: Compute implied predictions for *observed* cases
  - Check model fit — golems do make mistakes
  - Find model failures, stimulate new ideas
- Always average over the posterior distribution
  - Using only posterior mean leads to overconfidence
  - Embrace the uncertainty

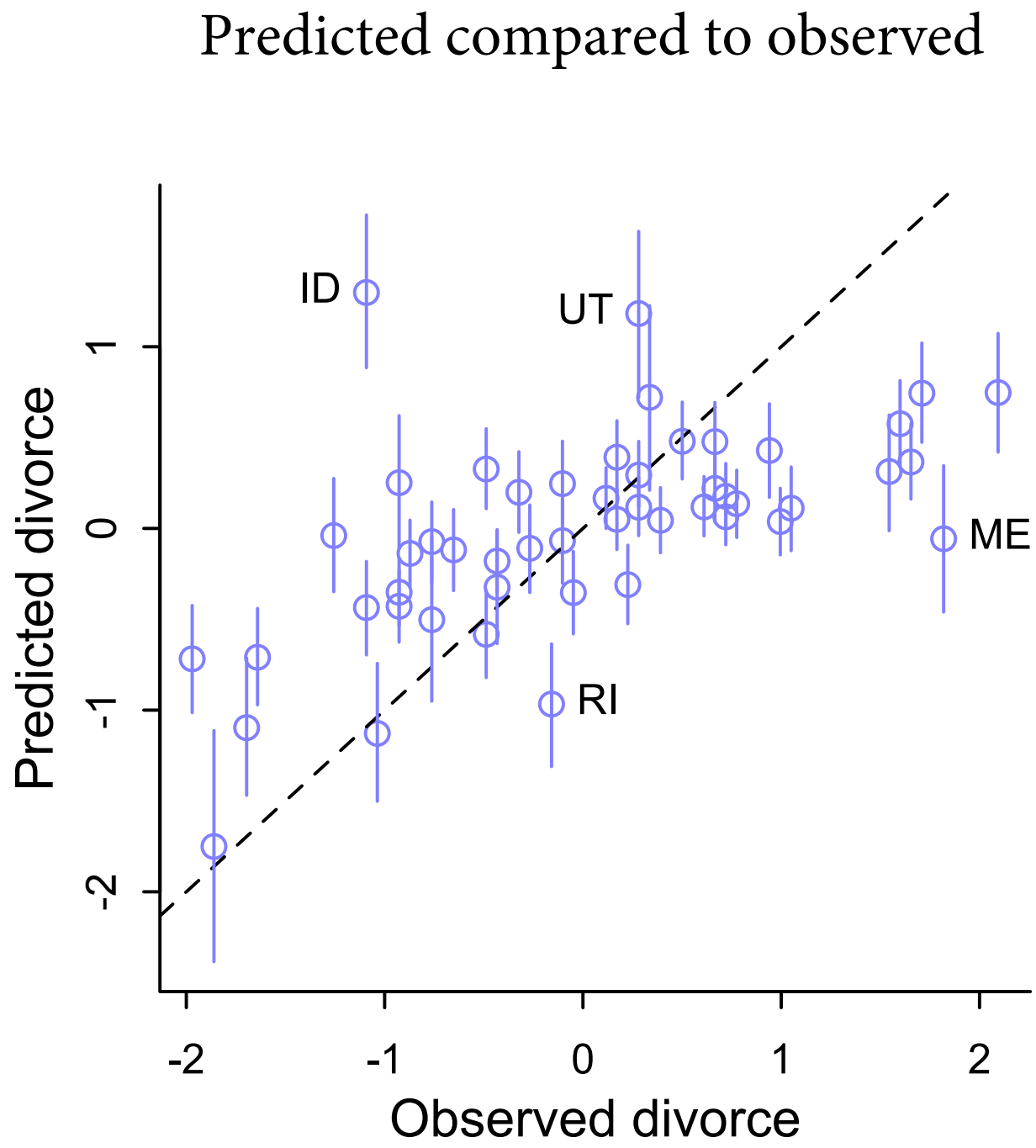

Predicted compared to observed
ID
UT
RI
ME
Predicted divorce
Observed divorce
-2
-1
0
1
2

Figure 5.6

Masked association

- Sometimes association between outcome and predictor masked by another variable
- Need both variables to see influence of either
- Tends to arise when
  - Another predictor associated with outcome *in opposite direction*
  - Both predictors associated with one another
- Noise in predictors can also mask association (*residual confounding*)

Milk and Brain

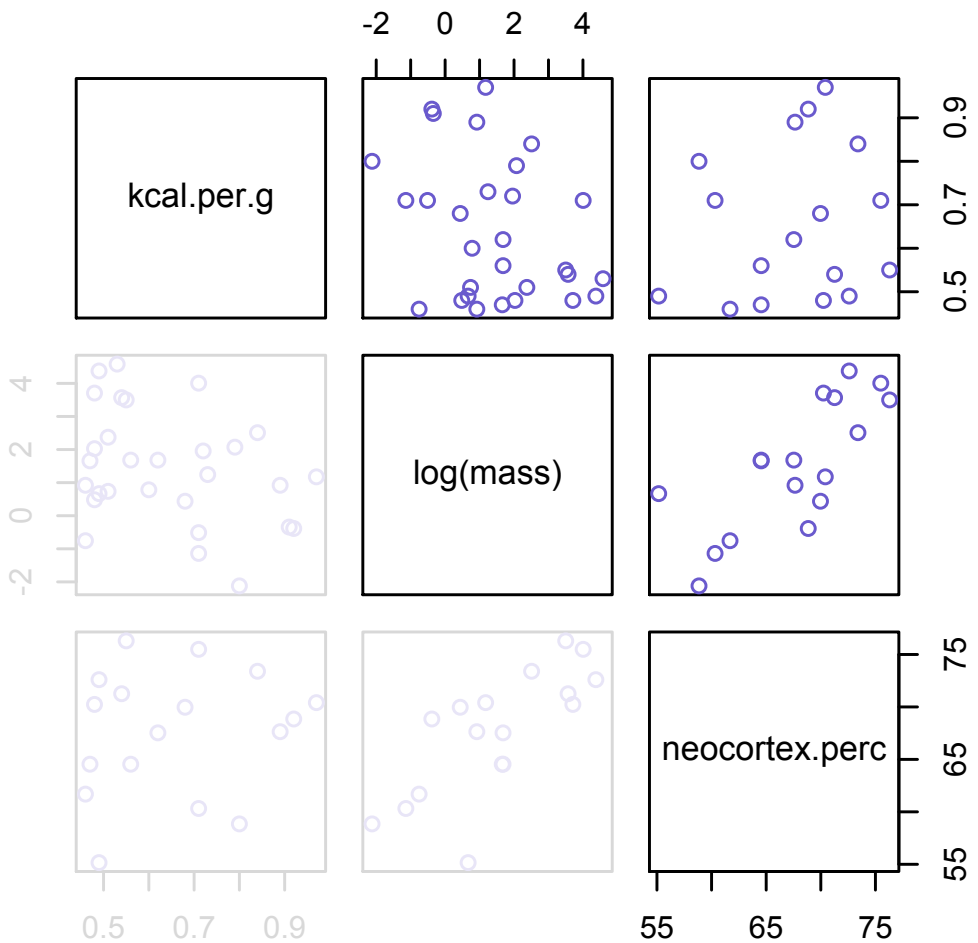
Eulemur fulvus
0.49 *kcal/g*
55% *neocortex*

Homo sapiens
0.71 *kcal/g*
75% *neocortex*

Cebus apella
0.89 *kcal/g*
68% *neocortex*

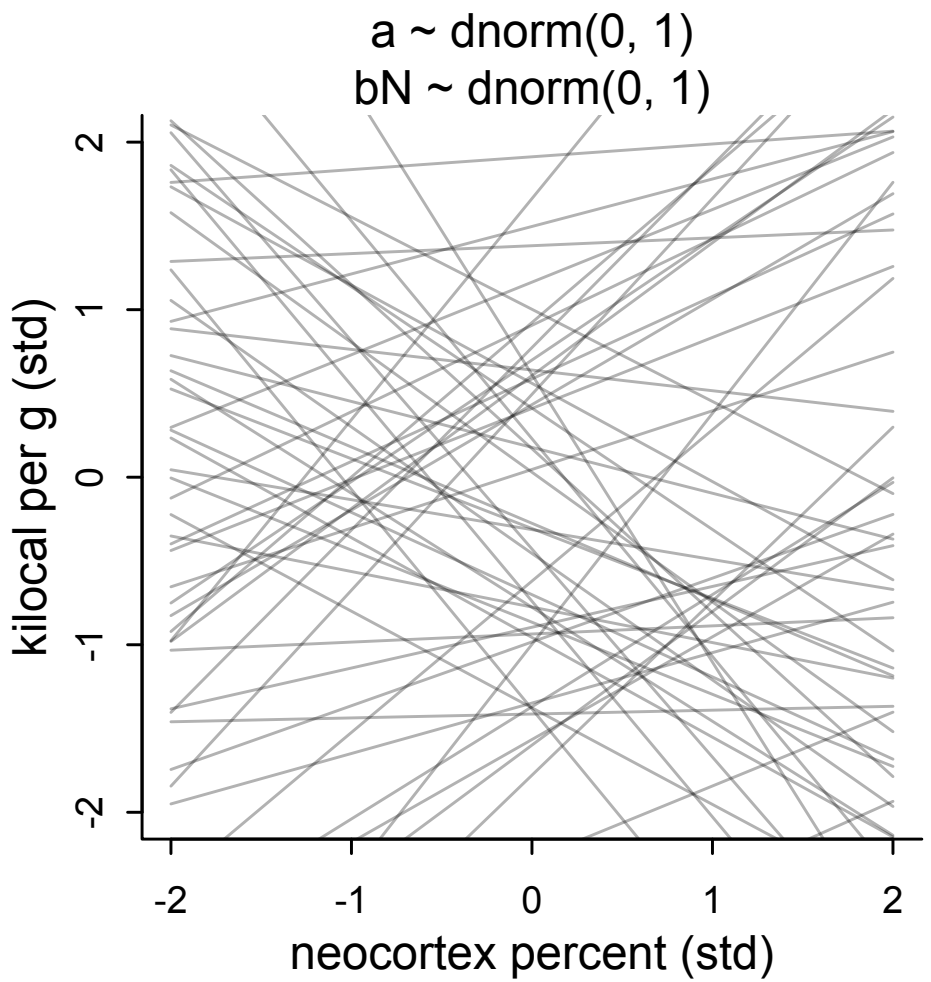
Masked influence

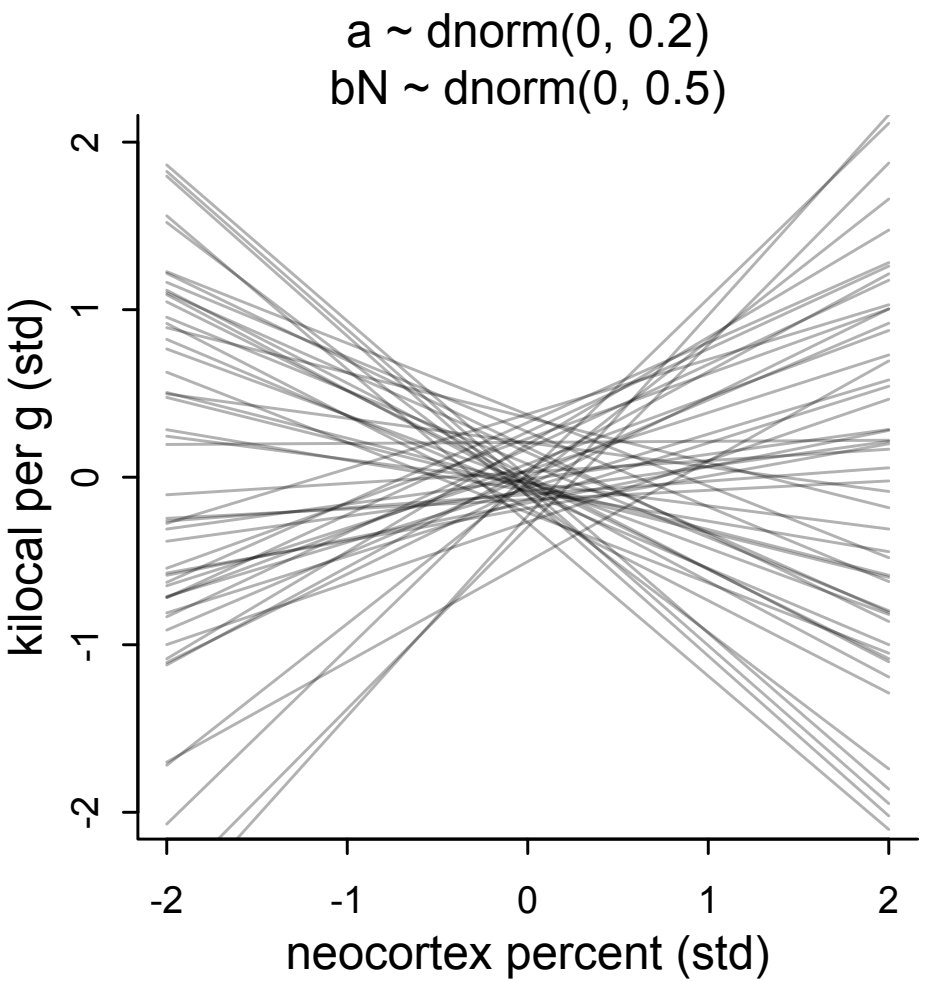
- Primate milk data



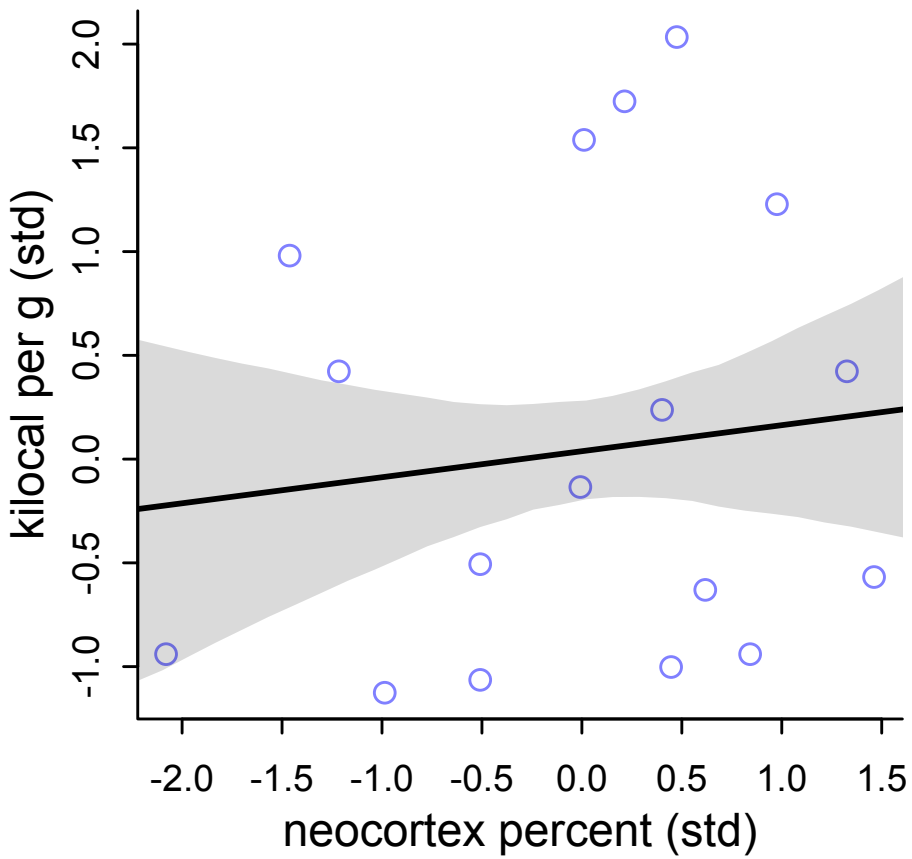

```
library(rethinking)
data(milk)
d <- milk
pairs(~kcal.per.g+log(mass)
    +neocortex.perc , data=d)
```

Necessary sermon on priors





Single predictor models



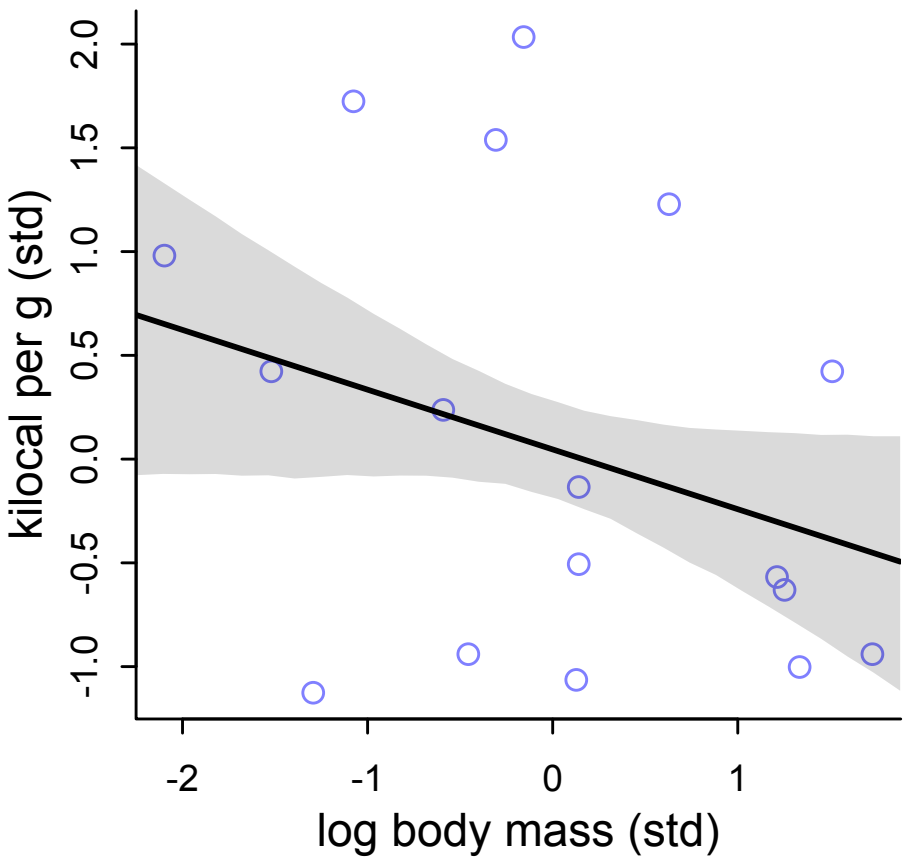


Figure 5.8

Multiple regression model

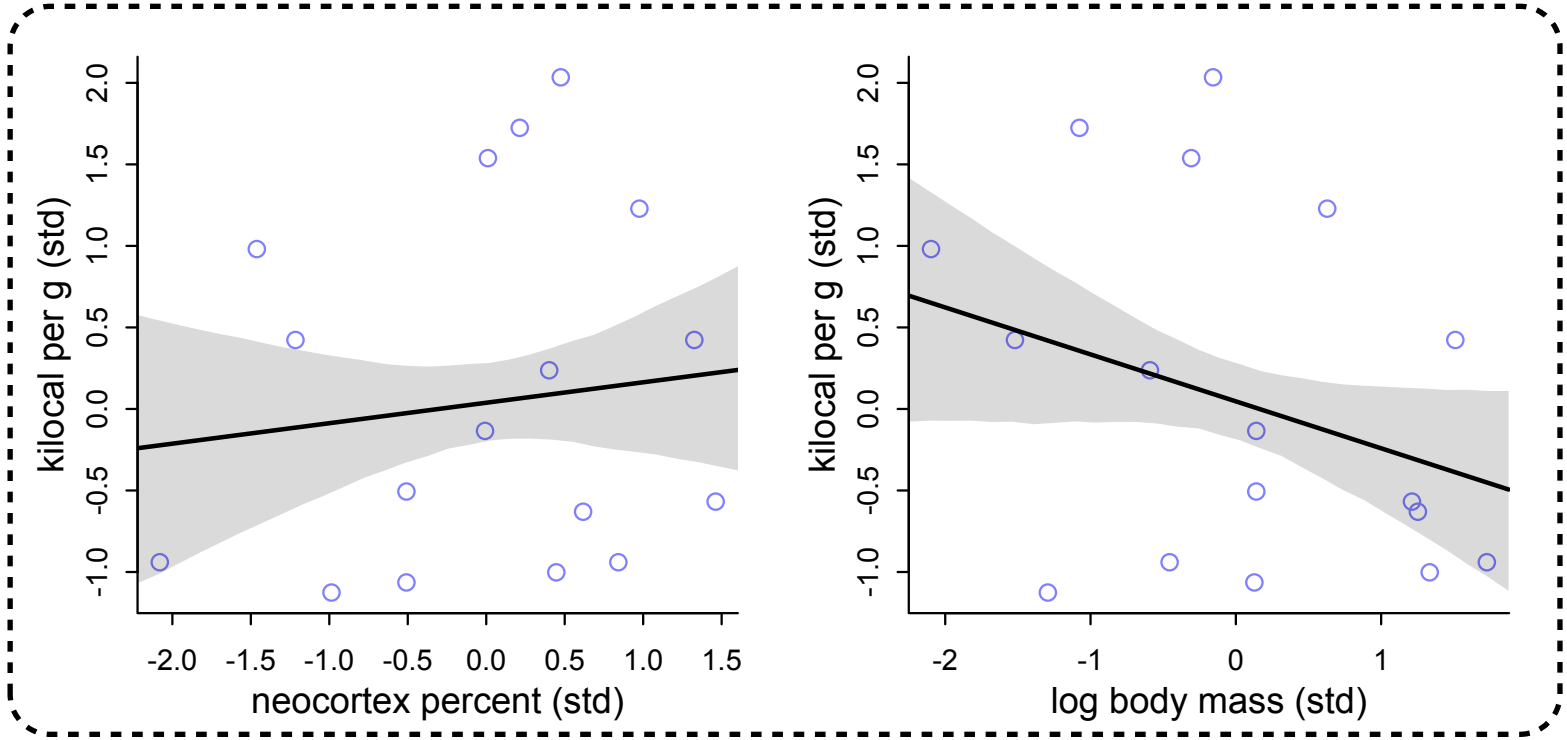
R code 5.29

```
m5.7 <- quap(
    alist(
        K ~ dnorm( mu , sigma ) ,
        mu <- a + bN*N + bM*M ,
        a ~ dnorm( 0 , 0.2 ) ,
        bN ~ dnorm( 0 , 0.5 ) ,
        bM ~ dnorm( 0 , 0.5 ) ,
        sigma ~ dexp( 1 )
    ) , data=dcc )
precis(m5.7)
```

```
        mean   sd  5.5% 94.5%
a       0.07 0.13 -0.15  0.28
bN      0.68 0.25  0.28  1.07
bM     -0.70 0.22 -1.06 -0.35
sigma  0.74 0.13  0.53  0.95
```

$$K_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_N N_i + \beta_M M_i$$
$$\alpha \sim \text{Normal}(0, 0.2)$$
$$\beta_n \sim \text{Normal}(0, 0.5)$$
$$\beta_m \sim \text{Normal}(0, 0.5)$$
$$\sigma \sim \text{Exponential}(1)$$

Bivariate



Multiple

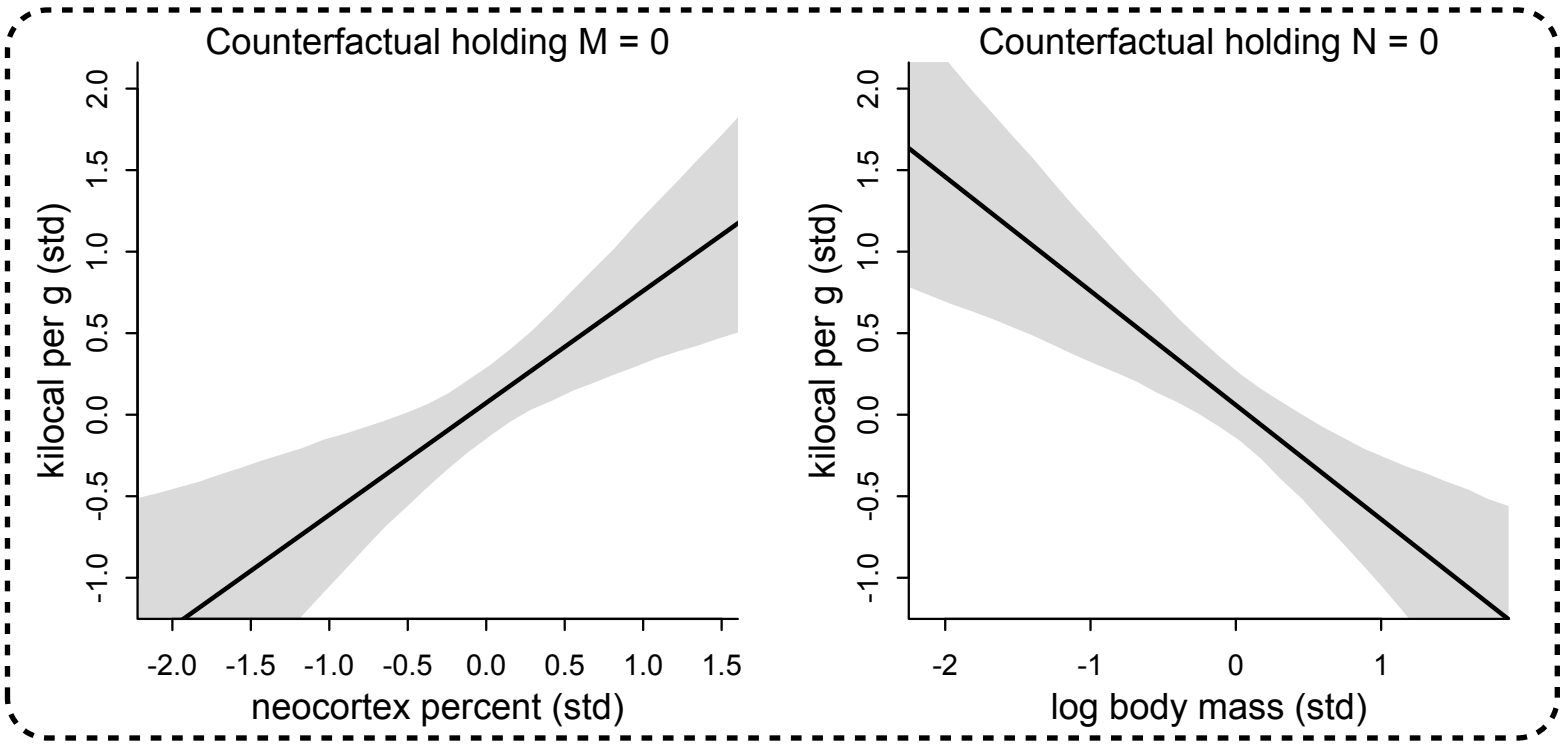
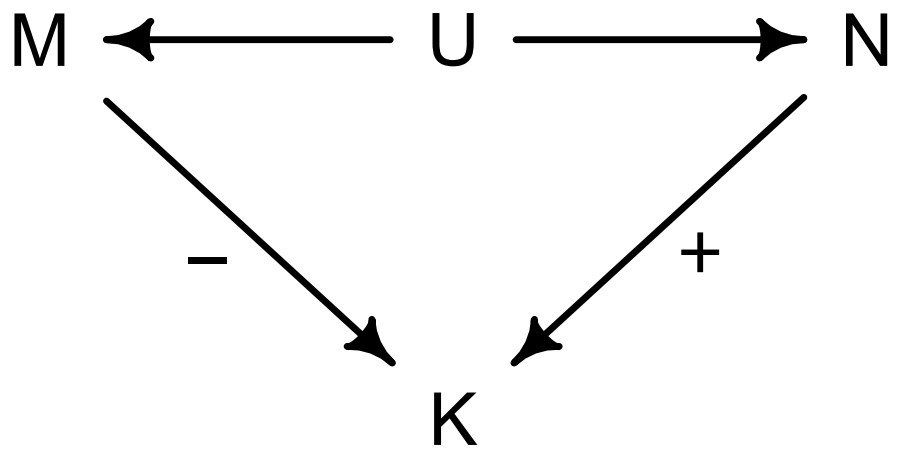


Figure 5.8

Synthetic masked association

```
# M -> K <- N
# M <- U -> N
n <- 100
U <- rnorm( n )
N <- rnorm( n , U )
M <- rnorm( n , U )
K <- rnorm( n , N - M )
d_sim3 <- data.frame(K=K,N=N,M=M)
```



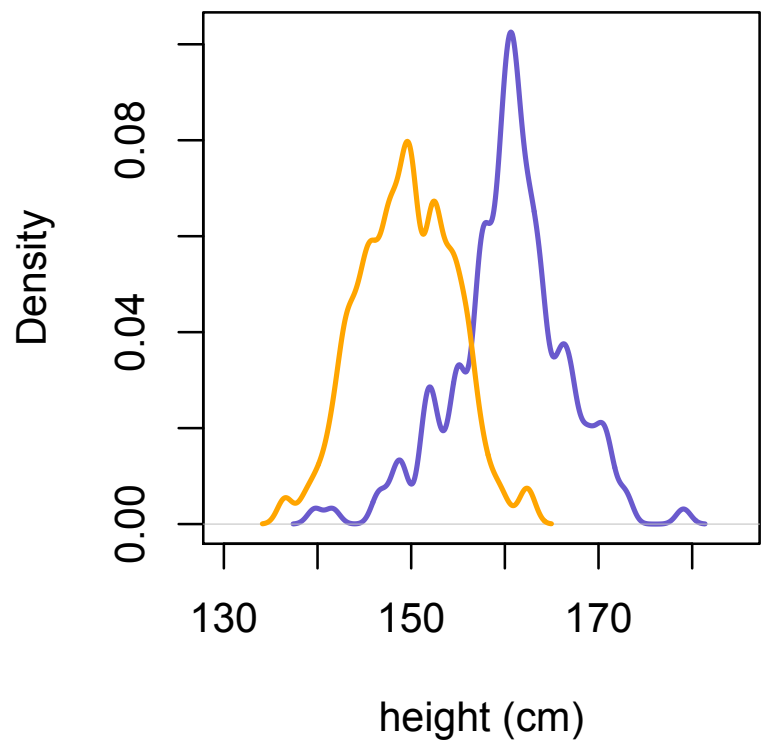
Categorical variables

- Many predictors are discrete, unordered categories
  - Gender, region, species
- How to use in regression?
- Two approaches
  - Use *dummy/indicator* variables
  - Use index variables
- Most automated software uses dummy variables
- Usually easier to think & code with index variables

Dummy (indicator) variables

- Variables that use 1 to indicate a category and 0 to indicate some other category

	height	weight	age	male
1	151.765	47.82561	63	1
2	139.700	36.48581	63	0
3	136.525	31.86484	65	0
4	156.845	53.04191	41	1
5	145.415	41.27687	51	0
6	163.830	62.99259	35	1
7	149.225	38.24348	32	0
8	168.910	55.47997	27	1
9	147.955	34.86988	19	0
10	165.100	54.48774	54	1
11	154.305	49.89512	47	0
12	151.130	41.22017	66	1



Dummy variables

- Dummy variables allow each category to have unique *intercept*
- Coefficient is the *difference* from baseline category

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

$$\mu_i = \alpha + \beta_m m_i$$

mean when $m_i = 0$

change in mean when $m_i = 1$

0/1 variable `male`

Problems with dummy variables

- For k categories, need k–1 dummy variables

```
  season spring summer fall
1 winter      0      0    0
2 spring      1      0    0
3 summer      0      1    0
4   fall      0      0    1
```

- Makes one of the categories a priori more uncertain than others

$$h_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta_m m_i$$
$$\alpha \sim \text{Normal}(178, 20)$$
$$\beta_m \sim \text{Normal}(0, 10)$$

Index variable

R code 5.36

```
d$sex <- ifelse( d$male==1 , 2 , 1 )
str( d$sex )
```

```
 num [1:544] 2 1 1 2 1 2 1 2 1 2 ...
```

$$
\begin{aligned}
h_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha_{\text{SEX}[i]} \\
\alpha_j &\sim \text{Normal}(178, 20) \quad \text{, for } j = 1..2 \\
\sigma &\sim \text{Uniform}(0, 50)
\end{aligned}
$$

Index variable

R code 5.37

```
m5.8 <- quap(
    alist(
        height ~ dnorm( mu , sigma ) ,
        mu <- a[sex] ,
        a[sex] ~ dnorm( 178 , 20 ) ,
        sigma ~ dunif( 0 , 50 )
    ) , data=d )
precis( m5.8 , depth=2 )
```

```
        mean   sd   5.5%  94.5%
a[1]  134.91 1.61 132.34 137.48
a[2]  142.58 1.70 139.86 145.29
sigma  27.31 0.83  25.98  28.63
```

Differences

```
post <- extract.samples(m5.8)
post$diff_fm <- post$a[,1] - post$a[,2]
precis( post , depth=2 )
```

R code 5.38

```
quap posterior: 10000 samples from m5.8
          mean   sd   5.5%  94.5%     histogram
sigma    27.29 0.84  25.95  28.63 ▁▁▁▂▇▇▇▃▂▁▁▁
a[1]    134.91 1.59 132.37 137.42 ▁▁▂▅▇▇▅▂▁▁▁
a[2]    142.60 1.71 139.90 145.35      ▁▂▅▇▃▁▁
diff_fm  -7.70 2.33 -11.41  -3.97    ▁▁▂▅▇▇▃▁▁
```

Difference and uncertainty