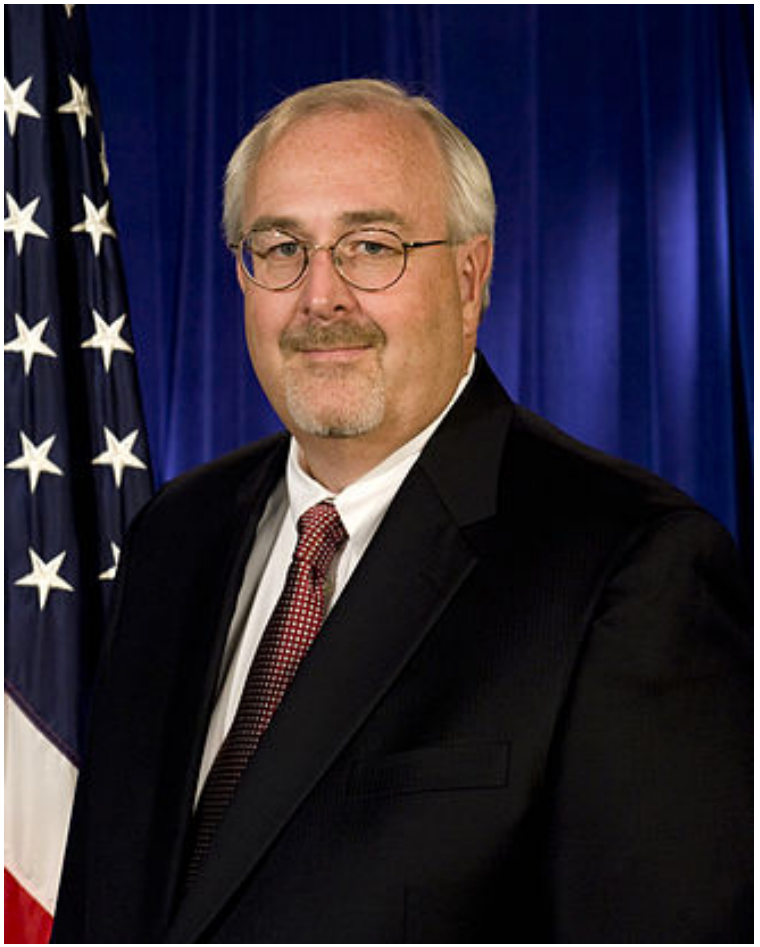
Statistical Rethinking

Week 3: Multivariate Models & The Causal Terror

Richard McElreath


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."



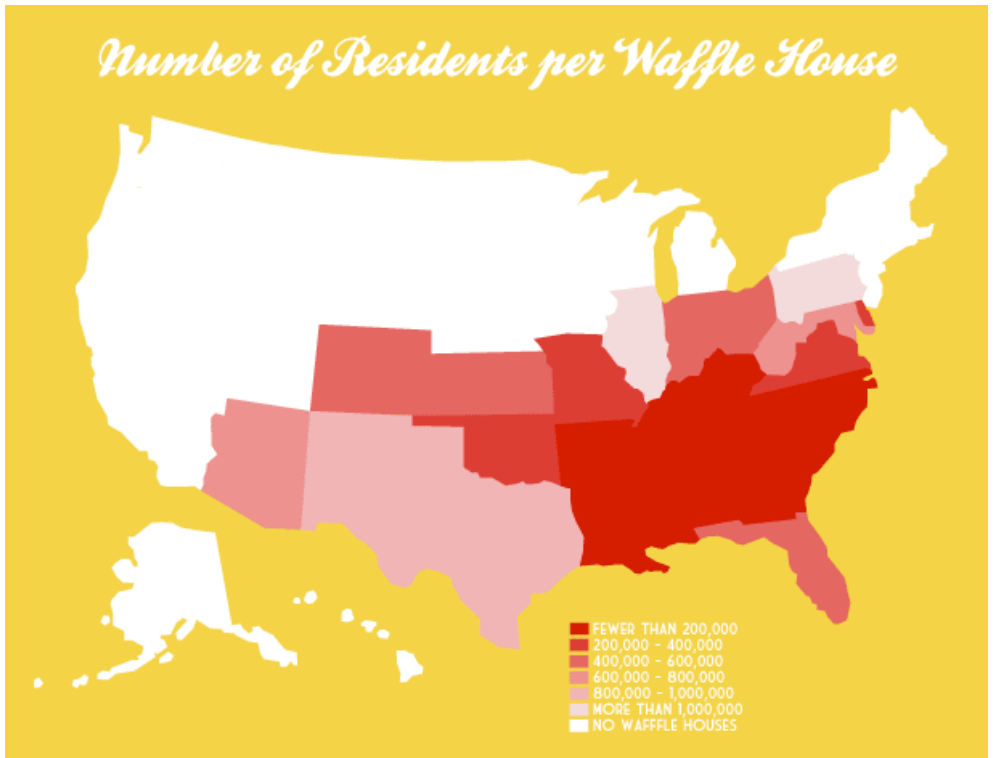
Craig Fugate, director (2009–2017)
USA Federal Emergency Management Agency (FEMA)

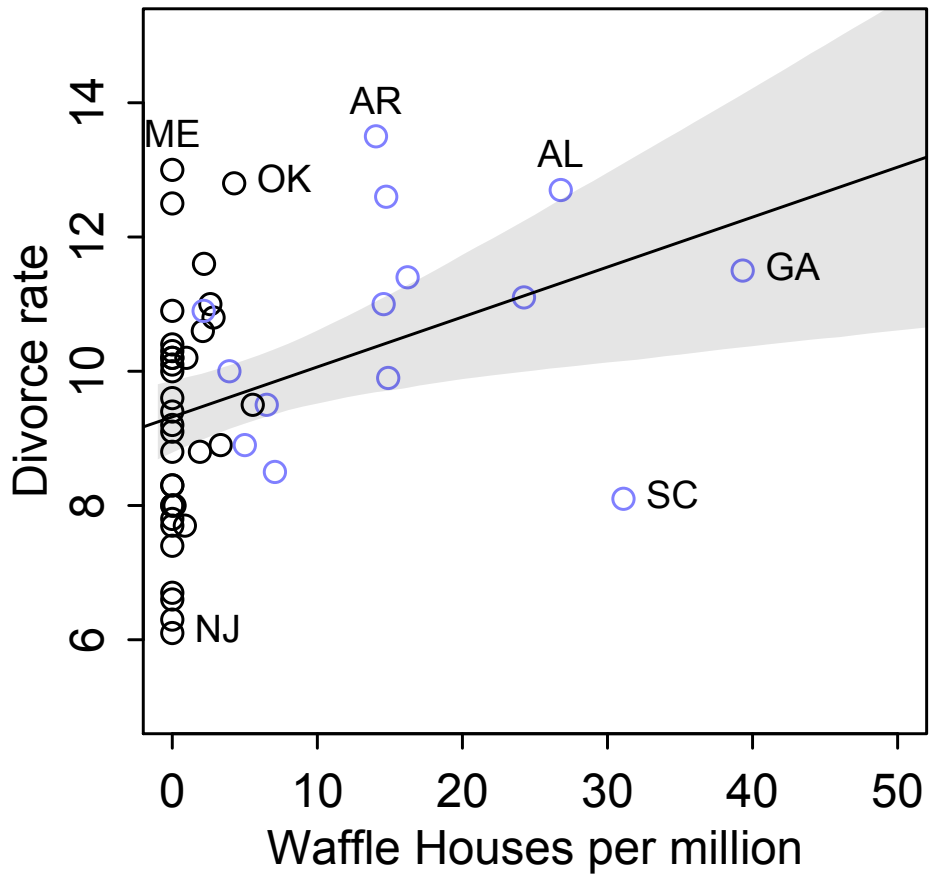
GREEN: Full menu – restaurant has power and damage is limited.

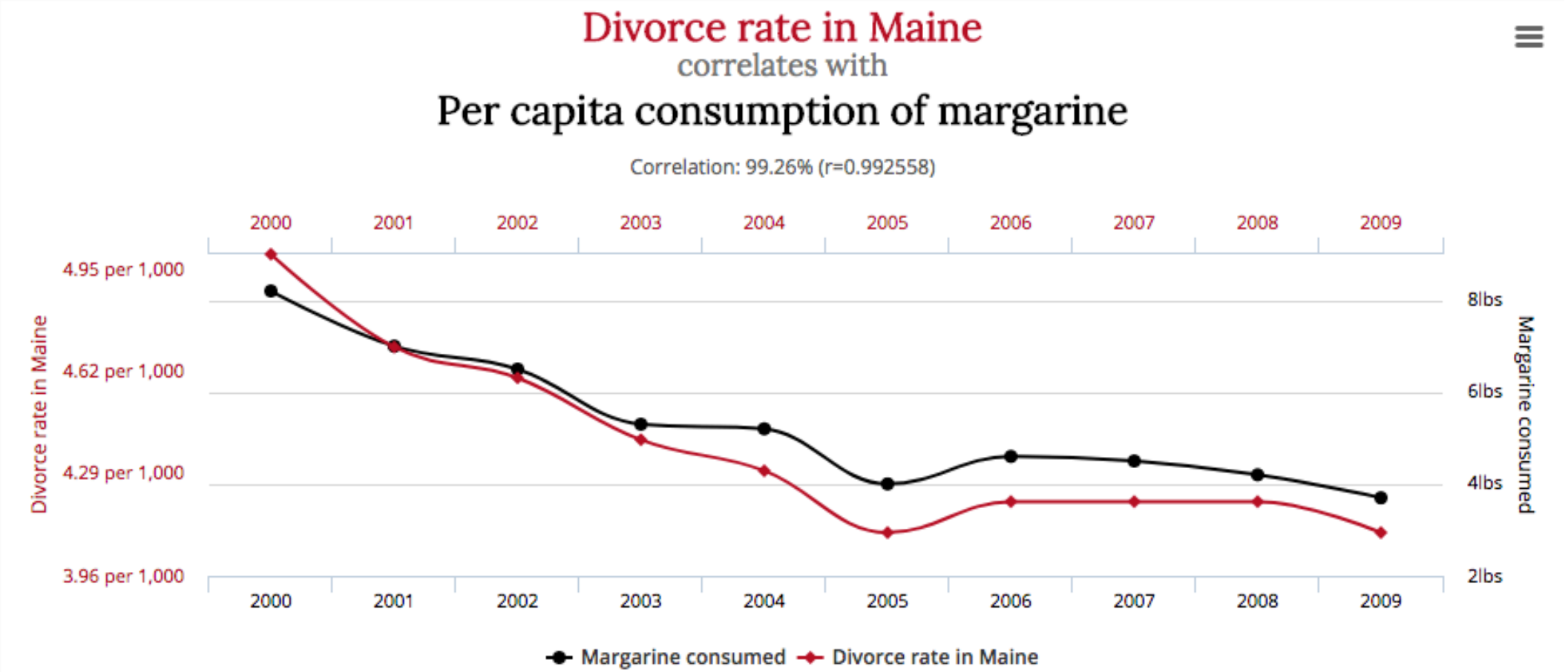
YELLOW: Limited menu – no power or only power from a generator, or food supplies may be low.

RED: Restaurant is closed – indicating severe damage.

Does Waffle House cause divorce?



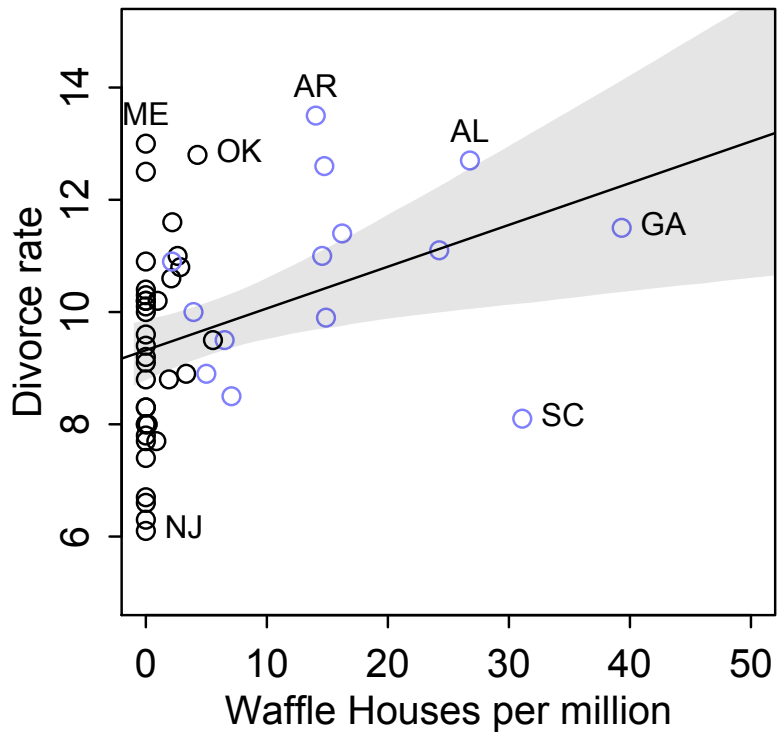




http://www.tylervigen.com/spurious-correlations

Goals this week

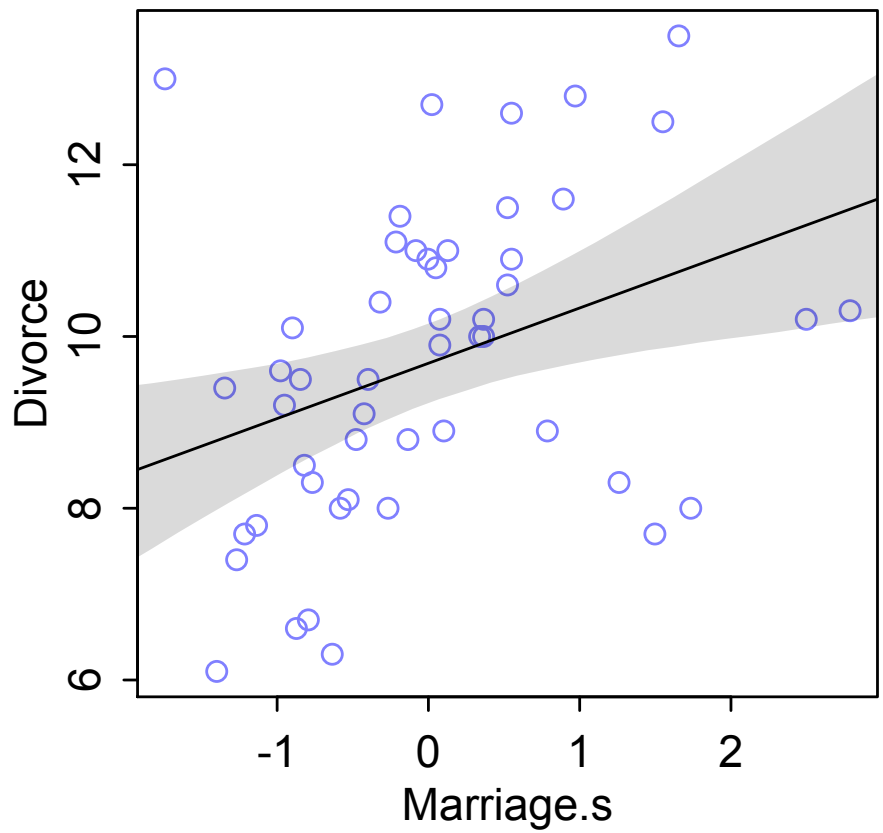
- Multivariate Gaussian models
- The good:
  - Reveal spurious correlation
  - Uncover masked association
- The bad:
  - *Cause* spurious correlation
  - Hide real associations



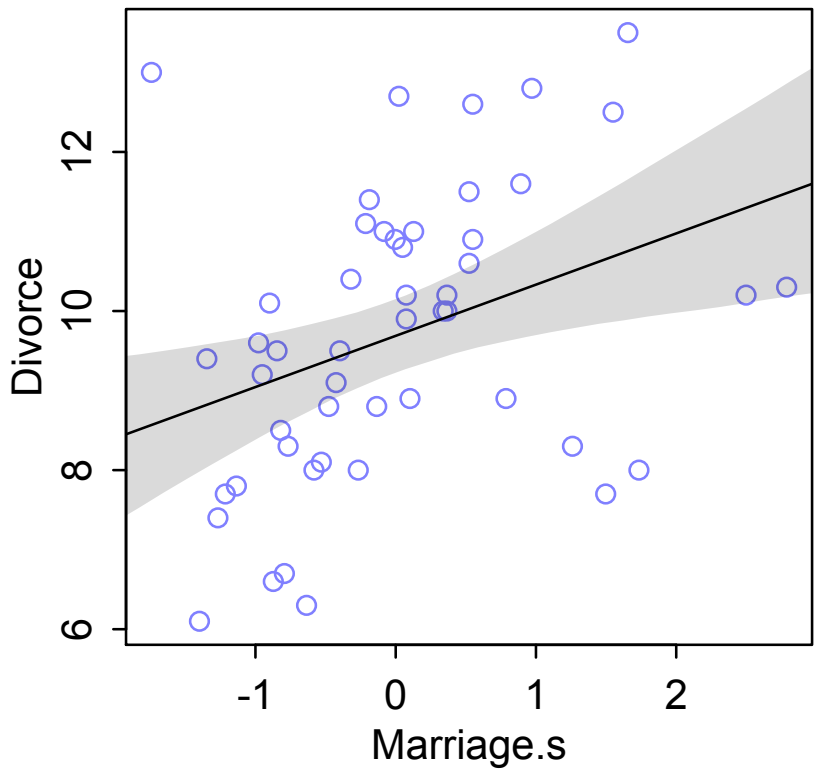
Spurious association

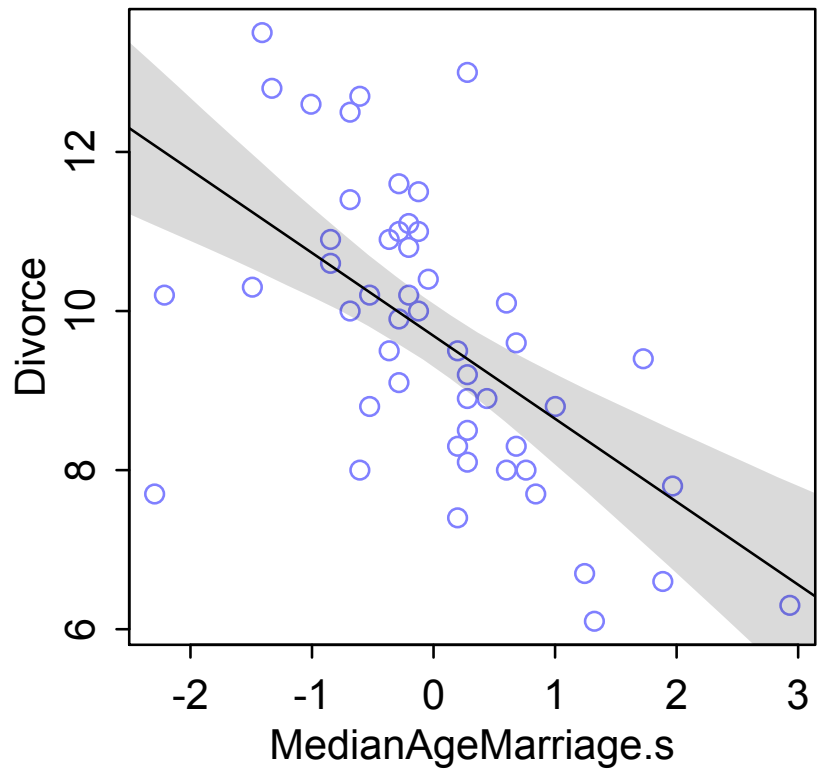
- Correlation does not imply causation
- Causation does not imply correlation
- Causation implies association, perhaps complex
- Need models

- Q: Does marriage cause divorce?



Spurious association



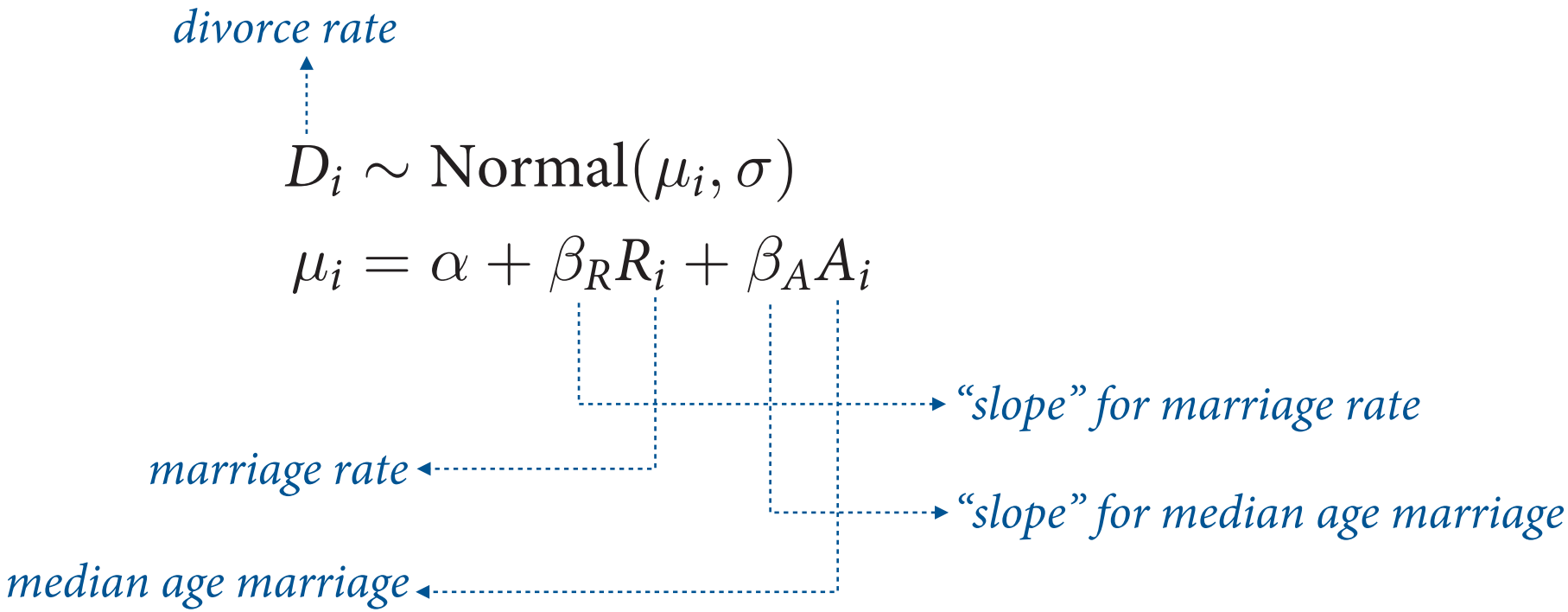


Multivariate divorce

- 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?

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

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


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

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

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

$$\alpha \sim \text{Normal}(10, 10) \quad \text{[prior for } \alpha\text{]}$$

$$\beta_R \sim \text{Normal}(0, 1) \quad \text{[prior for } \beta_R\text{]}$$

$$\beta_A \sim \text{Normal}(0, 1) \quad \text{[prior for } \beta_A\text{]}$$

$$\sigma \sim \text{Uniform}(0, 10) \quad \text{[prior for } \sigma\text{]}$$

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

```
Divorce ~ dnorm(mu,sigma)
```

$$\mu_i = \alpha + \beta_R R_i + \beta_A A_i$$

```
mu <- a+bR*Marriage.s+bA*MedianAgeMarriage.s
```

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

```
a ~ dnorm(10,10)
```

$$\beta_R \sim \text{Normal}(0, 1)$$

```
bR ~ dnorm(0,1)
```

$$\beta_A \sim \text{Normal}(0, 1)$$

```
bA ~ dnorm(0,1)
```

$$\sigma \sim \text{Uniform}(0, 10)$$

```
sigma ~ dunif(0,10)
```

R code 5.4

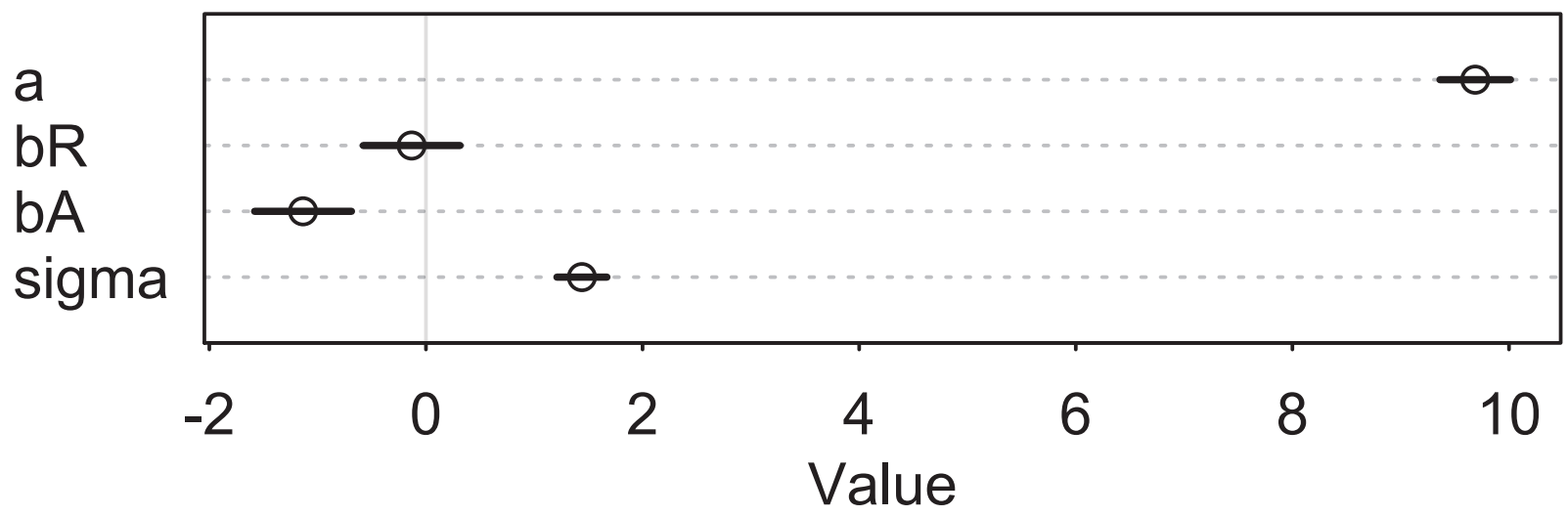
```
m5.3 <- map(
    alist(
        Divorce ~ dnorm( mu , sigma ) ,
        mu <- a + bR*Marriage.s + bA*MedianAgeMarriage.s ,
        a ~ dnorm( 10 , 10 ) ,
        bR ~ dnorm( 0 , 1 ) ,
        bA ~ dnorm( 0 , 1 ) ,
        sigma ~ dunif( 0 , 10 )
    ) ,
    data = d )
precis( m5.3 )
```

```
       Mean StdDev  5.5% 94.5%
a      9.69   0.20  9.36 10.01
bR    -0.13   0.28 -0.58  0.31
bA    -1.13   0.28 -1.58 -0.69
sigma  1.44   0.14  1.21  1.67
```

```
        Mean StdDev  5.5% 94.5%
a       9.69   0.20  9.36 10.01
bR     -0.13   0.28 -0.58  0.31
bA     -1.13   0.28 -1.58 -0.69
sigma   1.44   0.14  1.21  1.67
```

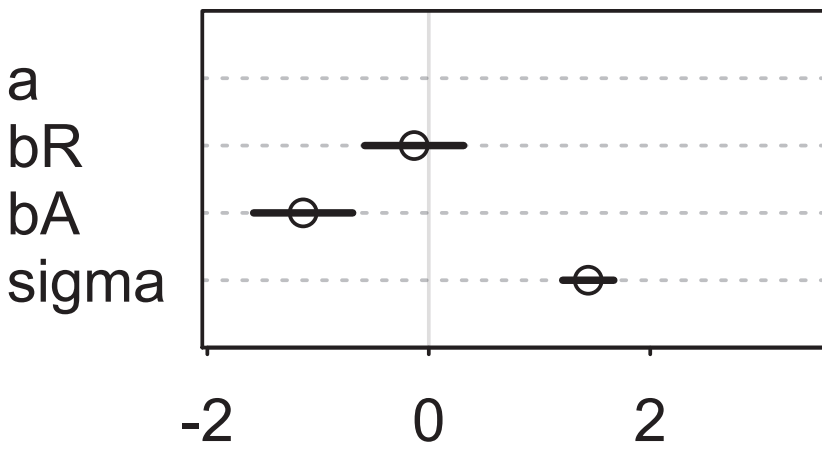
```
plot( precis(m5.3) )
```

R code 5.5



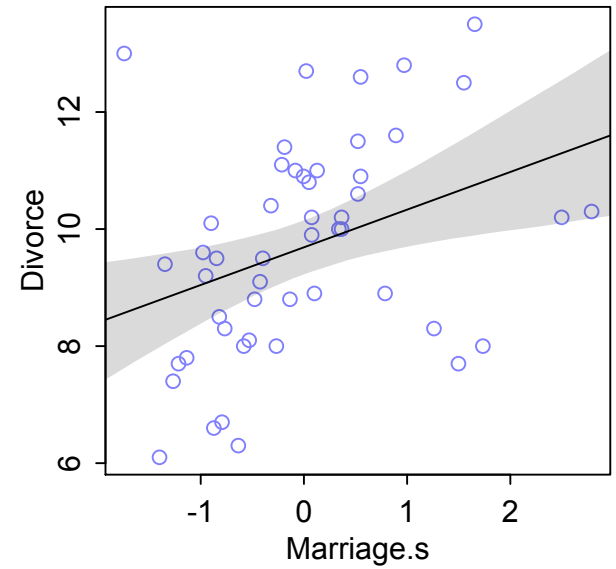
Multivariate divorce

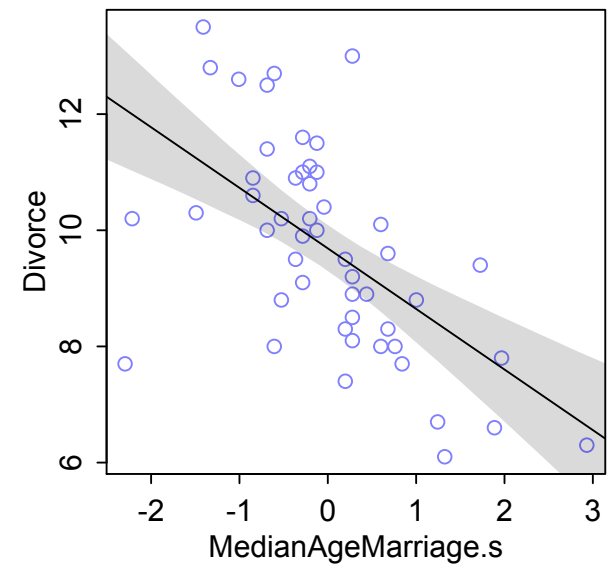
- Once we know median age marriage, little additional value in knowing marriage rate.
- Once we know marriage rate, still value in knowing median age marriage.
- If we *don't know* median age marriage, still useful to know marriage rate.



Plotting multivariate models

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





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

1. Predictor on predictor

- Regress *marriage rate* on *median age marriage*

```
m5.4 <- map(
    alist(
        Marriage.s ~ dnorm( mu , sigma ) ,
        mu <- a + b*MedianAgeMarriage.s ,
        a ~ dnorm( 0 , 10 ) ,
        b ~ dnorm( 0 , 1 ) ,
        sigma ~ dunif( 0 , 10 )
    ) ,
    data = d )
```

$$
\begin{aligned}
R_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta A_i \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta &\sim \text{Normal}(0, 1) \\
\sigma &\sim \text{Uniform}(0, 10)
\end{aligned}
$$

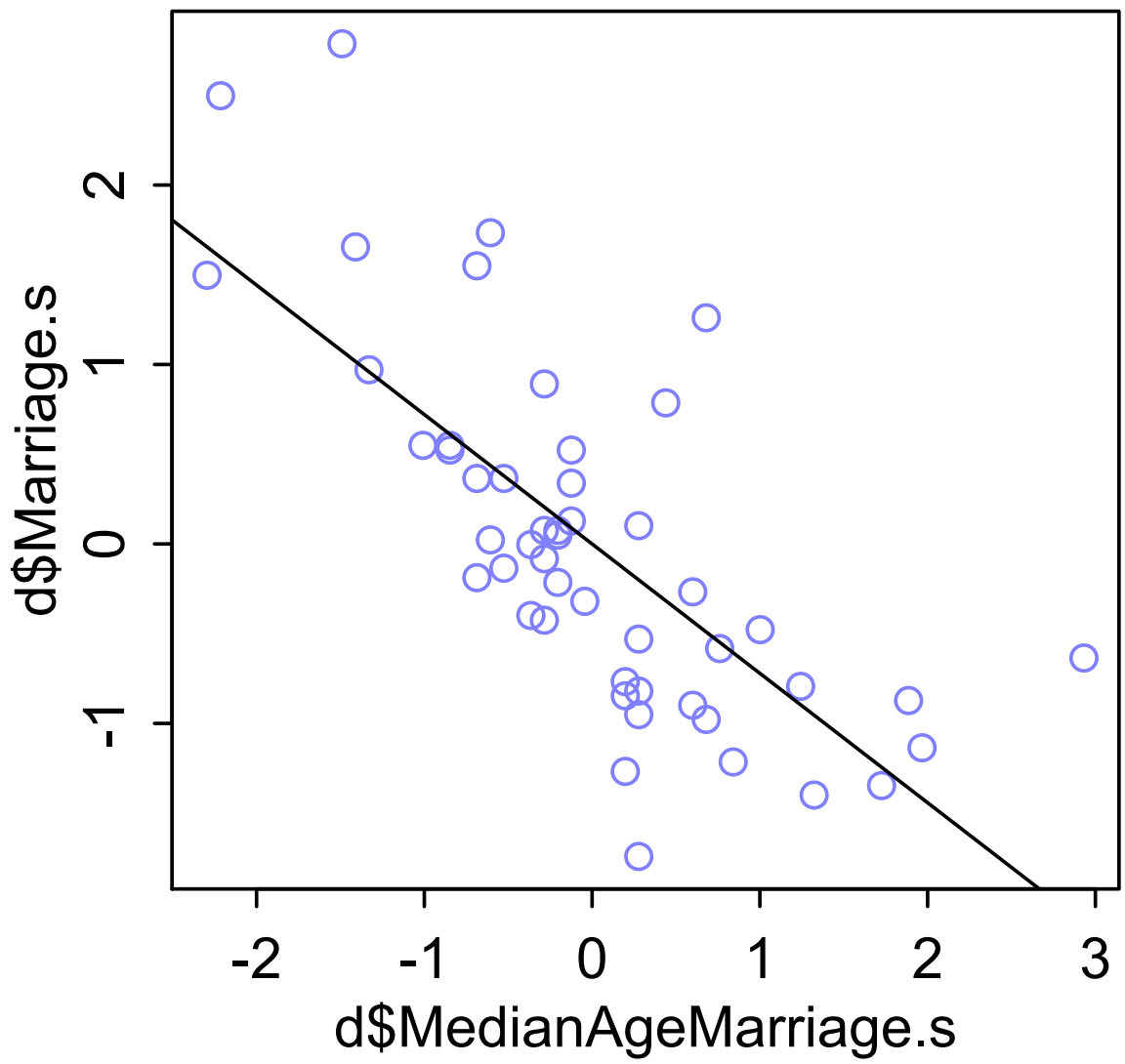
R code 5.6

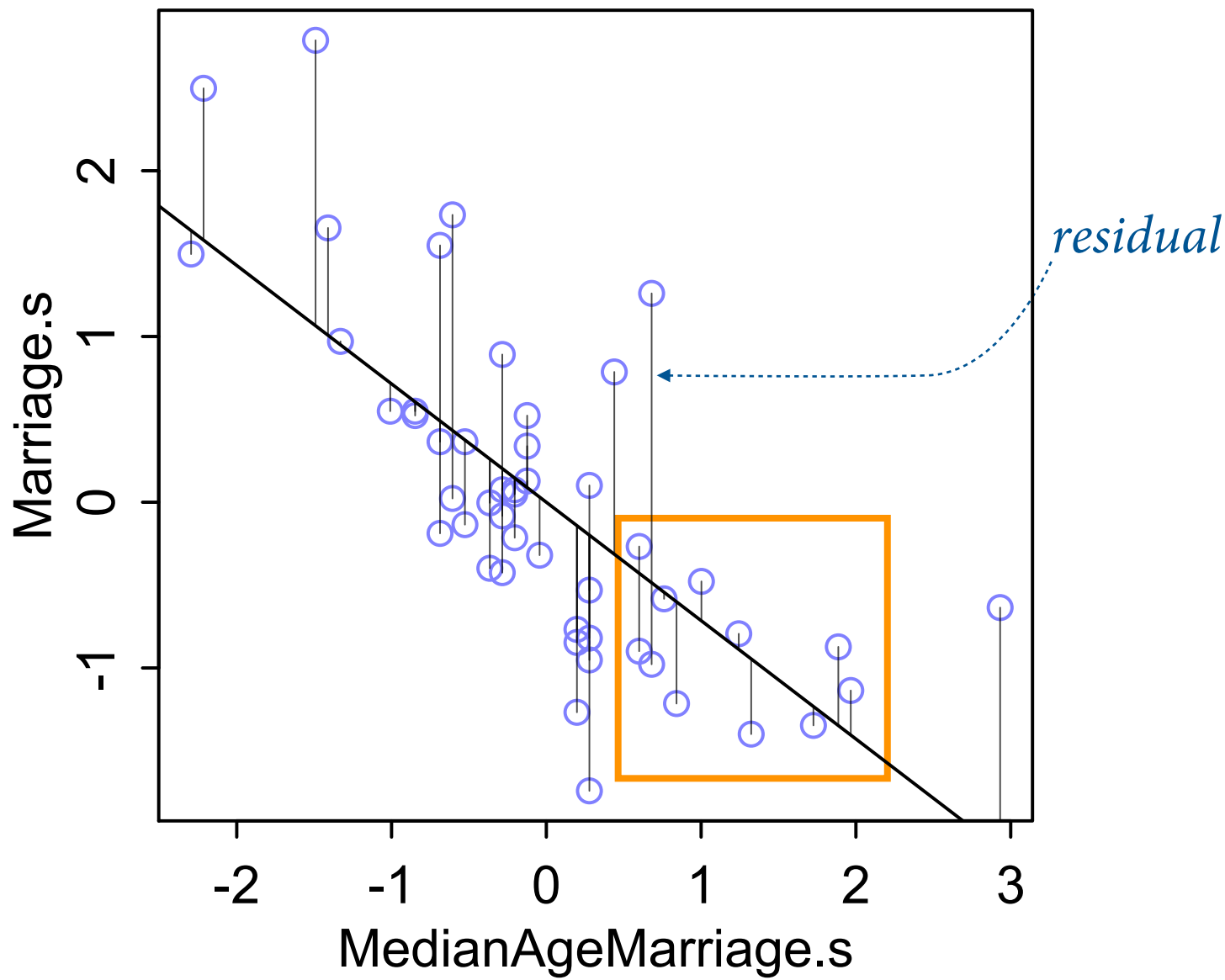
2. Compute residuals

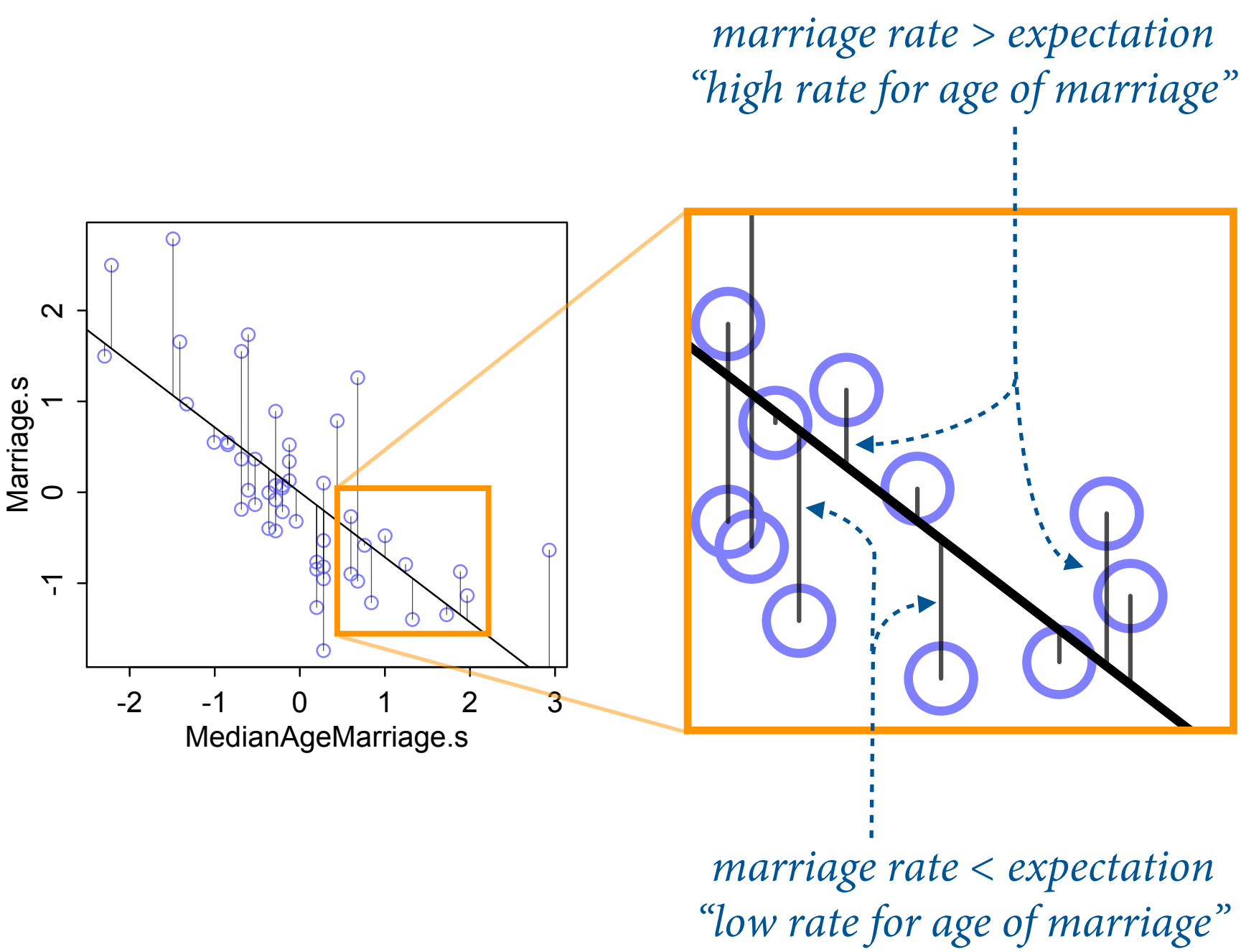
- *Residual*: distance of each outcome from expectation

```
# compute expected value at MAP, for each State
mu <- coef(m5.4)['a'] + coef(m5.4)['b']*d$MedianAgeMarriage.s
# compute residual for each State
m.resid <- d$Marriage.s - mu
```

R code 5.7

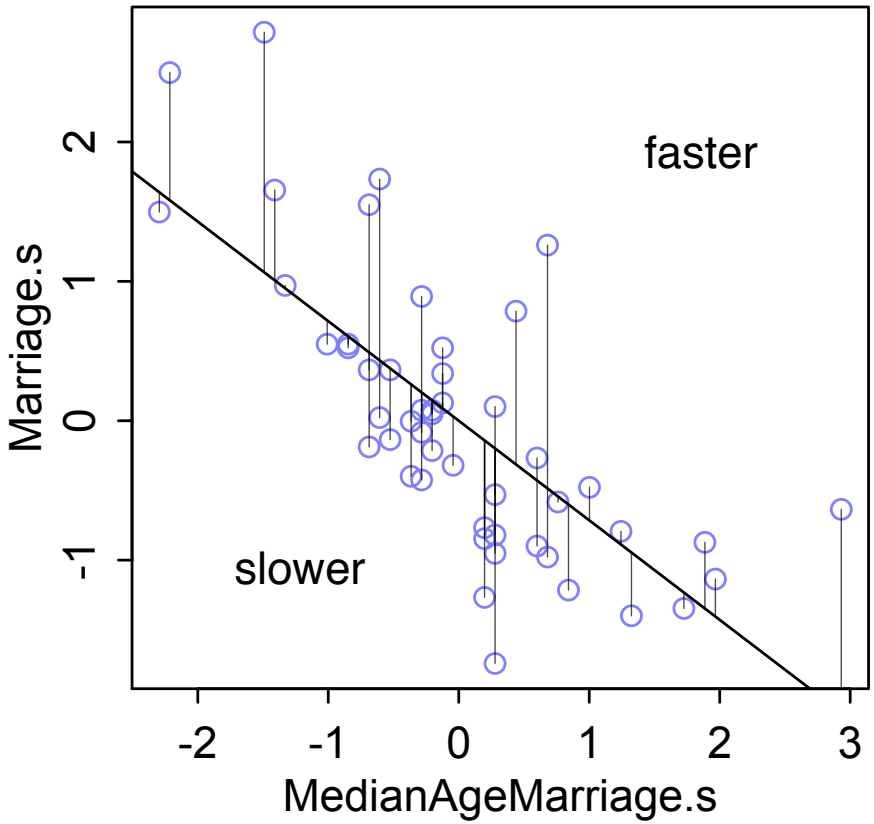

d$Marriage.s
2
1
0
-1
-2
-1
0
1
2
3
d$MedianAgeMarriage.s

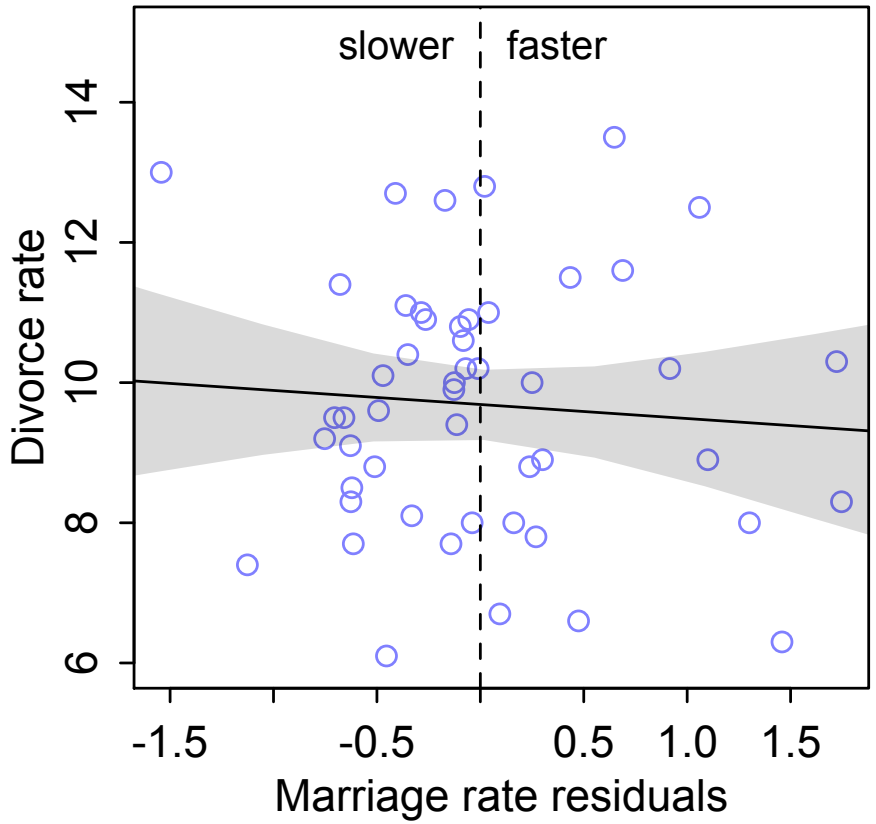

residual
Marriage.s
2
1
0
-1
-2
-1
0
1
2
3
MedianAgeMarriage.s


marriage rate > expectation
"high rate for age of marriage"
Marriage.s
2
1
0
-1
-2
-1
0
1
2
3
MedianAgeMarriage.s
marriage rate < expectation
"low rate for age of marriage"

3. Outcome on residuals

- How is divorce associated with residual marriage rate?

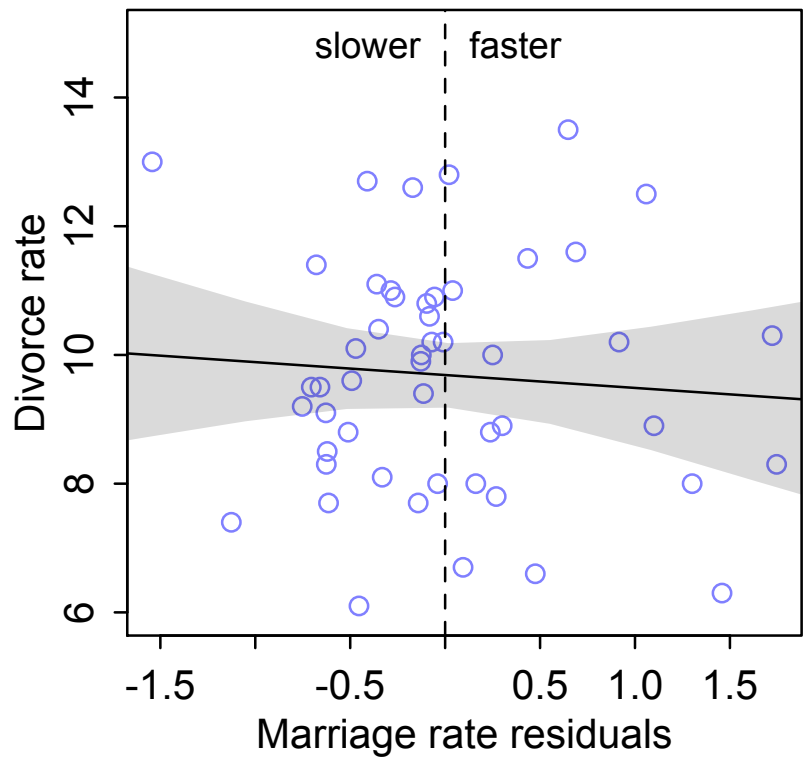

faster
slower
Marriage.s
MedianAgeMarriage.s
-2
-1
0
1
2
3
-1
0
1
2


slower
faster
Divorce rate
Marriage rate residuals
6
8
10
12
14
-1.5
-0.5
0.5
1.0
1.5

3. Outcome on residuals

- How is divorce associated with residual marriage rate?

States with fast/slow rates of marriage (for age of marriage) do not (on average) have fast/slow divorce rates



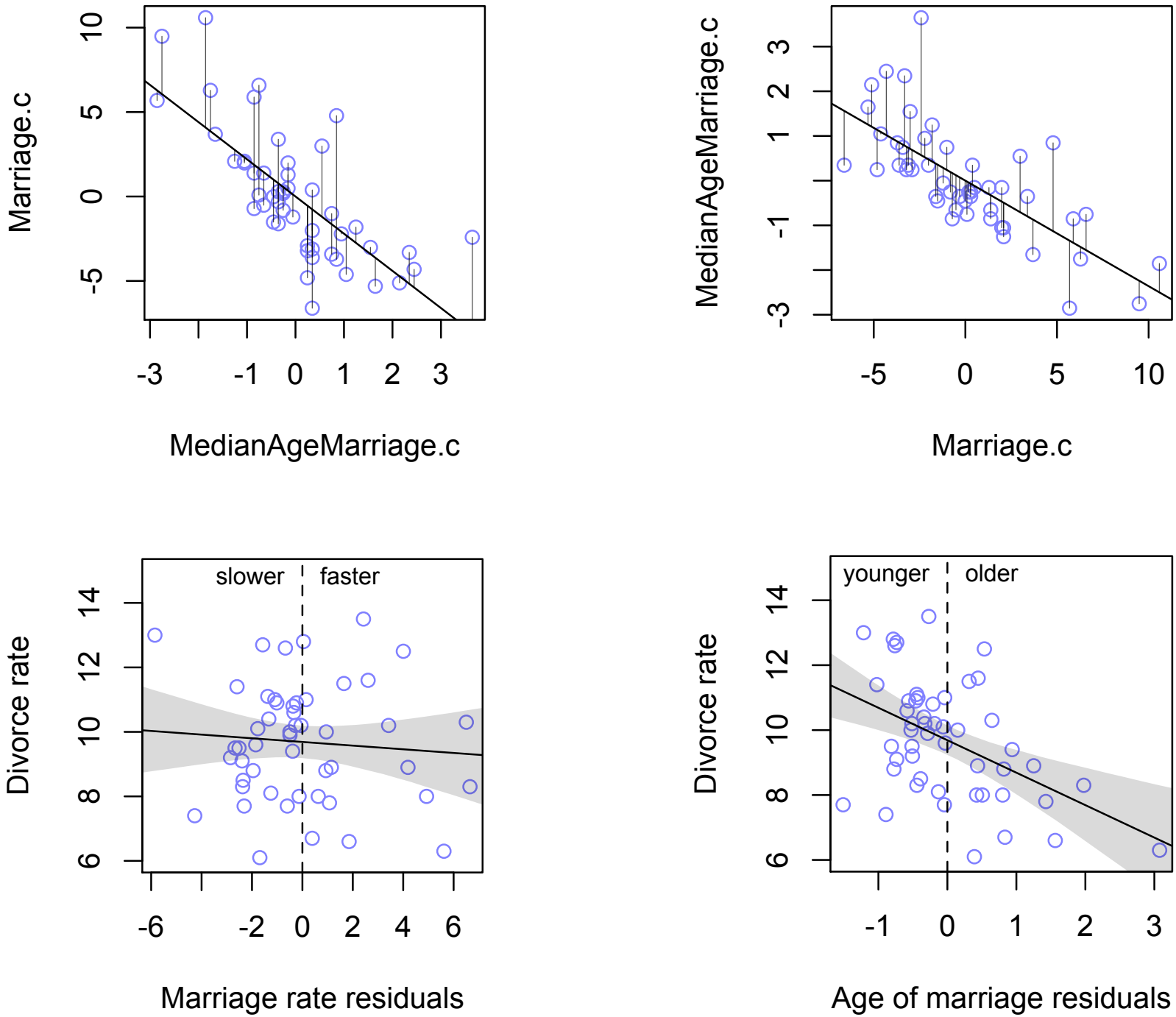

Marriage.c
MedianAgeMarriage.c
MedianAgeMarriage.c
Marriage.c
Divorce rate
slower
faster
Marriage rate residuals
Divorce rate
younger
older
Age of marriage residuals

Figure 5.4

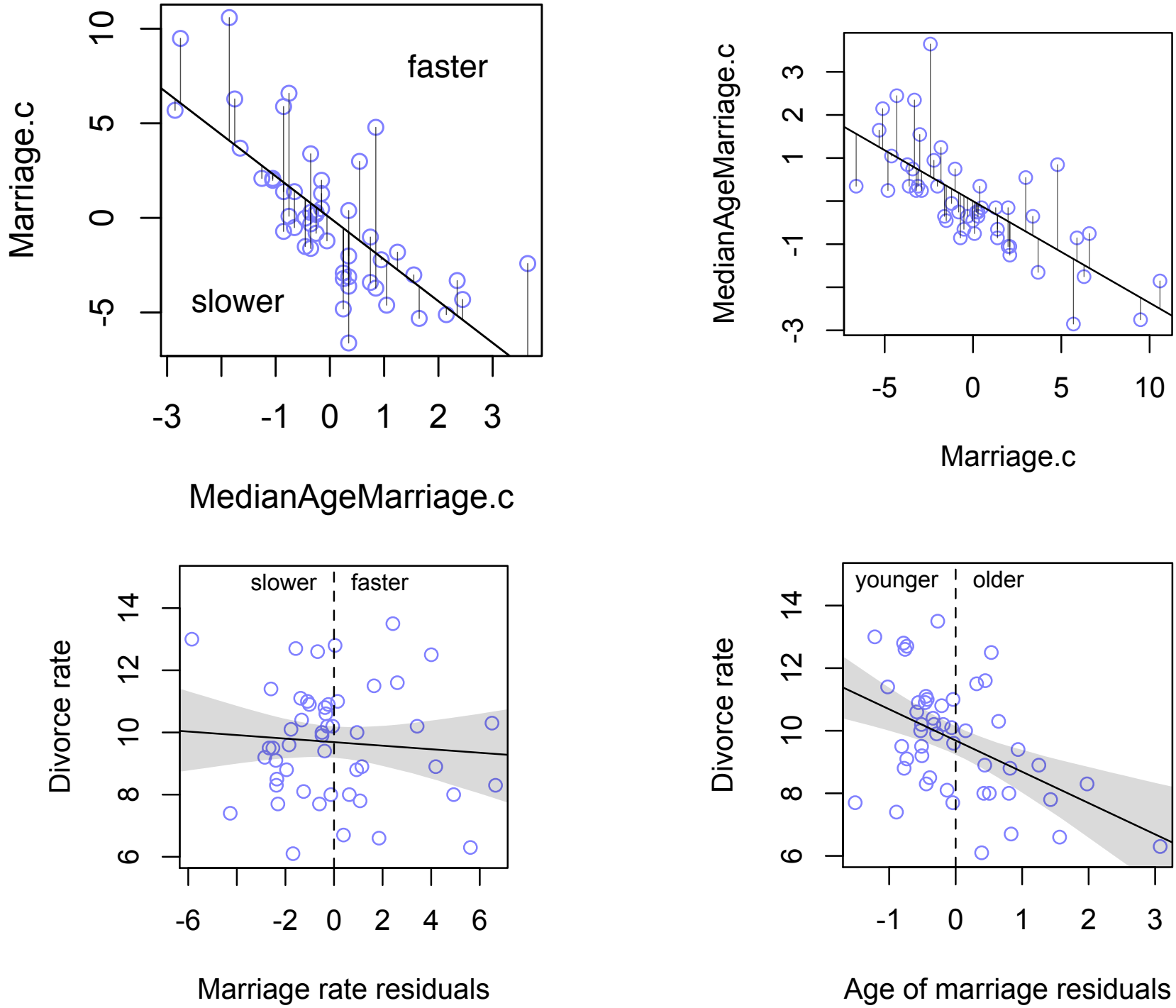

Marriage.c
faster
slower
MedianAgeMarriage.c
MedianAgeMarriage.c
Marriage.c
Divorce rate
slower
faster
Marriage rate residuals
Divorce rate
younger
older
Age of marriage residuals

Figure 5.4

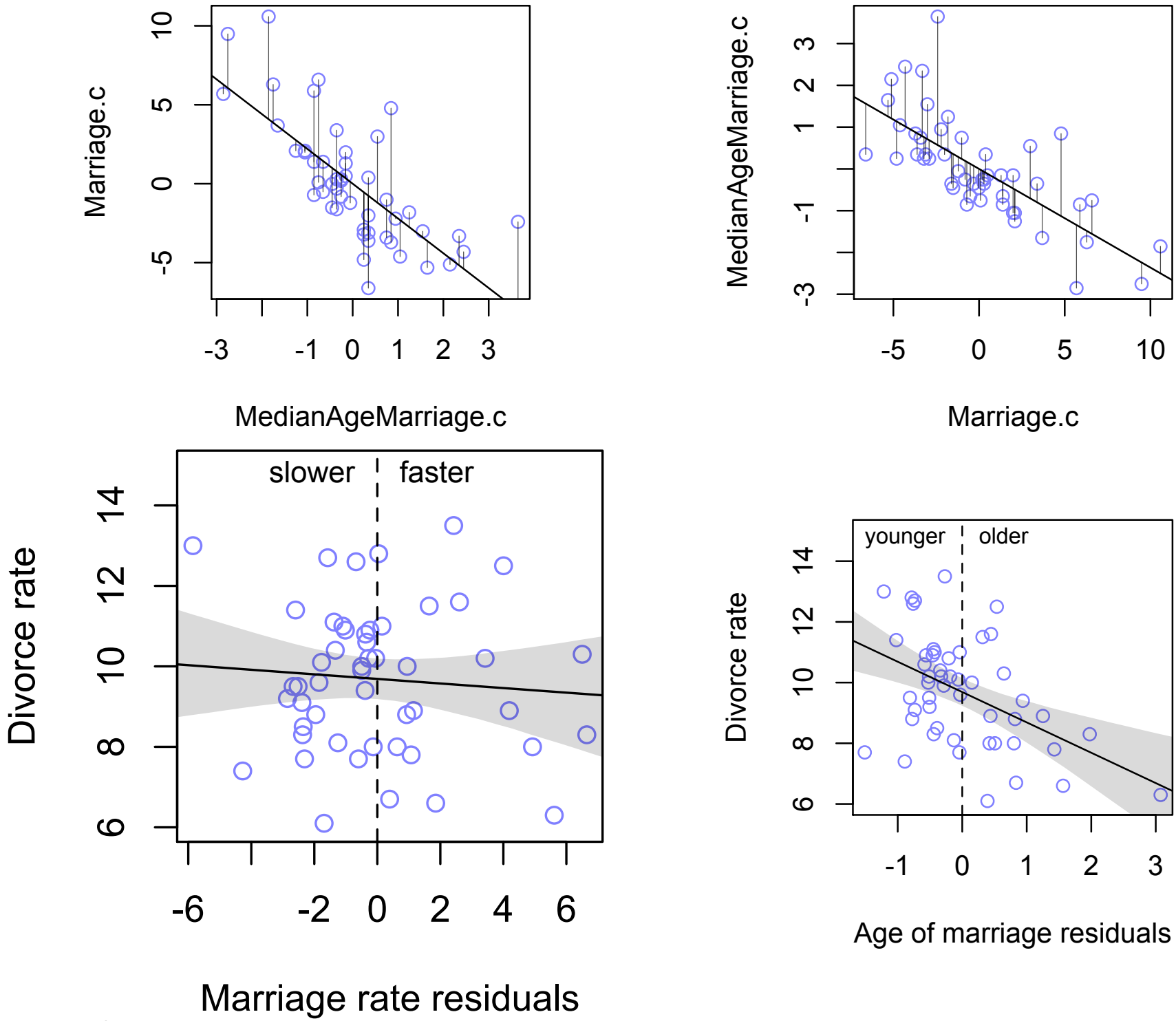
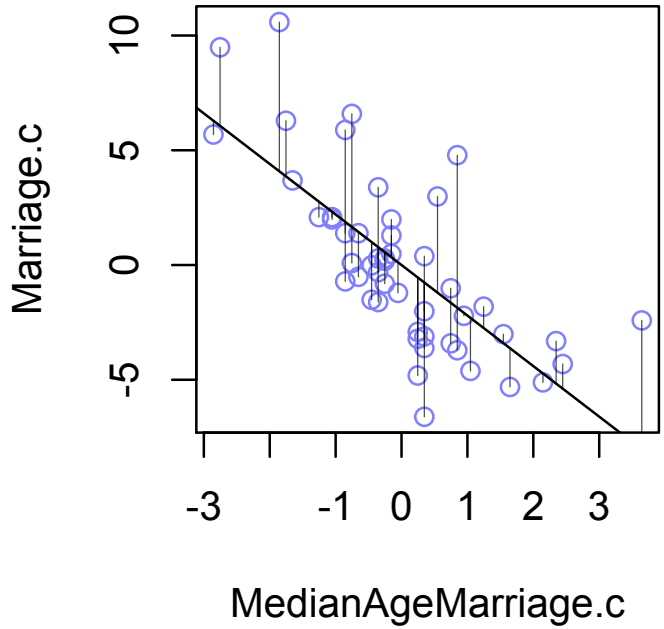
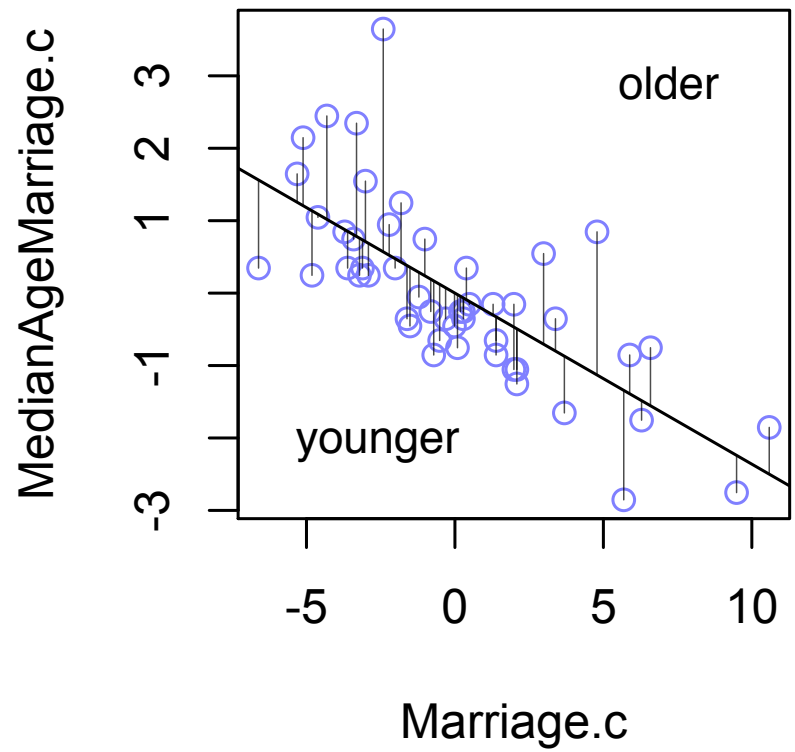
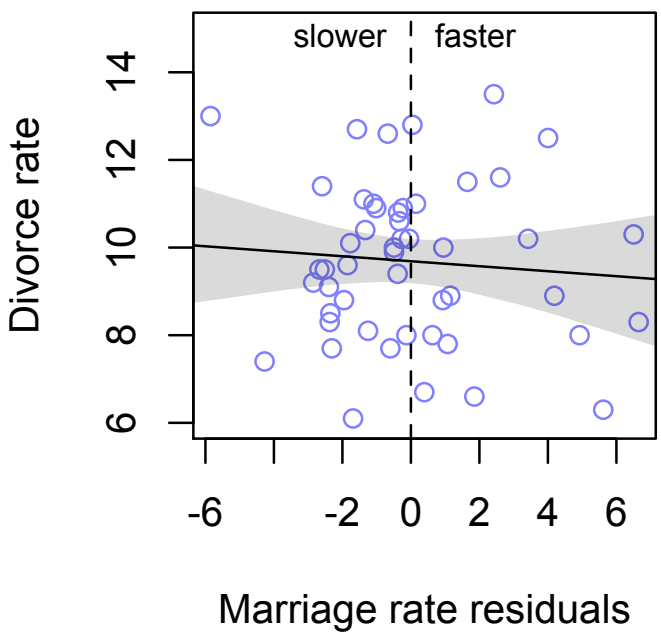

Marriage.c
MedianAgeMarriage.c
MedianAgeMarriage.c
Marriage.c
Divorce rate
slower
faster
Marriage rate residuals
Divorce rate
younger
older
Age of marriage residuals

Figure 5.4


Marriage.c
10
5
0
-5
-3
-1
0
1
2
3
MedianAgeMarriage.c


MedianAgeMarriage.c
3
2
1
-1
-3
older
younger
-5
0
5
10
Marriage.c


Divorce rate
14
12
10
8
6
slower
faster
-6
-2
0
2
4
6
Marriage rate residuals

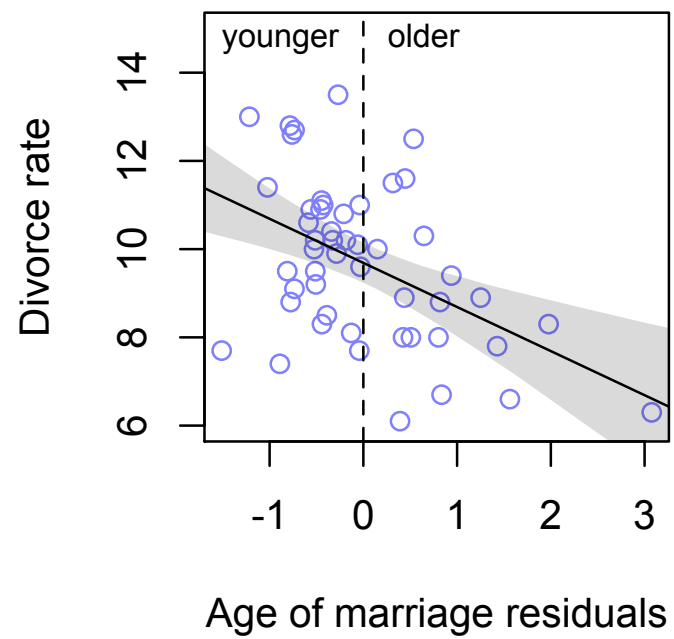

Divorce rate
14
12
10
8
6
younger
older
-1
0
1
2
3
Age of marriage residuals

Figure 5.4

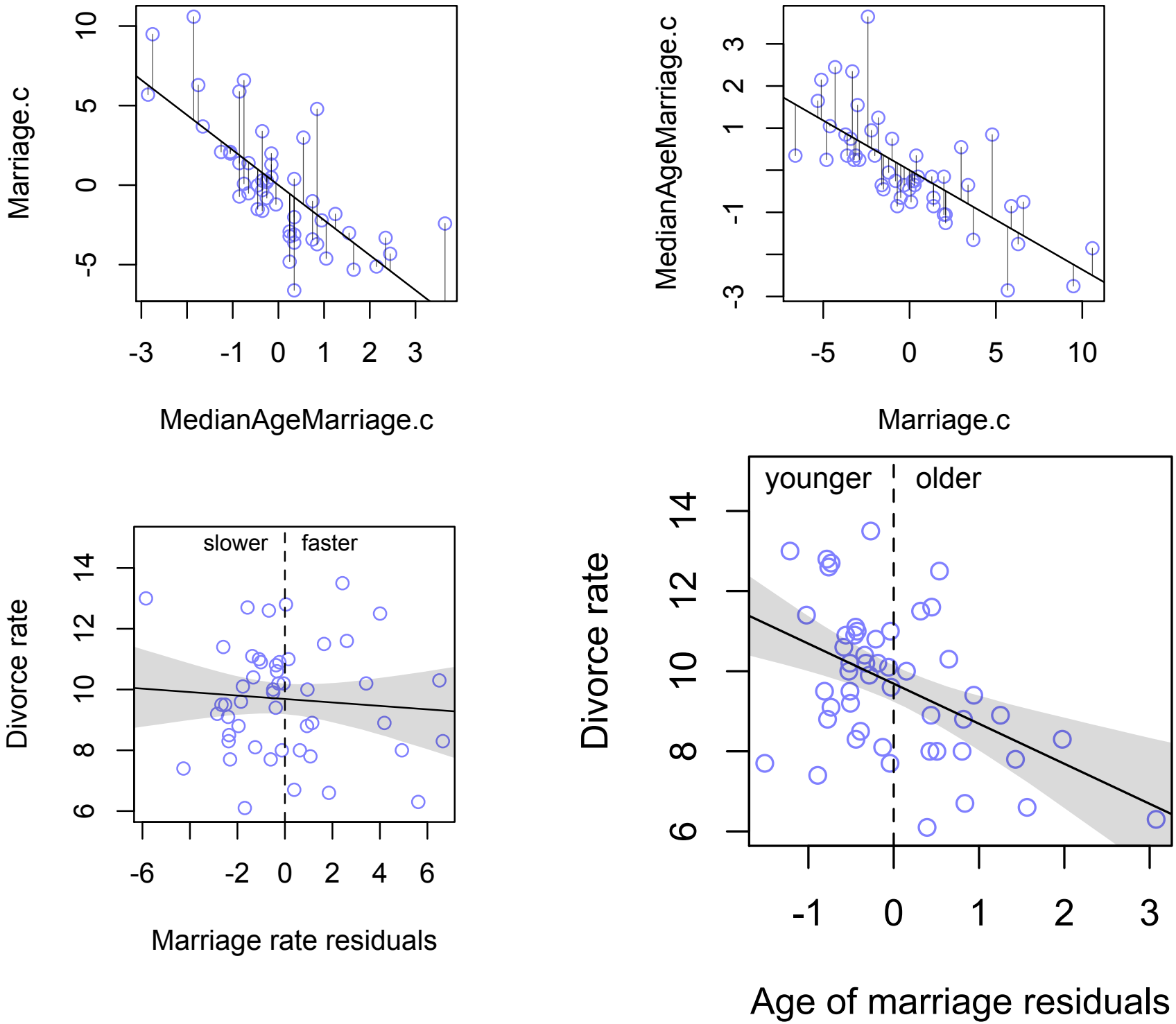
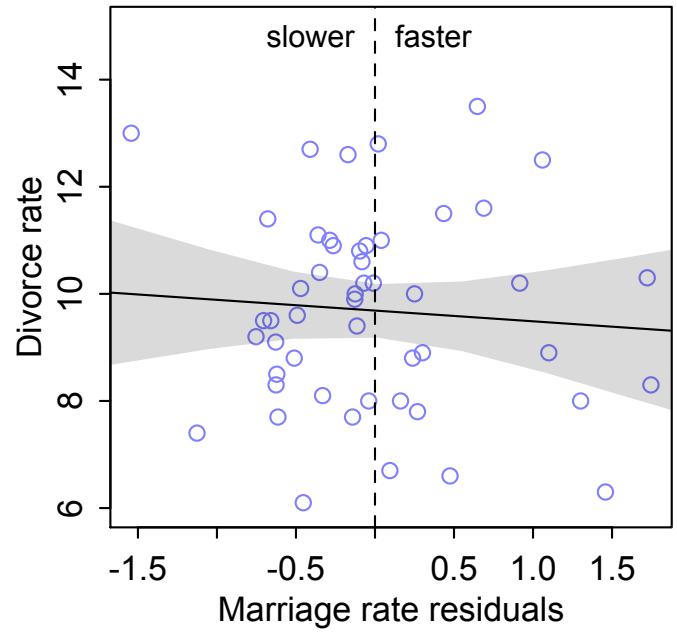

Marriage.c
MedianAgeMarriage.c
MedianAgeMarriage.c
Marriage.c
Divorce rate
slower
faster
Marriage rate residuals
Divorce rate
younger
older
Age of marriage residuals

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"

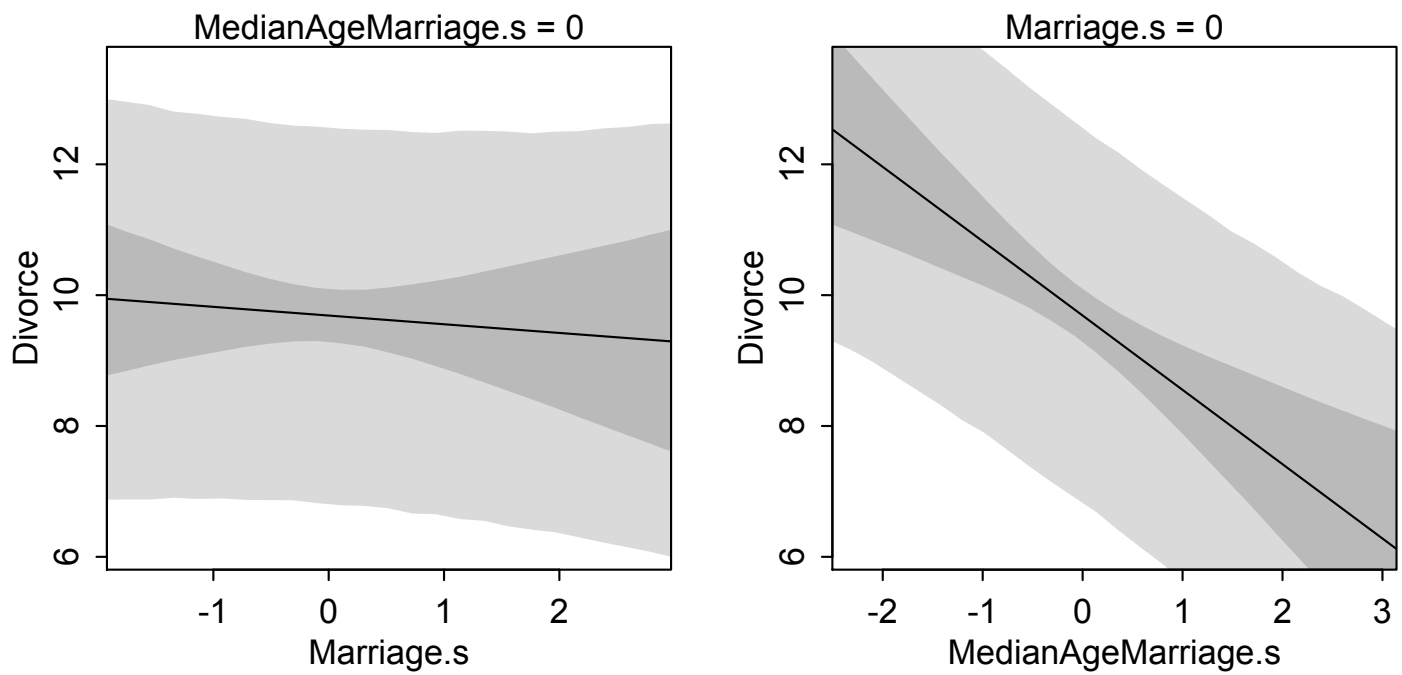
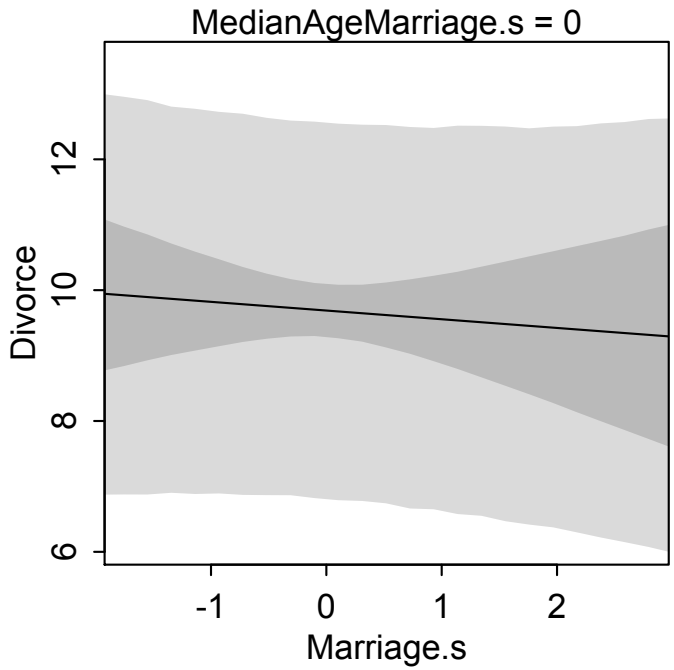
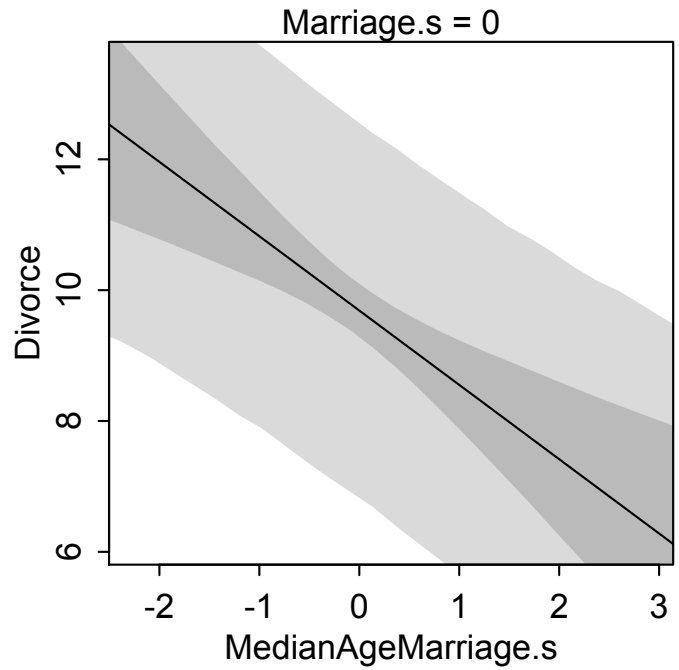


Figure 5.6



Change marriage rate, without changing median age marriage?



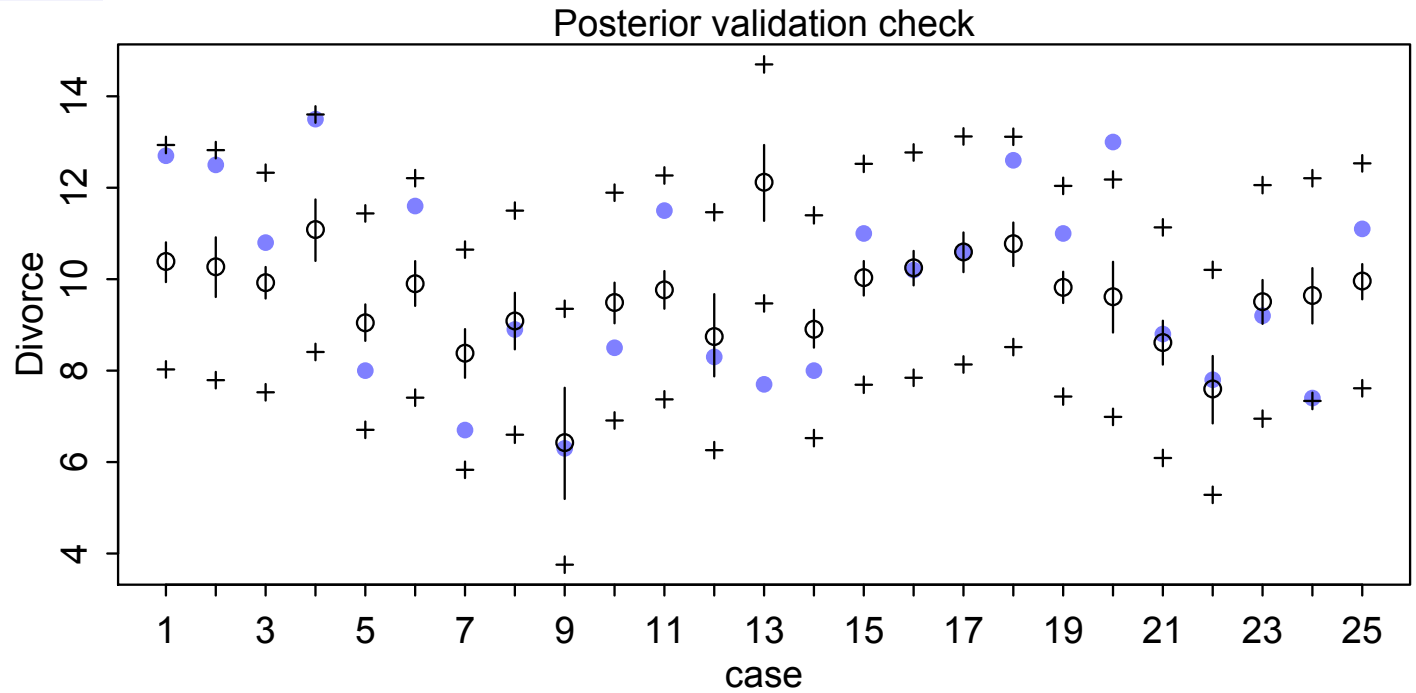
Change median age marriage, without changing marriage rate?

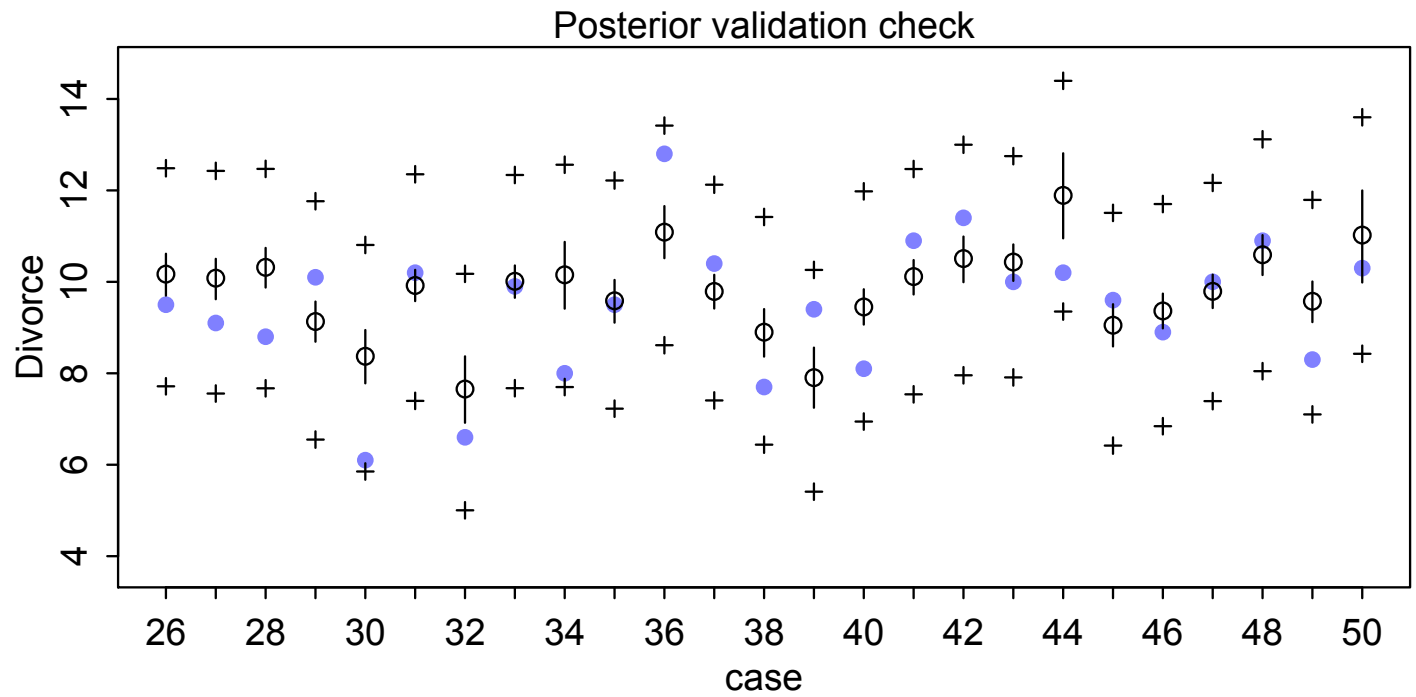
Figure 5.6

Posterior predictions

- 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 MAP leads to overconfidence
  - Embrace the uncertainty

```
postcheck(m5.3)
```





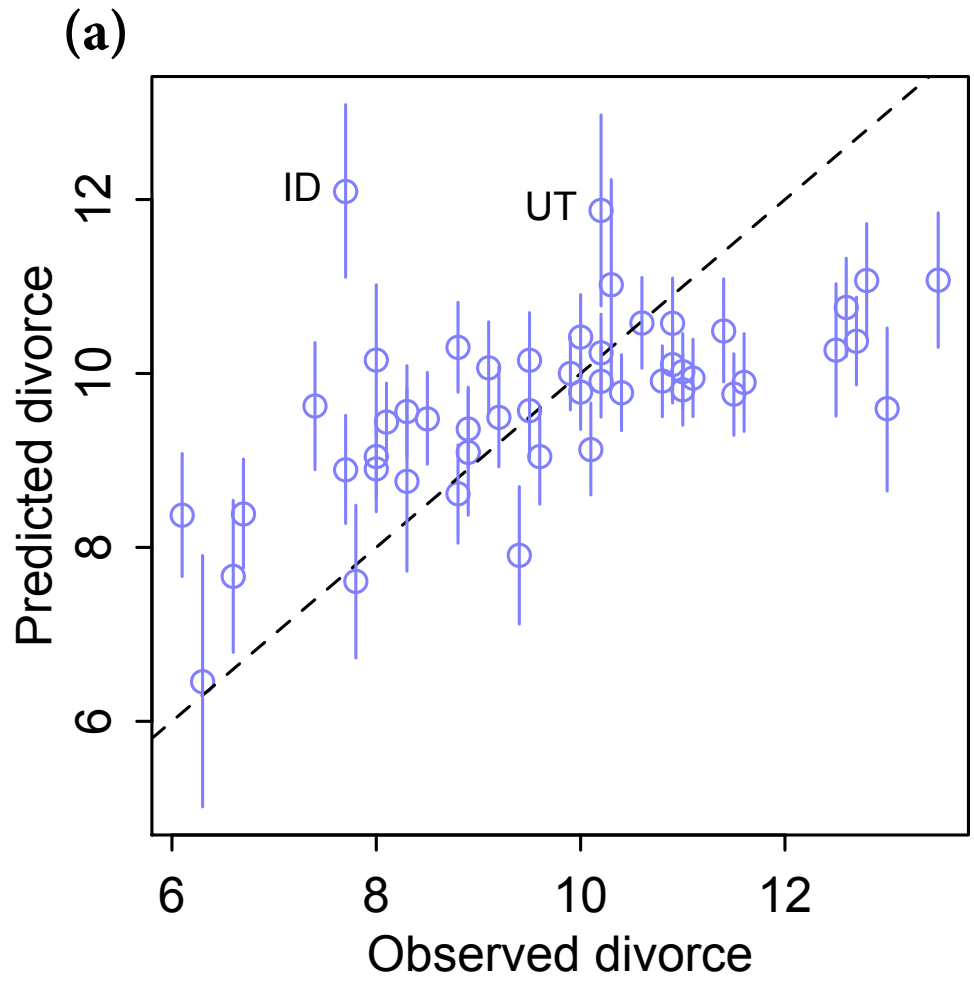

Predicted compared to observed
(a)
ID
UT
Predicted divorce
Observed divorce
6
8
10
12

Figure 5.6