STA 610L: Module 4.1

Measurement error

Dr. Olanrewaju Michael Akande

Measurement Error

We will switch gears a bit and see how to use the framework of hierarchical to account for measurement error.

Measurement error is the difference between a measured quantity and its true value.

It can be due to

- systematic bias (e.g., a scale is mis-calibrated by 1 pound for everyone)
- random error (e.g., some people take off their shoes, others are wearing coats, some may be dehydrated or have just eaten) that may be naturally occurring and may occur with any experiment.

Measurement error is often countered by tactics like taking the mean of multiple measurements or standardizing experimental conditions.

However, sometimes substantial sources of error are unavoidable.

Example: divorce and marriage rates

McElreath (2016) considers the relationship among divorce rate, marriage rate, and median age at marriage based on state-level data.

A good chunk of the code presented here follows directly from Section 14 of Statistical Rethinking with brms, ggplot2, and the tidyverse.

The material goes through this example in more detail, so you should definitely read it carefully.

Example: divorce and marriage rates

```
#devtools::install_github("wmurphyrd/fiftystater")
#library(fiftystater); #library(rethinking); #library(tidyverse)
data(WaffleDivorce)
d <- WaffleDivorce
rm(WaffleDivorce)
dim(d)
```

```
## [1] 50 13
```

```
head(d)
```

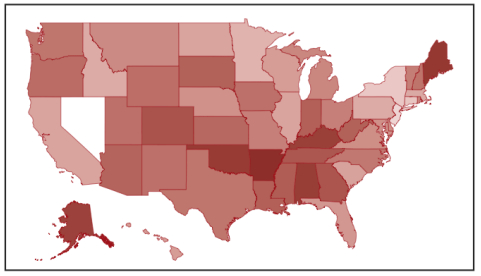
```
##     Location Loc Population MedianAgeMarriage Marriage Marriage.SE Divorce
## 1    Alabama  AL       4.78              25.3     20.2        1.27    12.7
## 2     Alaska  AK       0.71              25.2     26.0        2.93    12.5
## 3    Arizona  AZ       6.33              25.8     20.3        0.98    10.8
## 4   Arkansas  AR       2.92              24.3     26.4        1.70    13.5
## 5 California  CA      37.25              26.8     19.1        0.39     8.0
## 6   Colorado  CO       5.03              25.7     23.5        1.24    11.6
##   Divorce.SE WaffleHouses South Slaves1860 Population1860 PropSlaves1860
## 1       0.79          128     1     435080         964201           0.45
## 2       2.05            0     0          0              0           0.00
## 3       0.74           18     0          0              0           0.00
## 4       1.22           41     1     111115         435450           0.26
## 5       0.24            0     0          0         379994           0.00
## 6       0.94           11     0          0          34277           0.00
```

Example: divorce and marriage rates

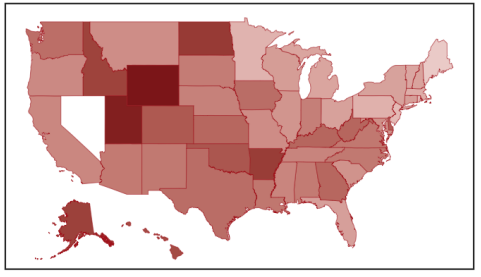
```
d %>%
  # first we'll standardize the three variables to put them all on the same scale
  mutate(Divorce_z = (Divorce - mean(Divorce)) / sd(Divorce),
         MedianAgeMarriage_z = (MedianAgeMarriage -
                                  mean(MedianAgeMarriage)) /sd(MedianAgeMarriage),
         Marriage_z = (Marriage - mean(Marriage))  / sd(Marriage),
         # need to make the state names lowercase to match with the map data
         Location            = str_to_lower(Location)) %>%
  # here we select the relevant variables and put them in the long format to facet with `f
  dplyr::select(Divorce_z:Marriage_z, Location) %>%
  gather(key, value, -Location) %>%
  ggplot(aes(map_id = Location)) +
  geom_map(aes(fill = value), map = fifty_states,
           color = "firebrick", size = 1/15) +
  expand_limits(x = fifty_states$long, y = fifty_states$lat) +
  scale_x_continuous(NULL, breaks = NULL) +
  scale_y_continuous(NULL, breaks = NULL) +
  scale_fill_gradient(low = "#f8eaea", high = "firebrick4") +
  coord_map() +
  theme_bw() +
  theme(panel.grid       = element_blank(),
        legend.position  = "none",
        strip.background = element_rect(fill = "transparent", color = "transparent")) +
  facet_wrap(~key)
```

Example: divorce and marriage rates

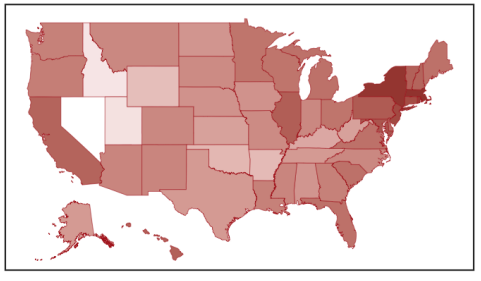
Divorce_z



Marriage_z



MedianAgeMarriage_z



Divorce and marriage rates

Note that data from Nevada are not included.

Is divorce associated with marriage? Well.....yes!

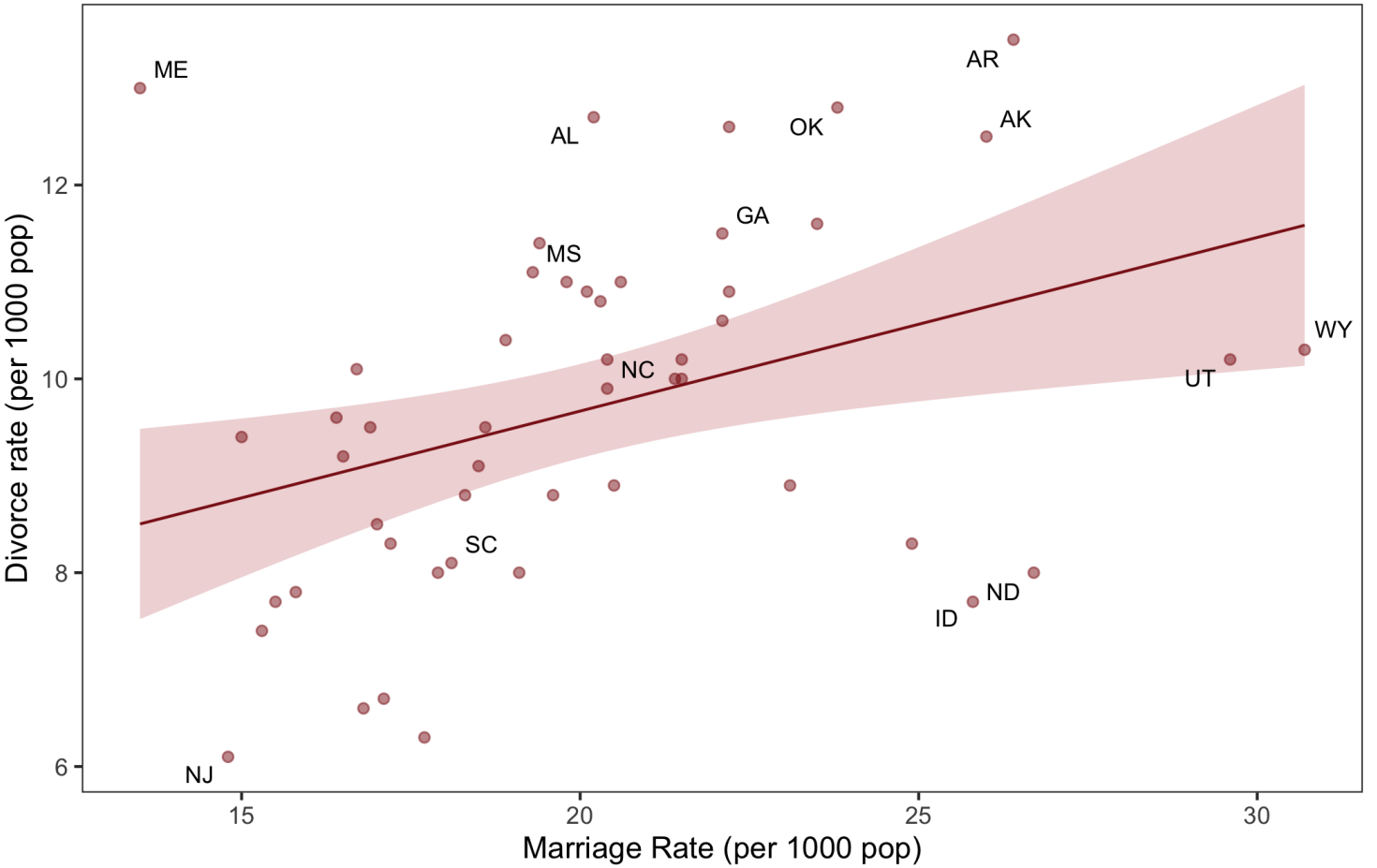
However, does a high marriage rate imply a high divorce rate?

How does median age at marriage affect divorce rates?

Divorce and marriage rates

```
#library(ggrepel)
d %>%
  ggplot(aes(x = Marriage, y = Divorce)) +
  stat_smooth(method = "lm", fullrange = T, size = 1/2,
              color = "firebrick4", fill = "firebrick", alpha = 1/5) +
  geom_point(size = 1.5, color = "firebrick4", alpha = 1/2) +
  geom_text_repel(data = d %>%
                    filter(Loc %in% c("ME", "OK", "AR", "AL", "GA","SC", "NJ",
                                      "NC","MS","UT","WY","AK","ID","ND")),
                  aes(label = Loc),
                  size = 3, seed = 1042) +  # this makes it reproducible
  xlab("Marriage Rate (per 1000 pop)")+
  ylab("Divorce rate (per 1000 pop)") +
  theme_bw() +
  theme(panel.grid = element_blank())
```

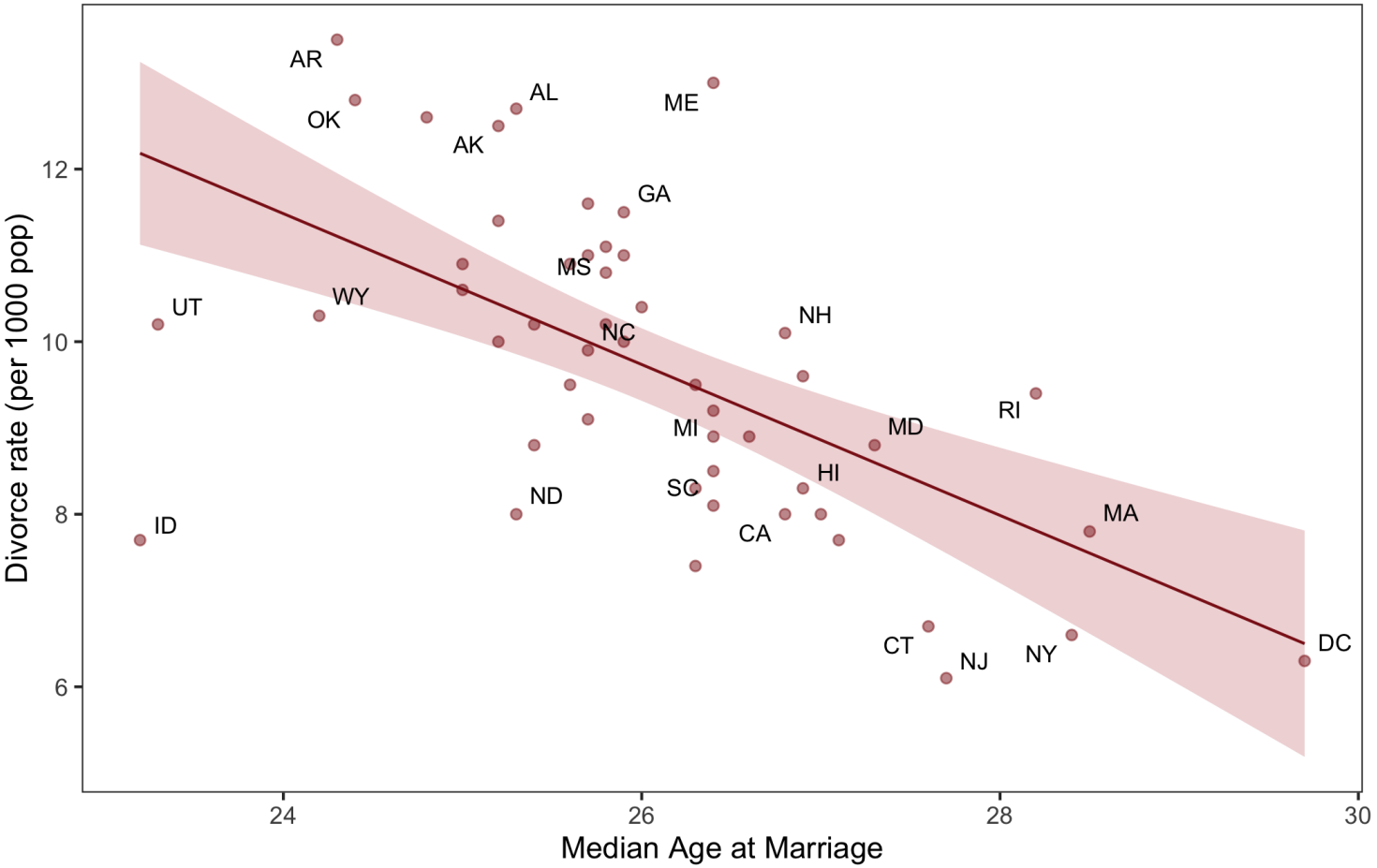
Divorce and marriage rates



Divorce rates vs median ages at marriage

```
d %>%
  ggplot(aes(x = MedianAgeMarriage, y = Divorce)) +
  stat_smooth(method = "lm", fullrange = T, size = 1/2,
              color = "firebrick4", fill = "firebrick", alpha = 1/5) +
  geom_point(size = 1.5, color = "firebrick4", alpha = 1/2) +
  geom_text_repel(data = d %>% filter(Loc %in% c("ME", "OK", "AR", "AL", "GA", "SC", "NJ",
                  aes(label = Loc),
                  size = 3, seed = 1042) +  # this makes it reproducible
  xlab("Median Age at Marriage")+
  ylab("Divorce rate (per 1000 pop)") +
  theme_bw() +
  theme(panel.grid = element_blank())
```

Divorce rates vs median ages at marriage

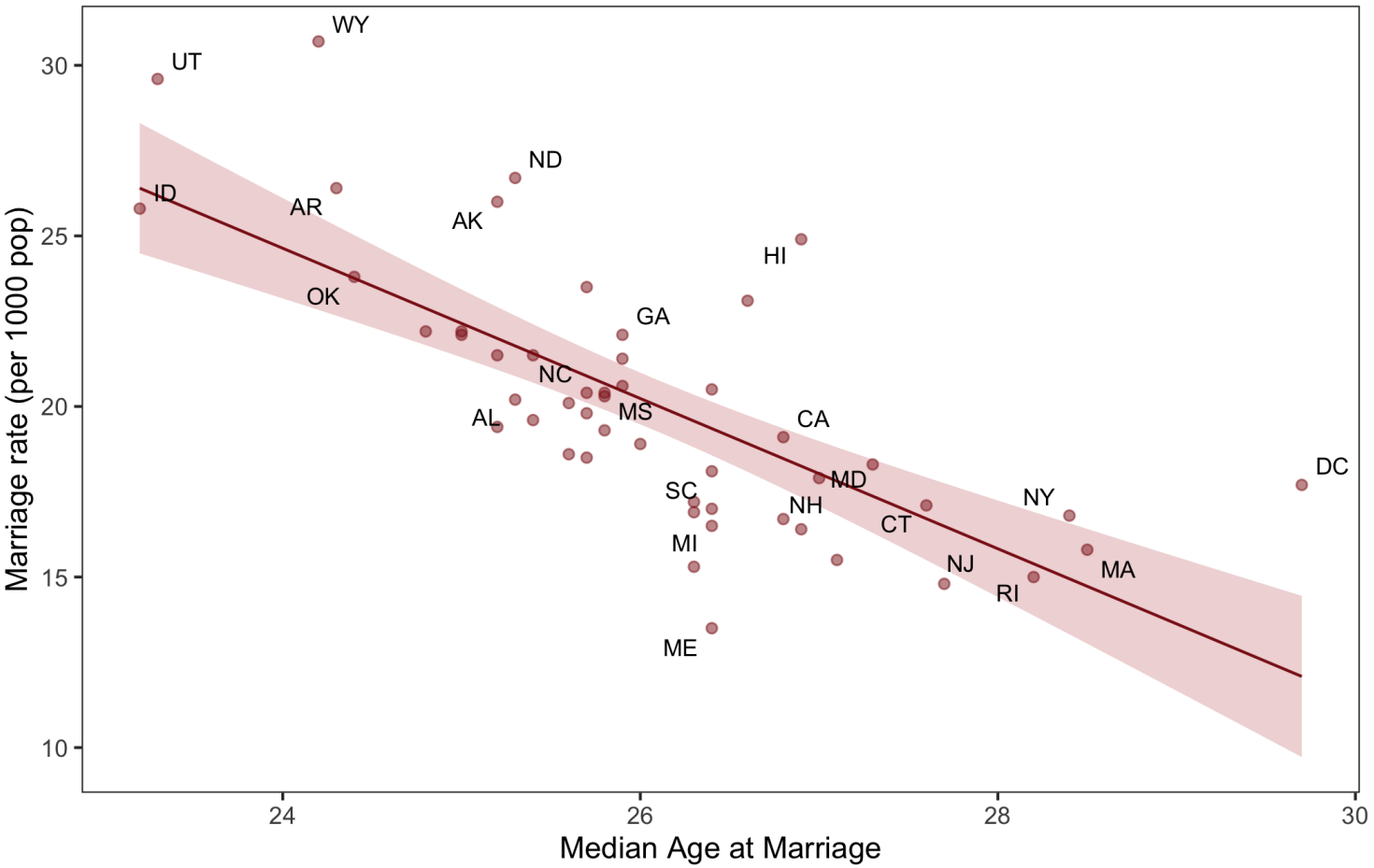


STA 610L

Marriage rates vs median ages at marriage

```
d %>%
  ggplot(aes(x = MedianAgeMarriage, y = Marriage)) +
  stat_smooth(method = "lm", fullrange = T, size = 1/2,
              color = "firebrick4", fill = "firebrick", alpha = 1/5) +
  geom_point(size = 1.5, color = "firebrick4", alpha = 1/2) +
  geom_text_repel(data = d %>% filter(Loc %in% c("ME", "OK", "AR", "AL", "GA", "SC", "NJ",
                  aes(label = Loc),
                  size = 3, seed = 1042) +  # this makes it reproducible
  xlab("Median Age at Marriage")+
  ylab("Marriage rate (per 1000 pop)") +
  theme_bw() +
  theme(panel.grid = element_blank())
```

Marriage rates vs median ages at marriage



STA 610L

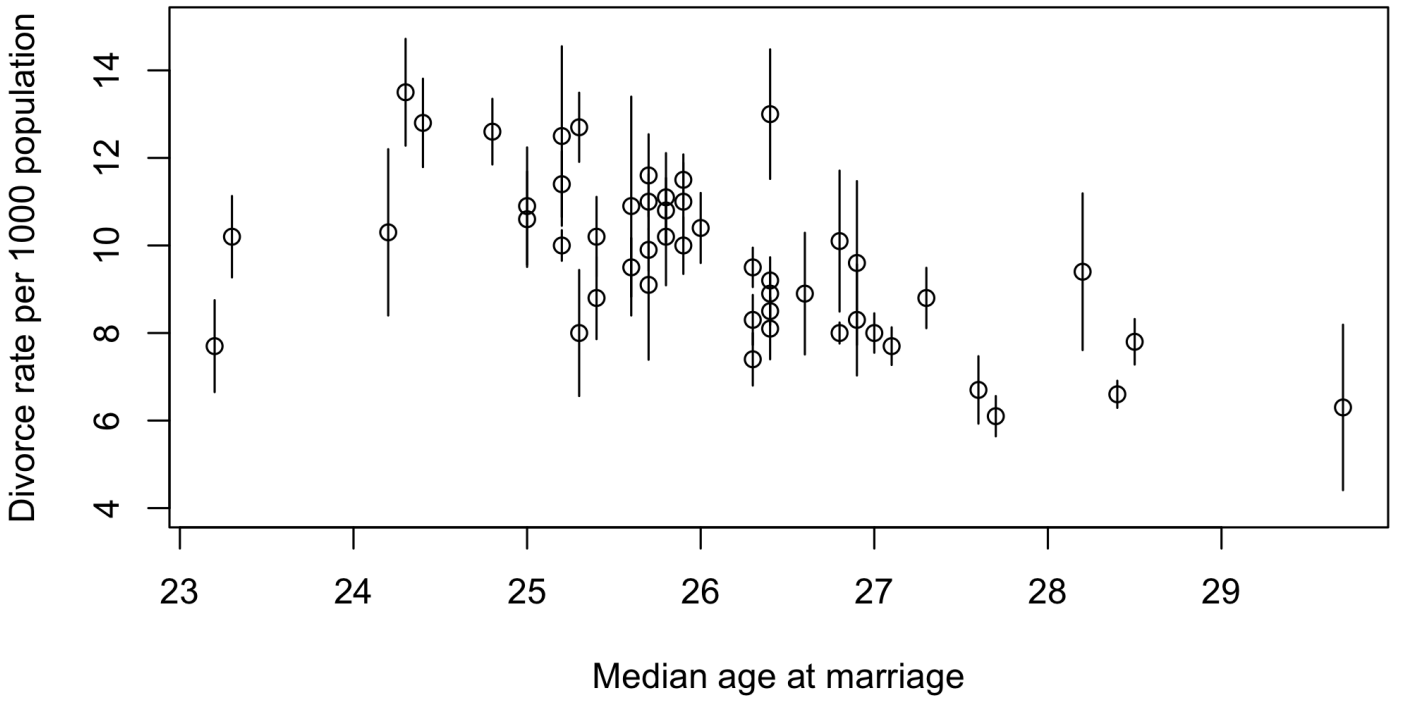
Divorce and marriage rates

One issue analyzing these data is that we have error involved in the measurement of both marriage rate and divorce rate.

First, we'll explore measurement error of our outcome, divorce rate.

```
plot(d$Divorce~d$MedianAgeMarriage,ylim=c(4,15),
     xlab="Median age at marriage",ylab="Divorce rate per 1000 population")
#add interval of 1 SE in each direction
for (i in 1:nrow(d)) {
  ci <- d$Divorce[i]+c(-1,1)*d$Divorce.SE[i]
  x <- d$MedianAgeMarriage[i]
  lines(c(x,x),ci)
}
```

Divorce and marriage rates



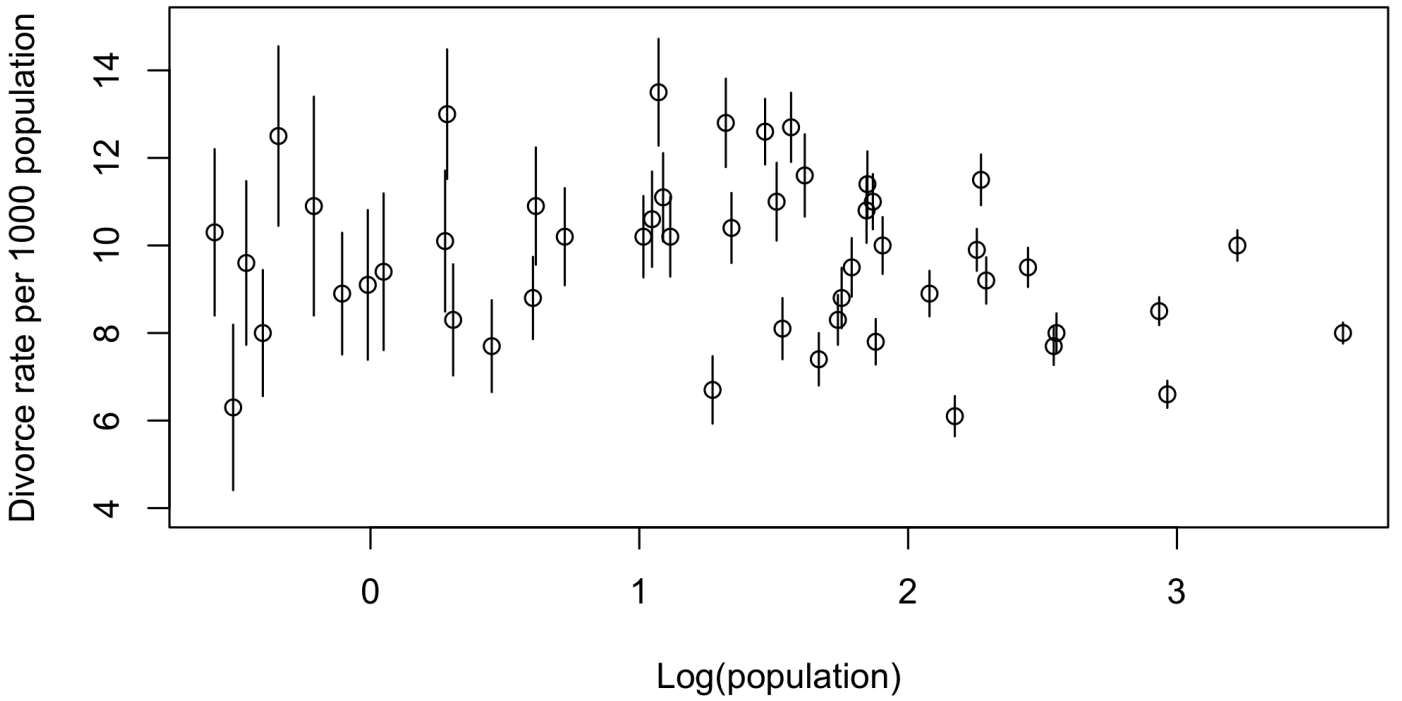
There is substantial variability in the certainty in the estimated divorce rates. Why?

Divorce and marriage rates

A hunch is that the size of the state's population may be involved.

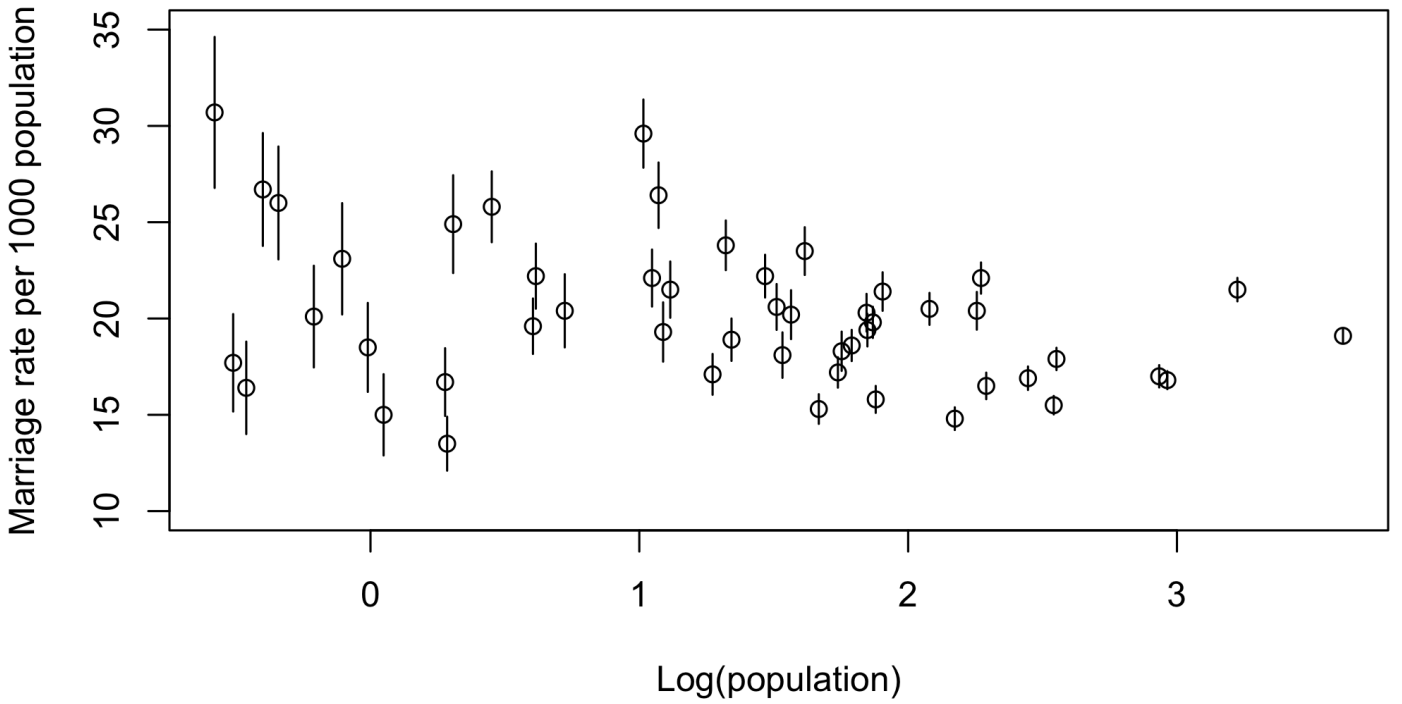
```
plot(d$Divorce~log(d$Population),ylim=c(4,15),
     xlab="Log(population)",ylab="Divorce rate per 1000 population")
#add interval of 1 SE in each direction
for (i in 1:nrow(d)) {
  ci <- d$Divorce[i]+c(-1,1)*d$Divorce.SE[i]
  x <- log(d$Population[i])
  lines(c(x,x),ci)
}
```

Divorce and marriage rates



Yes, there is a relationship between population size and certainty in the estimated rate!

Divorce and marriage rates



We also see this in marriage rates!

Handling measurement error

First, we focus on measurement error in our response, the divorce rate.

One reasonable approach is to use a hierarchical model.

Generally, the hierarchical model would include

- a model for the true unobserved/latent responses (conditional on the predictors); and
- a model that describes how the reported/observed responses are generated from the true responses.

Handling measurement error

For this example, we could do the following

- Define the parameter $D_{TRUE,i}$ to be the true (unknown) divorce rate for state i
- Define our observed outcome (subject to measurement error) as $D_{OBS,i}$ and its associated standard error (provided in the data) as $D_{SE,i}$
- Model $D_{OBS,i} \sim N\left(D_{TRUE,i},\ D^2_{SE,i}\right)$
- Here the observed divorce rates are centered on the true rates with the estimated measurement error treated as known (if unknown, treat as another parameter to be estimated).
- Define the covariates: let A_i be the median age at marriage and R_i be the marriage rate R_i.

Model

Now we can specify our desired model, for the true divorce rates, as follows.

$$
\begin{aligned}
D_{OBS,i} &\sim N\left(D_{TRUE,i}, D_{SE,i}^2\right) \\
D_{TRUE,i} &\sim N(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 A_i + \beta_2 R_i \\
\beta_0, \beta_1, \beta_2 &\sim N(0, 100) \\
\sigma &\sim \text{HalfCauchy}(0, 2.5)
\end{aligned}
$$

Model

First, we fit the model with no adjustment for measurement error, so that the outcome is just the observed (with error) divorce rate.

```
#library(brms)
#put data into a list
dlist <- list(
    div_obs = d$Divorce,
    div_sd  = d$Divorce.SE,
    R       = d$Marriage,
    A       = d$MedianAgeMarriage - mean(d$MedianAgeMarriage))

m1 <-
  brm(data = dlist, family = gaussian,
      #brm mean-centers by default when an intercept is included, which is OK!
      #however, if for some reason you prefer not to mean-center
      #but still want an intercept,use the command below
      div_obs ~ 0 + Intercept + R + A,
      prior = c(prior(normal(0, 10),class=b,coef=Intercept),
                prior(normal(0, 10), class = b),
                prior(cauchy(0, 2.5), class = sigma)),
      iter = 5000, warmup = 1000, chains = 4, cores = 4,
      seed = 14,control=list(adapt_delta=0.95))
```

Model

```
m1
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: div_obs ~ 0 + Intercept + R + A
##    Data: dlist (Number of observations: 50)
##   Draws: 4 chains, each with iter = 5000; warmup = 1000; thin = 1;
##          total post-warmup draws = 16000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    10.54      1.64     7.29    13.73 1.00     4357     5811
## R            -0.04      0.08    -0.20     0.12 1.00     4372     5765
## A            -0.97      0.25    -1.45    -0.48 1.00     5221     6730
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     1.52      0.16     1.24     1.87 1.00     7031     7607
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

While marriage rate is not significantly associated with divorce rate, conditional on median age at marriage, conditional on the marriage rate, a one-year higher median age at marriage is associated with an expected 0.97 fewer divorces per 1000 population (with 95% CI=(0.48,1.45)).

However, we may be concerned because of the error in our outcome.

Accounting for measurement error in response

```
# here we specify the initial (i.e., starting) values
inits      <- list(Yl = dlist$div_obs)
inits_list <- list(inits, inits)

m2 <-
  brm(data = dlist, family = gaussian,
      div_obs | mi(div_sd) ~ 0 + Intercept + R + A,
      prior = c(prior(normal(0, 10), class = b),
                prior(cauchy(0, 2.5), class = sigma)),
      iter = 5000, warmup = 1000, cores = 2, chains = 2,
      seed = 14,
      control = list(adapt_delta = 0.99,max_treedepth = 12),
      save_pars = save_pars(latent=TRUE),  # note this line for the `mi()` model
      inits = inits_list)
```

Results

```
m2
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: div_obs | mi(div_sd) ~ 0 + Intercept + R + A
##    Data: dlist (Number of observations: 50)
##   Draws: 2 chains, each with iter = 5000; warmup = 1000; thin = 1;
##          total post-warmup draws = 8000
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     9.26      1.71     5.97    12.64 1.00     2929     3929
## R             0.01      0.09    -0.16     0.18 1.00     2896     4123
## A            -0.97      0.25    -1.47    -0.47 1.00     3525     4987
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     1.07      0.19     0.72     1.49 1.00     2318     2874
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

The interpretation of this model is similar to what we saw before, though our estimate of σ is now lower (for comparison, it was 1.52).

Accounting for measurement error in predictor

Measurement error in the predictor, here marriage rate, can have an effect on estimation as well.

Here we allow the marriage rate to be measured with error as well by fitting the following model.

$$
\begin{aligned}
D_{OBS,i} &\sim N\left(D_{TRUE,i}, D_{SE,i}^2\right) \\
R_{OBS,i} &\sim N\left(R_{TRUE,i}, R_{SE,i}^2\right) \\
D_{TRUE,i} &\sim N(\mu_i, \sigma^2) \\
\mu_i &= \beta_0 + \beta_1 A_i + \beta_2 R_{TRUE,i} \\
\beta_0, \beta_1, \beta_2 &\sim N(0, 100) \\
\sigma &\sim \text{HalfCauchy}(0, 2.5)
\end{aligned}
$$

Accounting for measurement error in predictor

```
dlist <- list(
  div_obs = d$Divorce,
  div_sd  = d$Divorce.SE,
  mar_obs = d$Marriage,
  mar_sd  = d$Marriage.SE,
  A       = d$MedianAgeMarriage)
# the `inits`
inits      <- list(Yl = dlist$div_obs)
inits_list <- list(inits, inits)
# the model
m3 <-
  brm(data = dlist, family = gaussian,
      div_obs | mi(div_sd) ~ 0 + Intercept + me(mar_obs, mar_sd) + A,
      prior = c(prior(normal(0, 10), class = b),
                prior(cauchy(0, 2.5), class = sigma)),
      iter = 5000, warmup = 1000, cores = 2, chains = 2,
      seed = 1235,
      control = list(adapt_delta = 0.99,
                     max_treedepth = 12),
      save_pars = save_pars(latent=TRUE),
      inits = inits_list)
```

Results

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: div_obs | mi(div_sd) ~ 0 + Intercept + me(mar_obs, mar_sd) + A
##    Data: dlist (Number of observations: 50)
##   Draws: 2 chains, each with iter = 5000; warmup = 1000; thin = 1;
##          total post-warmup draws = 8000
##
## Population-Level Effects:
##                 Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept          15.53      6.78     2.07    28.66 1.00     2590     3768
## A                  -0.44      0.20    -0.83    -0.02 1.00     2914     4290
## memar_obsmar_sd     0.27      0.11     0.07     0.48 1.00     2431     4003
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     1.00      0.21     0.61     1.44 1.00     1793     2174
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Results

Now that we've accounted for measurement error in the exposure and outcome, we see substantial changes in effect estimates.

The interpretation of this model is that conditional on the marriage rate, a one-year higher median age at marriage is associated with an expected 0.44 fewer divorces per 1000 population (95% CI=(0.02,0.83)).

Conditional on the median age at marriage, an increase of the marriage rate by 1 per 1000 is associated with an expected increase in the divorce rate of 0.27 per 1000 (95% CI=(0.07, 0.48)).

Moral of the story: The moral of this story is that when you have error associated with a predictor or response (i.e., a distribution of responses), reducing the response to a single value -- discarding uncertainty -- can lead to spurious inference.

What's next?

Move on to the readings for the next module!