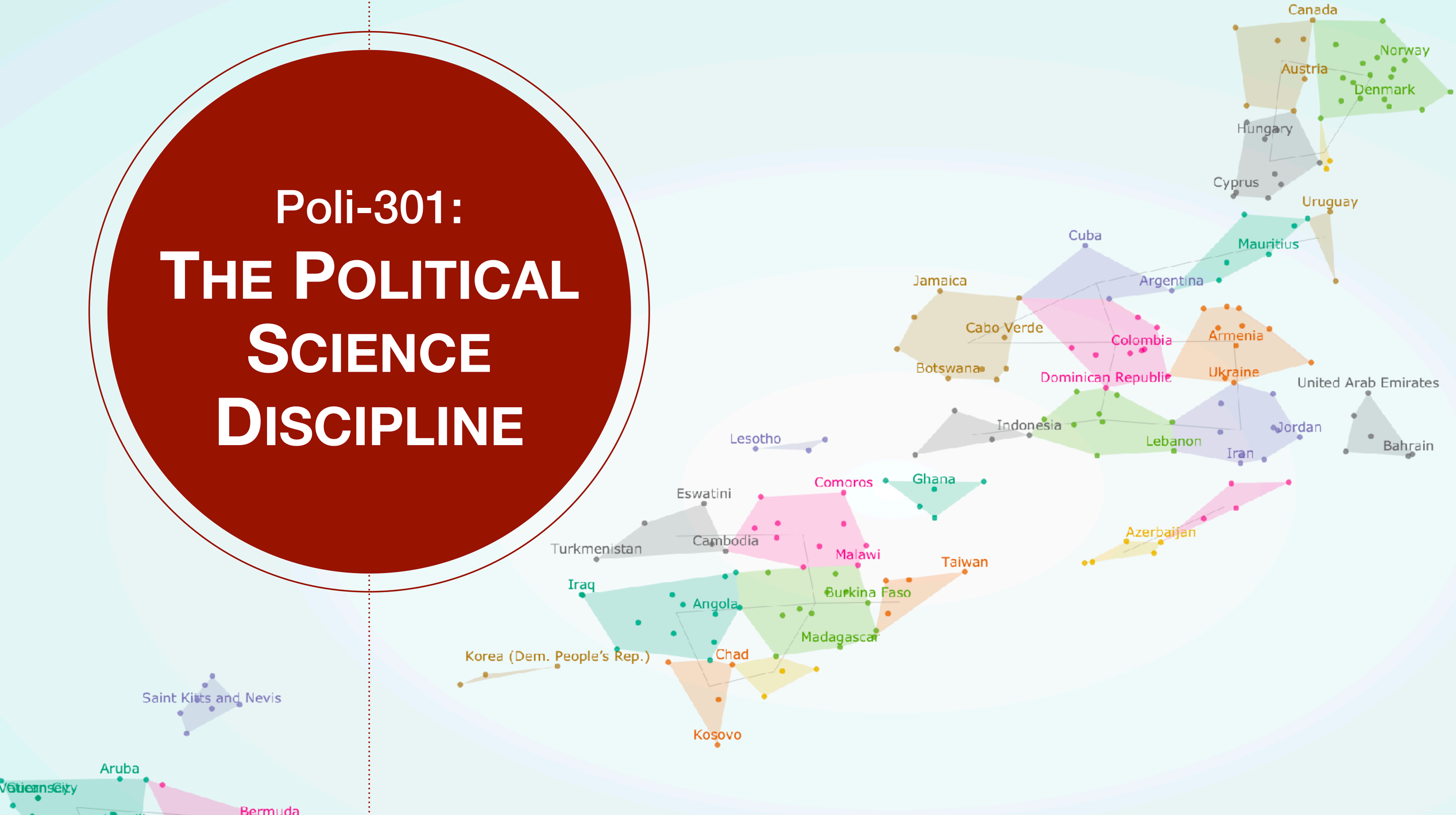
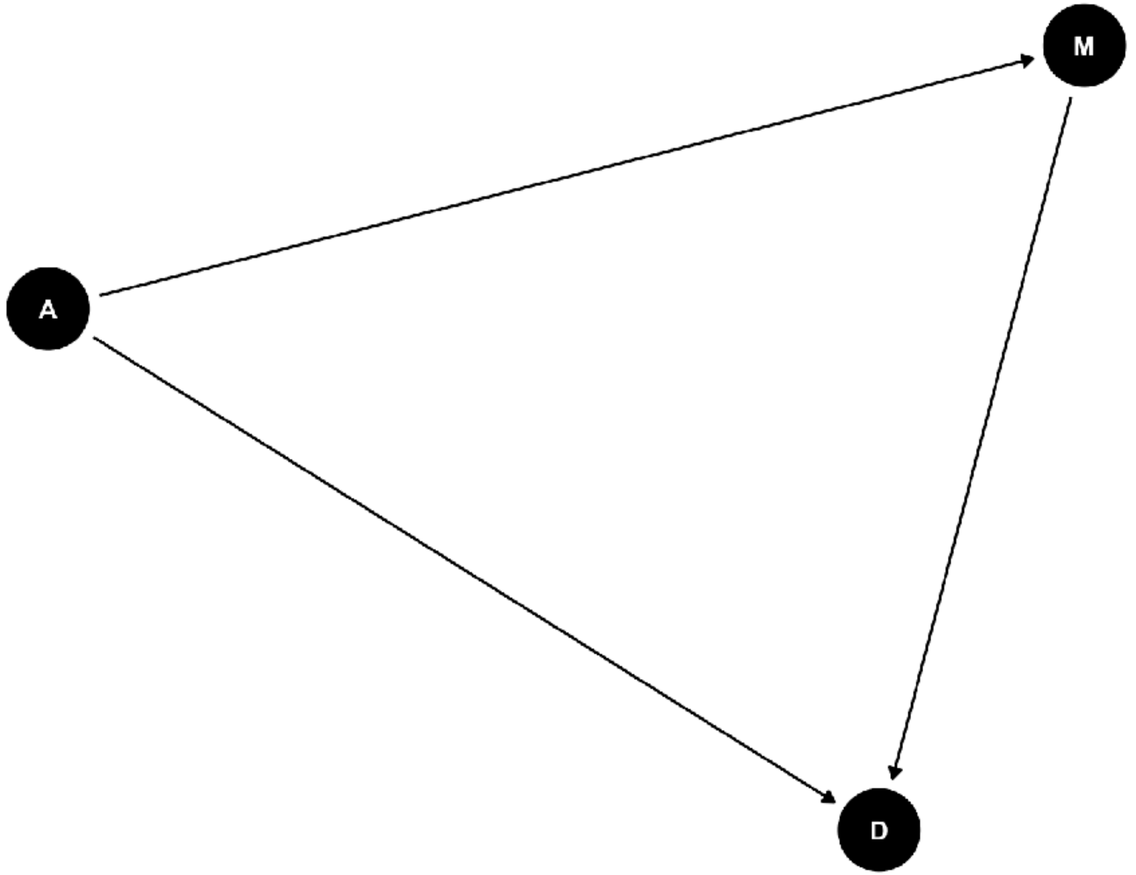
Poli-301:
The Political Science Discipline



MARRIAGE + DIVORCE (AGAIN)

	Loc	Marriage	MedianAgeMarriage	Divorce
1	AL	20.2	25.3	12.7
2	AK	26.0	25.2	12.5
3	AZ	20.3	25.8	10.8
4	AR	26.4	24.3	13.5
5	CA	19.1	26.8	8.0
6	CO	23.5	25.7	11.6
7	CT	17.1	27.6	6.7
8	DE	23.1	26.6	8.9
9	DC	17.7	29.7	6.3
10	FL	17.0	26.4	8.5
11	GA	22.1	25.9	11.5
12	HI	24.9	26.9	8.3
13	ID	25.8	23.2	7.7
14	IL	17.9	27.0	8.0
15	IN	19.8	25.7	11.0
16	IA	21.5	25.4	10.2



What's the effect of M on D, controlling for A?

MULTIPLE REGRESSION

$$Divorce = \alpha + \beta_1 Marriage + \beta_2 MedianAge + \epsilon$$

```r
# base model
m1 = lm(Divorce ~ Marriage, data = WaffleDivorce)
summary(m1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.08404    1.31337   4.632 2.78e-05 ***
Marriage     0.17918    0.06418   2.792  0.00751 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```r
# control for age at marriage
m2 = lm(Divorce ~ Marriage + MedianAgeMarriage,
        data = WaffleDivorce)
summary(m2)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       36.87665    7.66104   4.814 1.58e-05 ***
Marriage          -0.05686    0.08053  -0.706 0.483594
MedianAgeMarriage -0.99965    0.24593  -4.065 0.000182 ***
---
```

“Controlling for Median Age”/“all else equal”/“all other variables held constant”, a one-unit increase in Marriage decreases Divorce by -.06

EXPLAINING WITH VARIABLES

To understand multiple regression we need to know what it means to use one variable to *explain* another variable

"How does X explain Y" =
"what would I expect Y to look like given a specific value of X"

"How much of the variation in Divorce is explained by Marriage? How much is **not** explained?"

POVERTY IN MIDWEST

```
# A tibble: 437 x 3
   state county    percbelowpoverty
   <chr> <chr>                <dbl>
 1 IL    ADAMS                13.2
 2 IL    ALEXANDER            32.2
 3 IL    BOND                 12.1
 4 IL    BOONE                 7.21
 5 IL    BROWN                13.5
 6 IL    BUREAU               10.4
 7 IL    CALHOUN              15.1
 8 IL    CARROLL              11.7
 9 IL    CASS                 13.9
10 IL    CHAMPAIGN            15.6
```

STATE AS EXPLANATION

How much of the poverty rate can be **explained** by what state someone is in?

```
# let's get average rate by state
midwest %>% group_by(state) %>%
  # add column that equals average poverty rate in state
  mutate(avebystate = mean(percbelowpoverty)) %>%
  select(state, county, percbelowpoverty, avebystate) %>%
  # calculate residual
  mutate(residual = percbelowpoverty - avebystate)
```

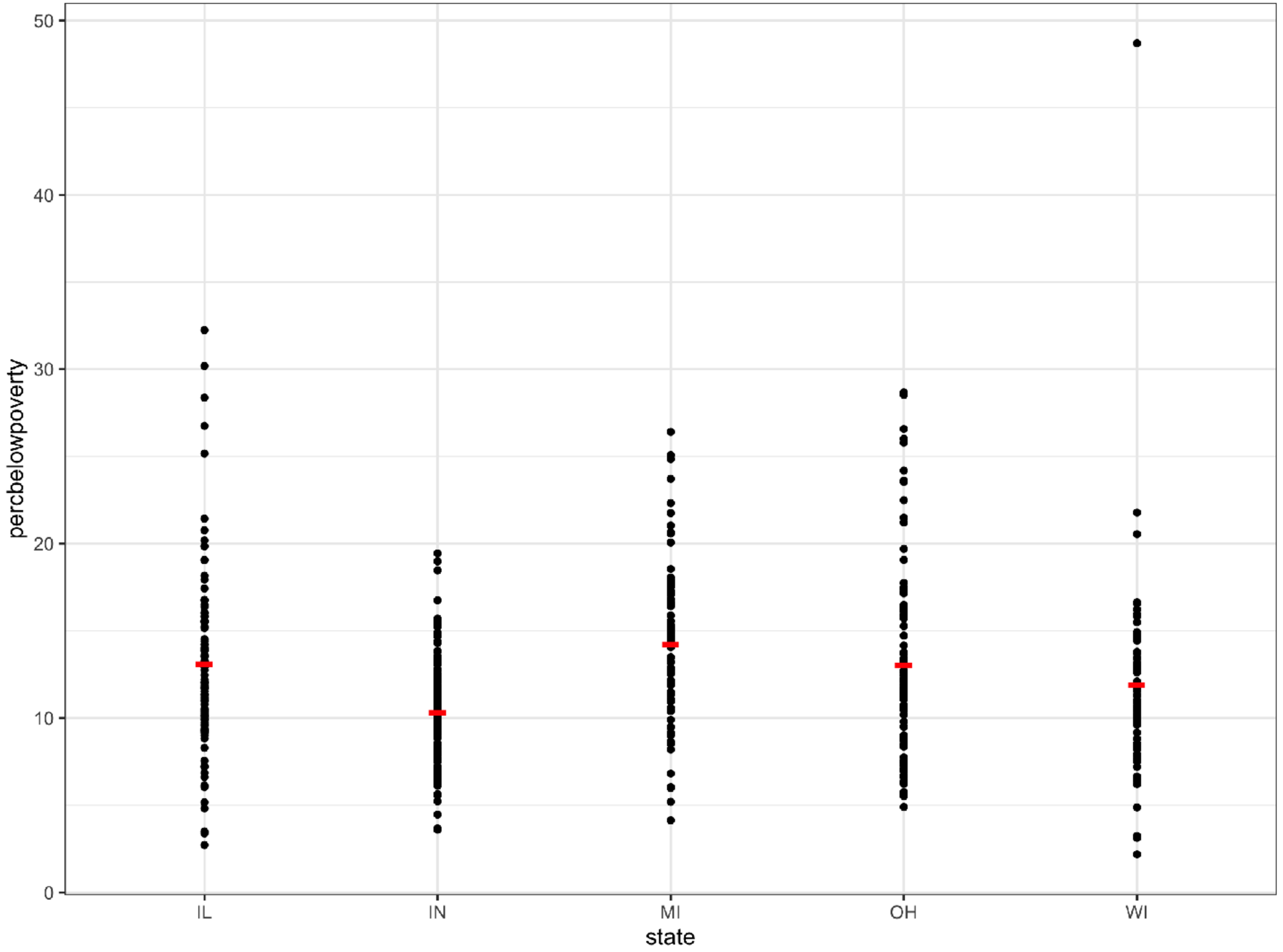
```
# A tibble: 437 x 5
# Groups:   state [5]
   state county    percbelowpoverty avebystate residual
   <chr> <chr>                <dbl>      <dbl>    <dbl>
 1 IL    ADAMS                13.2        13.1   0.0750
 2 IL    ALEXANDER            32.2        13.1  19.2
 3 IL    BOND                 12.1        13.1  -1.01
 4 IL    BOONE                 7.21       13.1  -5.87
 5 IL    BROWN                13.5        13.1   0.444
 6 IL    BUREAU               10.4        13.1  -2.68
 7 IL    CALHOUN              15.1        13.1   2.07
 8 IL    CARROLL              11.7        13.1  -1.37
 9 IL    CASS                 13.9        13.1   0.799
10 IL    CHAMPAIGN            15.6        13.1   2.50
```

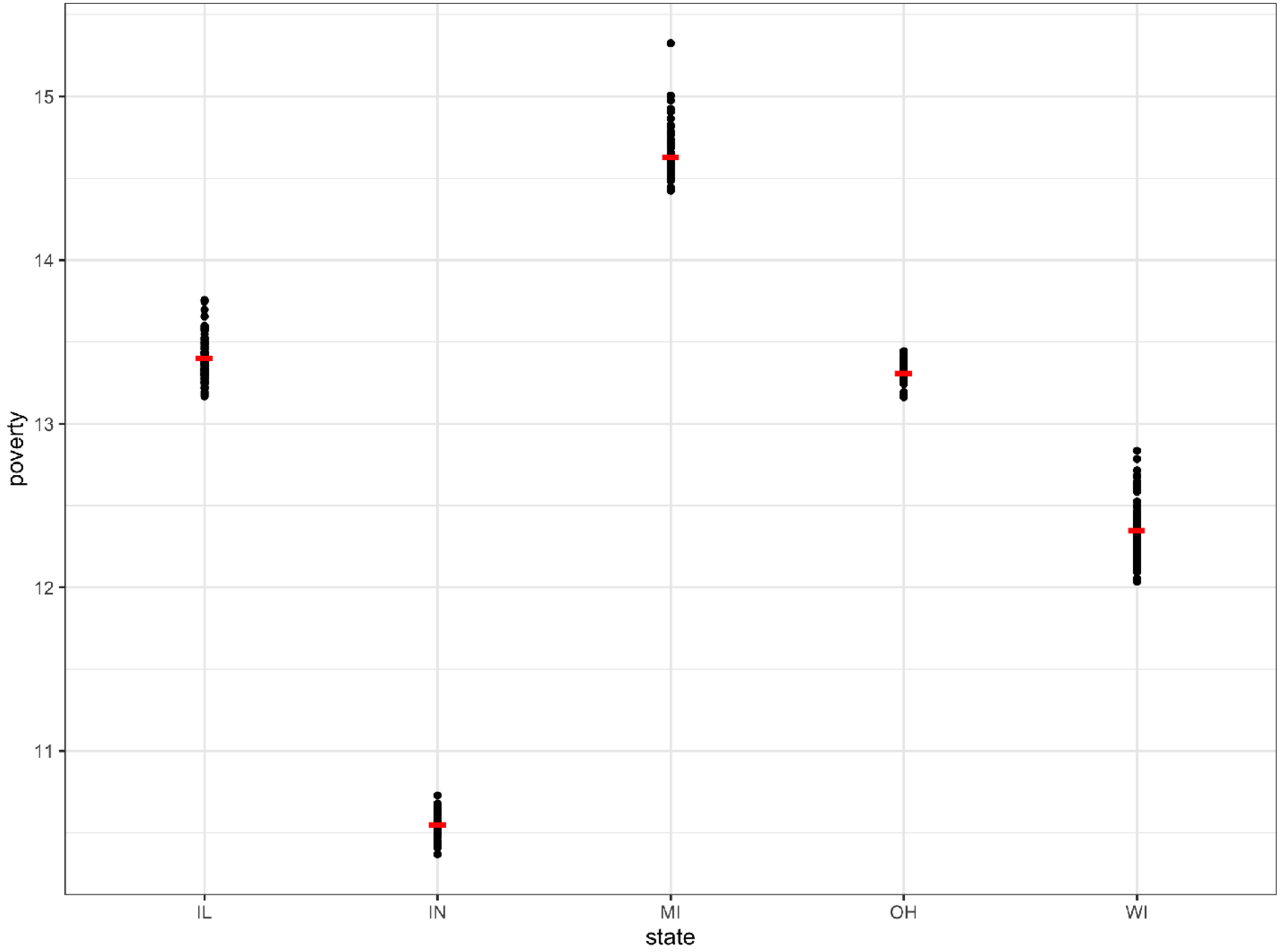
RESIDUALS

Residual =
part of poverty-by-county **not explained** by what state you're in (what's "left over")

The **smaller** the residuals, the **better** state explains poverty

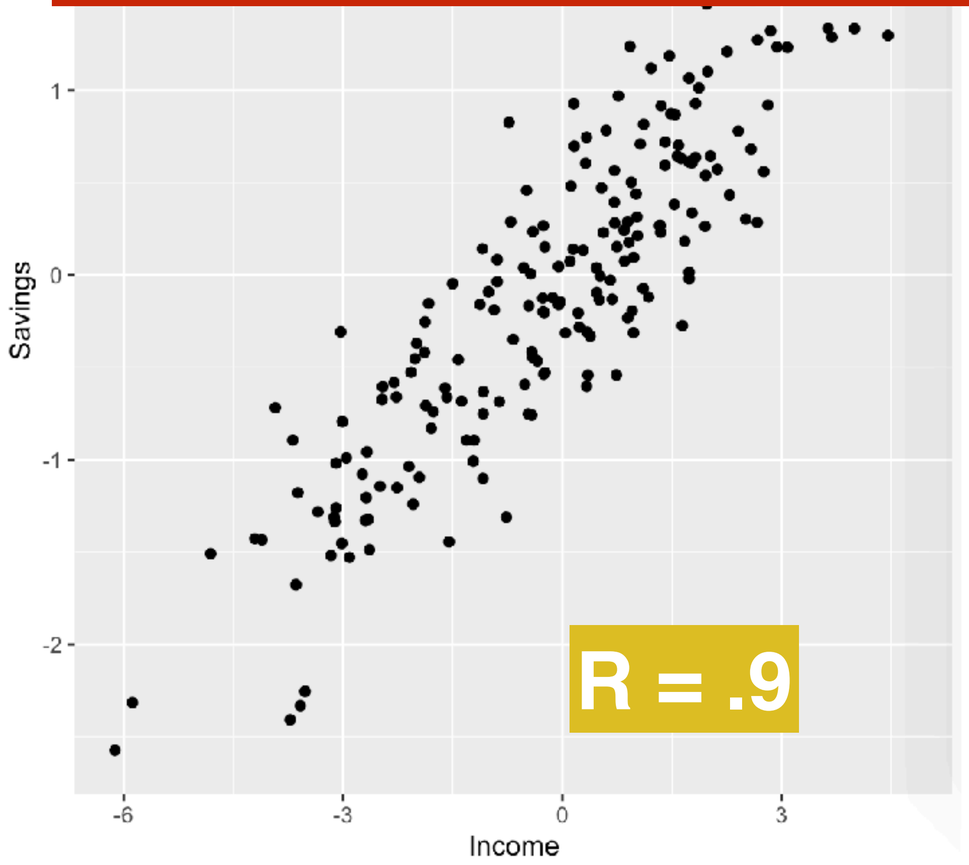
```
# A tibble: 437 x 5
# Groups:   state [5]
   state county    percbelowpoverty avebystate residual
   <chr> <chr>                <dbl>      <dbl>    <dbl>
 1 IL    ADAMS                 13.2       13.1   0.0750
 2 IL    ALEXANDER             32.2       13.1  19.2
 3 IL    BOND                  12.1       13.1  -1.01
 4 IL    BOONE                  7.21      13.1  -5.87
 5 IL    BROWN                 13.5       13.1   0.444
 6 IL    BUREAU                10.4       13.1  -2.68
 7 IL    CALHOUN               15.1       13.1   2.07
 8 IL    CARROLL               11.7       13.1  -1.37
 9 IL    CASS                  13.9       13.1   0.799
10 IL    CHAMPAIGN             15.6       13.1   2.50
```

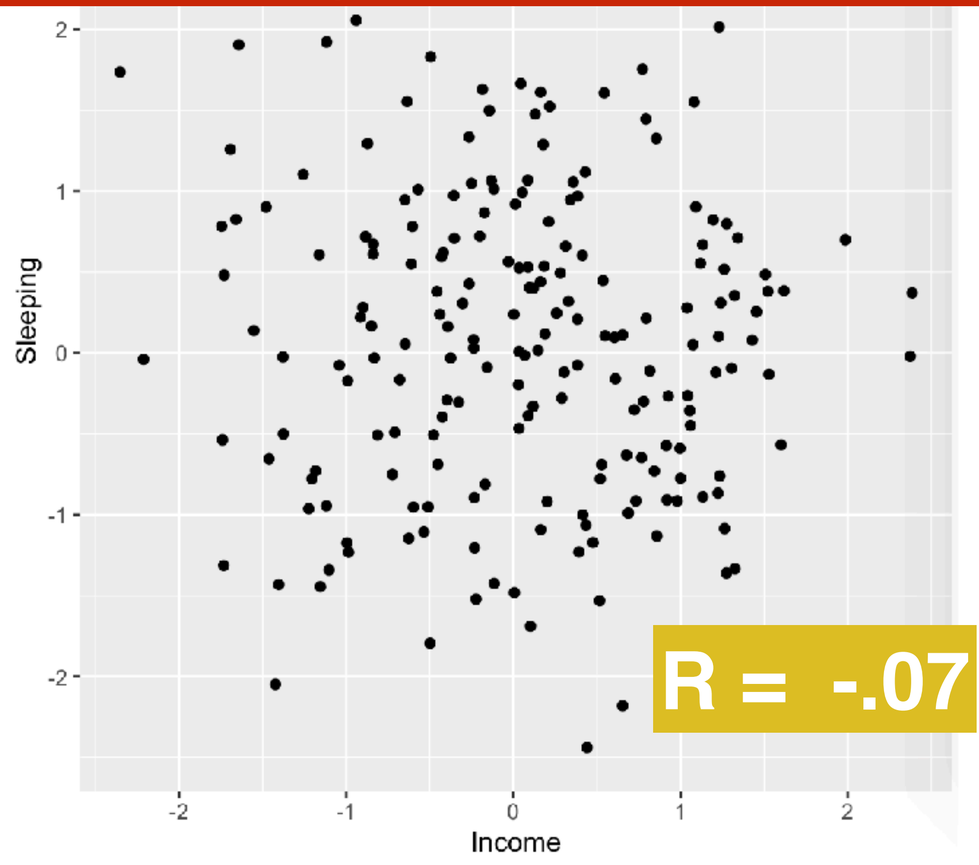

50
40
30
20
10
0
percbelowpoverty
IL
IN
MI
OH
WI
state


15
14
13
12
11
poverty
IL
IN
MI
OH
WI
state

Relationships between variables

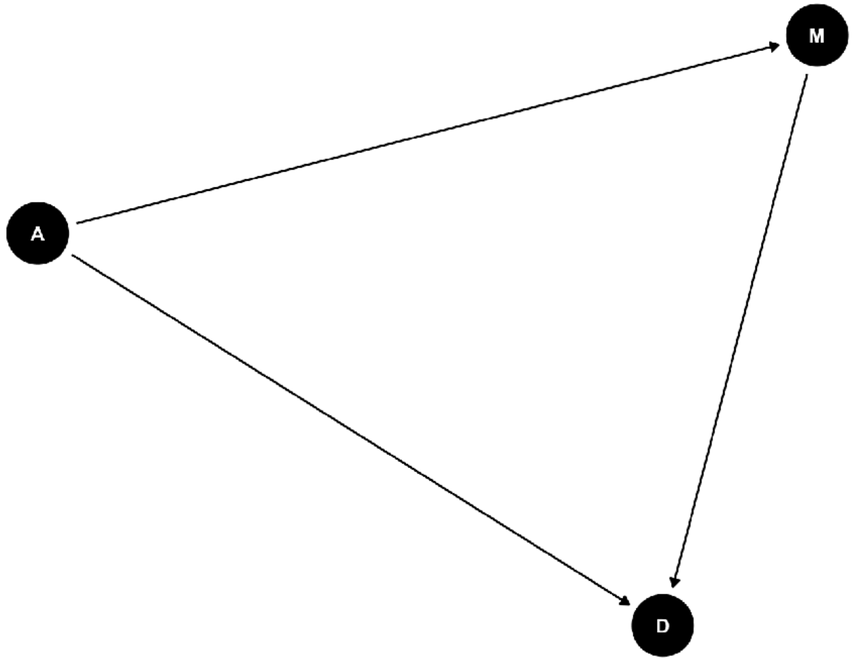
The stronger the relationship between X and Y, the less of Y there is “left over” to explain





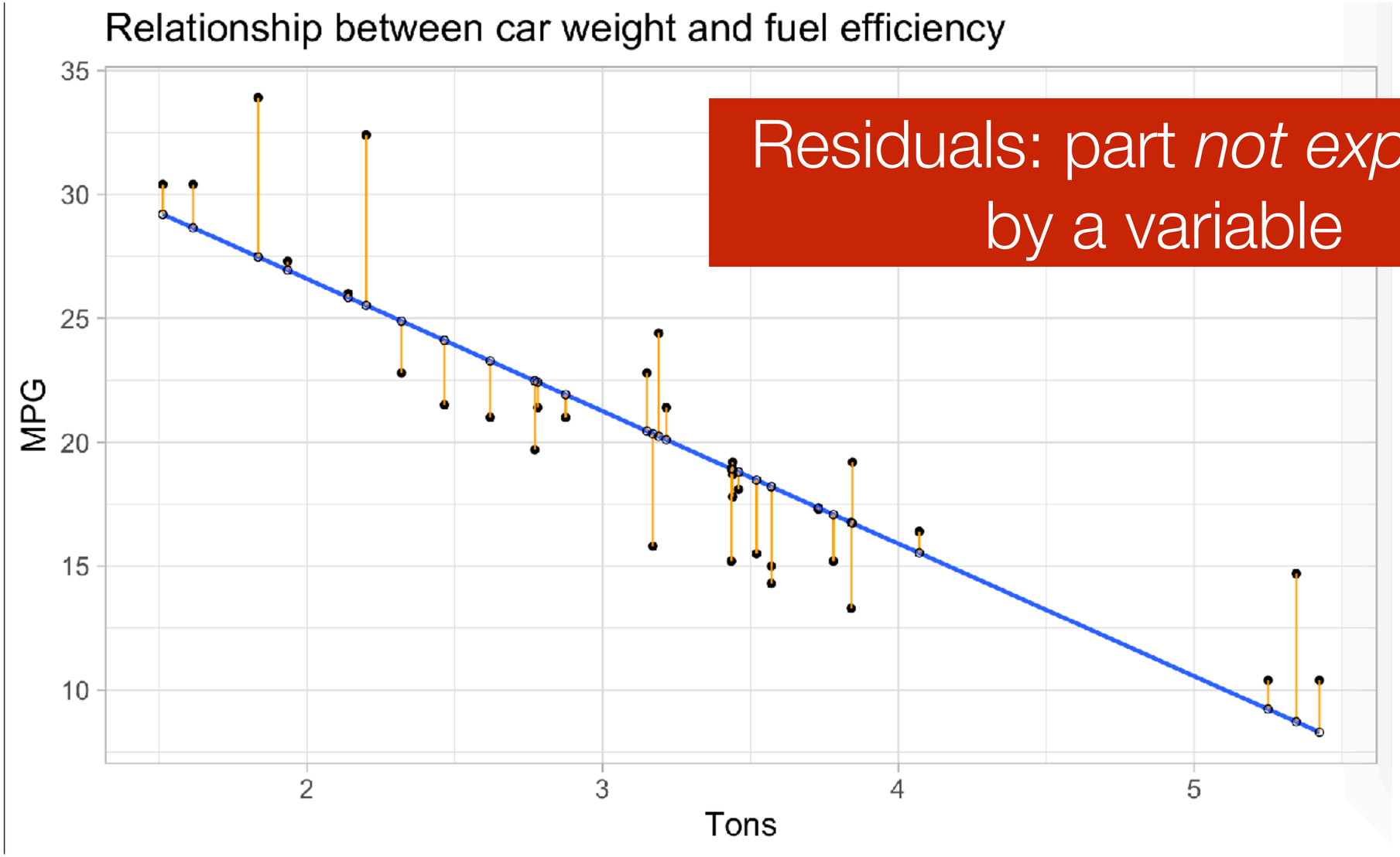
INTUITION

We want to know relationship between M and D **controlling for** A



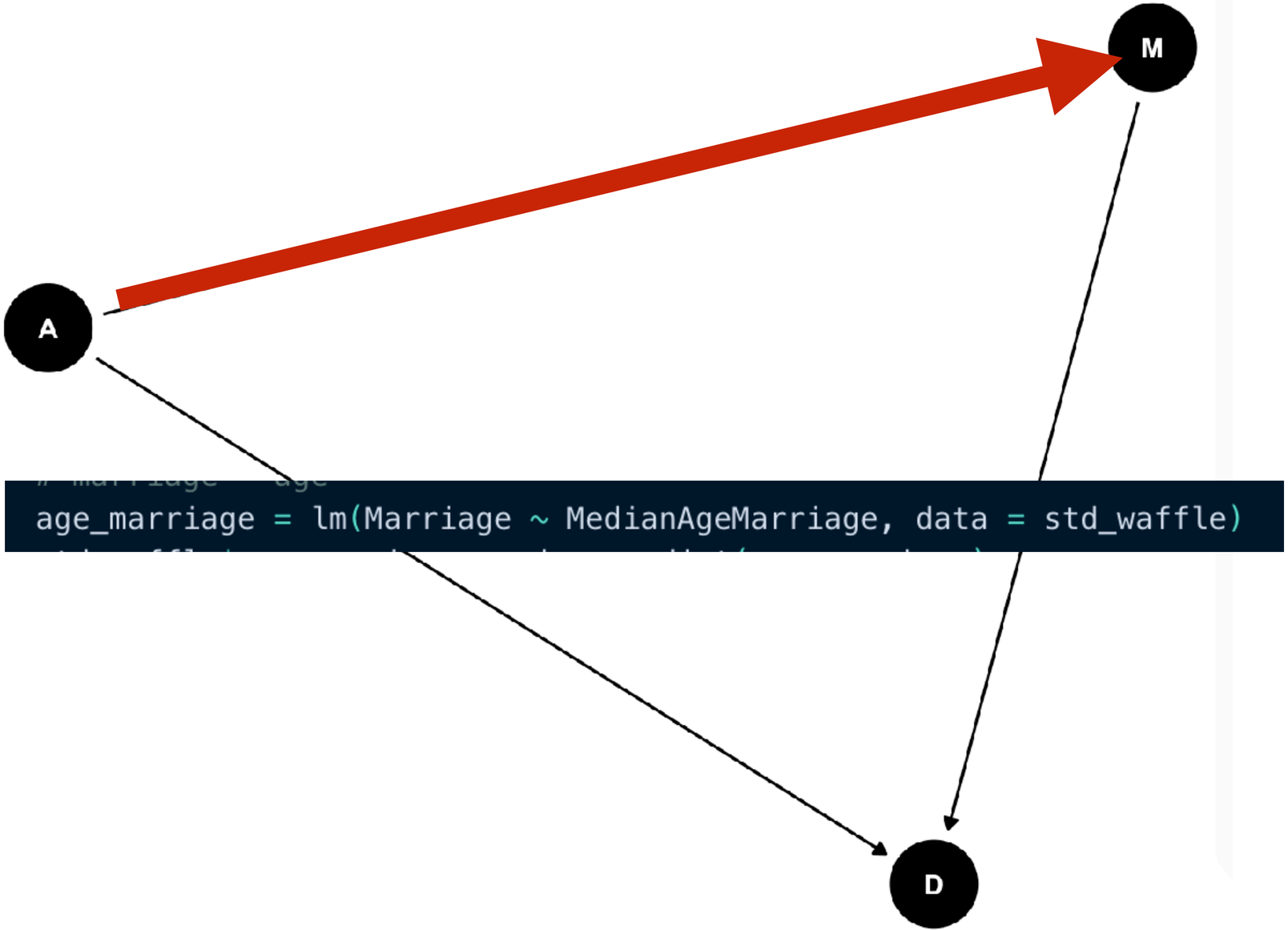
We want to remove the influence of A on **both** M and D

INTUITION

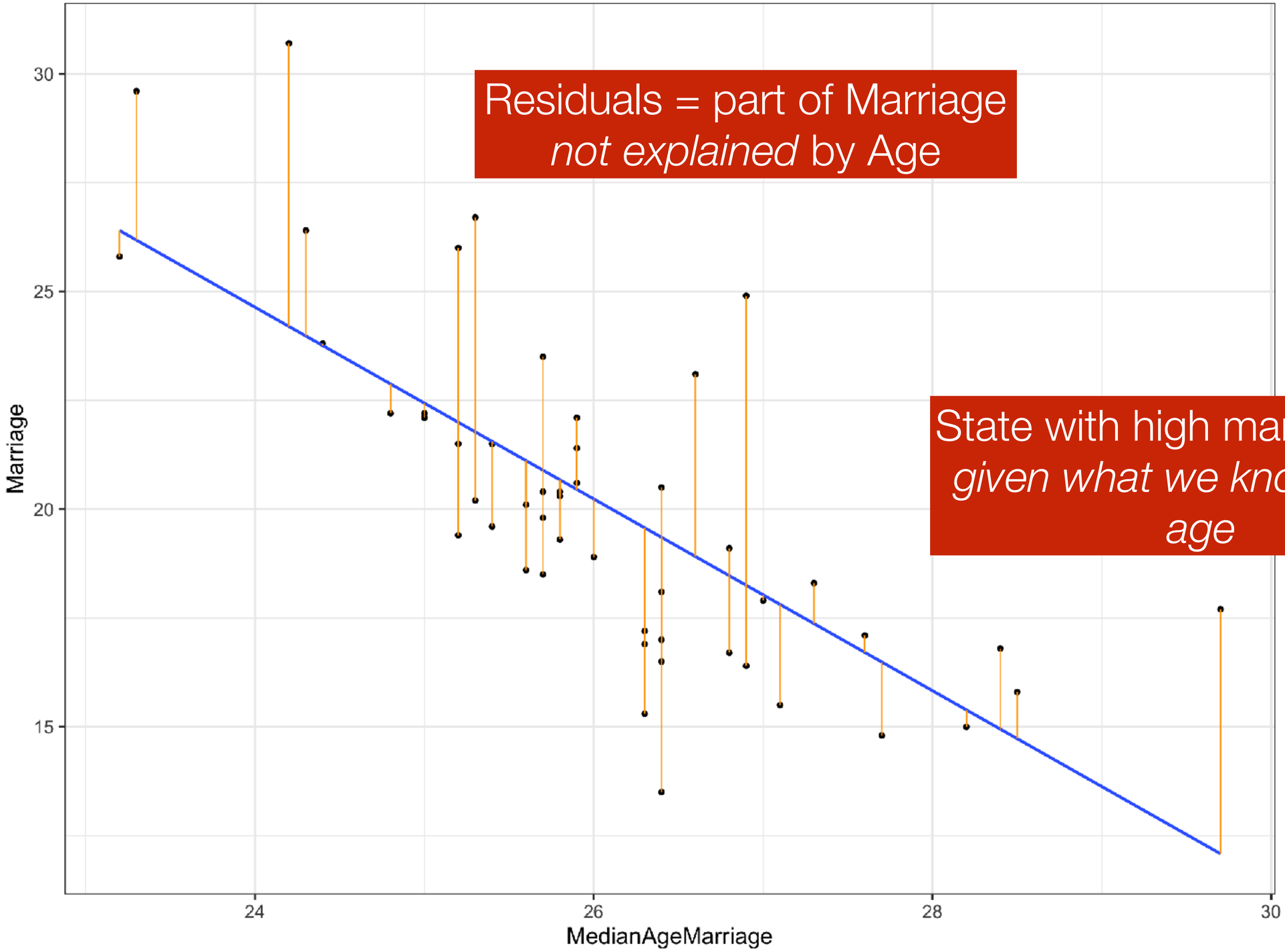


Residuals: part *not explained* by a variable

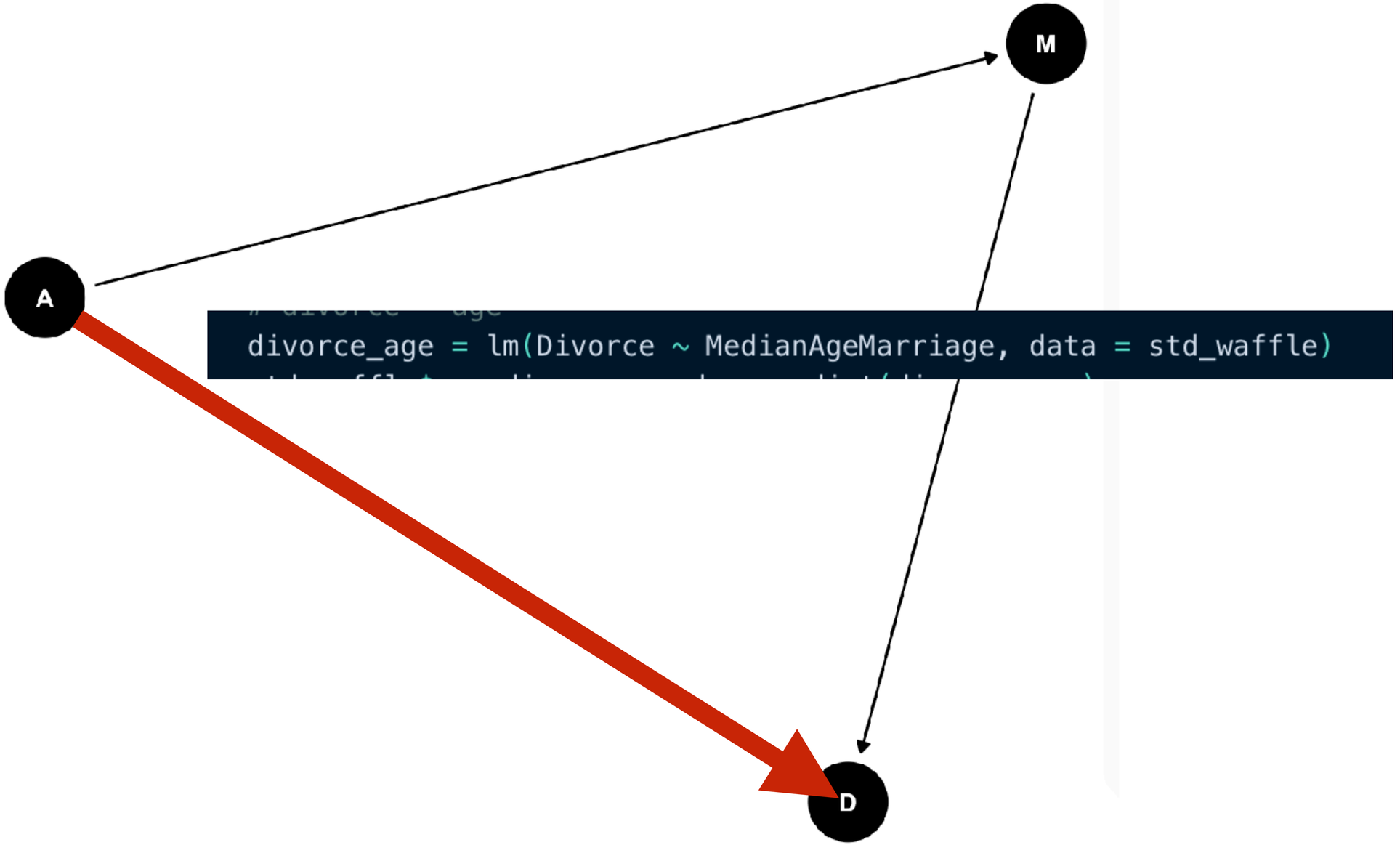
First: let's figure out how A *explains* M, and keep part of M *not explained* by A



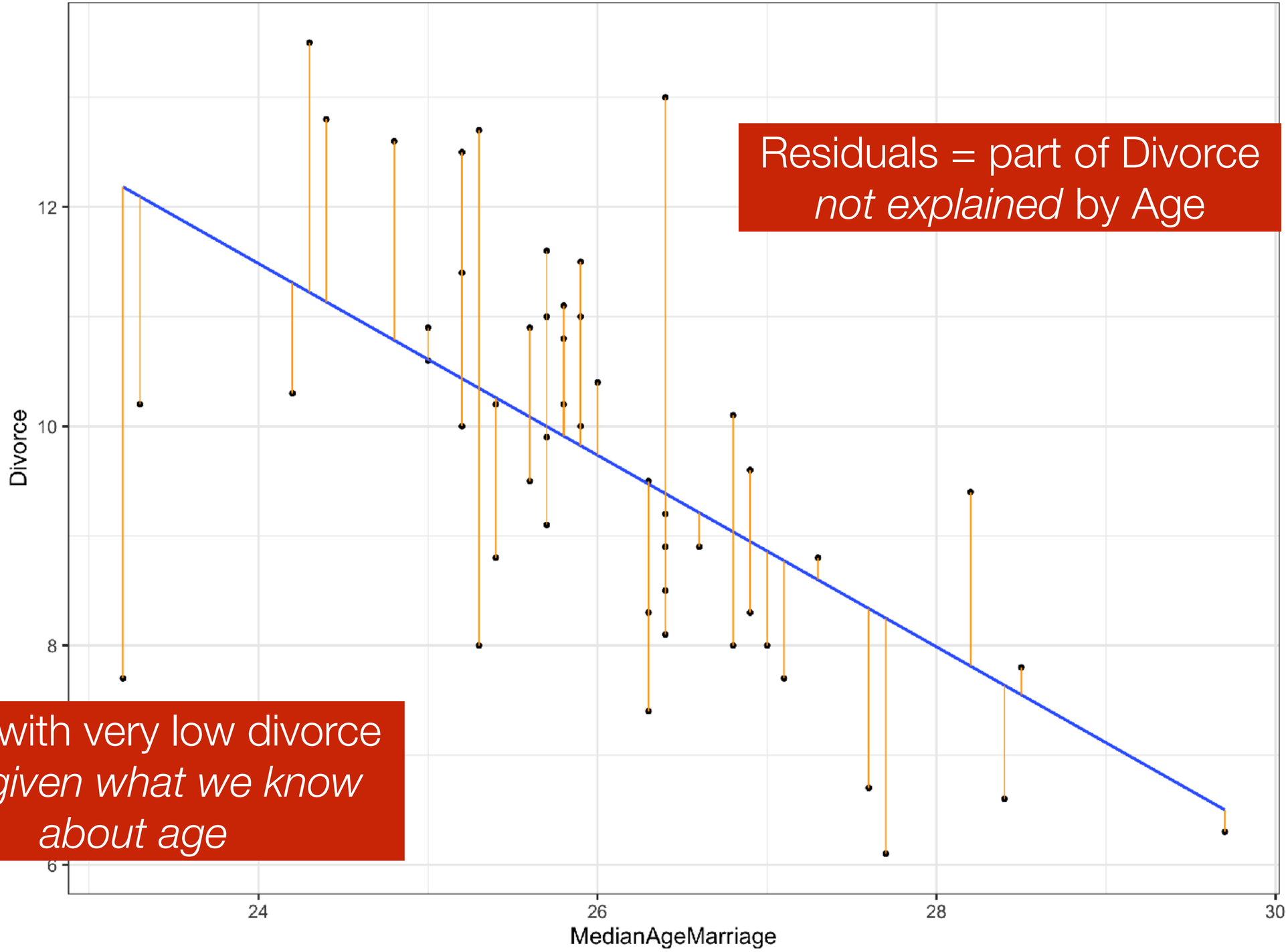

```
age_marriage = lm(Marriage ~ MedianAgeMarriage, data = std_waffle)
```


Residuals = part of Marriage
not explained by Age
State with high marriage rate
given what we know about age
Marriage
30
25
20
15
24
26
28
30
MedianAgeMarriage

Second: let's figure out how A *explains* D, and keep part of D *not explained* by A



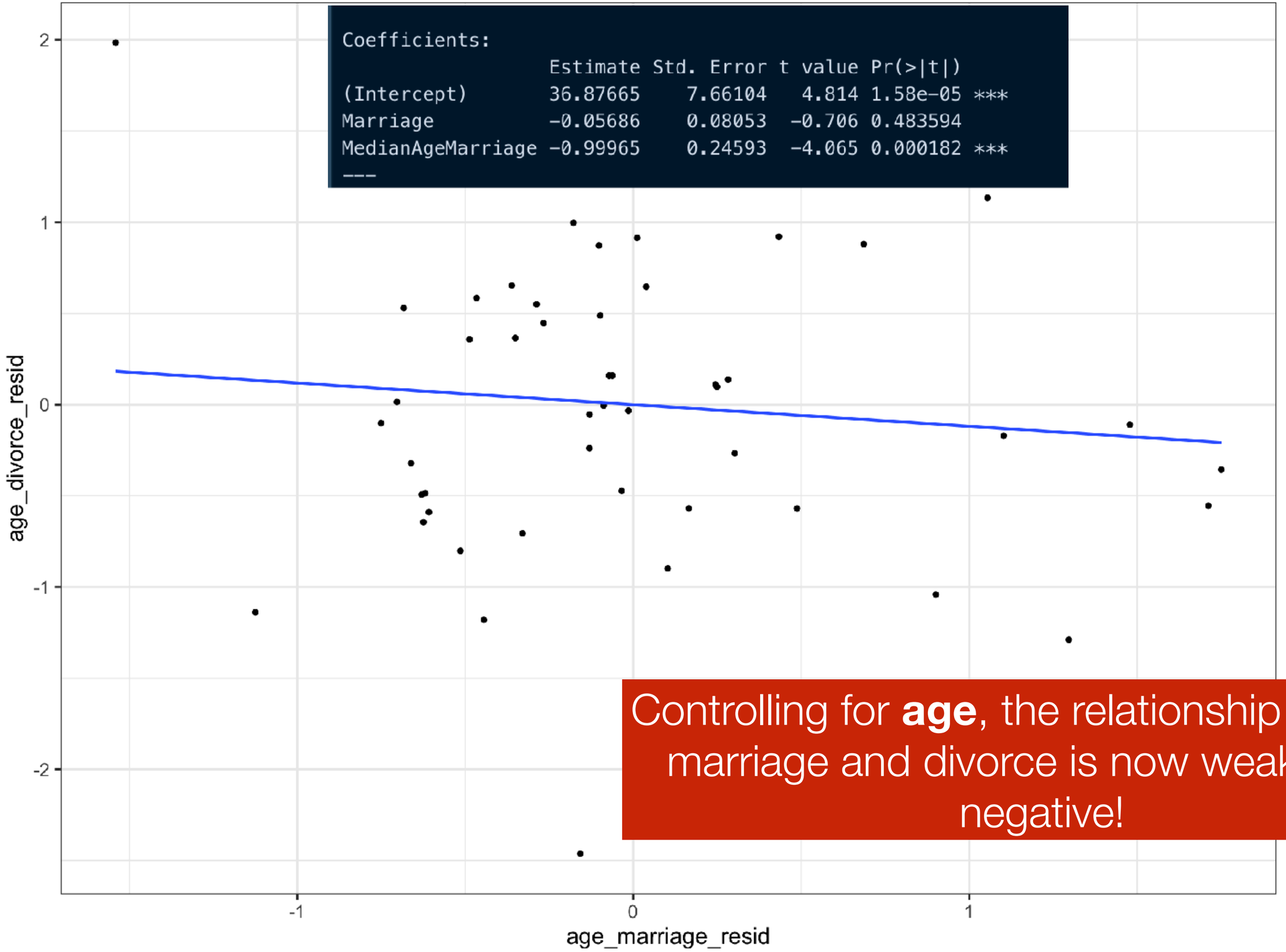

```
divorce_age = lm(Divorce ~ MedianAgeMarriage, data = std_waffle)
```


Residuals = part of Divorce
not explained by Age
State with very low divorce rate *given what we know about age*
Divorce
12
10
8
6
24
26
28
30
MedianAgeMarriage

	Loc	Marriage	MedianAgeMarriage	Divorce	age_marriage_resid	age_divorce_resid
1	AL	20.2	25.3	12.7	-1.5744174	2.352684
2	AK	26.0	25.2	12.5	4.0053681	2.065241
3	AZ	20.3	25.8	10.8	-0.3733448	0.889896
4	AR	26.4	24.3	13.5	2.4234376	2.278259
5	CA	19.1	26.8	8.0	0.6288002	-1.035679

Part of Marriage *unexplained by age*

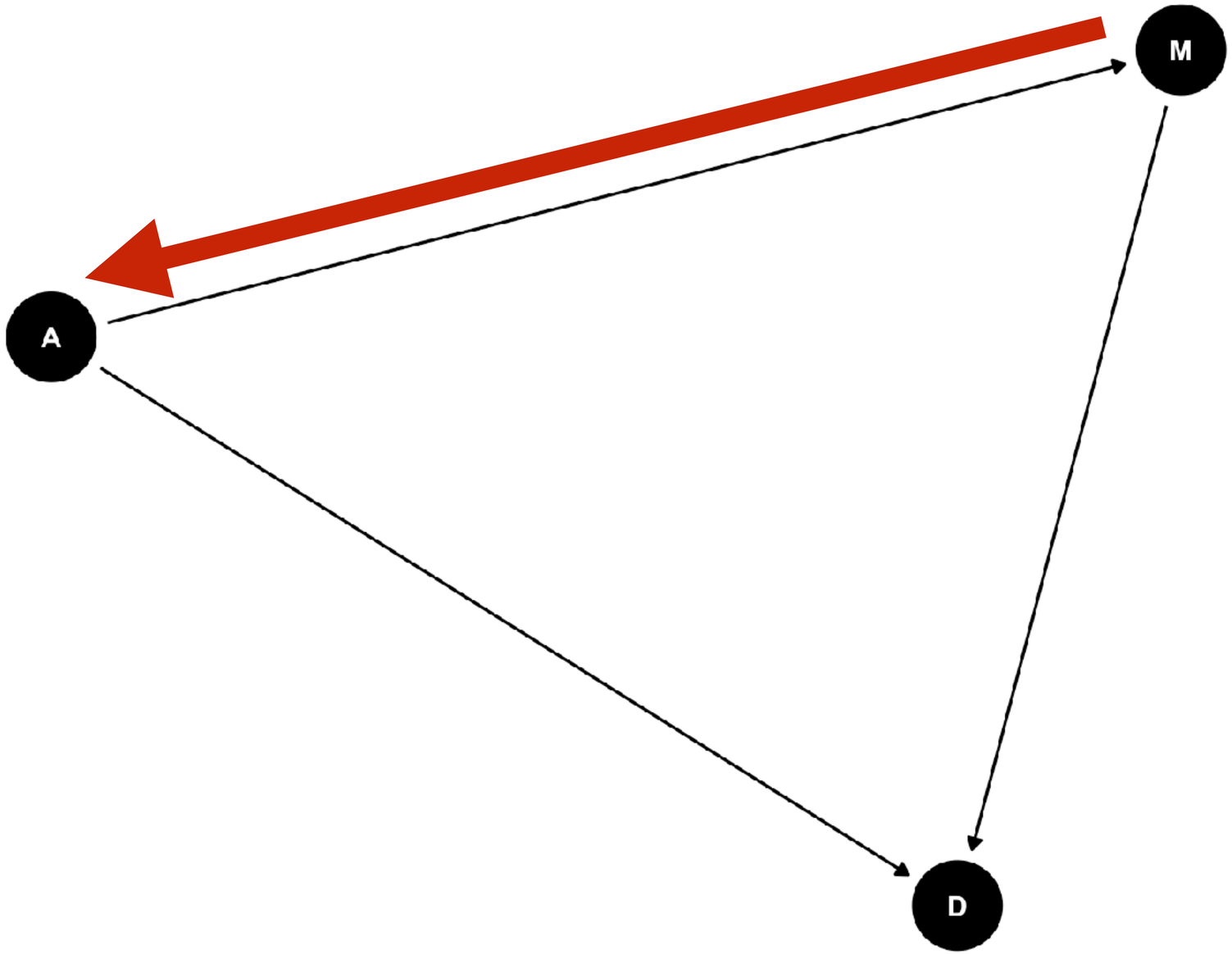
Part of Divorce *unexplained by age*

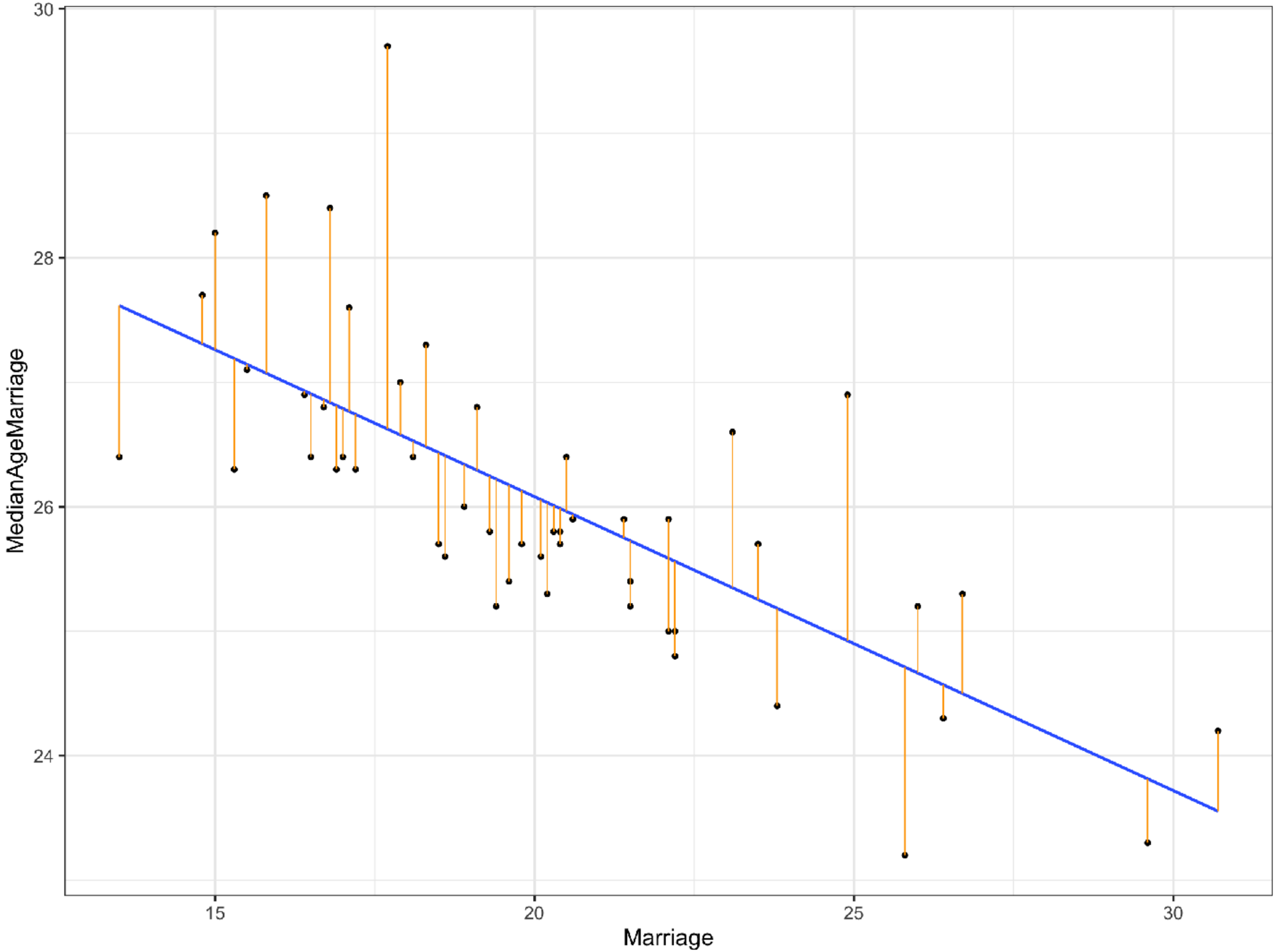


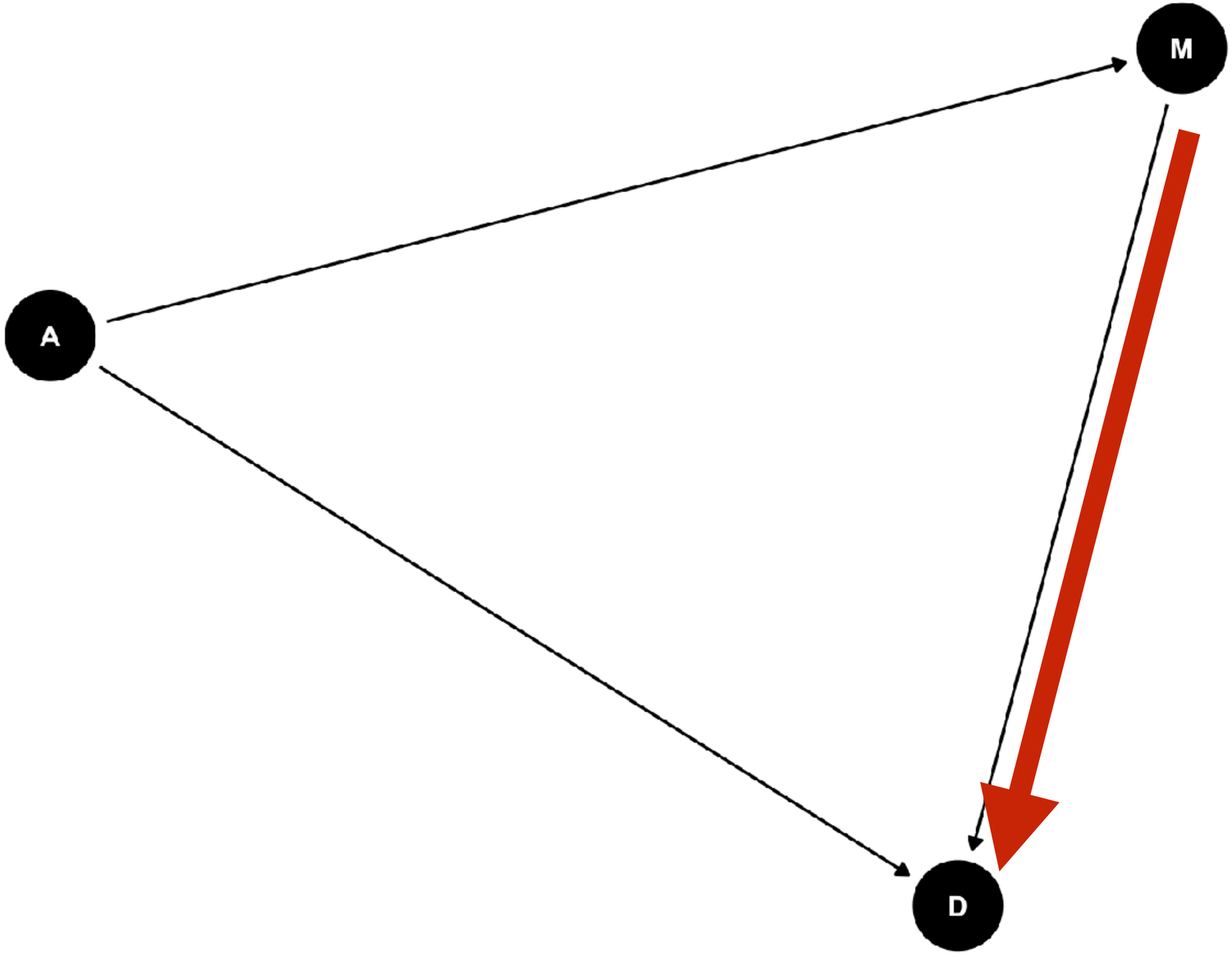
Controlling for **age**, the relationship between marriage and divorce is now weaker and negative!

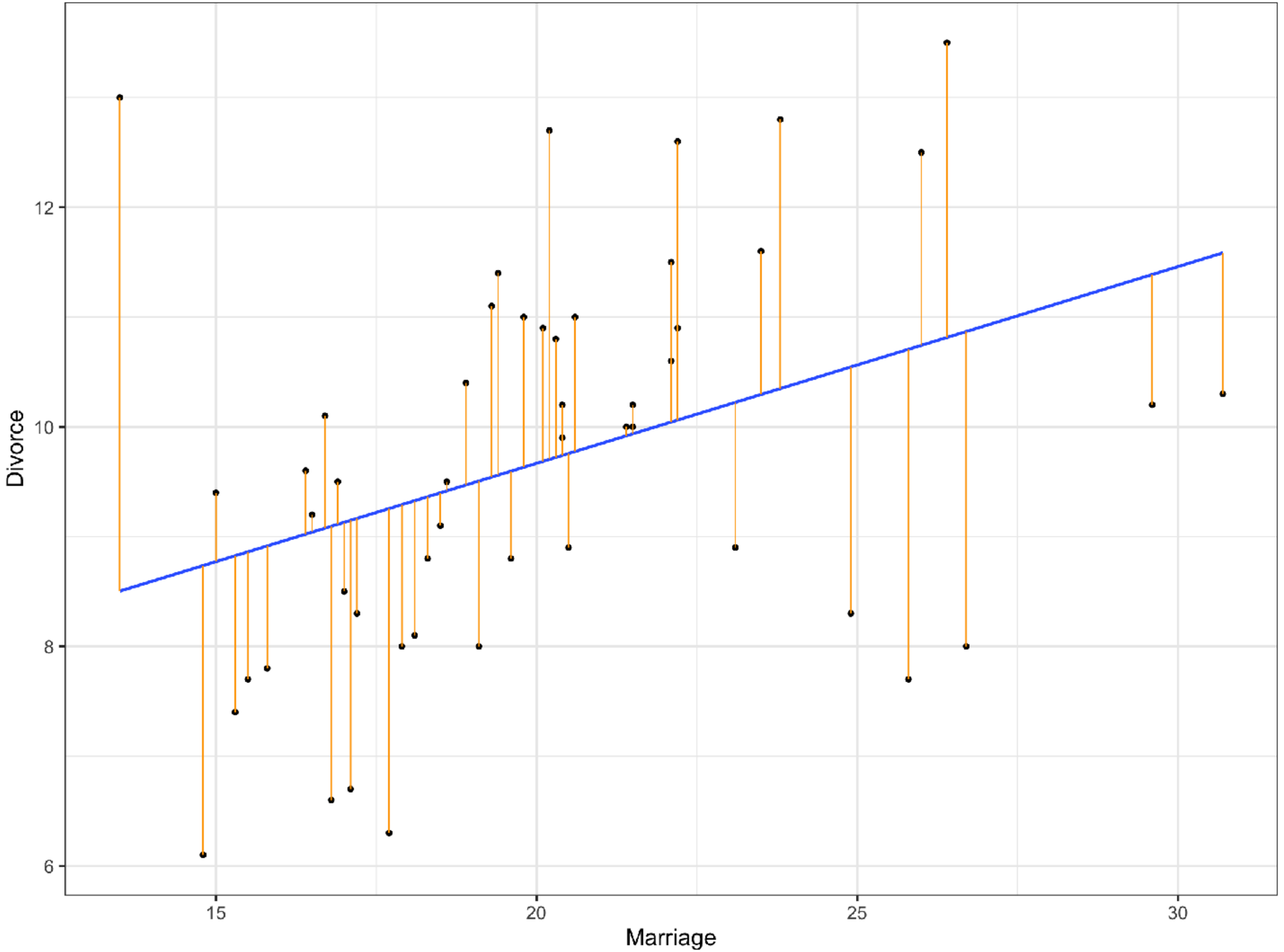
```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       36.87665    7.66104   4.814 1.58e-05 ***
Marriage          -0.05686    0.08053  -0.706 0.483594
MedianAgeMarriage -0.99965    0.24593  -4.065 0.000182 ***
---
```

Now let’s do the same for Age

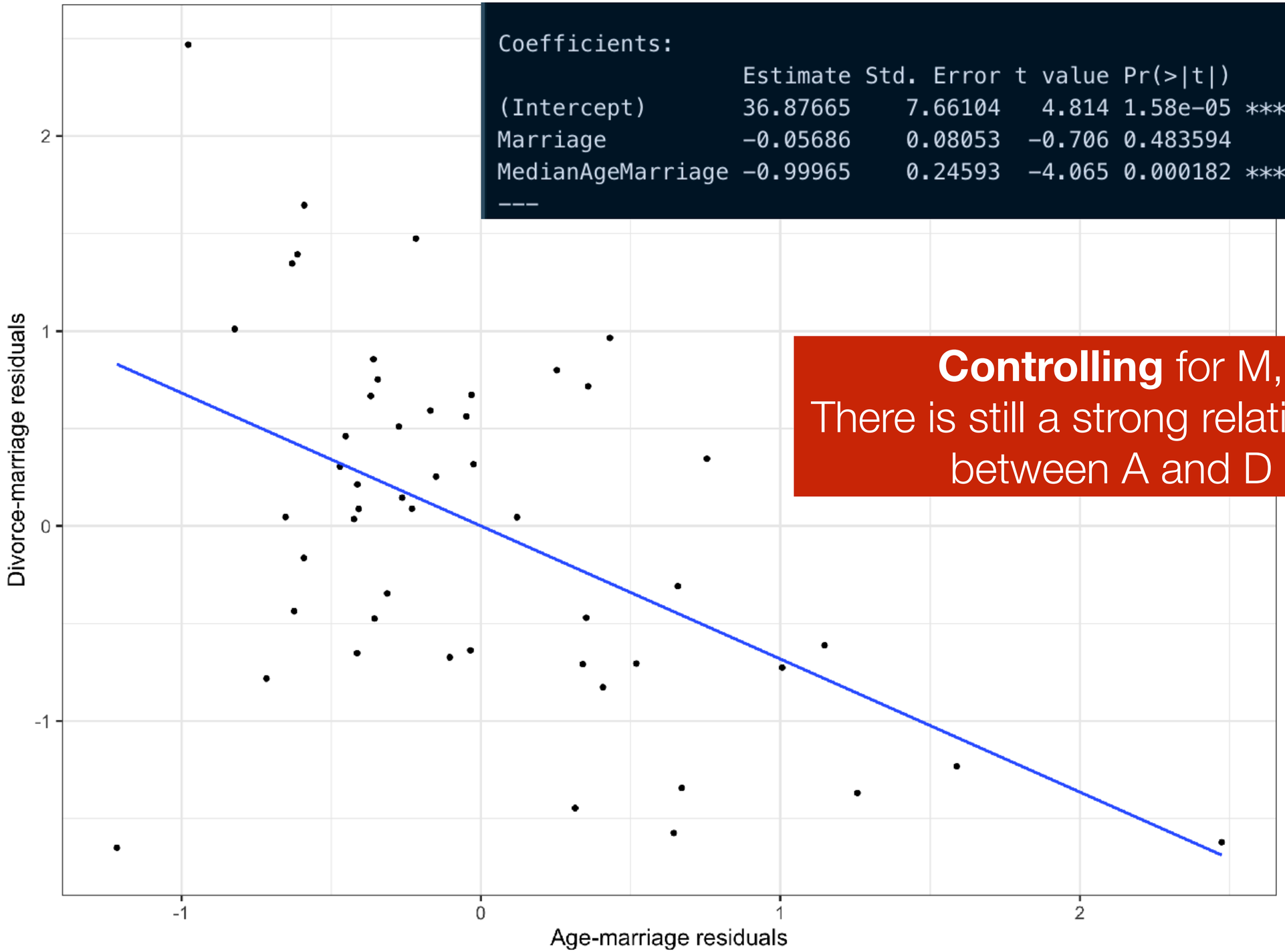

M
A
D


30
28
26
24
MedianAgeMarriage
15
20
25
30
Marriage


M
A
D


Divorce
12
10
8
6
15
20
25
30
Marriage

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       36.87665    7.66104   4.814 1.58e-05 ***
Marriage          -0.05686    0.08053  -0.706 0.483594
MedianAgeMarriage -0.99965    0.24593  -4.065 0.000182 ***
---
```



Controlling for M,
There is still a strong relationship between A and D

SLEEP AND GRADES

OK, but how do we **know** true effect of M on D is -.06?

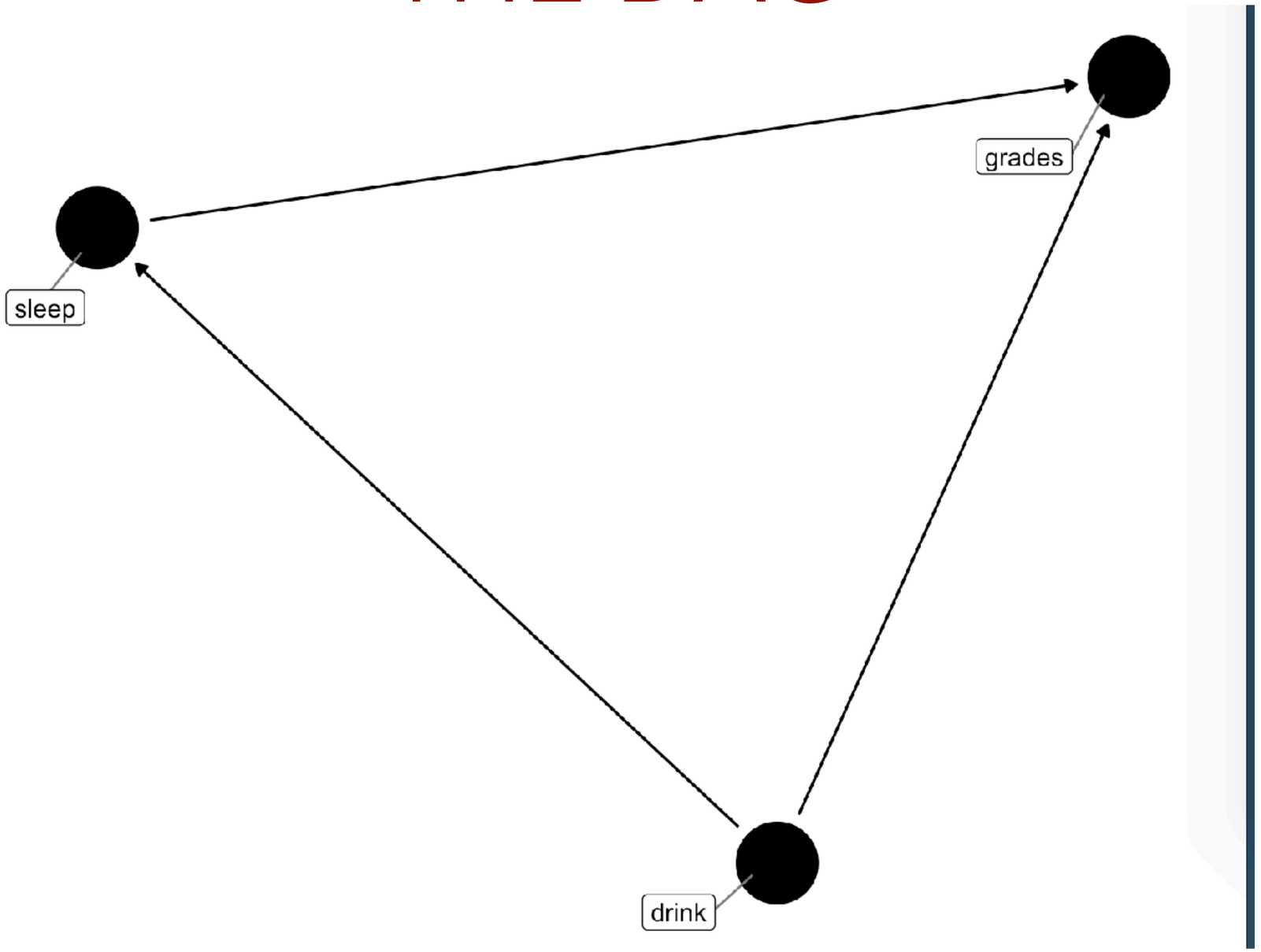
We can't know, so let's see an example where we can

What is the effect of getting a good night's sleep on grades?

I'll make up some data where I already know the answer to this question

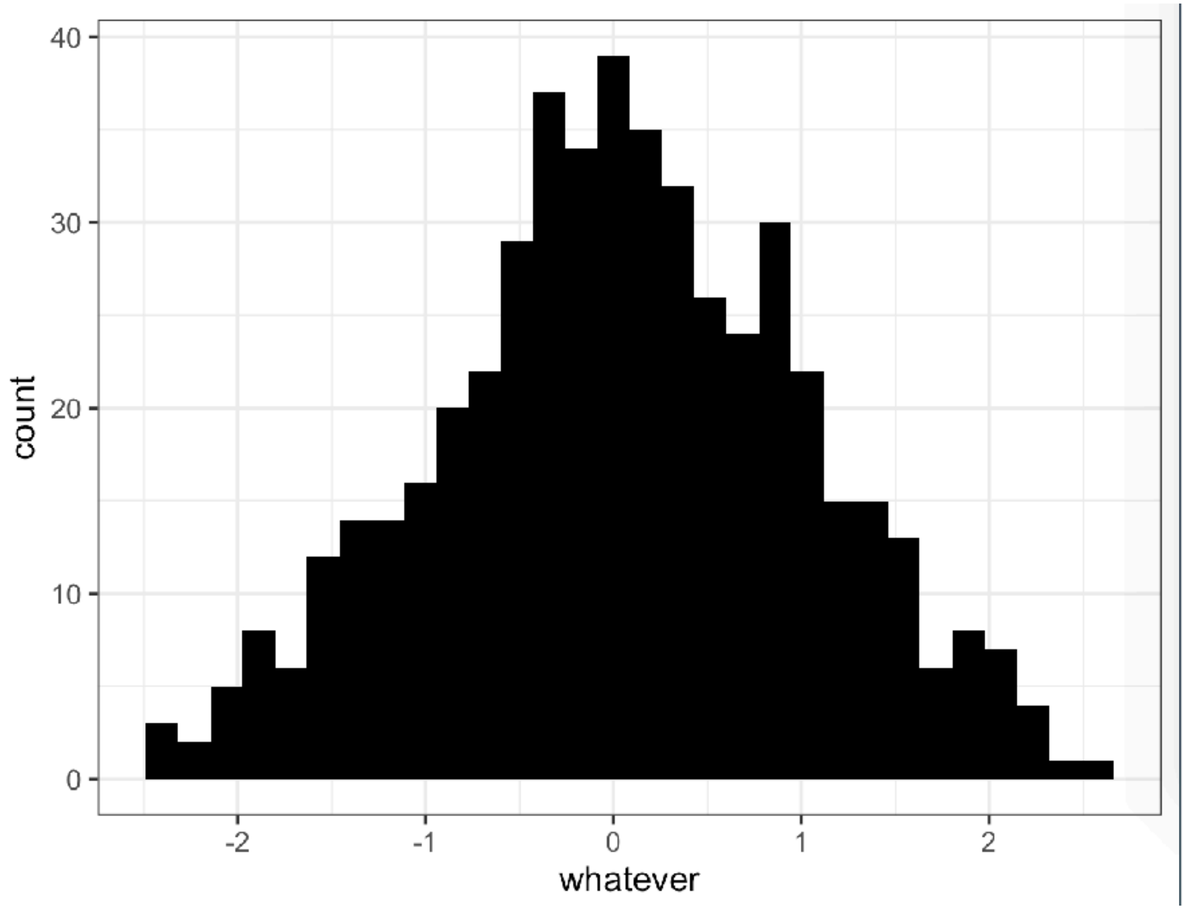
$$Grades = \alpha + 3 \times Sleep$$

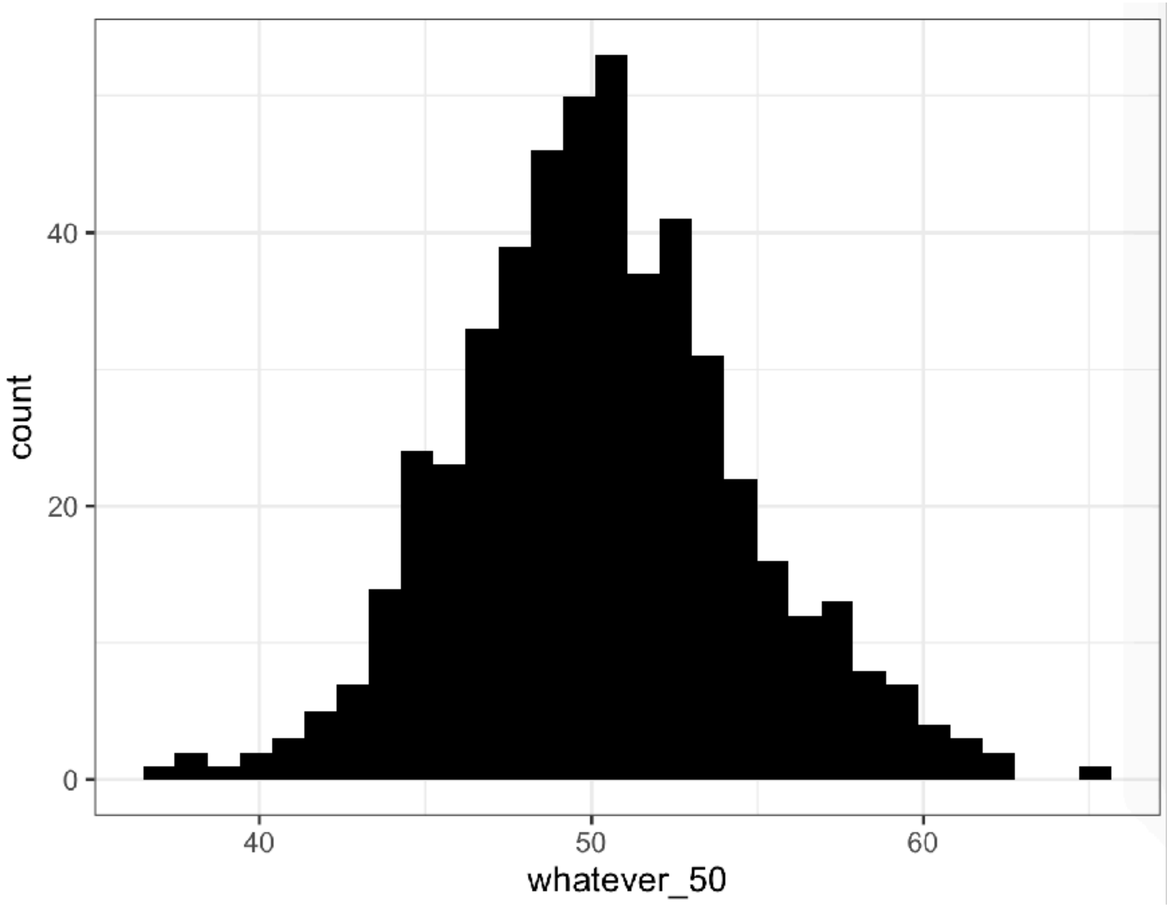
THE DAG



MAKING UP DATA

rnorm(n = 500): pick 500 random numbers that look like this





MAKING UP DATA

Make up data on how much people are **drinking**

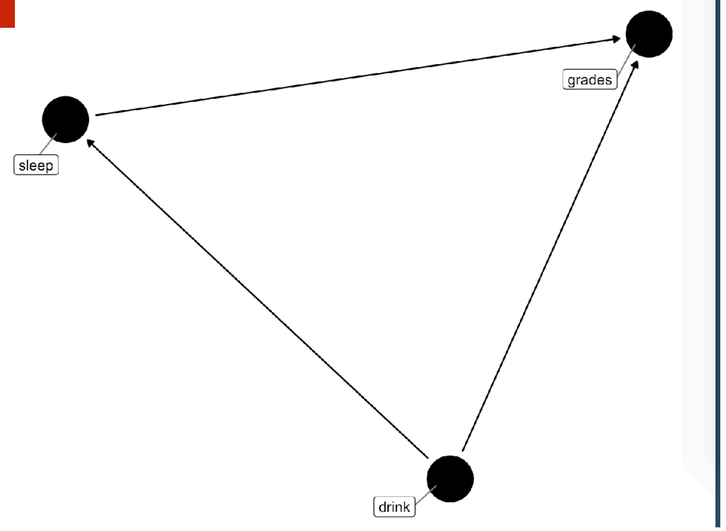
```
drink = rnorm(n = 500, mean = 3),
```

Make up data on how much people are **sleeping**

```
sleep = rnorm(500) - drink,
```

Make up data on people's **grades**

```
grades = sleep*3 + drink*2 + rnorm(500)
```



REGRESSION

```r
# fit model without controls
m1 = lm(grades ~ sleep, data = df)
summary(m1)
```

```
Coefficients:
             Estimate
(Intercept)  3.14277
sleep        2.05697
```

```r
# fit model with controls
m2 = lm(grades ~ sleep + drink, data = df)
summary(m2)
```

```
Coefficients:
             Estimate
(Intercept)  0.09897
sleep        3.01762
drink        1.98586
```

```r
grades = sleep*3 + drink*2 + rnorm(500)
```

RECAP

We want to close backdoors (Z) from X to Y

Statistical “controls” allow us to do this by removing the part of X and Y that can be explained by Z

Interpretation template: “all else equal”/“controlling for A, B, C, a one unit increase in X produces _____ in Y”

HOMEWORK:
Making DAGs in R + very basic multiple regression