Poli-301:

The Political Science Discipline

TODAY'S AGENDA

1 Controls and categorical variables

2 Prediction and practice

Last week

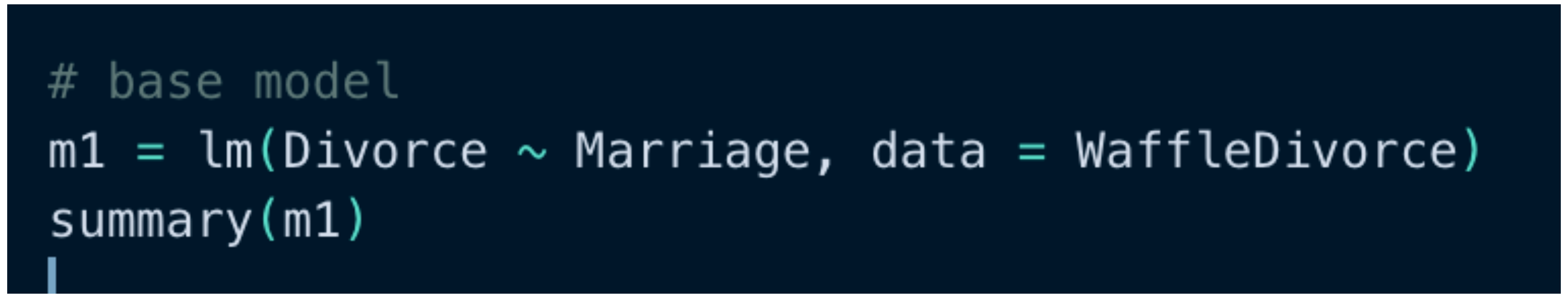
We want to close backdoors (Z) from X to Y

“controls” subtract the part of X and Y that can be explained by Z

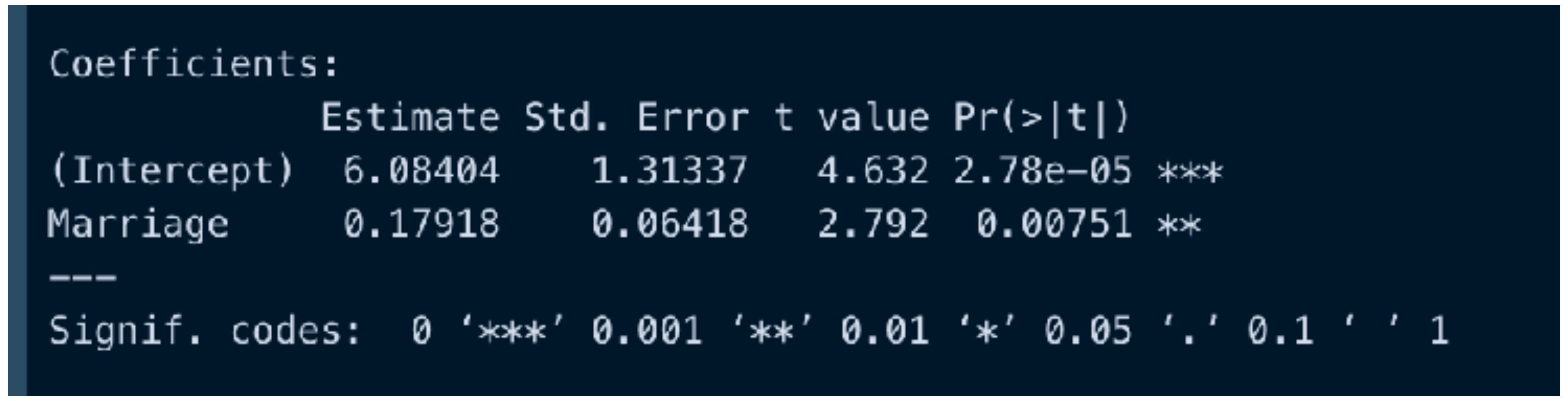
Interpretation template: “all else equal”/“controlling for A, B, C, a one unit increase in X produces _____ in Y”

On their own

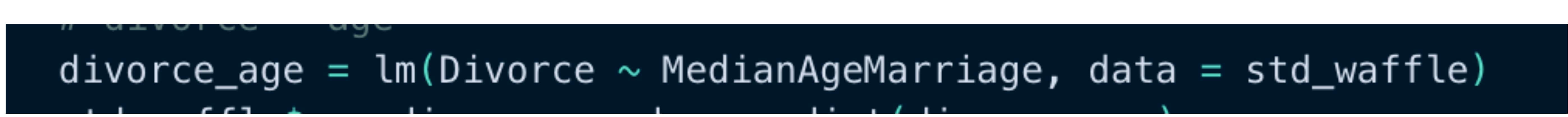
Marriage rate and Median Age at Marriage both explain some variation in Divorce rate



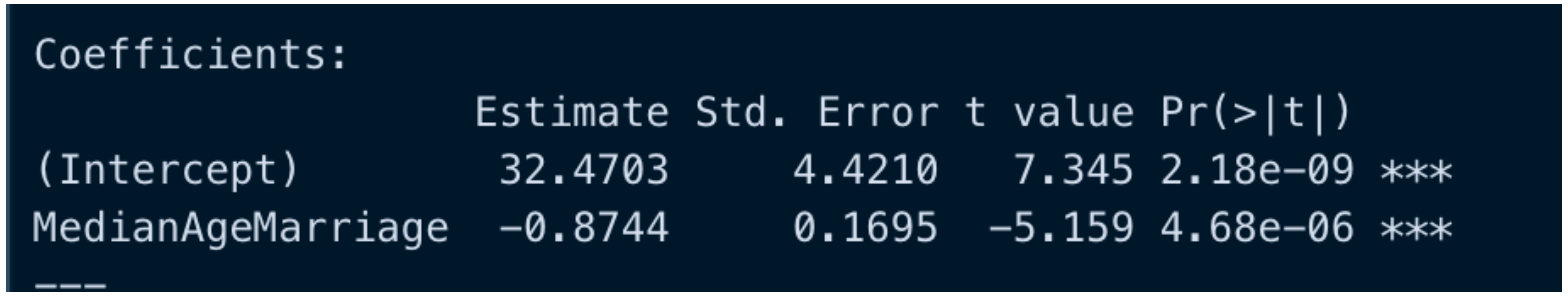
```
# base model
m1 = lm(Divorce ~ Marriage, data = WaffleDivorce)
summary(m1)
```



```
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
```



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



```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        32.4703     4.4210   7.345 2.18e-09 ***
MedianAgeMarriage  -0.8744     0.1695  -5.159 4.68e-06 ***
---
```

Some of that explanation is shared!

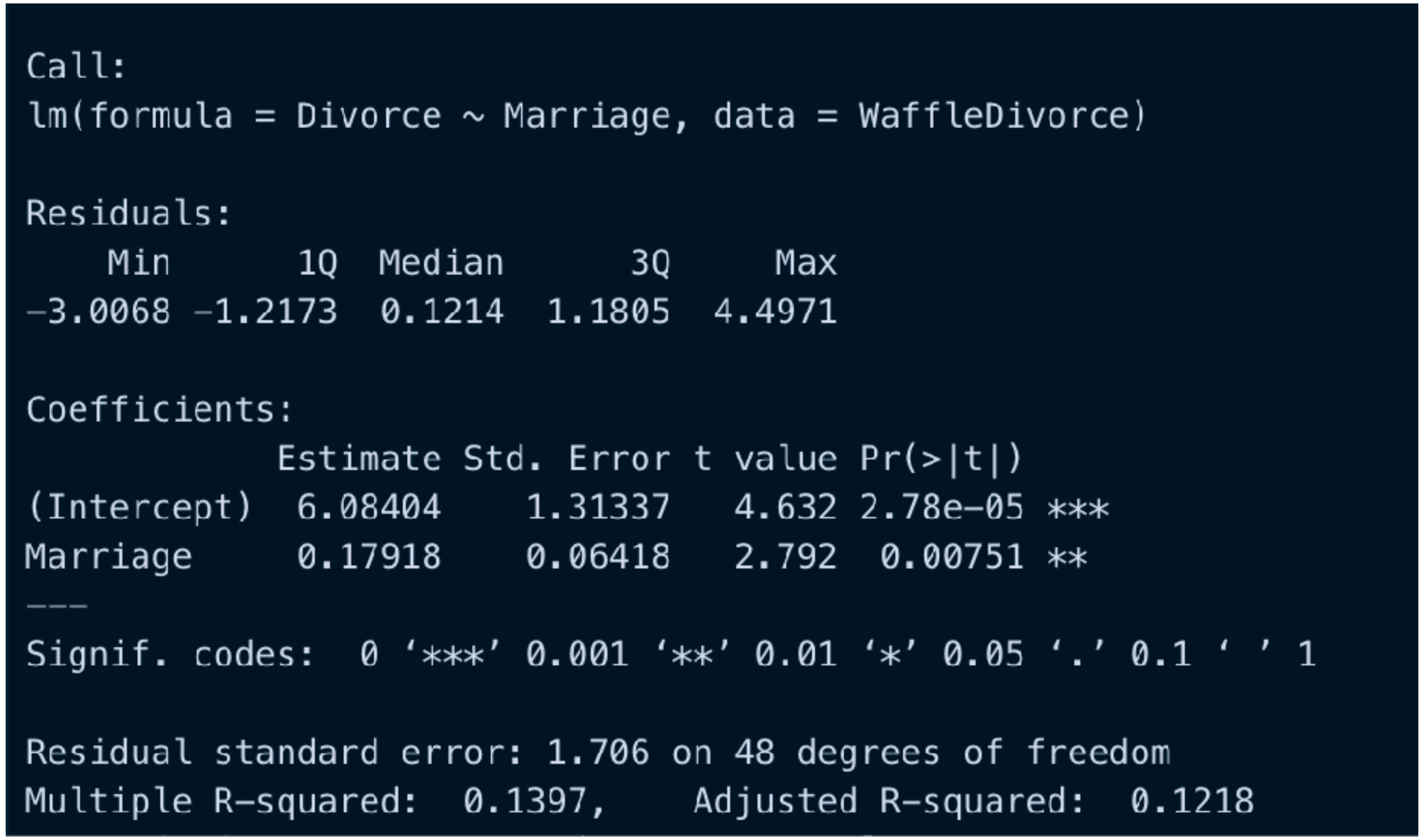
R^2

R^2

How much variation in Y is explained by X

0-1 scale; a %

Higher = more variation



```
Call:
lm(formula = Divorce ~ Marriage, data = WaffleDivorce)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0068 -1.2173  0.1214  1.1805  4.4971 

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

Residual standard error: 1.706 on 48 degrees of freedom
Multiple R-squared:  0.1397,    Adjusted R-squared:  0.1218
```

Adjusted R^2

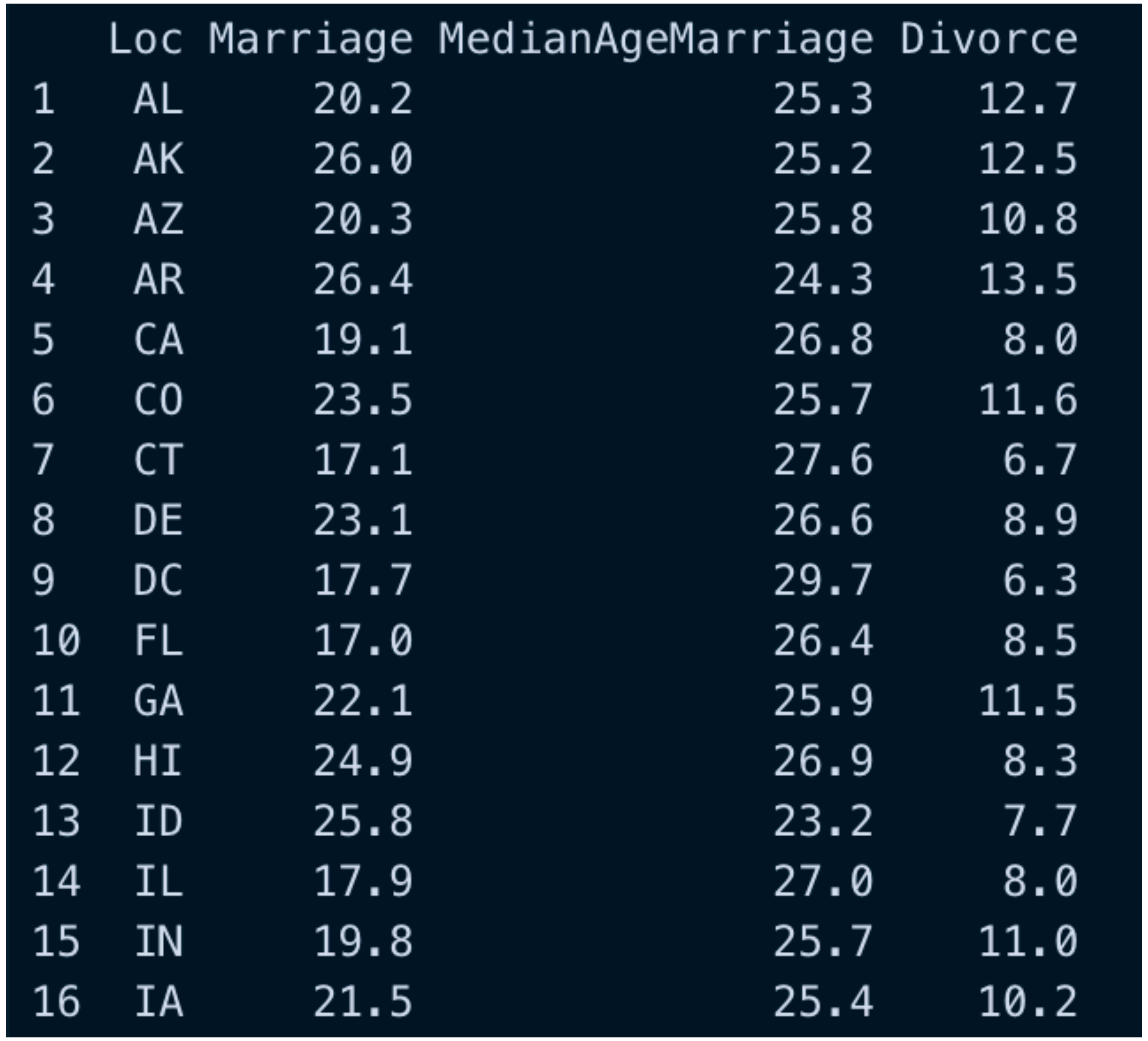
Technically, the more Xs you add the “better” you explain Y

Adjusted R^2 calculates differently to account for htis, penalizes adding terms

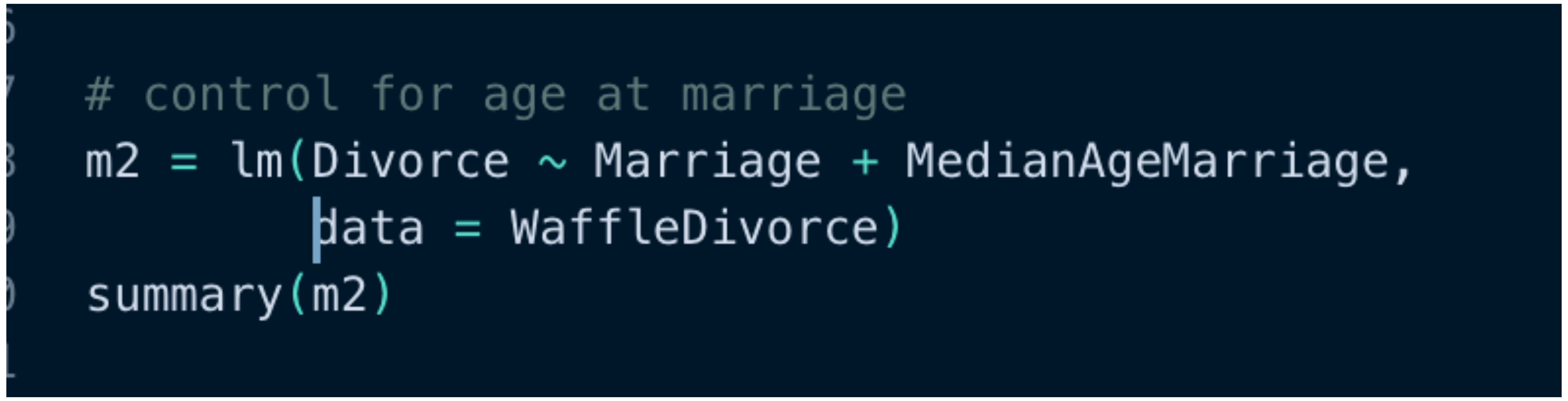
Not super relevant for us, and there’s a lot of skepticism around R^2 today!

But just so you know…

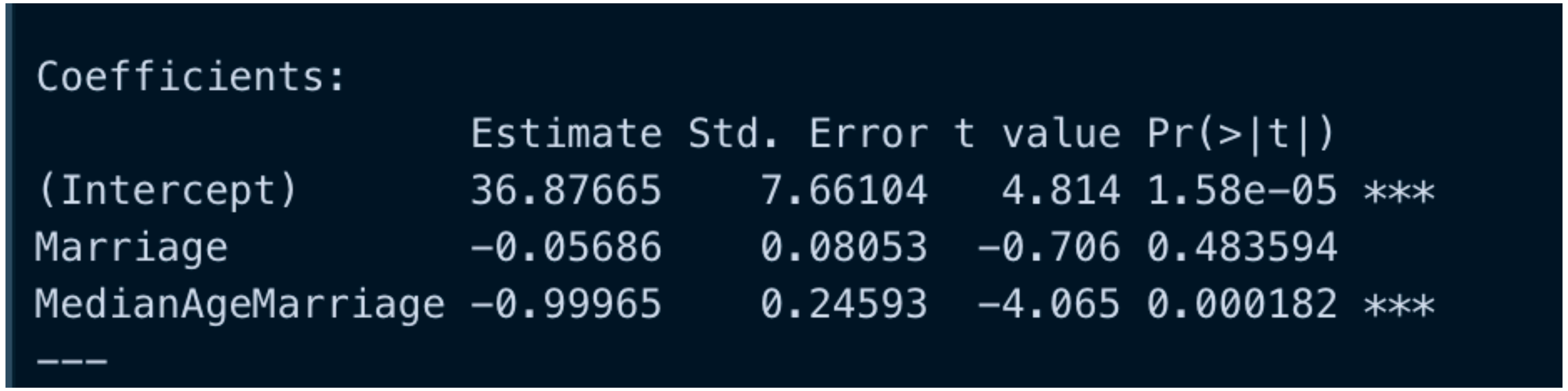
Interpretation



```
   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
```



```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 ***
---
```

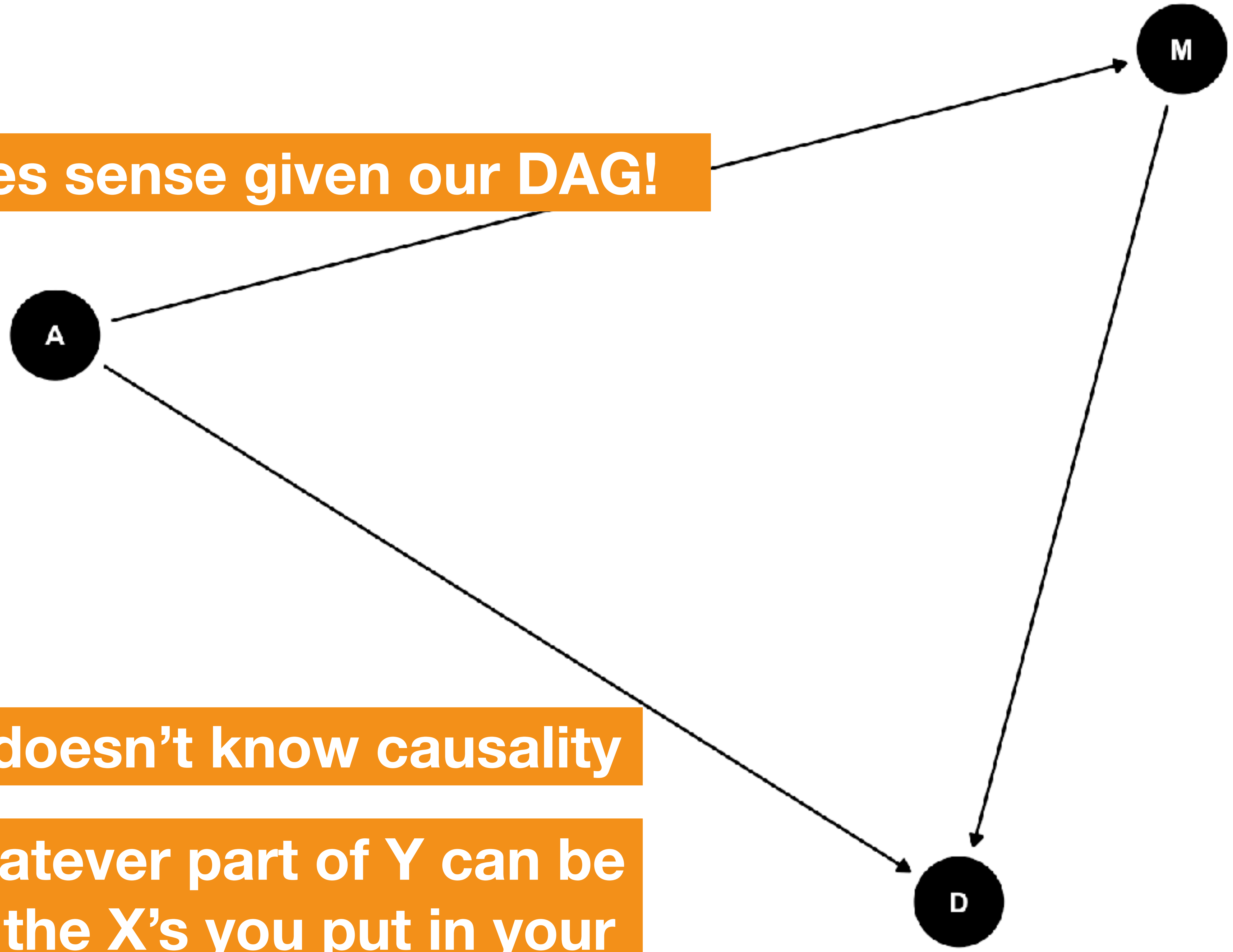
Explaining variation

Each X in a model “explains” some portion of the variation in Y

This means regression coefficients change depending on what’s in the model

Taking all other variables in the model into account, a one unit increase in xn is associated with a βn increase (or decrease) in y, on average

Controlling for the median age at marriage, a increase in the rate of 1 person per thousand who is married is associated with a decrease of .05 in the divorce rate, on average



This makes sense given our DAG!

But again, R doesn’t know causality

R will find whatever part of Y can be explained by the X’s you put in your model, even if those X’s are causally unrelated to Y

Interpretation

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

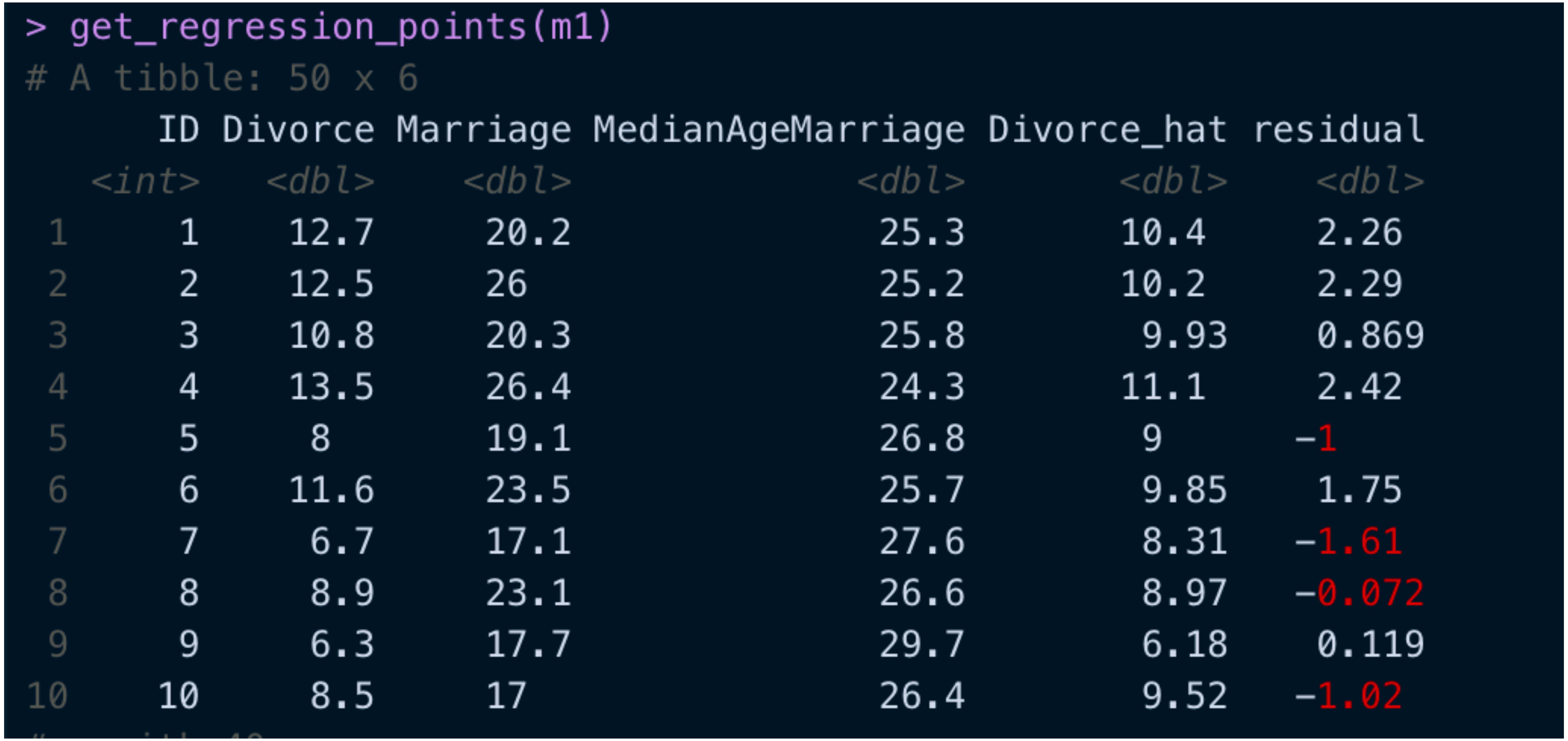
$$Divorce = 36 + -.06 \times Marriage + -.1 \times MedianAge + \epsilon$$

What's our best guess for…

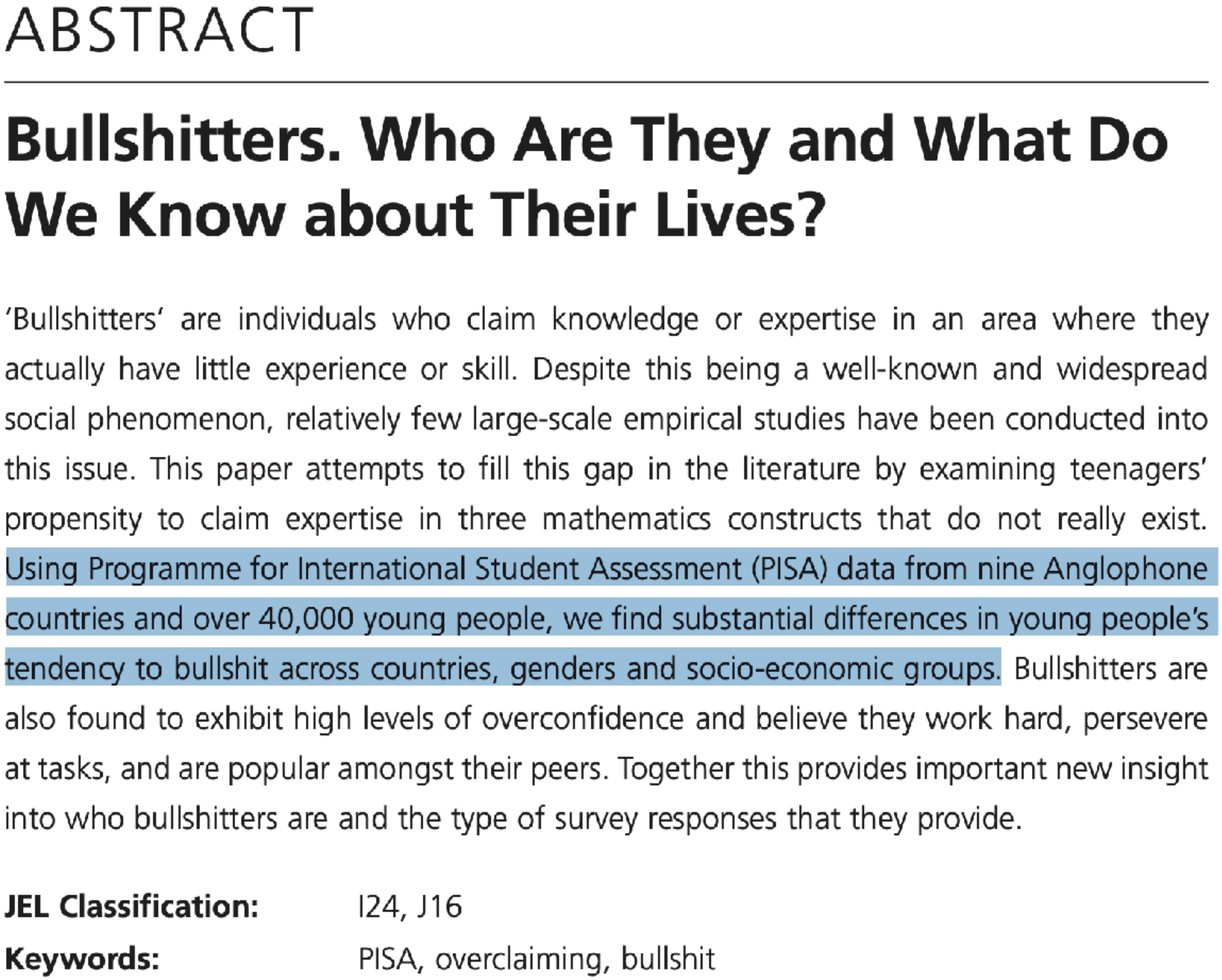
A state where people get married at 30 and where the marriage rate is 20 per thousand?

get_regression_points()

$$Divorce = 36 + -.06 \times Marriage + -.1 \times MedianAge + \epsilon$$



```
> get_regression_points(m1)
# A tibble: 50 x 6
      ID Divorce Marriage MedianAgeMarriage Divorce_hat residual
   <int>   <dbl>    <dbl>             <dbl>       <dbl>    <dbl>
 1     1    12.7     20.2              25.3       10.4     2.26
 2     2    12.5     26                25.2       10.2     2.29
 3     3    10.8     20.3              25.8        9.93    0.869
 4     4    13.5     26.4              24.3       11.1     2.42
 5     5     8       19.1              26.8        9      -1
 6     6    11.6     23.5              25.7        9.85    1.75
 7     7     6.7     17.1              27.6        8.31   -1.61
 8     8     8.9     23.1              26.6        8.97   -0.072
 9     9     6.3     17.7              29.7        6.18    0.119
10    10     8.5     17                26.4        9.52   -1.02
```



ABSTRACT

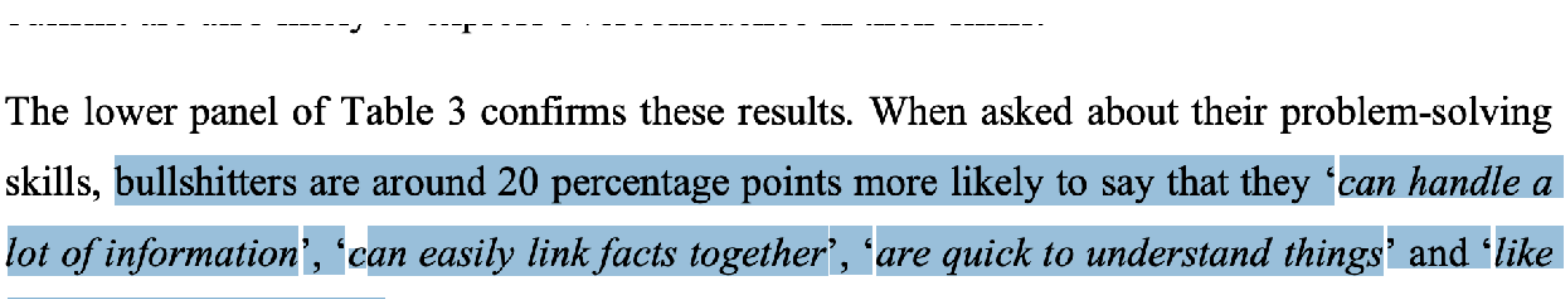
Bullshitters. Who Are They and What Do We Know about Their Lives?

'Bullshitters' are individuals who claim knowledge or expertise in an area where they actually have little experience or skill. Despite this being a well-known and widespread social phenomenon, relatively few large-scale empirical studies have been conducted into this issue. This paper attempts to fill this gap in the literature by examining teenagers' propensity to claim expertise in three mathematics constructs that do not really exist. Using Programme for International Student Assessment (PISA) data from nine Anglophone countries and over 40,000 young people, we find substantial differences in young people's tendency to bullshit across countries, genders and socio-economic groups. Bullshitters are also found to exhibit high levels of overconfidence and believe they work hard, persevere at tasks, and are popular amongst their peers. Together this provides important new insight into who bullshitters are and the type of survey responses that they provide.

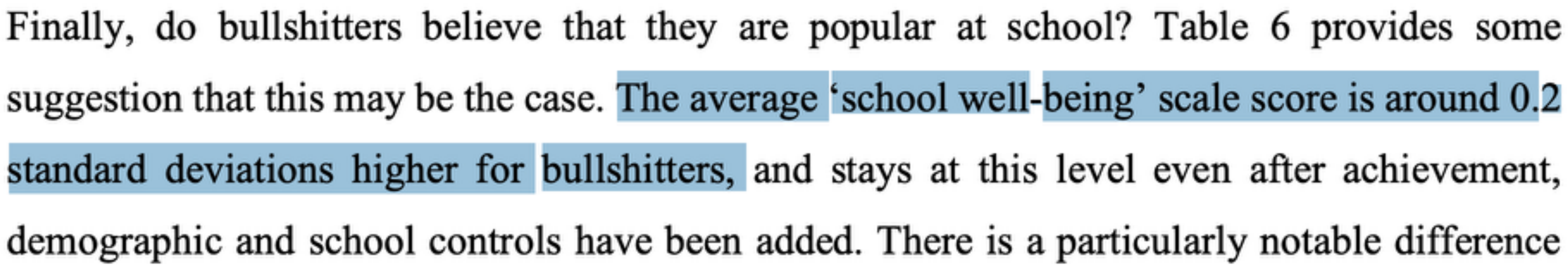
JEL Classification: I24, J16

Keywords: PISA, overclaiming, bullshit

The psychology of bullshit



The lower panel of Table 3 confirms these results. When asked about their problem-solving skills, bullshitters are around 20 percentage points more likely to say that they '*can handle a lot of information*', '*can easily link facts together*', '*are quick to understand things*' and '*like*



Finally, do bullshitters believe that they are popular at school? Table 6 provides some suggestion that this may be the case. The average 'school well-being' scale score is around 0.2 standard deviations higher for bullshitters, and stays at this level even after achievement, demographic and school controls have been added. There is a particularly notable difference

What's the effect of… on bullshit?

Self-efficacy?

Perseverance?

self-esteem?

What could go wrong?

What if we regress
`lm(bullshit ~ psych_trait)`

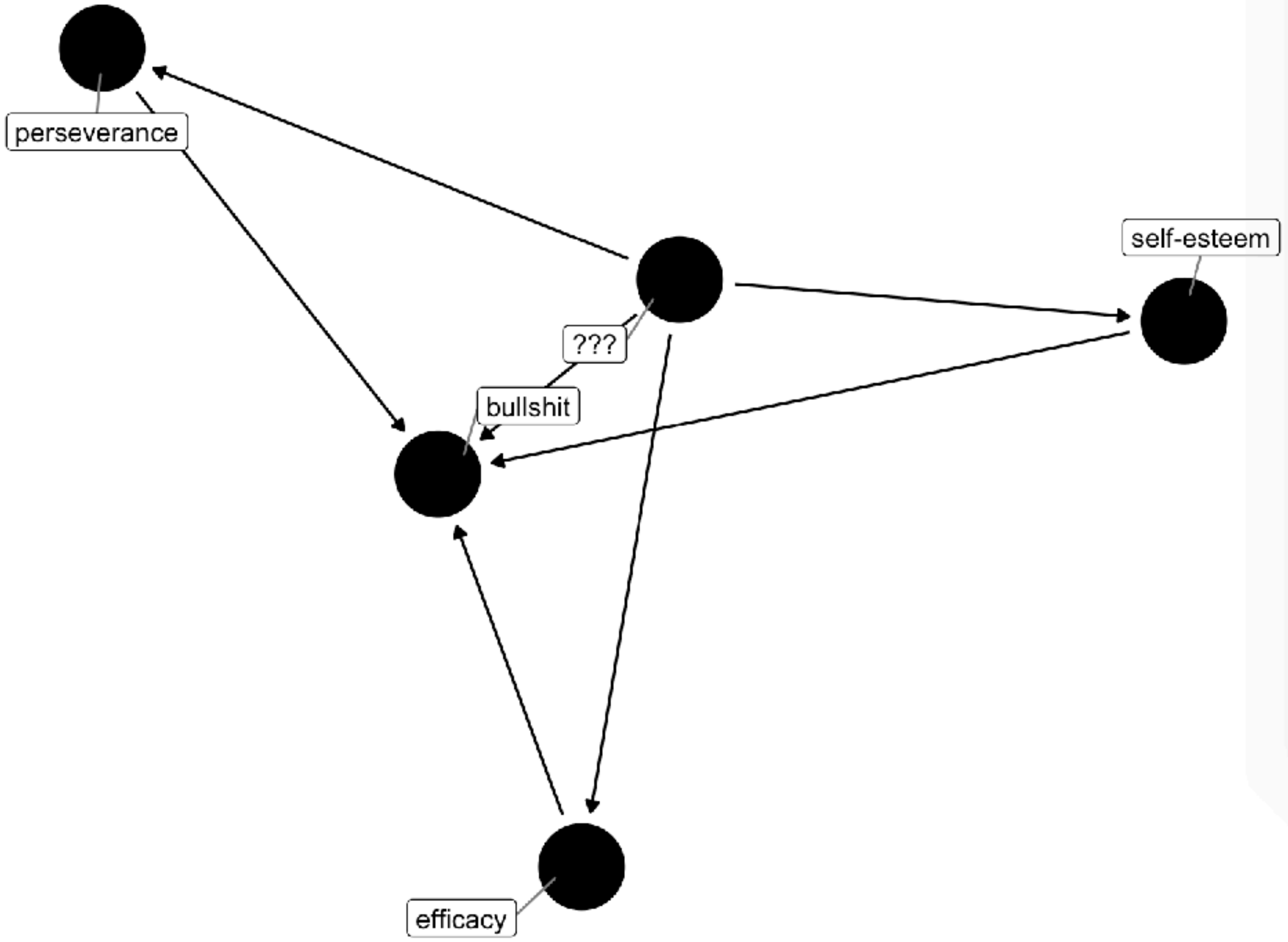
What might we need to control for?

Age

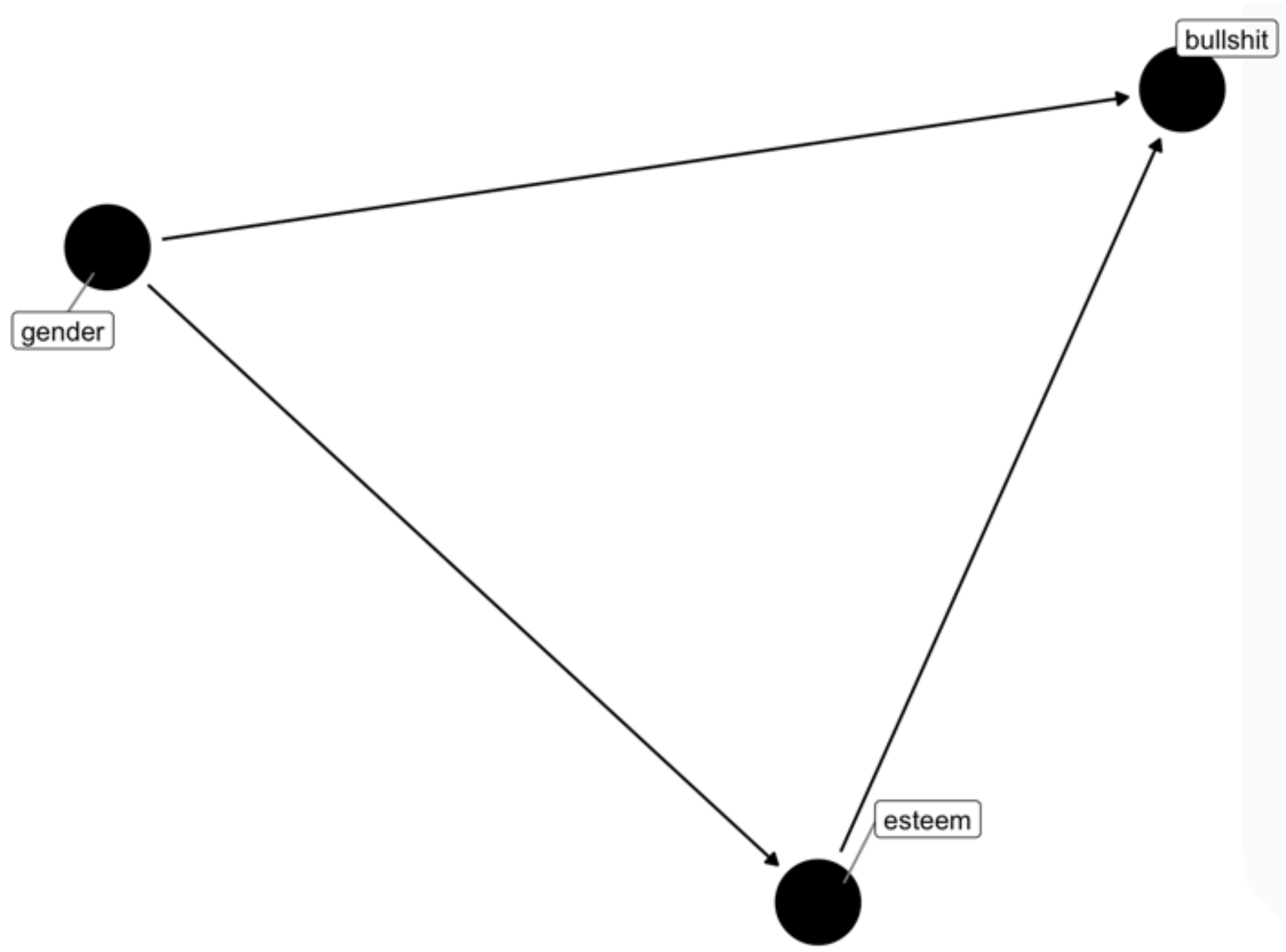
Wealth

Culture

Gender



Gender, bullshit, esteem

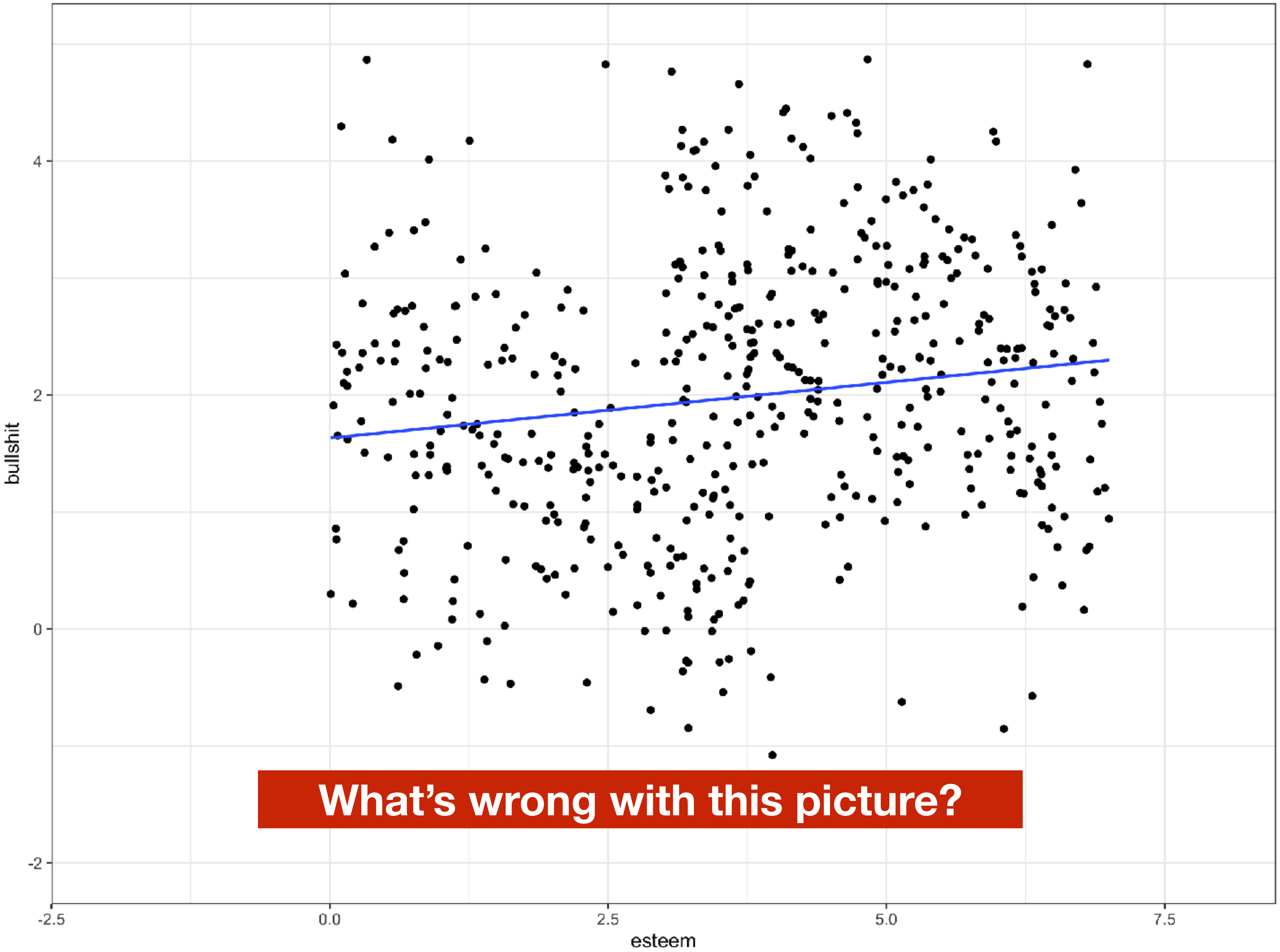


Let's make up some data where we already know the results

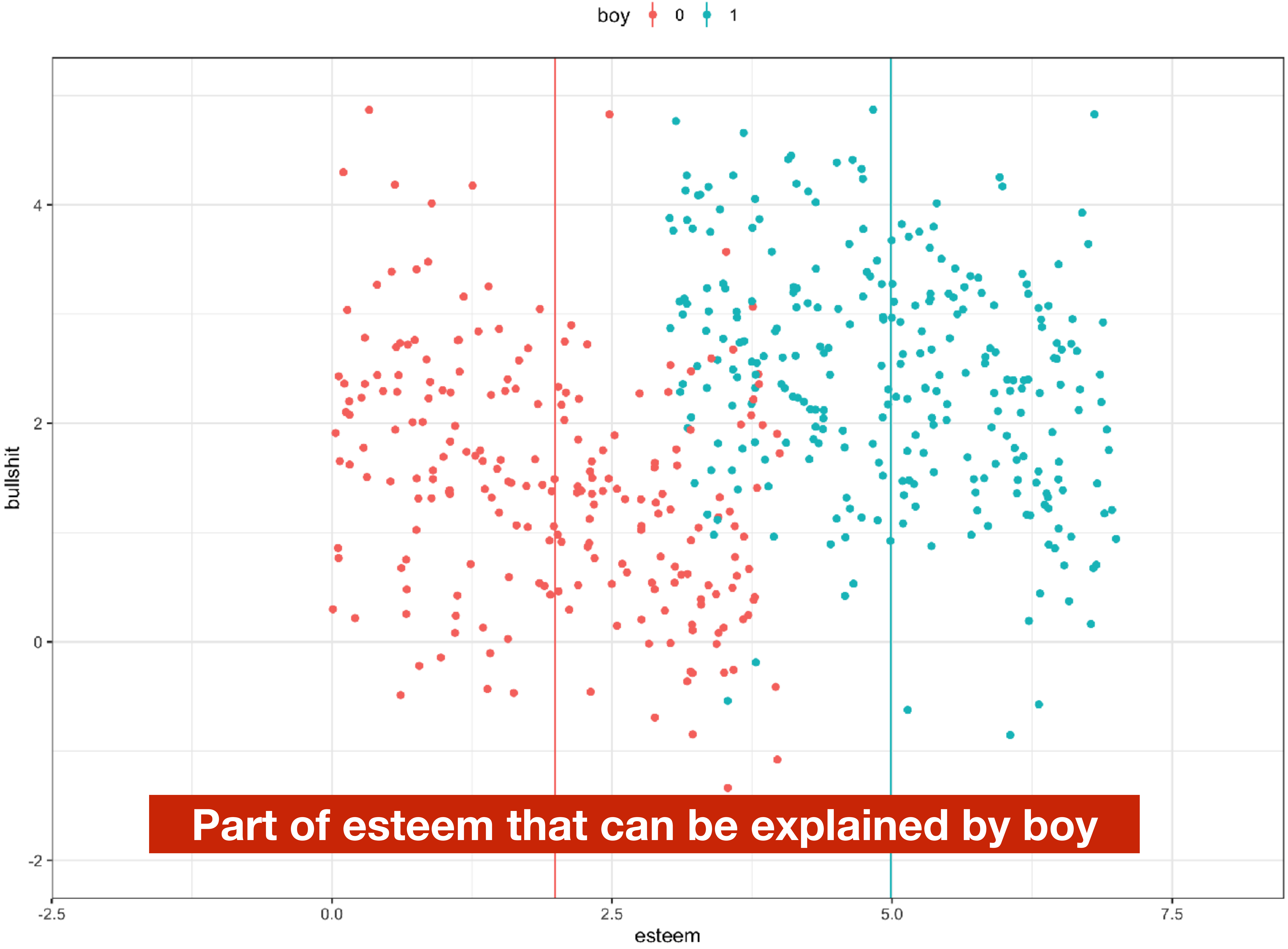
The truth is that people with higher, self-esteem on average bullshit less (.3 less in fact)

Boys have, on average, 3 higher score on self-esteem (out of 4)

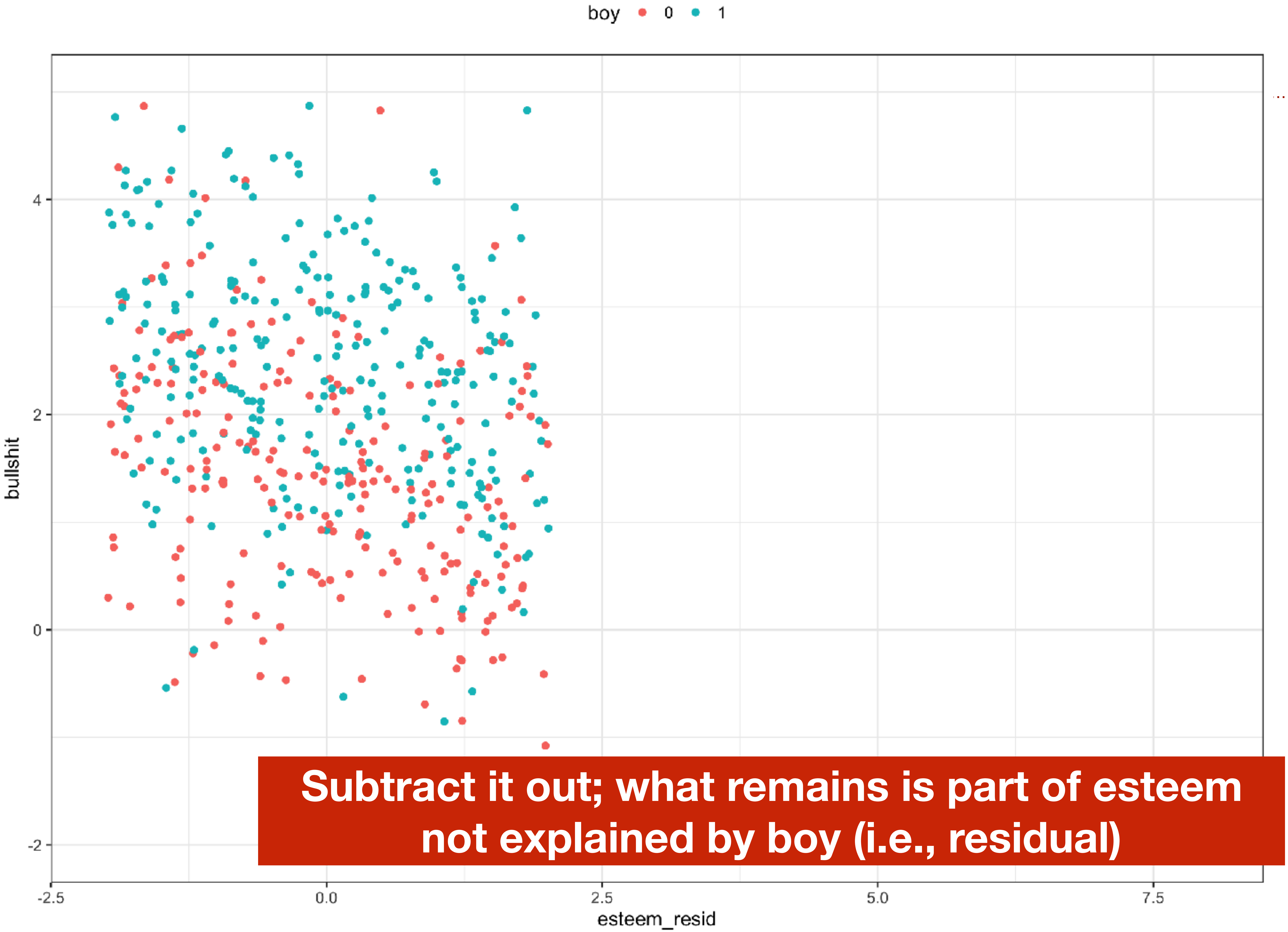
Boys bullshit, on average, 2 points more



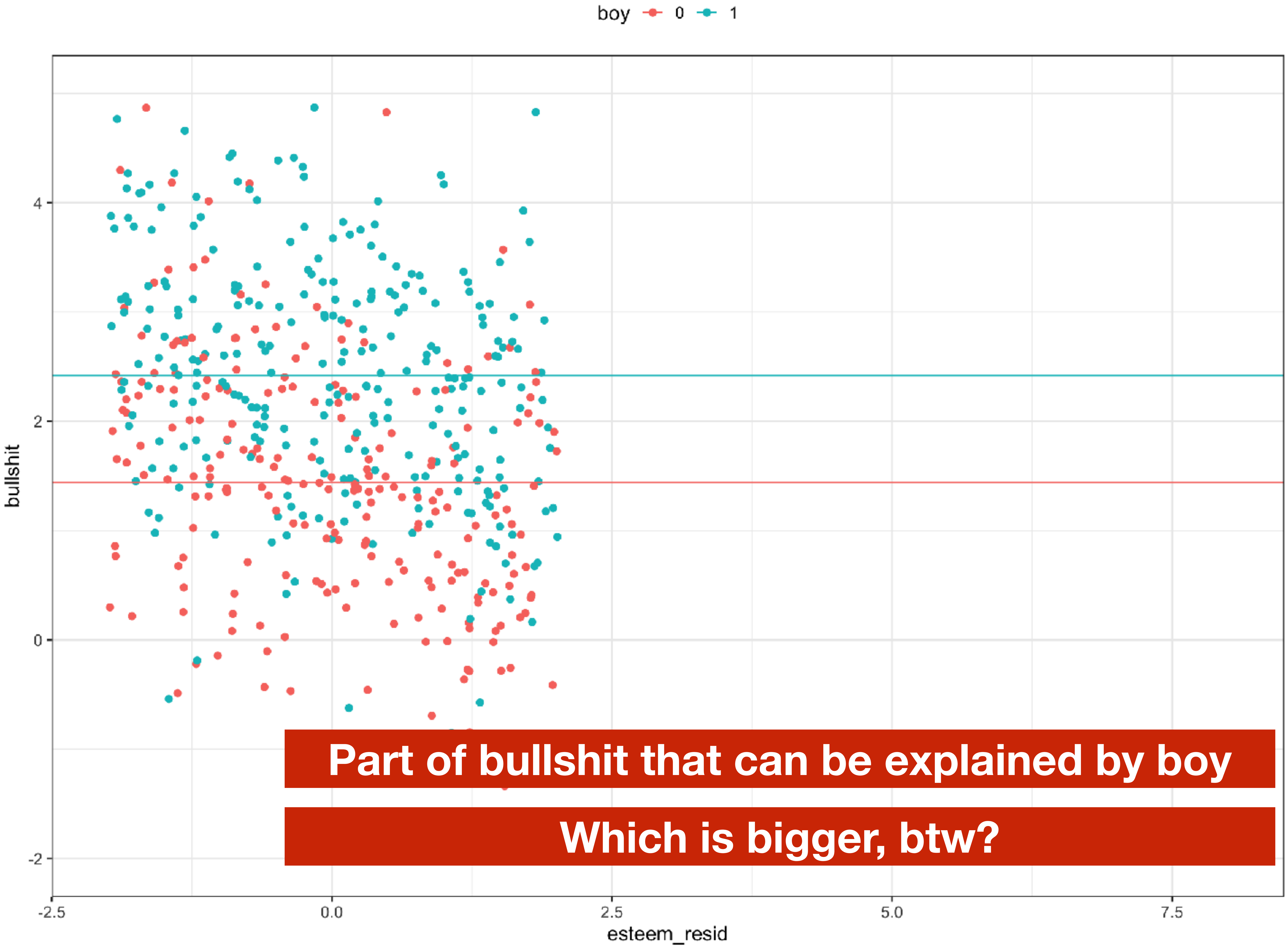
What’s wrong with this picture?



Part of esteem that can be explained by boy

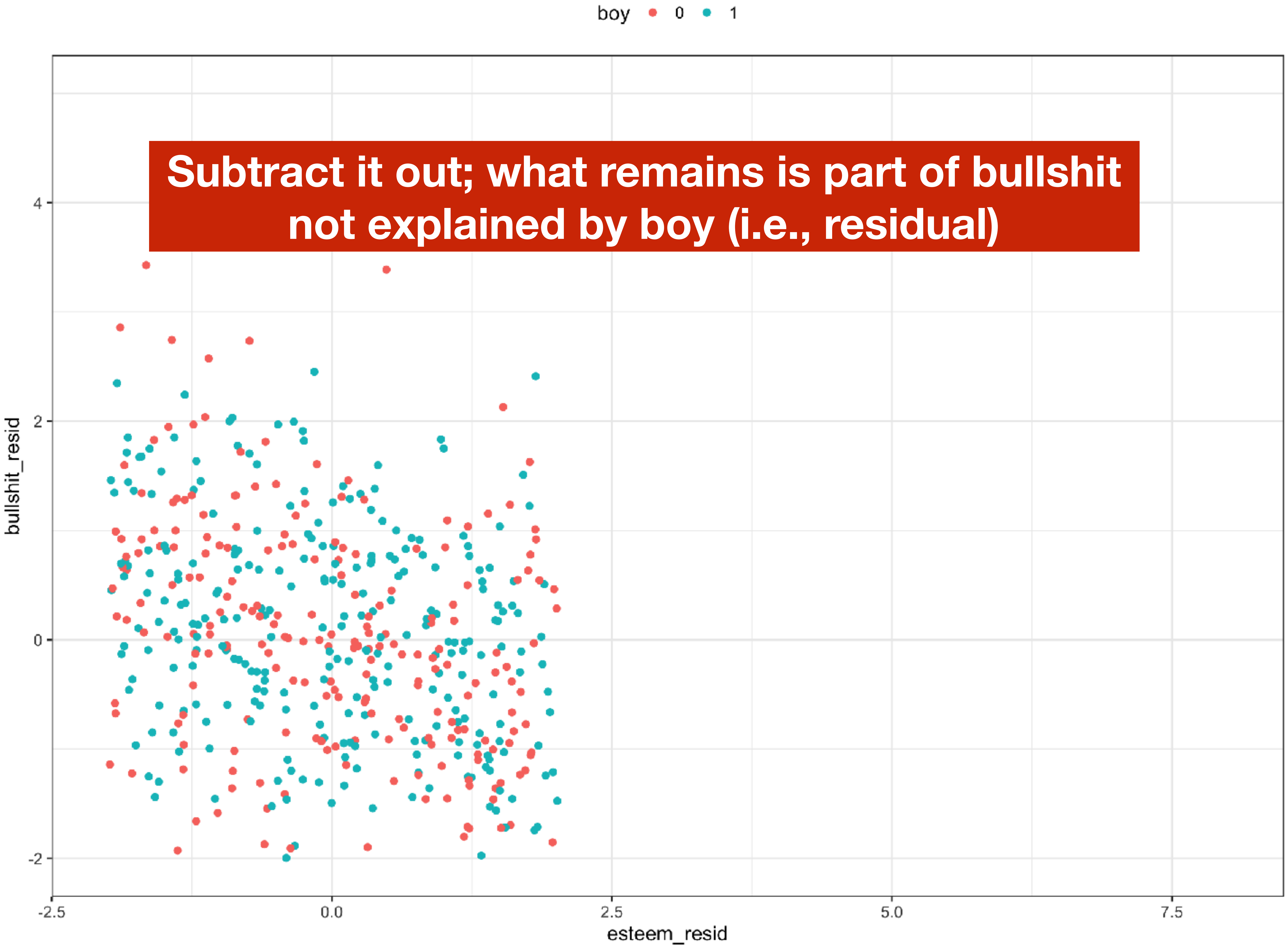


Subtract it out; what remains is part of esteem not explained by boy (i.e., residual)

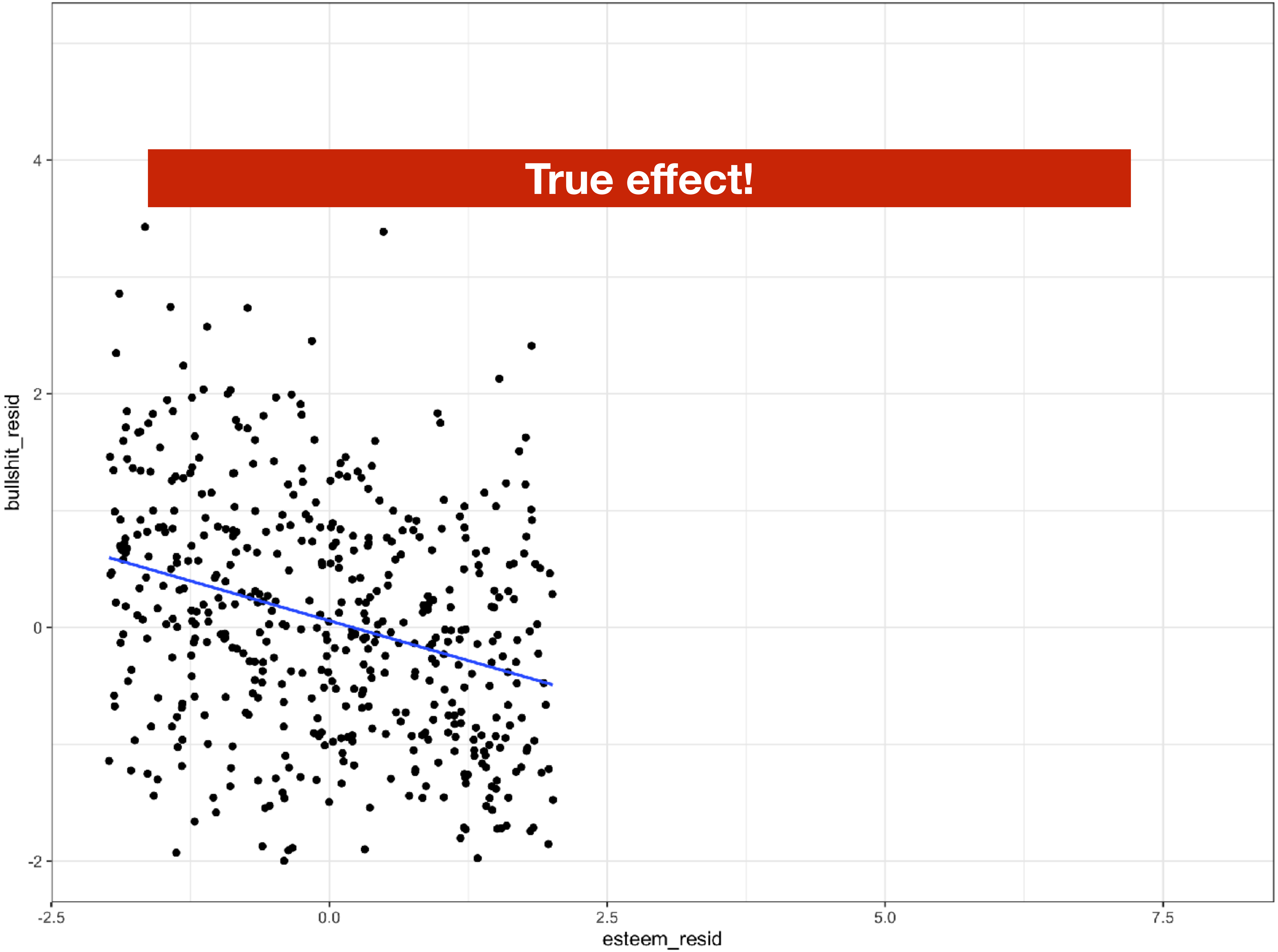


Part of bullshit that can be explained by boy

Which is bigger, btw?



Subtract it out; what remains is part of bullshit not explained by boy (i.e., residual)



True effect!

Follow along in R

More practice

Install “socviz”

Look at ?gss_sm

You want to explain who voted for Obama in 2012

obama coded (0/1)

This means coefficients = percent change in probability of voting for Obama!

Should use logit instead of OLS but that’s a different class

Pick a variable, regress against Obama, interpret

Pick a new variable, regress against Obama, interpret

Discussion

Did coefficients change much? Examples?

Does this mean we are de-confounding some true relationship?

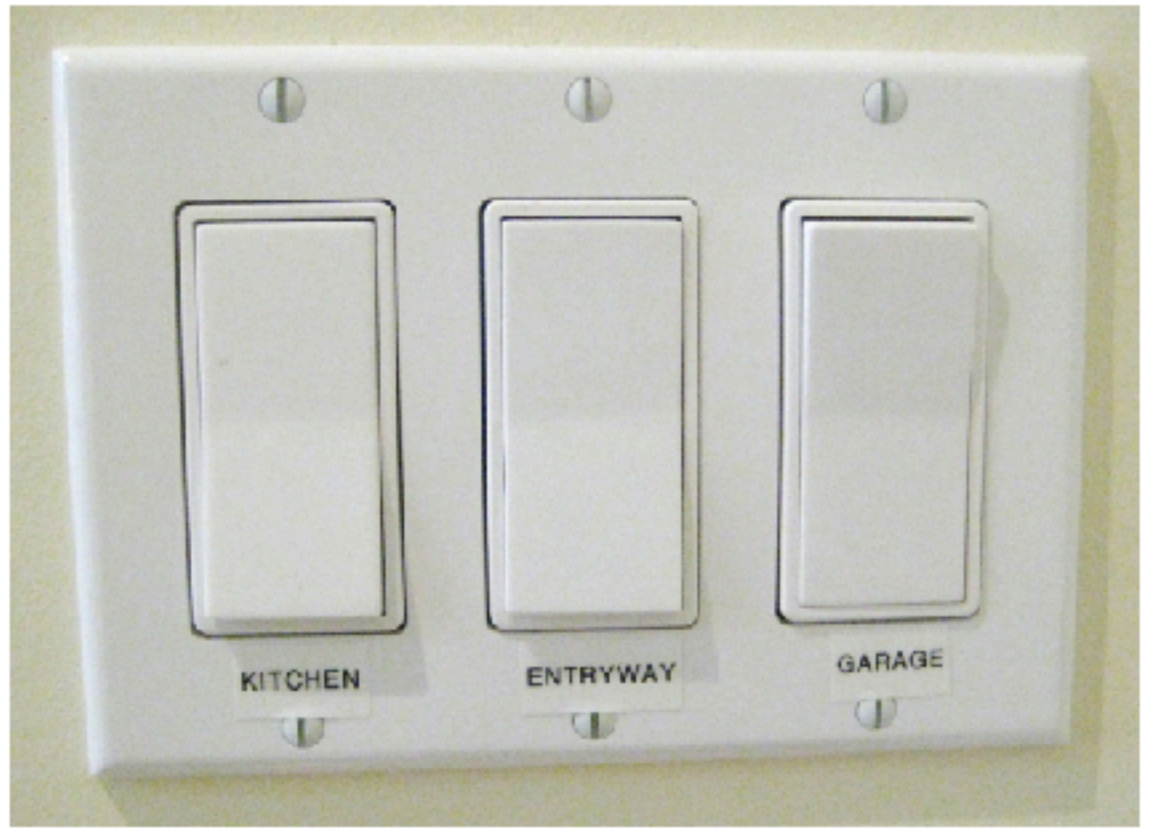
No! You add X’s to lm() and R will find out what part of Y can be explained by X

R is just following orders; you are *telling* R that X1, X2, etc. all belong in the model

Remember

categorical variables are switches

What is the beta coefficient on “boy”?



numeric variables are sliders

What is the beta coefficient on “esteem”?

One last things: prediction

You can use the *augment* function from the package *broom* to "predict" new observations

Let's you look at predicted, or "fitted", values based on equation

$$\text{bullshit} = 2.04 - 0.3(\text{esteem}) + 1.88(\text{boy}) + \epsilon$$

See class code