REGRESSION DIAGNOSTICS AND PREDICTIONS

MPA 630: Data Science for Public Management

October 25, 2018

Fill out your reading report on Learning Suite

PLAN FOR TODAY

Miscellanea

What does it mean to control for things?

How do we know if a model is good?

Interpretation practice

Making predictions

MISCELLANEA

UPCOMING THINGS

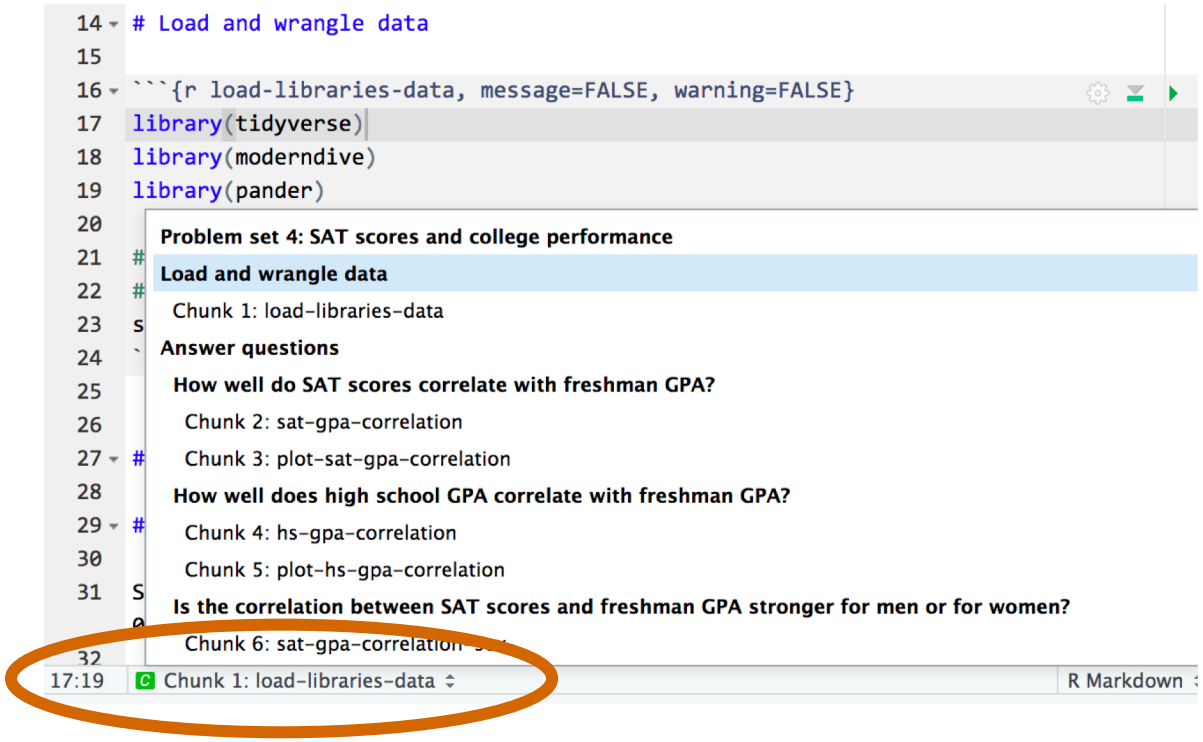
Problem set 4

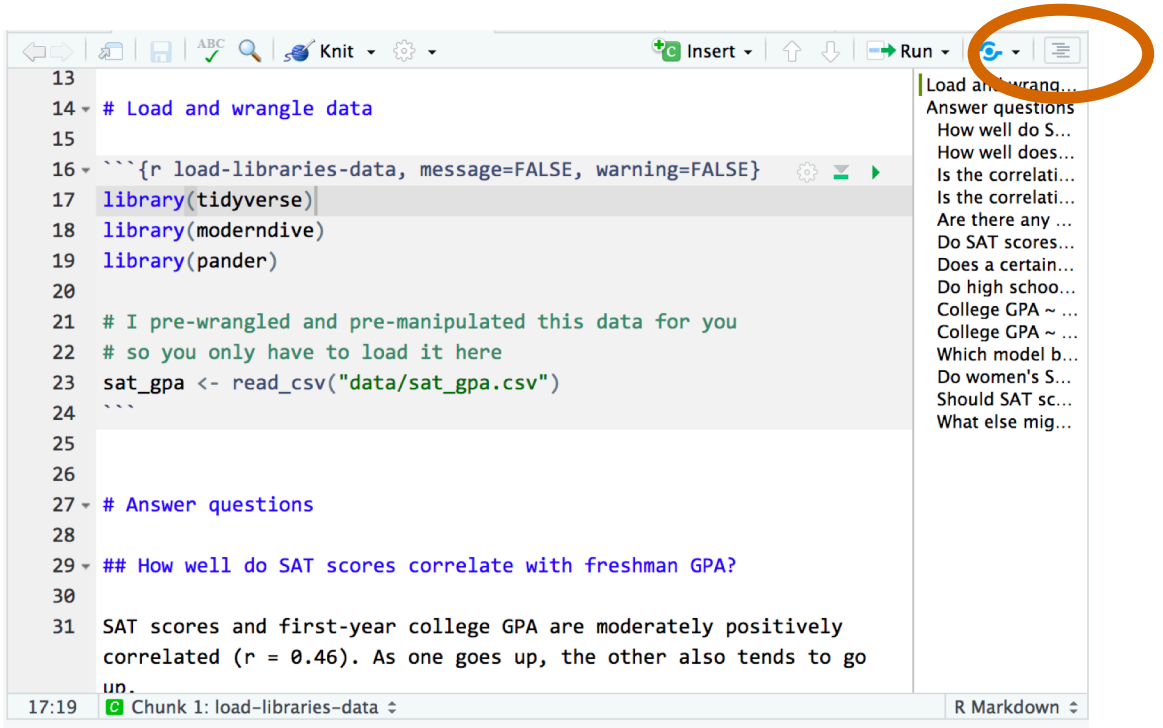
Exam 2

Final project

Code-through

NAVIGATING R MARKDOWN



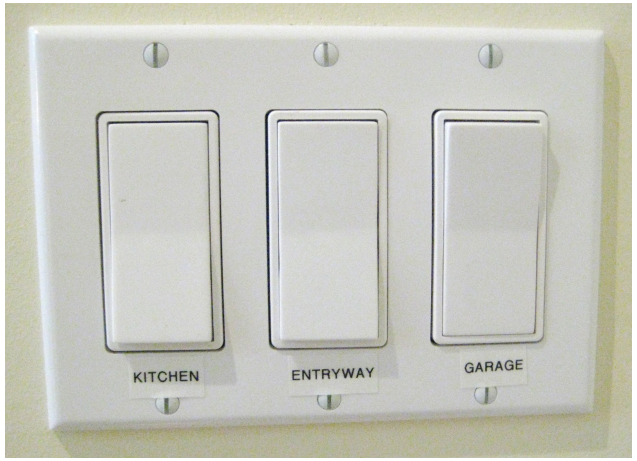


Dollar signs

WHAT DOES IT MEAN TO CONTROL FOR THINGS?

SLIDERS AND SWITCHES

$$\widehat{\text{property taxes}} = \beta_0 + \beta_1 \text{home values} + \epsilon$$



$$\widehat{\text{property taxes}} = \beta_0 + \beta_1 \text{California} +$$
$$\beta_2 \text{Idaho} + \beta_3 \text{Nevada} + \beta_4 \text{Utah} + \epsilon$$

ALL AT ONCE!

$$\widehat{\text{property taxes}} = \beta_0 + \beta_1 \text{home values} + \beta_2 \% \text{ houses with kids} +$$
$$\beta_3 \text{California} + \beta_4 \text{Idaho} + \beta_4 \text{Nevada} + \beta_6 \text{Utah} + \epsilon$$

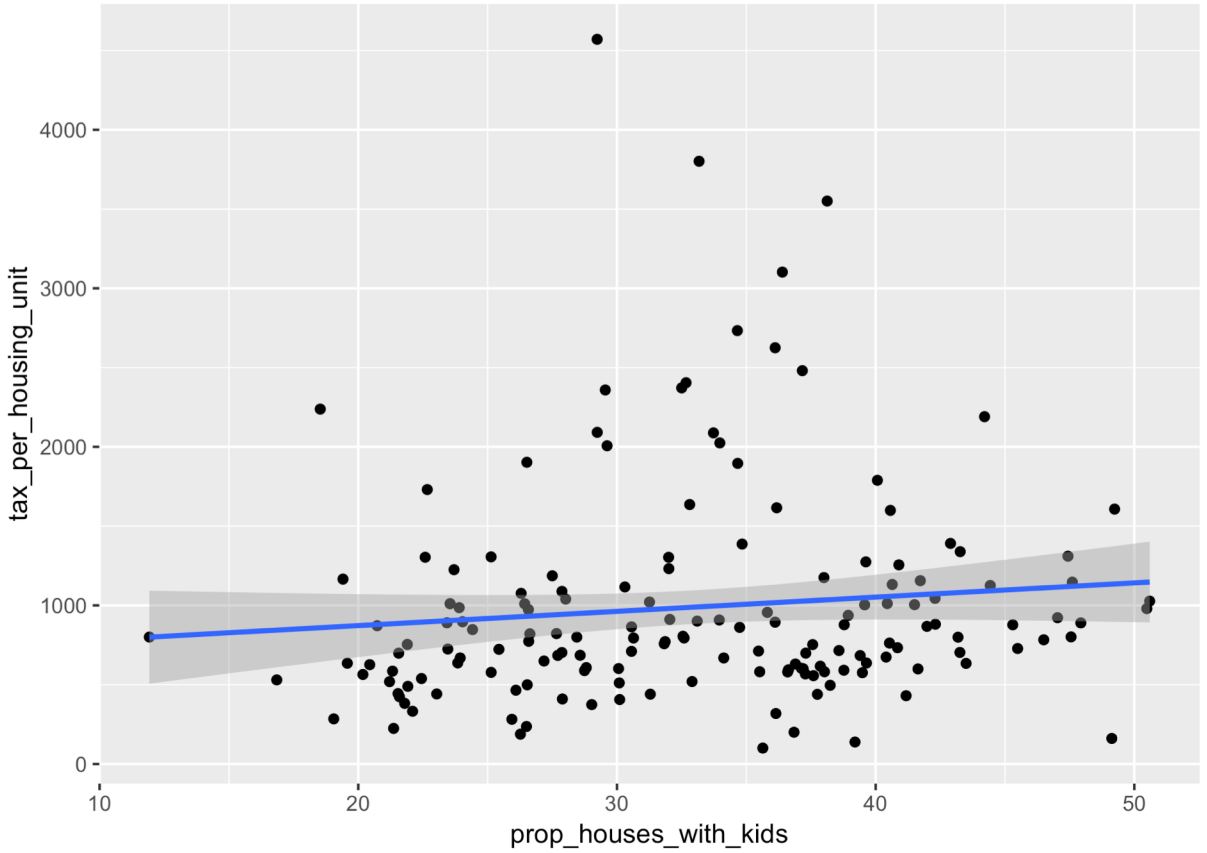
FILTERING OUT VARIATION

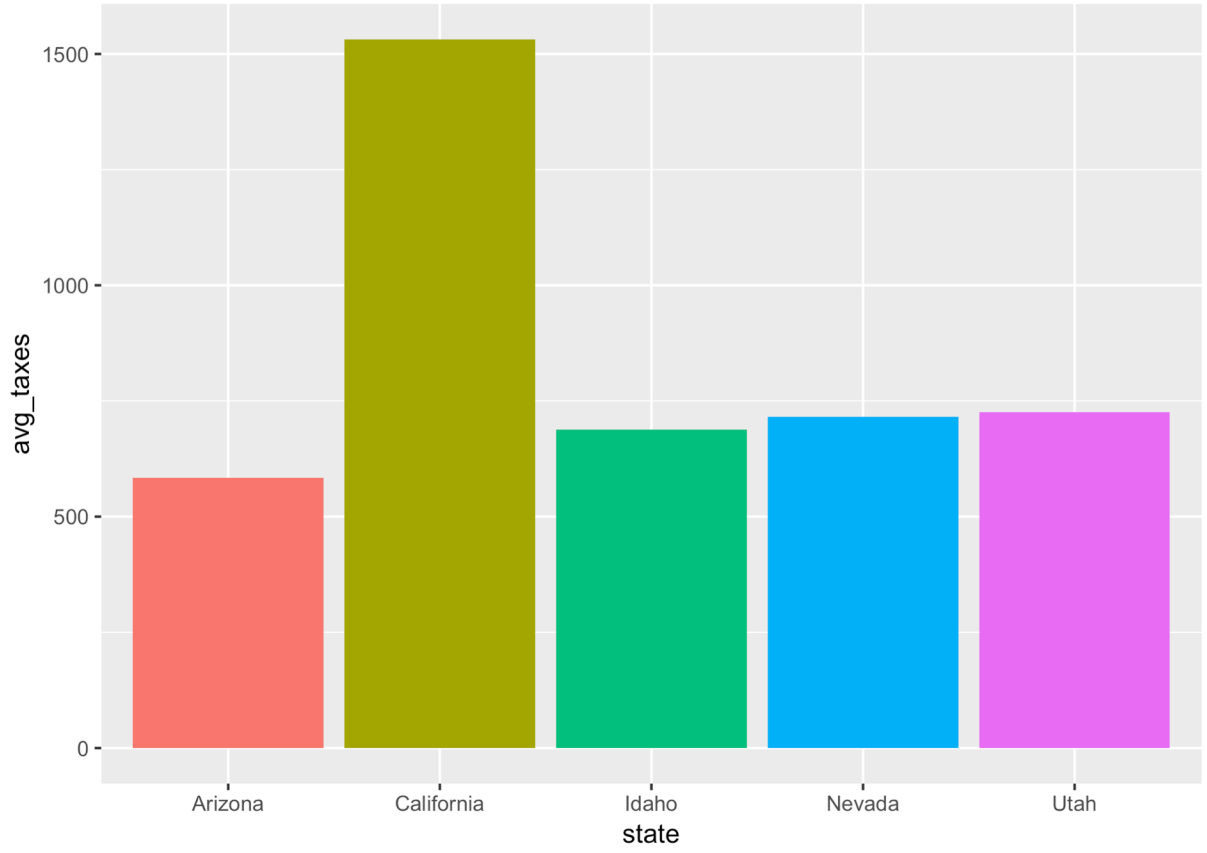
Each x in the model explains some portion of the variation in y

This will often change the simple regression coefficients

Interpretation is a little trickier, since you can only ever move **one** switch or slider (or variable)

TAXES ~ KIDS & TAXES ~ STATE





BOTH AT THE SAME TIME

Kids and states both explain some variation in property tax rates

On its own, a 1% increase in the number of households with kids in them is associated with a $X increase in per-household taxes, on average

On its own, being in State X is associated with $X higher/lower per-household property taxes compared to Arizona, on average

Some of that explanation is shared!

WHY CONTROL?

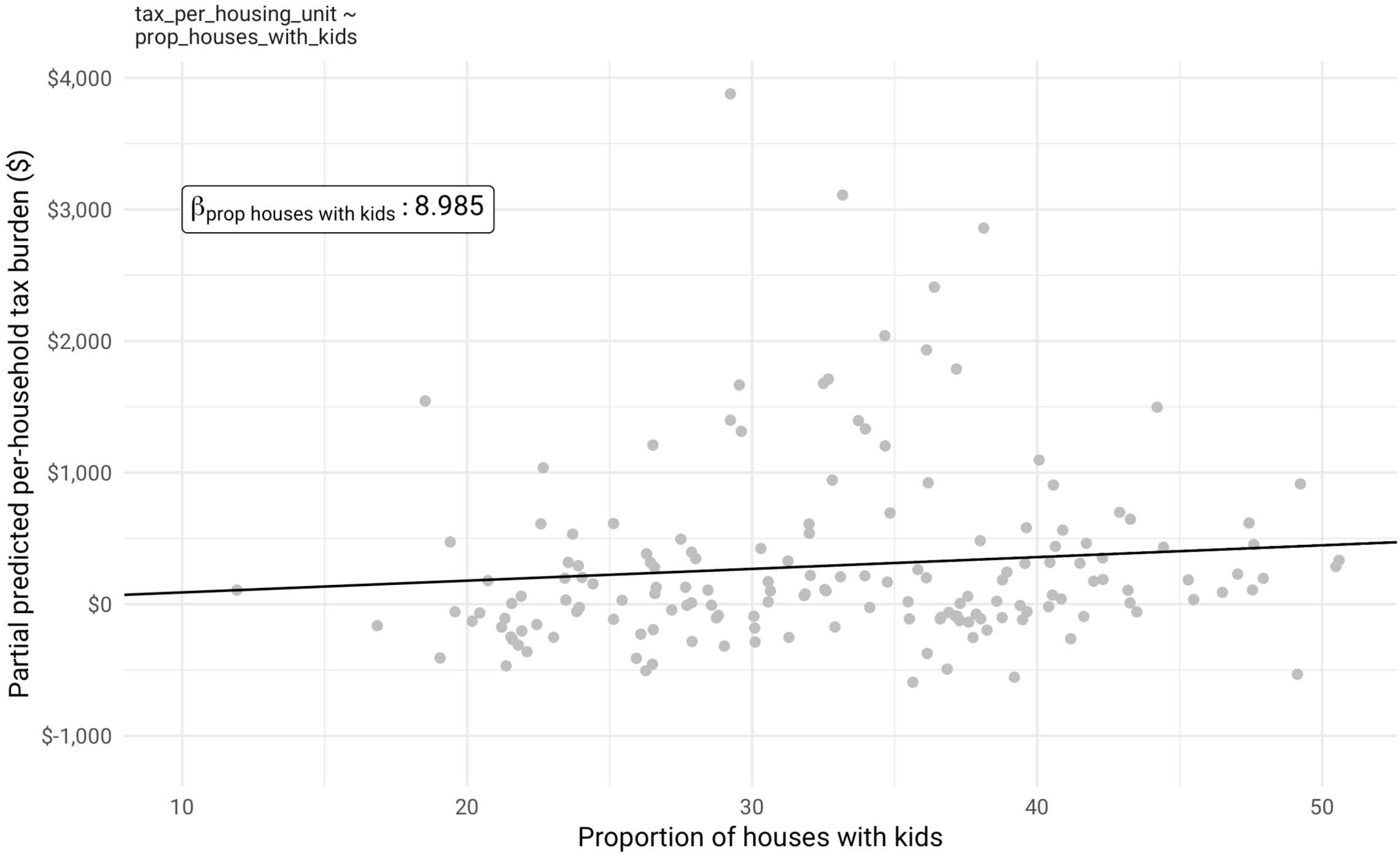
“Taking into account” or “controlling for” essentially means filtering out the effects of other variables

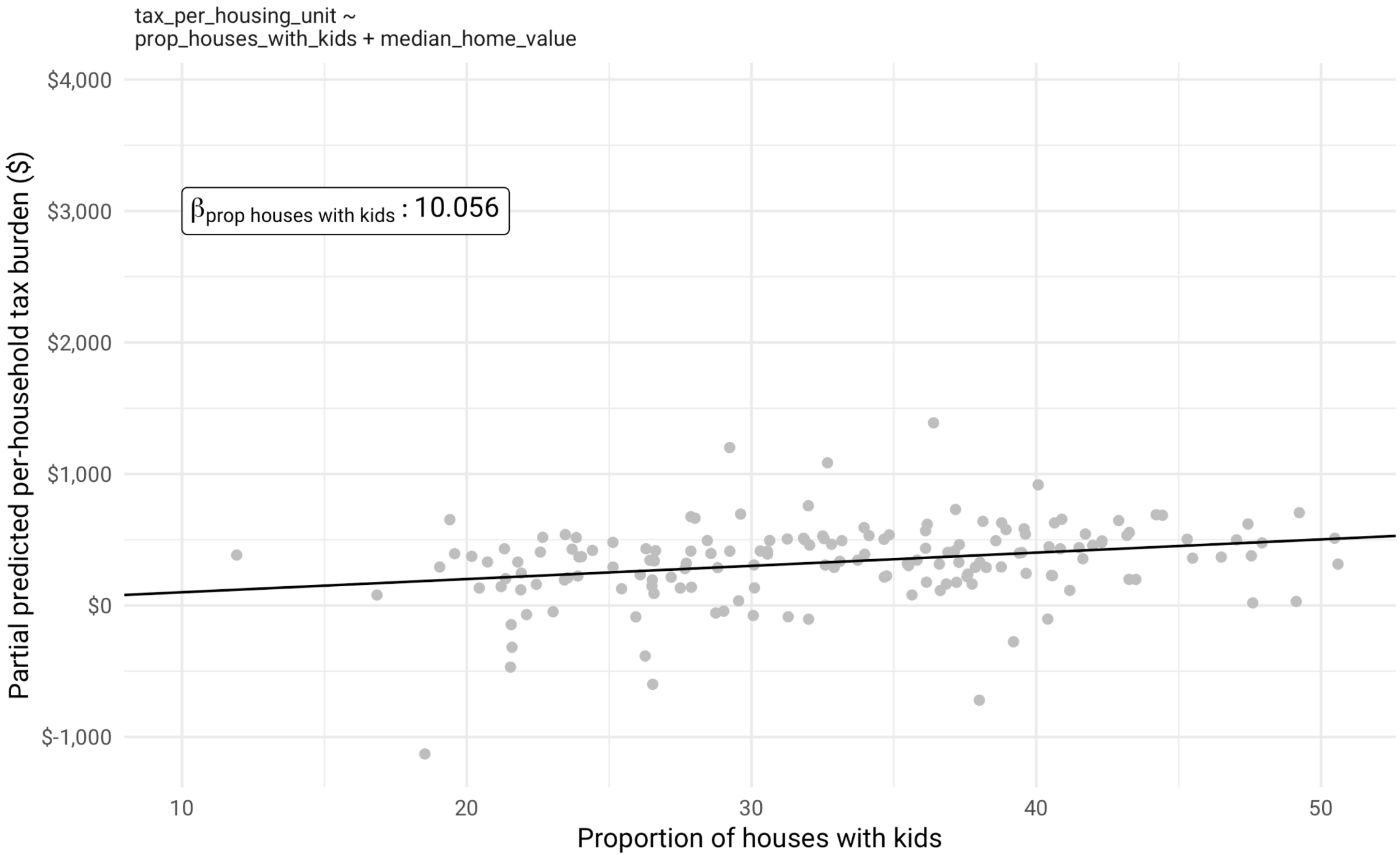
It lets you isolate the effect of specific levers/switches/sliders/Xs

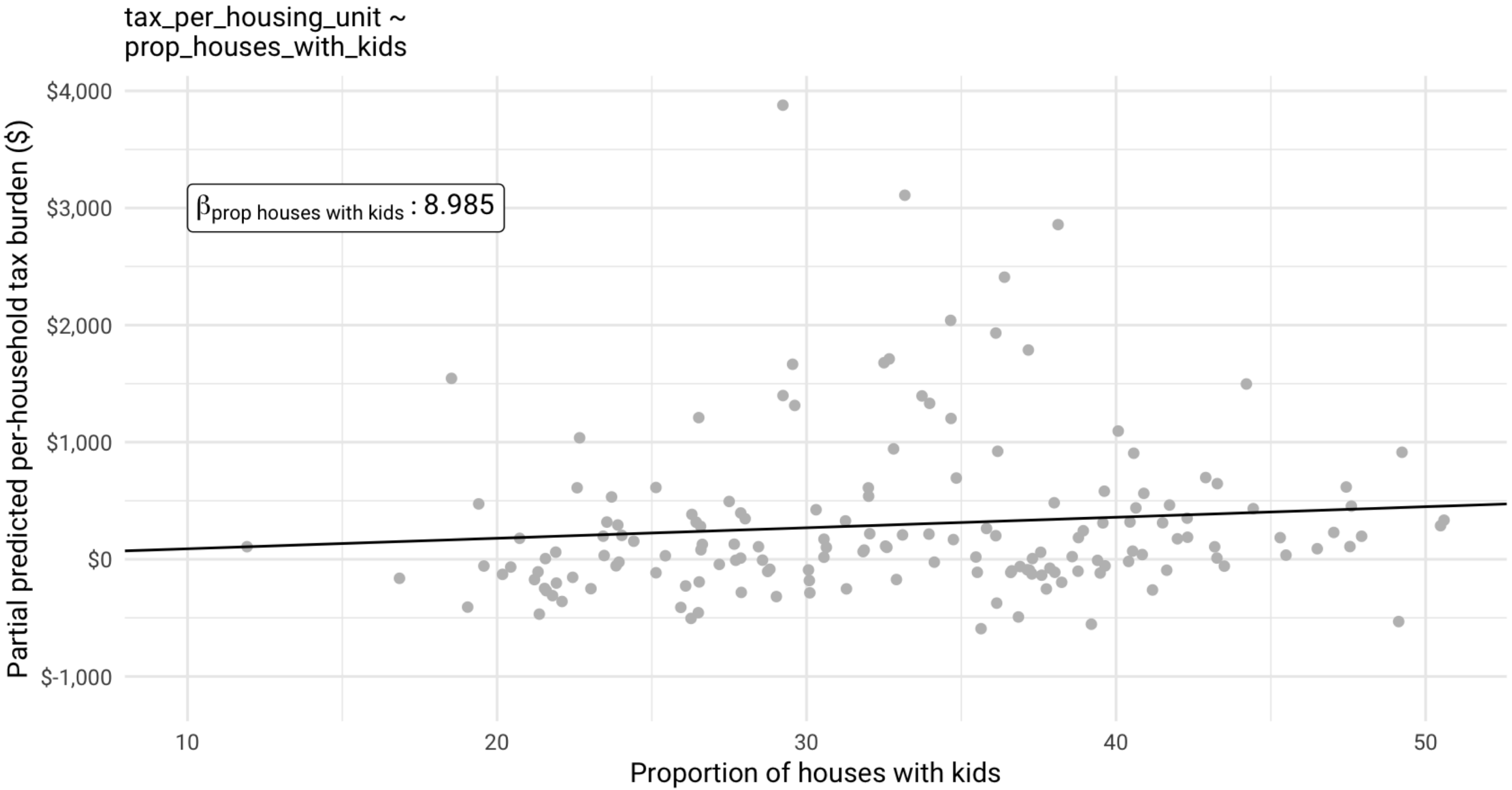
```
model4 <- lm(tax_per_housing_unit ~
               median_home_value + prop_houses_with_kids + state,
             data = world_happiness)
```

term	estimate	std_error	statistic	p_value
intercept	-412.5	118.1	-3.493	0.001
median_home_value	0.004	0	21.99	0
prop_houses_with_kids	14.09	2.853	4.941	0
stateCalifornia	123.3	88.22	1.397	0.164
stateIdaho	9.526	82.74	0.115	0.908
stateNevada	102.5	98.25	1.043	0.299
stateUtah	-213.2	91.21	-2.337	0.021

Utah has high per capita taxes compared to the other states in the region. If we control for the number of households with kids, though, Utah is actually substantially undertaxed. Lots of the reason that Utah's taxes are so high is because there are so many kids.


tax_per_housing_unit ~
prop_houses_with_kids
$\beta_{\text{prop houses with kids}}$: 8.985
Partial predicted per-household tax burden ($)
$4,000
$3,000
$2,000
$1,000
$0
$-1,000
10
20
30
40
50
Proportion of houses with kids


tax_per_housing_unit ~
prop_houses_with_kids + median_home_value
$\beta_{\text{prop houses with kids}}$: 10.056
$4,000
$3,000
$2,000
$1,000
$0
$-1,000
Partial predicted per-household tax burden ($)
10
20
30
40
50
Proportion of houses with kids


tax_per_housing_unit ~
prop_houses_with_kids
$\beta_{\text{prop houses with kids}}$: 8.985
Partial predicted per-household tax burden ($)
$4,000
$3,000
$2,000
$1,000
$0
$-1,000
10
20
30
40
50
Proportion of houses with kids

HOW DO WE KNOW IF A MODEL IS GOOD?

Or, how do we know what to control for?

WHICH VARIABLES TO INCLUDE?

Explanation

Your goal is to explain what specific levers (Xs) do to Y.

You need to have some theoretical reason to include each variable.

Prediction

Your goal is to make the best prediction of Y.

Include whatever

Basically

WHAT COUNTS AS "BEST"?

R^2

How much variation in Y is explained by X

0–1 scale; represents %

Higher = better fit

TEMPLATE FOR R^2

This model explains X% of the variation in Y

HOW TO FIND IT

```
model1 <- lm(tax_per_housing_unit ~ prop_houses_with_kids,
             data = taxes)
get_regression_summaries(model1)
```

r_squared	adj_r_squared	mse	rmse	sigma	statistic	p_value	df
0.011	0.005	464890	681.8	686	1.851	0.176	2

CORRELATION AND R^2

Remember how the letter for correlation is r?

This is the same r!

$R^2 = correlation^2$

LIMITS OF R^2

Correlation only works for y ~ x

What happens when a model has multiple Xs?

We can't use the regular R^2

ADJUSTED R^2

$$R^2_{adj} = R^2 \times \frac{\text{number of observations} - 1}{\text{number of observations} - \text{number of variables in model} - 1}$$

Almost always lowers the R^2

Penalizes you for small data and lots of variables

TEMPLATE FOR ADJUSTED R^2

This model explains X% of the variation in Y

HOW TO FIND IT

```
model5 <- lm(tax_per_housing_unit ~
               median_home_value + prop_houses_with_kids +
               median_income + population + state,
             data = taxes)
get_regression_summaries(model5)
```

r_squared	adj_r_squared	mse	rmse	sigma	statistic	p_value	df
0.854	0.846	68846	262.4	269.9	112.2	0	9

MODEL SELECTION

In general, the higher a model's adjusted R^2, the better its fit

R^2 is not the *best* measure for model fit, but it's good enough for this class. It's intuitive.

r_squared	adj_r_squared	mse	rmse	sigma	statistic	p_value	df
0.854	0.846	68846	262.4	269.9	112.2	0	9

logLik	AIC	BIC	deviance	df.residual
-1139	2298	2329	11221939	154

GENERAL GUIDELINES

If your model has one explanatory variable (x), use R^2

If your model has more than one explanatory variable (x), use the adjusted R^2

Higher is better

No magic threshold for good or bad number; depends on domain

	(1)	(2)	(3)	(4)	(5)
(Intercept)	692.926 **	583.392 ***	261.149	-412.485 ***	-595.561 ***
prop_houses_ with_kids	8.985		10.314	14.094 ***	9.934 **
stateCalifornia		948.197 ***	932.986 ***	123.282	160.820
stateIdaho		104.530	101.385	9.526	32.713
stateNevada		132.498	160.949	102.450	4.885
stateUtah		142.387	67.274	-213.191 *	-241.628 **
median_home_ value				0.004 ***	0.003 ***
median_income					0.010 **
population					0.000
N	163	163	163	163	163
R2	0.011	0.350	0.363	0.845	0.854
logLik	-1294.826	-1260.678	-1259.023	-1144.053	-1139.167
AIC	2595.652	2533.357	2532.046	2304.105	2298.334

CHOOSING VARIABLES

Forwards

Add variables 1–2 at a time and see if they help or hurt

Better for explanatory work where you care about the x variables

Backwards

Start with a kitchen sink model, remove unhelpful variables

Better for predictive work where you don't care about the x variables

```
step(name_of_giant_model)
```

INTERPRETATION PRACTICE

ELECTIONS

2016

Clinton vs. **Trump**

Brexit

Stay vs. **Leave**

Labour

UKIP

FOLLOW ALONG IN R

MAKING PREDICTIONS

HOW TO PREDICT

Plug in values for all the Xs, get a predicted Y

$$\widehat{\text{property taxes}} = \beta_0 + \beta_1 \text{home values} + \beta_2 \% \text{ houses with kids} + \beta_3 \text{California} + \beta_4 \text{Idaho} + \beta_4 \text{Nevada} + \beta_6 \text{Utah} + \epsilon$$

term	estimate	std_error	statistic	p_value
intercept	-412.5	118.1	-3.493	0.001
median_home_value	0.004	0	21.99	0
prop_houses_with_kids	14.09	2.853	4.941	0
stateCalifornia	123.3	88.22	1.397	0.164
stateIdaho	9.526	82.74	0.115	0.908
stateNevada	102.5	98.25	1.043	0.299
stateUtah	-213.2	91.21	-2.337	0.021

$$\widehat{\text{property taxes}} = -412.5 + (0.004 \times \text{median home value}) + (14.09 \times \% \text{ houses with kids}) + (123.3 \times \text{California}) + (9.526 \times \text{Idaho}) + (102.5 \times \text{Nevada}) + (-213.2 \times \text{Utah}) + \epsilon$$

What's the predicted median per-household property tax rate for a county in Nevada where the median home value is $155,000 and 30% of the houses have kids?

$$\widehat{\text{property taxes}} = -412.5 + (0.004 \times \text{median home value}) + (14.09 \times \% \text{ houses with kids}) + (123.3 \times \text{California}) + (9.526 \times \text{Idaho}) + (102.5 \times \text{Nevada}) + (-213.2 \times \text{Utah}) + \epsilon$$

$$\widehat{\text{property taxes}} = -412.5 + (0.004 \times 150,000) + (14.09 \times 30) + (123.3 \times 0) + (9.526 \times 0) + (102.5 \times 1) + (-213.2 \times 0) + \epsilon$$

$$\widehat{\text{property taxes}} = 741.04$$

```r
model_thing <- lm(tax_per_housing_unit ~
                    median_home_value + prop_houses_with_kids + state,
                  data = taxes)

imaginary_county <- data_frame(prop_houses_with_kids = 30,
                               median_home_value = 155000,
                               state = "Nevada")


predict(model_thing, imaginary_county)
#> 741.0414

predict(model_thing, imaginary_county, interval = "prediction")
#>    fit       lwr      upr
#> 1 741.0414 179.2417 1302.841
```