

Causal Modeling in R: Whole Game

Malcolm Barrett

2021-09-01

We'll focus on the broader ideas behind each step and what they look like all together; we don't expect you to fully digest each idea. We'll spend the rest of the workshop taking up each step in detail

Do people who quit smoking gain weight?

```
1 library(causaldata)
2 nhefs_complete_uc <- nhefs_complete |>
3   filter(censored == 0)
4 nhefs_complete_uc
```

A tibble: 1,566 × 67

	seqn	qsmk	death	yrdth	modth	dadth	sbp	dbp	sex
	<dbl>	<fct>							
1	233	0	0	NA	NA	NA	175	96	0
2	235	0	0	NA	NA	NA	123	80	0
3	244	0	0	NA	NA	NA	115	75	1
4	245	0	1	85	2	14	148	78	0
5	252	0	0	NA	NA	NA	118	77	0
6	257	0	0	NA	NA	NA	141	83	1
7	262	0	0	NA	NA	NA	132	69	1
8	266	0	0	NA	NA	NA	100	53	1
9	419	0	1	84	10	13	163	79	0
10	420	0	1	86	10	17	184	106	0
..

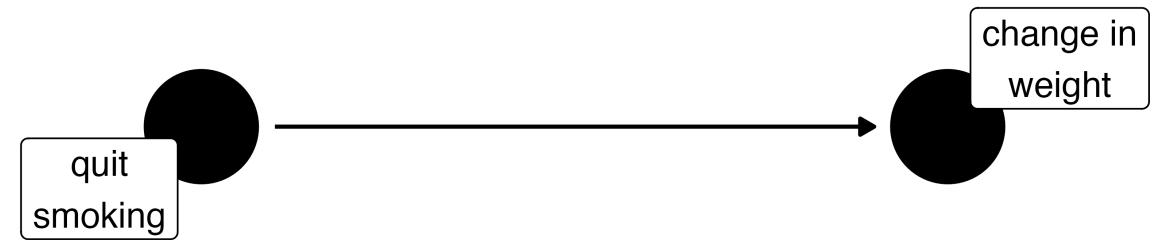
Did those who quit smoking gain weight?

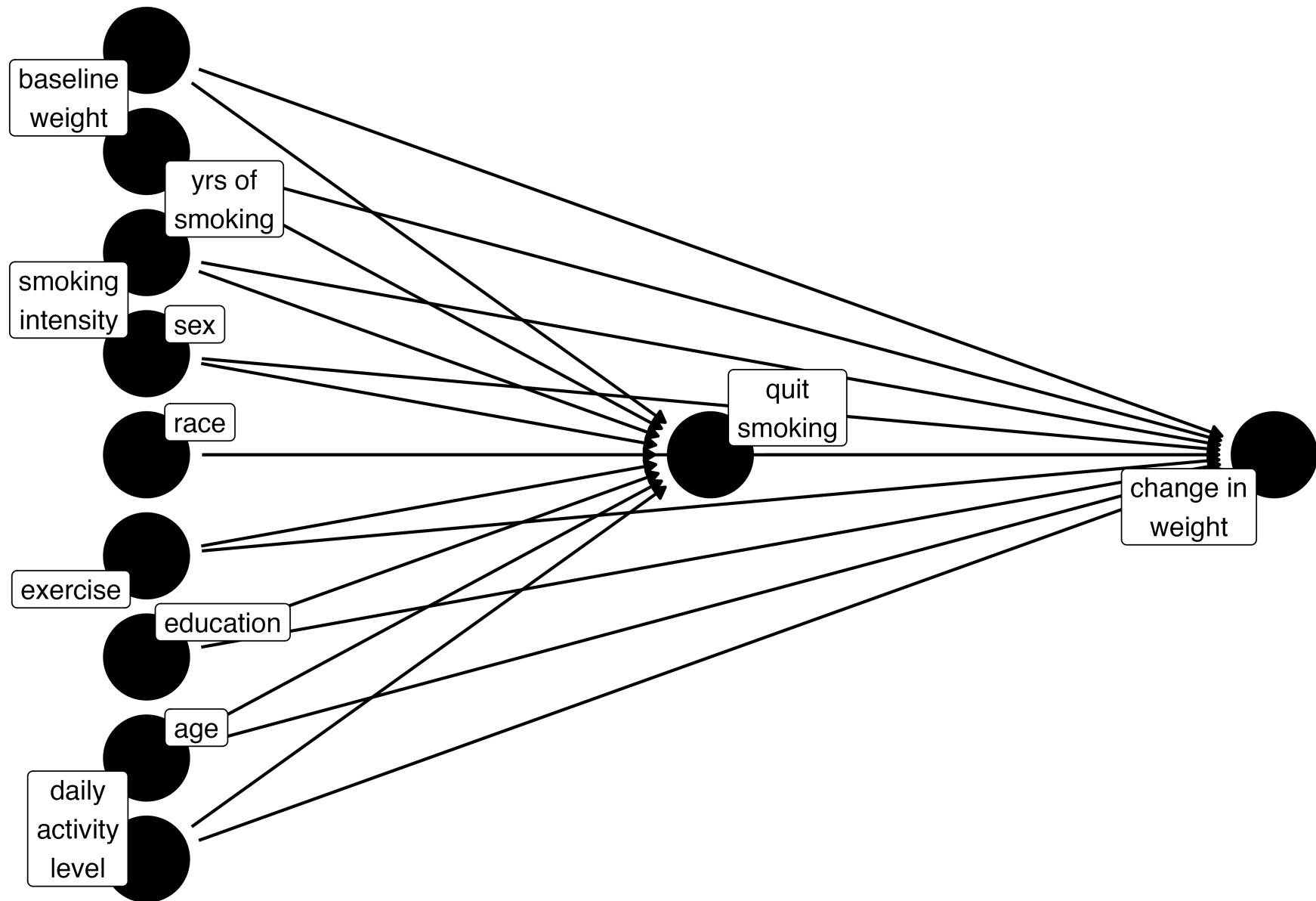
Did those who quit smoking gain weight?

```
1 # ~2.5 KGs gained for quit vs. not quit
2 nhefs_complete_uc |>
3   group_by(qsmk) |>
4   summarize(
5     mean_weight_change = mean(wt82_71),
6     sd = sd(wt82_71),
7     .groups = "drop"
8   )
```

```
# A tibble: 2 × 3
  qsmk mean_weight_change     sd
  <dbl>             <dbl>   <dbl>
1     0                 1.98    7.45
2     1                 4.53    8.75
```

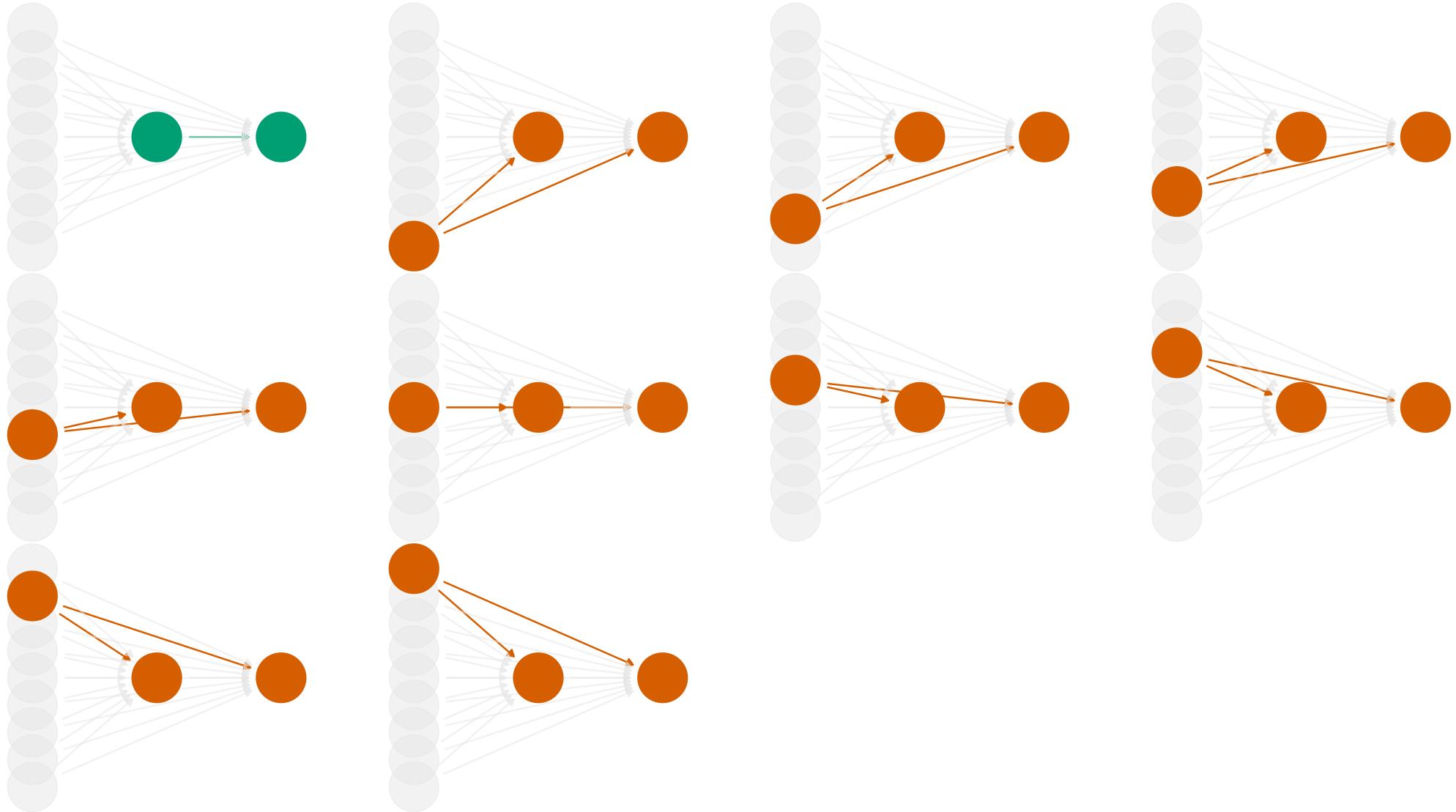
draw your assumptions



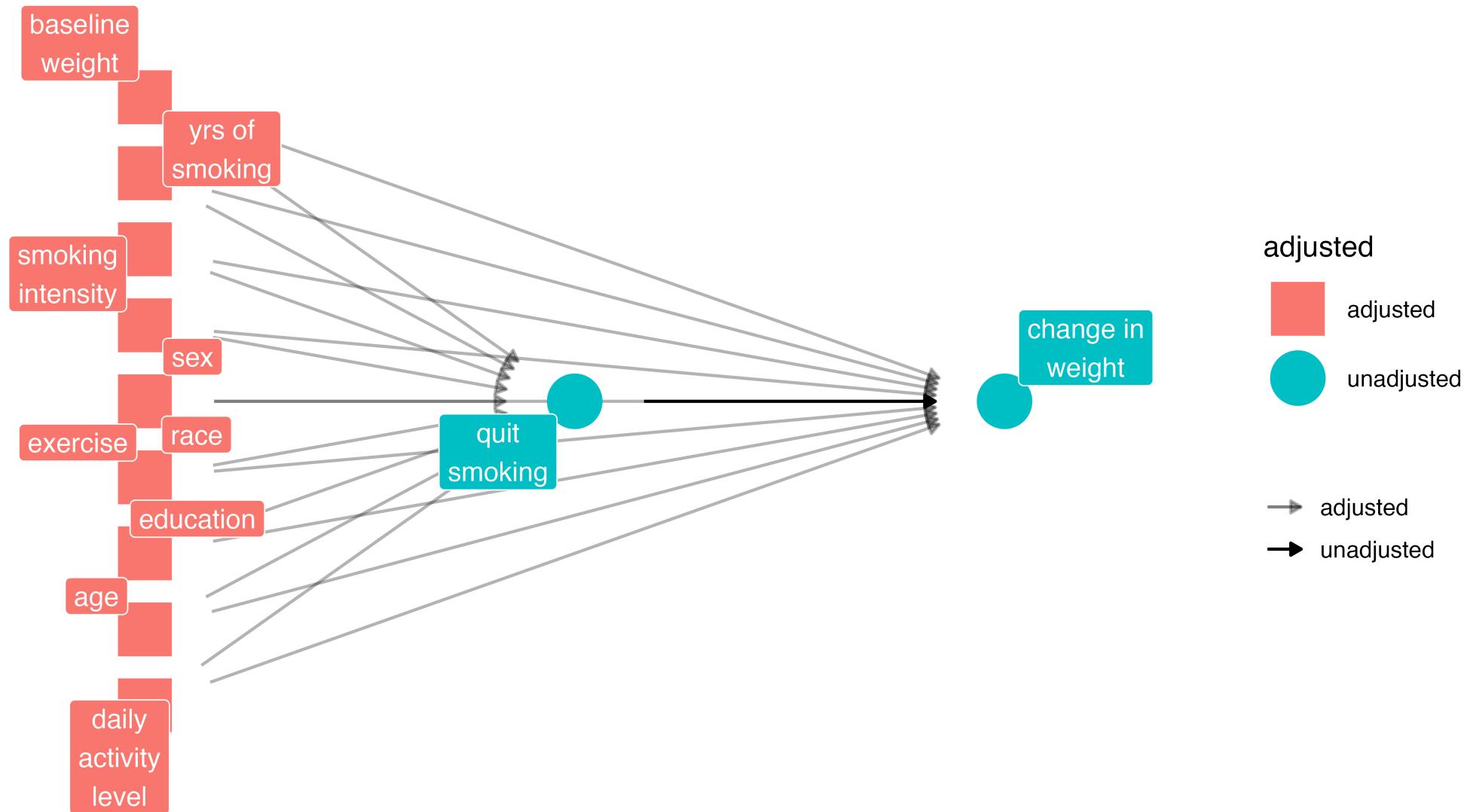


What do I need to control for?

■ true effect ■ confounding effect



{active, age, education, exercise,
race, sex, smokeintensity, smokeyrs,
wt71}



Multivariable regression: what's the association?

```
1 lm(  
2   wt82_71 ~ qsmk + sex +  
3     race + age + I(age^2) + education +  
4     smokeintensity + I(smokeintensity^2) +  
5     smokeyrs + I(smokeyrs^2) + exercise + active +  
6     wt71 + I(wt71^2),  
7   data = nhefs_complete_uc  
8 ) |>  
9 tidy(conf.int = TRUE) |>  
10 filter(term == "qsmk")
```

```
# A tibble: 1 × 7  
  term estimate std.error statistic p.value conf.low conf.high  
  <chr>    <dbl>     <dbl>     <dbl>    <dbl>    <dbl>    <dbl>  
1 qsmk      3.46      0.438      7.90 5.36e-15      2.60  
# i 1 more variable: conf.high <dbl>
```

model your assumptions

**counterfactual: what if everyone quit smoking
vs. what if no one quit smoking**

Fit propensity score model

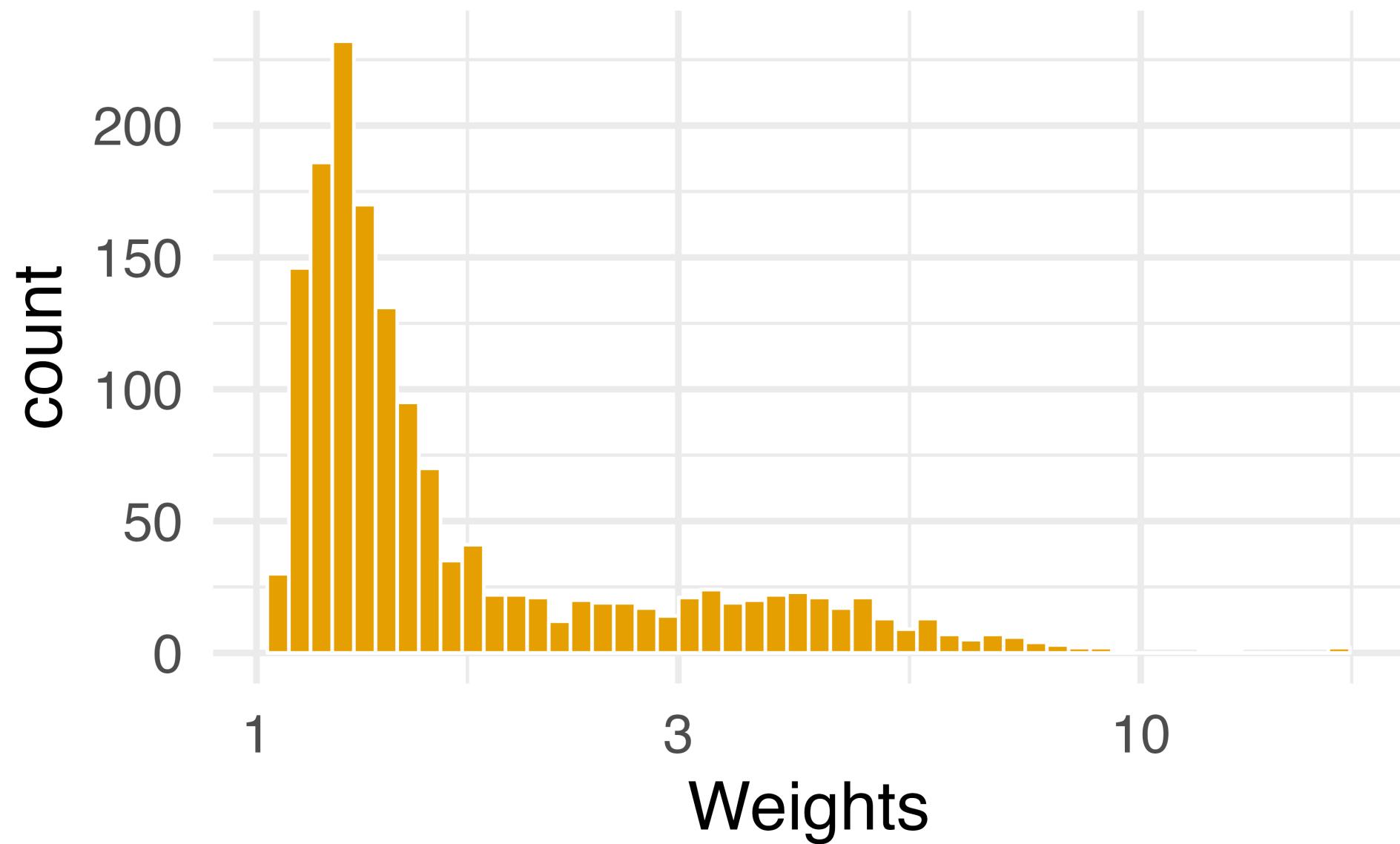
```
1 propensity_model <- glm(  
2   qsmk ~ sex +  
3     race + age + I(age^2) + education +  
4     smokeintensity + I(smokeintensity^2) +  
5     smokeyrs + I(smokeyrs^2) + exercise + active +  
6     wt71 + I(wt71^2),  
7   family = binomial(),  
8   data = nhefs_complete_uc  
9 )
```

Calculate inverse probability weights

```
1 library(propensity)
2 nhefs_complete_uc <- propensity_model |>
3   # predict whether quit smoking
4   augment(type.predict = "response", data = nhefs_complete_uc) |>
5   # calculate inverse probability
6   mutate(wts = wt_ate(.fitted, qsmk))
```

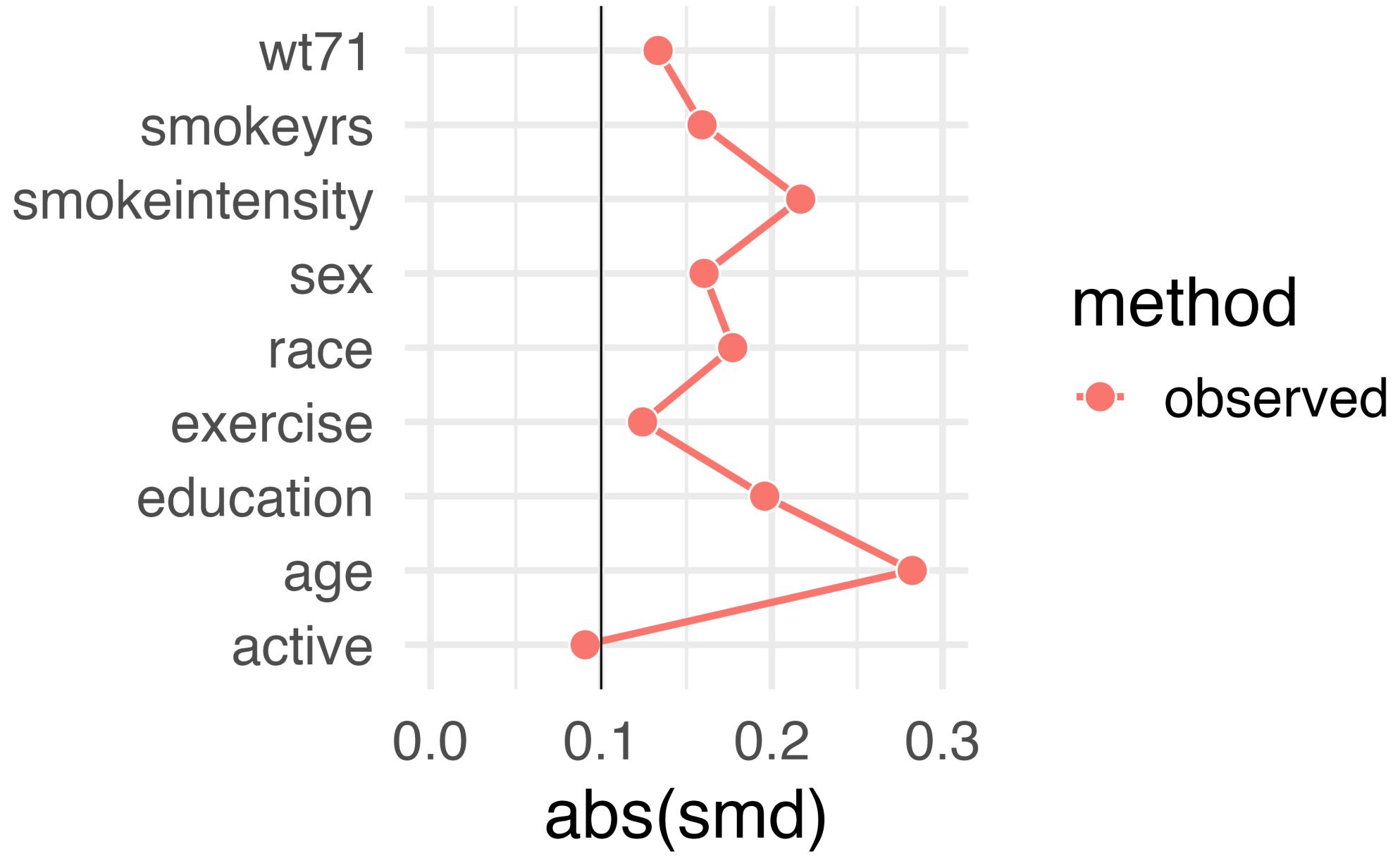
diagnose your model assumptions

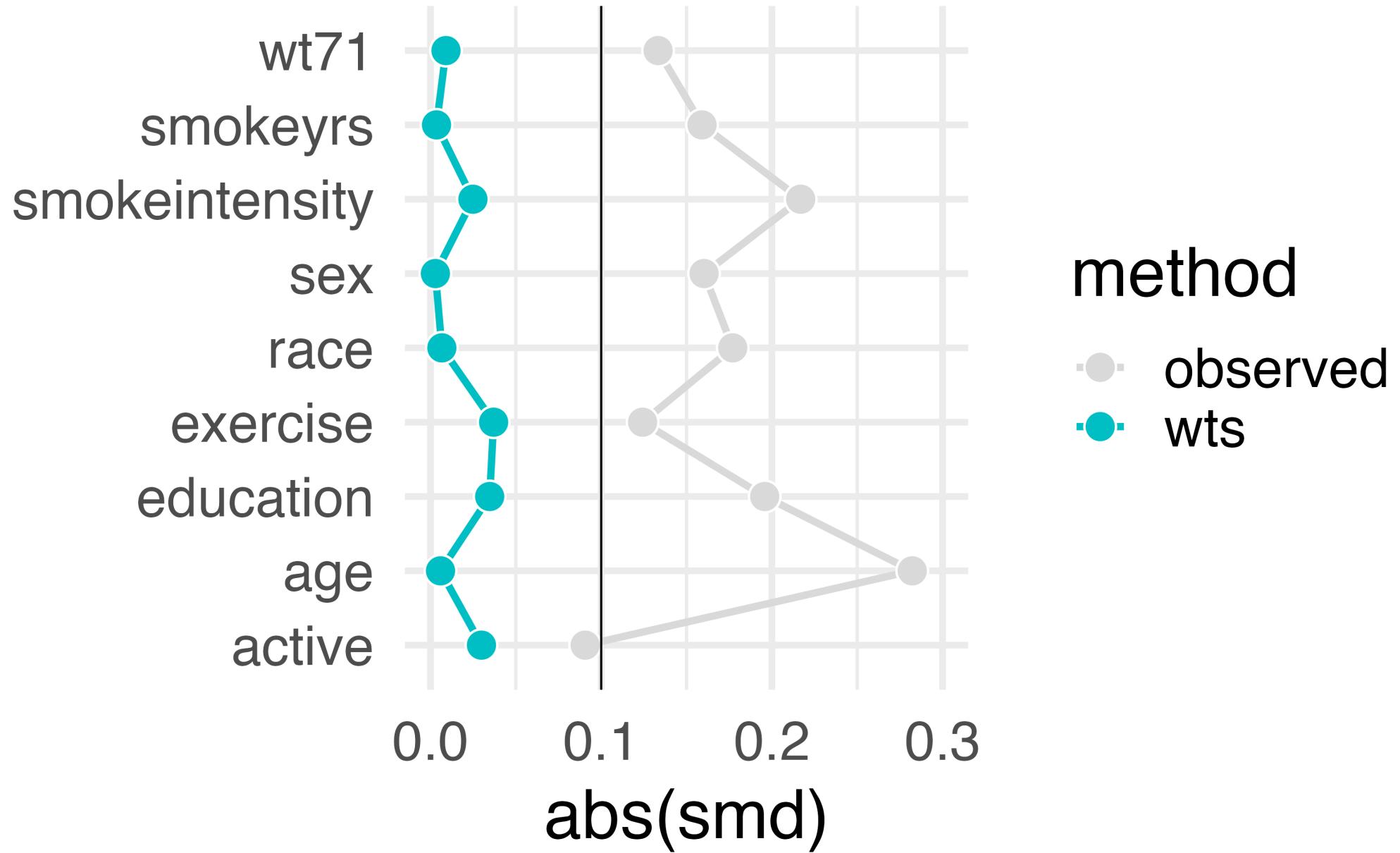
What's the distribution of weights?



What are the weights doing to the sample?

What are the weights doing to the sample?





estimate the causal effects

Estimate causal effect with IPW

```
1 ipw_model <- lm(  
2   wt82_71 ~ qsmk,  
3   data = nhefs_complete_uc,  
4   weights = wts  
5 )  
6  
7 ipw_estimate <- ipw_model |>  
8   tidy(conf.int = TRUE) |>  
9   filter(term == "qsmk")
```

Estimate causal effect with IPW

```
1 ipw_estimate
```

```
# A tibble: 1 × 7
  term    estimate std.error statistic p.value conf.low conf.high
  <chr>     <dbl>     <dbl>     <dbl>    <dbl>    <dbl>    <dbl>
1 qsmk      3.44     0.408     8.43 7.47e-17     2.64
# i 1 more variable: conf.high <dbl>
```

Let's fix our confidence intervals with robust SEs!

```
1 # also see robustbase, survey, gee, and others
2 library(estimatr)
3 ipw_model_robust <- lm_robust(
4   wt82_71 ~ qsmk,
5   data = nhefs_complete_uc,
6   weights = wts
7 )
8
9 ipw_estimate_robust <- ipw_model_robust |>
10 tidy(conf.int = TRUE) |>
11 filter(term == "qsmk")
```

Let's fix our confidence intervals with robust SEs!

```
1 as_tibble(ipw_estimate_robust)

# A tibble: 1 × 9
  term estimate std.error statistic p.value conf.low
  <chr>    <dbl>     <dbl>     <dbl>    <dbl>    <dbl>
1 qsmk      3.44      0.526     6.54 8.57e-11      2.41
# i 3 more variables: conf.high <dbl>, df <dbl>,
#   outcome <chr>
```

Let's fix our confidence intervals with the bootstrap!

```
1 # fit ipw model for a single bootstrap sample
2 fit_ipw_not Quite_rightly <- function(split, ...) {
3   # get bootstrapped data sample with `rsample::analysis()`
4   .df <- analysis(split)
5
6   # fit ipw model
7   lm(wt82_71 ~ qsmk, data = .df, weights = wts) |>
8     tidy()
9 }
```

```

1 fit_ipw <- function(split, ...) {
2   .df <- analysis(split)
3
4   # fit propensity score model
5   propensity_model <- glm(
6     qsmk ~ sex +
7       race + age + I(age^2) + education +
8       smokeintensity + I(smokeintensity^2) +
9       smokeyrs + I(smokeyrs^2) + exercise + active +
10      wt71 + I(wt71^2),
11      family = binomial(),
12      data = .df
13    )
14
15   # calculate inverse probability weights
16   .df <- propensity_model |>
17     augment(type.predict = "response", data = .df) |>
18     mutate(wts = wt_ate(.fitted, qsmk))
19
20   # fit correctly bootstrapped ipw model
21   lm(wt82_71 ~ qsmk, data = .df, weights = wts) |>
22     tidy()
23

```

Using {rsample} to bootstrap our causal effect

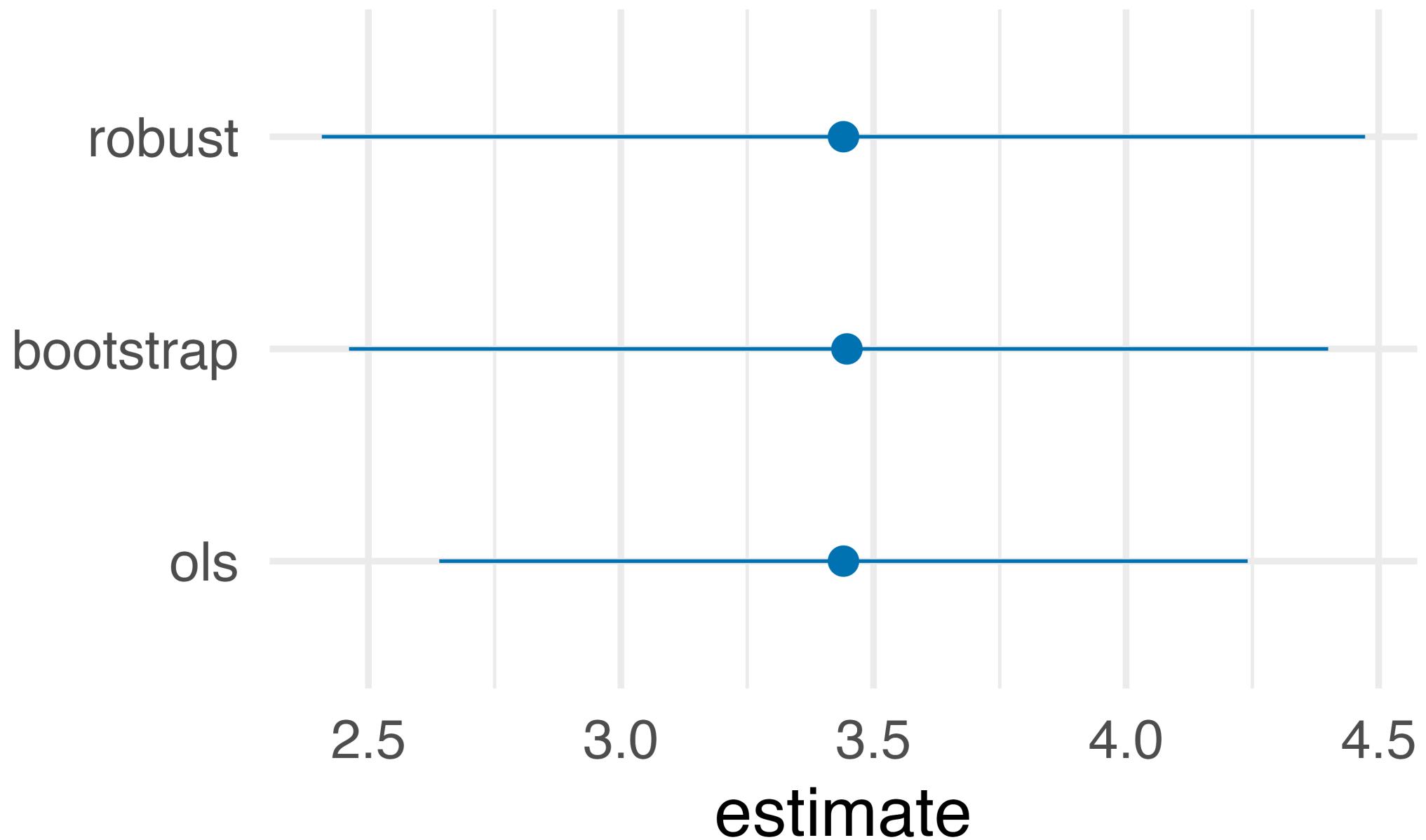
```
1 # fit ipw model to bootstrapped samples  
2 ipw_results <- bootstraps(nhefs_complete_uc, 1000, apparent = TRUE  
3   mutate(results = map(splits, fit_ipw))
```

Using {rsample} to bootstrap our causal effect

```
1 # get t-statistic-based CIs
2 boot_estimate <- int_t(ipw_results, results) |>
3   filter(term == "qsmk")
4
5 boot_estimate
```

Using {rsample} to bootstrap our causal effect

```
# A tibble: 1 × 6
  term    .lower  .estimate  .upper  .alpha  .method
  <chr>    <dbl>      <dbl>    <dbl>    <dbl>   <chr>
1 qsmk     2.46      3.45    4.40    0.05 student-t
```



*Our causal effect estimate: 3.5 kg (95% CI 2.4 kg,
4.4 kg)*

Review the Quarto file... later!

Resources

Causal Inference: Comprehensive text on causal inference. Free online.

Bootstrap confidence intervals with {rsample}

R-causal: Our GitHub org with R packages and examples

