

# Causal Modeling in R: Whole Game

Malcolm Barrett

RStudio, PBC

2021-09-01 (updated: 2022-10-16)

- 1 Specify causal question**
- 2 Draw assumptions (causal diagram)**
- 3 Model assumptions (e.g. propensity score)**
- 4 Analyze propensities (diagnostics)**
- 5 Estimate causal effects (e.g. IPW)**
- 6 Sensitivity analysis (more later!)**

**Do people who quit smoking  
gain weight?**

```
library(causaldata)
nhefs_complete_uc <- nhefs_complete %>%
  filter(censored == 0)
nhefs_complete_uc
```

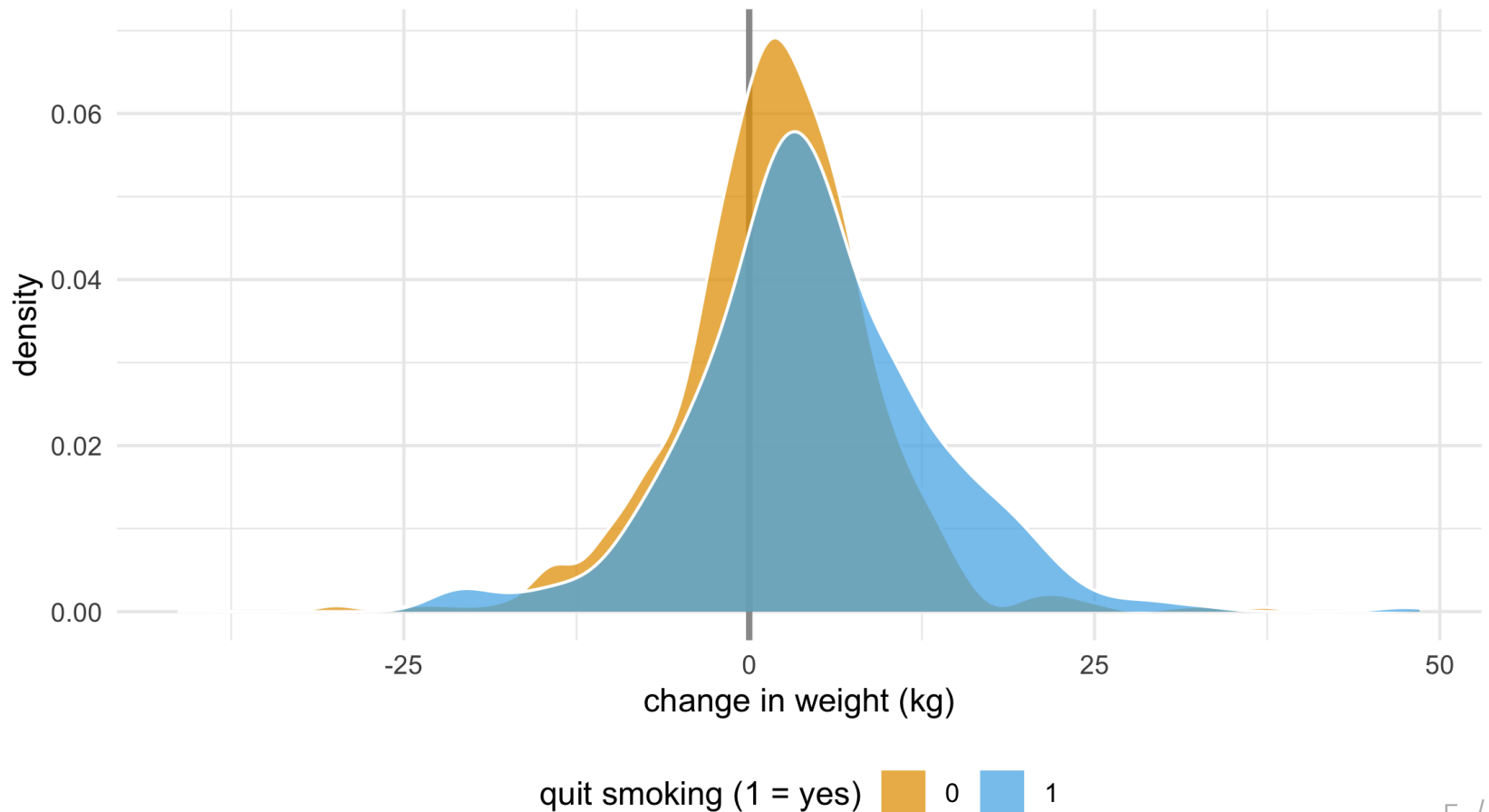
```
## # A tibble: 1,566 × 67
```

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

|      | sex   | age   | race  | income | marital | school | education |
|------|-------|-------|-------|--------|---------|--------|-----------|
|      | <fct> | <dbl> | <fct> | <dbl>  | <dbl>   | <dbl>  | <fct>     |
| ## 1 | 0     | 42    | 1     | 19     | 2       | 7      | 1         |
| ## 2 | 0     | 36    | 0     | 18     | 2       | 9      | 2         |
| ## 3 | 1     | 56    | 1     | 15     | 3       | 11     | 2         |
| ## 4 | 0     | 68    | 1     | 15     | 3       | 5      | 1         |
| ## 5 | 0     | 40    | 0     | 18     | 2       | 11     | 2         |

# Did those who quit smoking gain weight?



# Did those who quit smoking gain weight?

```
# ~2.5 KGs gained for quit vs. not quit
nhefs_complete_uc %>%
  group_by(qsmk) %>%
  summarize(
    mean_weight_change = mean(wt82_71),
    sd = sd(wt82_71),
    .groups = "drop"
  )
```

```
## # 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?**



# Multivariable regression: what's the association?

```
lm(
  wt82_71~ qsmk + sex +
    race + age + I(age^2) + education +
    smokeintensity + I(smokeintensity^2) +
    smokeyrs + I(smokeyrs^2) + exercise + active +
    wt71 + I(wt71^2),
  data = nhefs_complete_uc
) %>%
  tidy(conf.int = TRUE) %>%
  filter(term == "qsmk")
```

# Multivariable regression: what's the association?

```
lm(
  wt82_71~ qsmk + sex +
    race + age + I(age^2) + education +
    smokeintensity + I(smokeintensity^2) +
    smokeyrs + I(smokeyrs^2) + exercise + active +
    wt71 + I(wt71^2),
  data = nhefs_complete_uc
) %>%
  tidy(conf.int = TRUE) %>%
  filter(term == "qsmk")
```

```
## # A tibble: 1 × 7
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>      <dbl>      <dbl>    <dbl>
## 1 qsmk        3.46        0.438        7.90 5.36e-15
##   conf.low conf.high
##   <dbl>      <dbl>
## 1      2.60      4.32
```

**model your assumptions**

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

# Fit propensity score model

```
propensity_model <- glm(  
  qsmk ~ sex +  
    race + age + I(age^2) + education +  
    smokeintensity + I(smokeintensity^2) +  
    smokeyrs + I(smokeyrs^2) + exercise + active +  
    wt71 + I(wt71^2),  
  family = binomial(),  
  data = nhefs_complete_uc  
)
```

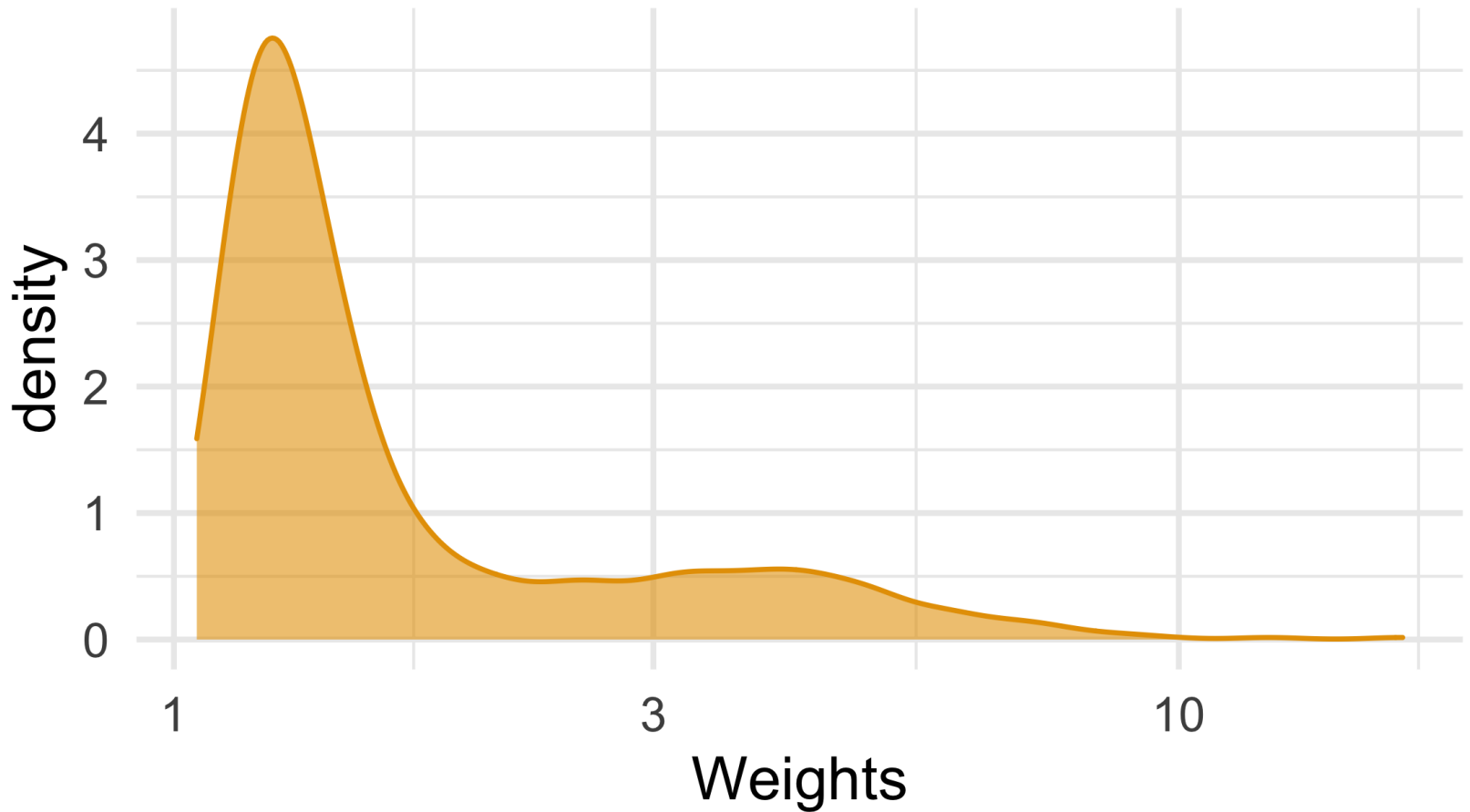


# Calculate inverse probability weights

```
nhefs_complete_uc <- propensity_model %>%  
  # predict whether quit smoking  
  augment(type.predict = "response", data = nehs_complete_uc) %>%  
  # calculate inverse probability  
  mutate(wts = 1 / ifelse(qsmk == 0, 1 - .fitted, .fitted))
```

**diagnose your model  
assumptions**

# What's the distribution of weights?







**estimate the causal effects**

# Estimate causal effect with IPW

```
ipw_model <- lm(  
  wt82_71 ~ qsmk,  
  data = nhfs_complete_uc,  
  weights = wts  
)  
  
ipw_estimate <- ipw_model %>%  
  tidy(conf.int = TRUE) %>%  
  filter(term == "qsmk")
```

# Estimate causal effect with IPW

```
ipw_estimate
```

```
## # A tibble: 1 × 7
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>      <dbl>      <dbl>    <dbl>
## 1 qsmk        3.44        0.408        8.43 7.47e-17
##   conf.low conf.high
##   <dbl>      <dbl>
## 1      2.64      4.24
```



# Let's fix our confidence intervals with robust SEs!

```
# also see robustbase, survey, gee, and others
```

```
library(estimatr)
```

```
ipw_model_robust <- lm_robust(
```

```
  wt82_71 ~ qsmk,
```

```
  data = nhfs_complete_uc,
```

```
  weights = wts
```

```
)
```

```
ipw_estimate_robust <- ipw_model_robust %>%
```

```
  tidy(conf.int = TRUE) %>%
```

```
  filter(term == "qsmk")
```

# Let's fix our confidence intervals with robust SEs!

```
as_tibble(ipw_estimate_robust)
```

```
## # A tibble: 1 × 9
##   term estimate std.error statistic  p.value
##   <chr>      <dbl>      <dbl>      <dbl>    <dbl>
## 1 qsmk      3.44      0.526      6.54 8.57e-11
##   conf.low conf.high    df outcome
##   <dbl>      <dbl> <dbl> <chr>
## 1 2.41      4.47 1564 wt82_71
```

# Let's fix our confidence intervals with the bootstrap!

```
# fit ipw model for a single bootstrap sample
fit_ipw_not_quite_rightly <- function(split, ...) {
  # get bootstrapped data sample with 'rsample::analysis()'
  .df <- analysis(split)

  # fit ipw model
  lm(wt82_71 ~ qsmk, data = .df, weights = wts) %>%
    tidy()
}
```

```

fit_ipw <- function(split, ...) {
  .df <- analysis(split)

  # fit propensity score model
  propensity_model <- glm(
    qsmk ~ sex +
      race + age + I(age^2) + education +
      smokeintensity + I(smokeintensity^2) +
      smokeyrs + I(smokeyrs^2) + exercise + active +
      wt71 + I(wt71^2),
    family = binomial(),
    data = .df
  )

  # calculate inverse probability weights
  .df <- propensity_model %>%
    augment(type.predict = "response", data = .df) %>%
    mutate(wts = 1 / ifelse(qsmk == 0, 1 - .fitted, .fitted))

  # fit correctly bootstrapped ipw model
  lm(wt82_71 ~ qsmk, data = .df, weights = wts) %>%
    tidy()
}

```

# Using {rsample} to bootstrap our causal effect

```
# fit ipw model to bootstrapped samples  
ipw_results <- bootstraps(nhefs_complete, 1000, apparent = TRUE) %>%  
  mutate(results = map(splits, fit_ipw))
```

# Using {rsample} to bootstrap our causal effect

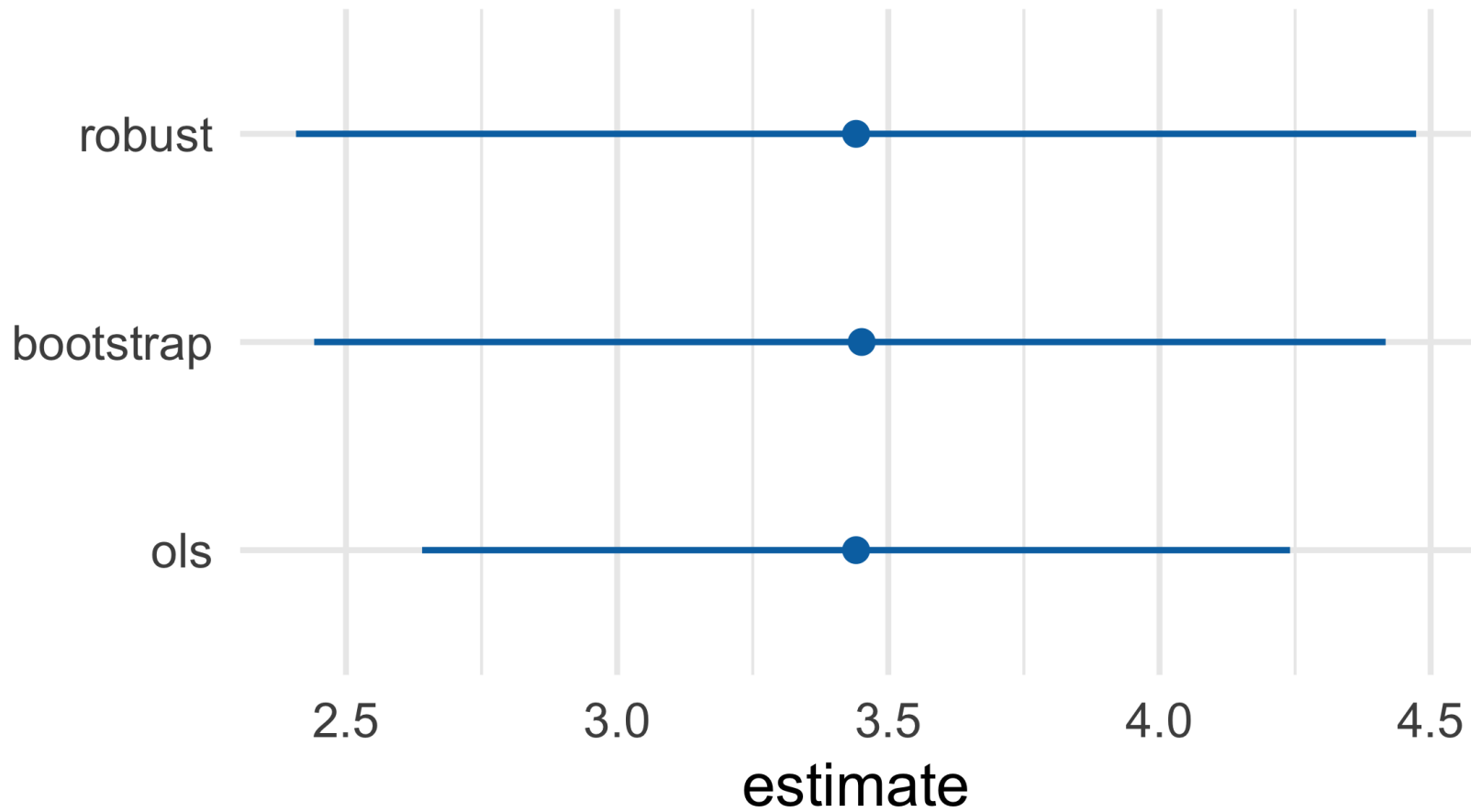
```
# get t-statistic-based CIs  
boot_estimate <- int_t(ipw_results, results) %>%  
  filter(term == "qsmk")  
  
boot_estimate
```

# Using {rsample} to bootstrap our causal effect

```
# get t-statistic-based CIs
boot_estimate <- int_t(ipw_results, results) %>%
  filter(term == "qsmk")

boot_estimate
```

```
## # A tibble: 1 × 6
##   term   .lower .estimate .upper .alpha .method
##   <chr> <dbl>      <dbl> <dbl> <dbl> <chr>
## 1 qsmk    2.44      3.45    4.42  0.05 student-t
```





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

**Review the R Markdown file...  
later!**

# Resources

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

**Causal Inference Notebook:** R code to go along with Causal Inference

**Bootstrap confidence intervals with {rsample}**