Causal Inference

Octavio Mesner

Association vs Causation

- What is causation? Why is it important?
- Causal questions:
    - Which cancer treatments are best for which patients?
    - Will a national gun law result in fewer homicides?
    - Does increasing minimum wage reduce job openings?
    - Will changing `X` also change `Y` ?

Simpson's paradox

- How can statistics be misleading on causation?
- New infectious disease with high mortality rate
- Scientists develop an experimental treatment and give it to doctors to try out

```
##        treatment symptoms mortality
## 687 experimental   severe     alive
## 239 experimental   severe      died
## 530      placebo     mild      died
## 846      placebo     mild     alive
## 346 experimental     mild      died
## 634      placebo     mild     alive
## 509      placebo     mild     alive
## 475 experimental   severe     alive
## 213      placebo   severe      died
## 431      placebo     mild     alive
## 632 experimental     mild     alive
## 630 experimental   severe     alive
## 318      placebo   severe      died
## 991      placebo     mild     alive
## 648      placebo   severe     alive
## 880      placebo     mild     alive
## 136 experimental   severe      died
## 655 experimental   severe     alive
## 497      placebo     mild     alive
## 358      placebo     mild     alive
```

```
##
## experimental      placebo
##          487          513
```

```
##
##   mild severe
##    514    486
```

```
##
## alive  died
##   754   246
```

- The mortality rate for each group is below:

```
table(df$treatment, df$mortality)[,2]/table(df$treatment)
```

```
##
## experimental      placebo
##    0.2648871    0.2280702
```

- The death rate is higher in the experimental treatment group than the placebo group
- Taking a closer look, we stratify death rate by symptom severity:

```
table(df$treatment, df$symptoms, df$mortality)[,,2]/table(df$treatment, df$symptoms)
```

```
##
##                      mild     severe
##   experimental 0.07352941 0.33903134
##   placebo      0.16402116 0.40740741
```

- Someone with severe symptoms is much more likely to die than someone with mild symptoms
- But, in both groups, the experimental treatment was associated with fewer deaths
- How can this happen?

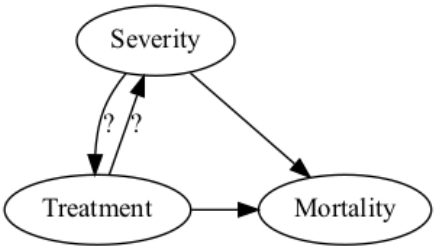
```
table(df$treatment, df$symptoms)
```

```
##
##                mild severe
##   experimental  136    351
##   placebo       378    135
```

```
chisq.test(df$treatment, df$symptoms)
```

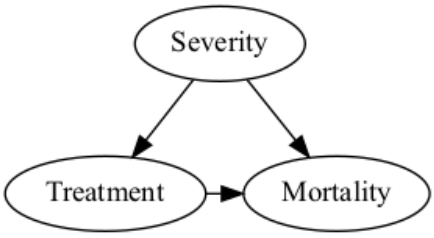
```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$treatment and df$symptoms
## X-squared = 207.58, df = 1, p-value < 2.2e-16
```

- Those with severe symptoms were more likely to be on the experimental treatment
- Timing matters here! Was severity taken before or after treatment?
- Neither scenario below is conclusive from the data, but knowing the time ordering can help rule one out



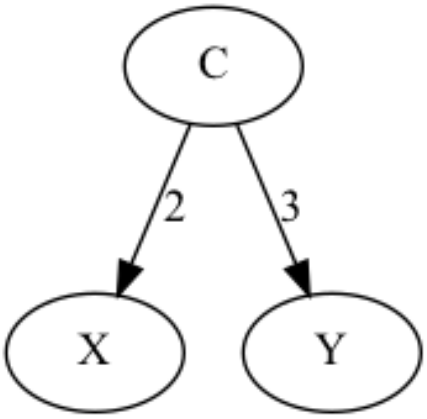
Confounding

- Example: Sleeping with shoes on is associated with waking up with a headache
- Example: Yellow teeth and cancer are confounded by smoking
- Confounding (https://en.wikipedia.org/wiki/Confounding): `X` and `Y` are confounded if they are both influenced by a third variable
- Example from last section: Assume that severity was taken before treatment was given



- There are two “open” paths from treatment to mortality:
    - `treatment <- severity -> mortality`
    - `treatment -> mortality`
- Looking at the total association between treatment and mortality (without severity) will use include assocations from both paths
- Controlling for severity “blocks” the `treatment <- severity -> mortality` pathway
- We’re left with the direct causal path `treatment -> mortality`

Confounding Simulations



- Diagram above: `A` and `B` are confounded by `C` and have no direct causal relationship
- $X = 2C + \epsilon_X \Rightarrow C = 0.5X + 0.5\epsilon_X$, so $Y = 0.5 \cdot 3X + \text{noise}$

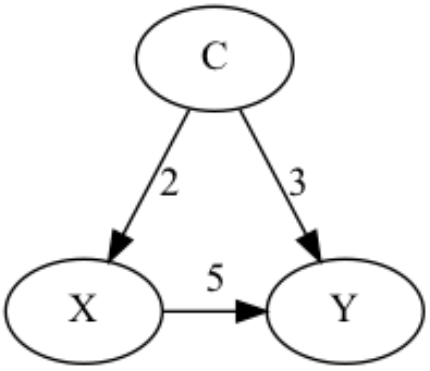
```
size <- 500
C <- 10*runif(size)
X <- 2*C + rnorm(size)
Y <- 3*C + rnorm(size)

summary(lm(Y~X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6491 -1.2526 -0.0418  1.2540  4.1460
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.6102     0.1522    4.01 7.01e-05 ***
## X             1.4342     0.0131  109.52  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.74 on 498 degrees of freedom
## Multiple R-squared:  0.9601, Adjusted R-squared:  0.9601
## F-statistic: 1.199e+04 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
summary(lm(Y~X+C))
```

```
##
## Call:
## lm(formula = Y ~ X + C)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.81847 -0.63492  0.02683  0.65114  2.43640
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.03361    0.08757   0.384    0.701
## X           -0.03018    0.04536  -0.665    0.506
## C            3.04750    0.09315  32.717   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.981 on 497 degrees of freedom
## Multiple R-squared:  0.9874, Adjusted R-squared:  0.9873
## F-statistic: 1.941e+04 on 2 and 497 DF,  p-value: < 2.2e-16
```



- Diagram above: `X` influences `Y` but the effect is confounded by `Z`. That is, failing to account for `Z` in a regression will lead to an incorrect causal parameter estimate
- parameter = (counfounding bias) + (causal effect) = 1.5 + 5 = 6.5

```
Y <- 3*C + 5*X + rnorm(size)

summary(lm(Y~X))
```

```
##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.261 -1.139 -0.012  1.183  5.458
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.65110    0.14861   4.381 1.44e-05 ***
## X            6.43596    0.01279 503.296  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.7 on 498 degrees of freedom
## Multiple R-squared:  0.998,  Adjusted R-squared:  0.998
## F-statistic: 2.533e+05 on 1 and 498 DF,  p-value: < 2.2e-16
```

```
summary(lm(Y~X+C))
```

```
##
## Call:
## lm(formula = Y ~ X + C)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2887 -0.6351 -0.0185  0.6507  2.9811
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09275    0.08704   1.066    0.287
## X            5.01797    0.04509 111.295   <2e-16 ***
## C            2.95101    0.09258  31.875   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.975 on 497 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9994
## F-statistic: 3.853e+05 on 2 and 497 DF,  p-value: < 2.2e-16
```

- Note: the covariate causal effect math only works here because the simulated data is linear. In the real world, we typically can't assume linearity. This math doesn't work without if the true data aren't linear.
- In most real-world datasets, there will always be the possibility of latent confounding

Sampling Bias

- Sampling Bias (https://en.wikipedia.org/wiki/Sampling_bias) occurs when sample is collected in such a way that some members of the intended population have a lower or higher sampling probability than others.
- Sampling bias can lead to incorrect estimates when variables of interest influence sampling
- Example: In 1936, the American Literary Digest sent out two million surveys to its readers and predicted that Alf Landon would beat incumbent president, Franklin Roosevelt, by a landslide, but the opposite happened. This was because readers over-represented Republicans.
- Example (https://catalogofbias.org/biases/collider-bias/): A researcher analysed data from 257 hospitalized individuals and detected an association between locomotor disease and respiratory disease (odds ratio 4.06). The researcher repeated the analysis in a sample of 2783 individuals from the general population and found no association (odds ratio 1.06)

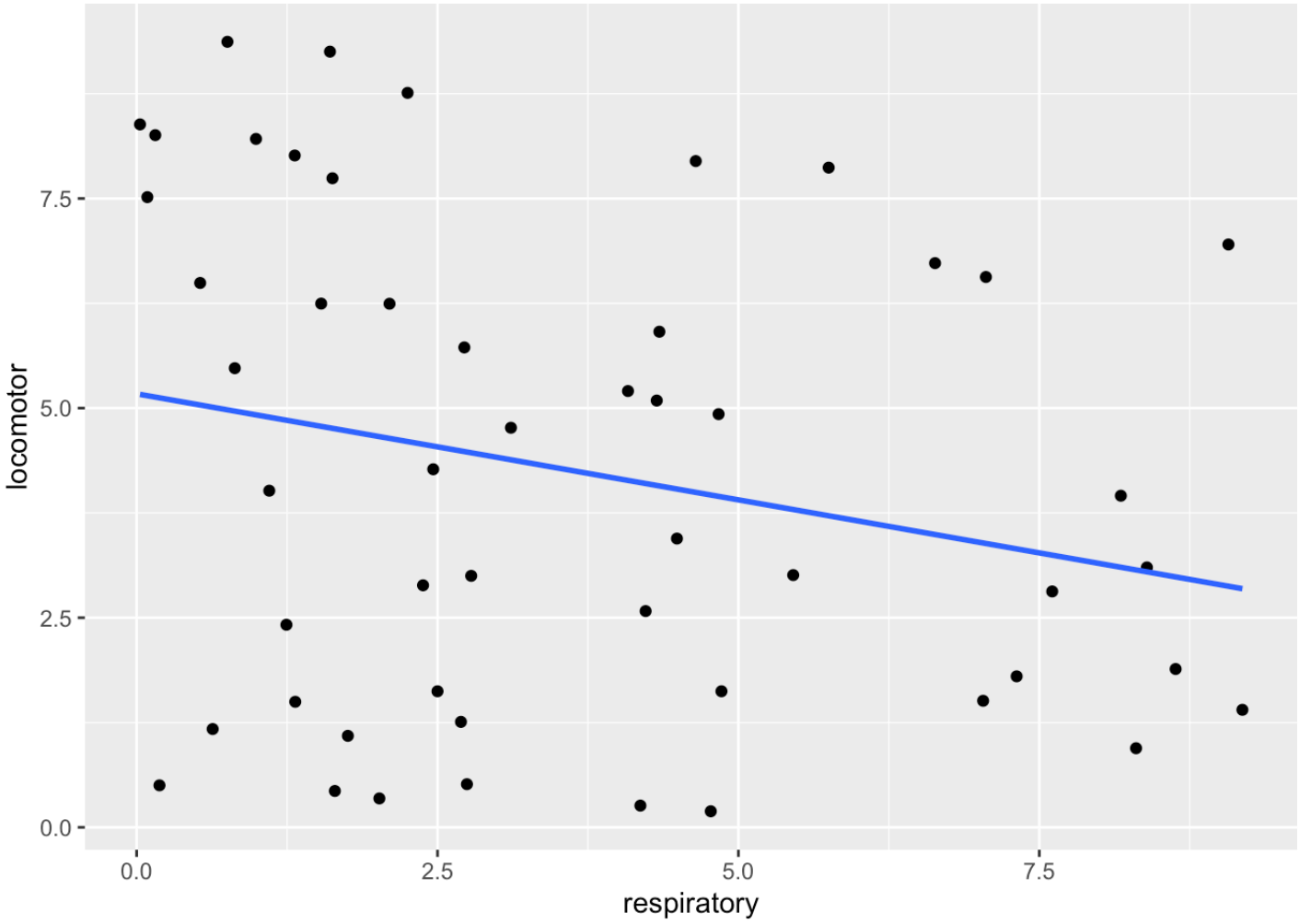
```
size <- 100
locomotor <- 10*runif(size)
respiratory <- 10*runif(size)
hospitalized <- rbinom(size, 1, (locomotor+respiratory)/20)==0
df_all <- data.frame(locomotor, respiratory, hospitalized)
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:latticeExtra':
##
##     layer
```

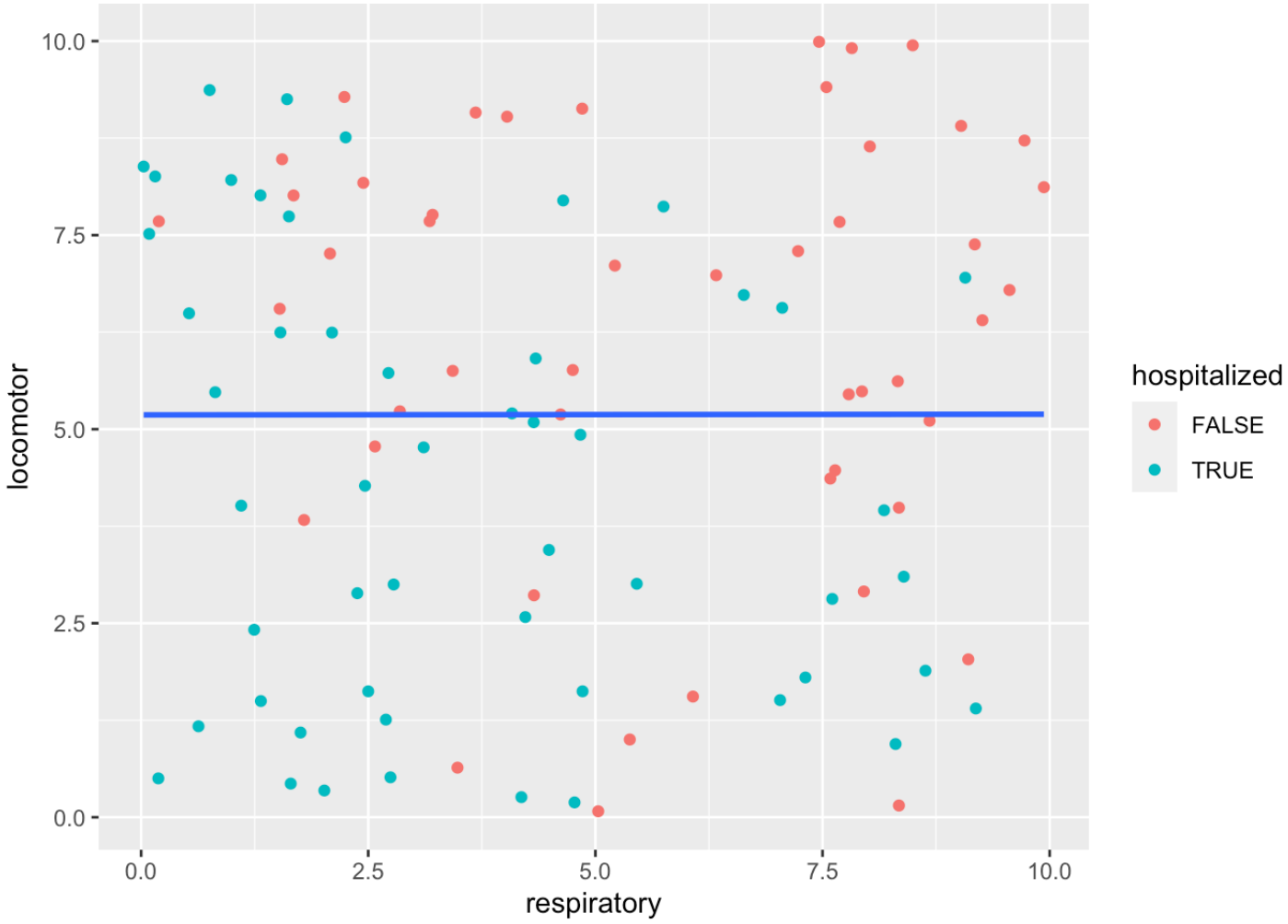
```
ggplot(df_all[hospitalized==TRUE,], aes(respiratory, locomotor)) + geom_point() +
  geom_smooth(method = lm, se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

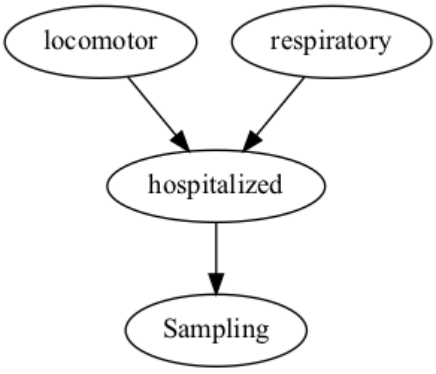



```
ggplot(df_all, aes(respiratory, locomotor)) + geom_point(aes(color=hospitalized)) +
  stat_smooth(method=lm, se=FALSE)
```

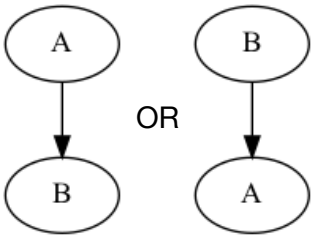
```
## `geom_smooth()` using formula 'y ~ x'
```

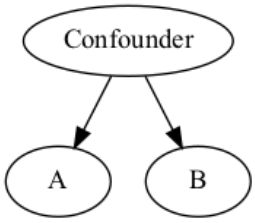


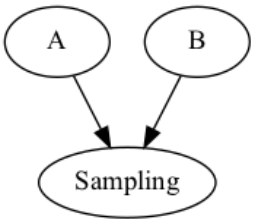
- Sampling bias is an example of collider bias
- Here we are conditioning on hospitalization
- Even though locomotor and respiratory are independent, conditioning on a collider induces an association
- Conditioning on a collider “open” an association pathway



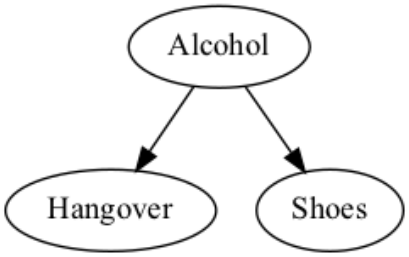
Statistical Dependency and Causation

- If A and B are random variables and associated (dependent) $A \not\perp\!\!\!\perp B$, it's thought that there are 3 ways this can happen
  - Direct causation: 

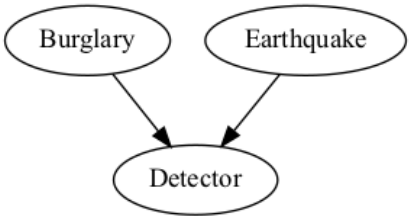
  - A and B are unmeasured confounding: 

  - Collider bias: A and B both influence sampling 

  - If we only have data for A and B, we cannot distinguish between these case without more assumptions
- If there are more variables, we figure out more using conditional independence
  - Thinking about if knowing one variable gives information another
  - Variables: Drank too much alcohol last night, Woke up with a hangover, Woke up with shoes on



  - Hangover⊥Shoes?
  - Hangover ⊥Shoes |Alcohol?
    - Given that someone drank too much alcohol last night, does knowing that they woke up with their shoes on give any information about if they have a hangover?
  - Variables: Burglary, Earthquake, Home motion detector

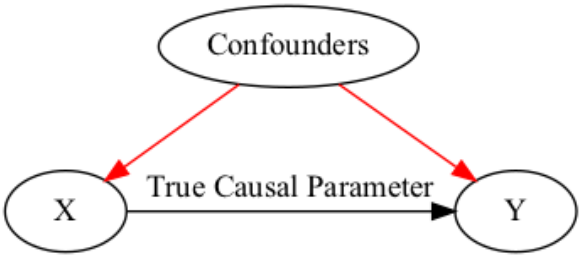


- Burglary⊥Earthquake
- Burglary ⊥Earthquake |Detector?
  - Given that the home motion detector alerted, does knowing that there was no earthquake give any information about a burglary?

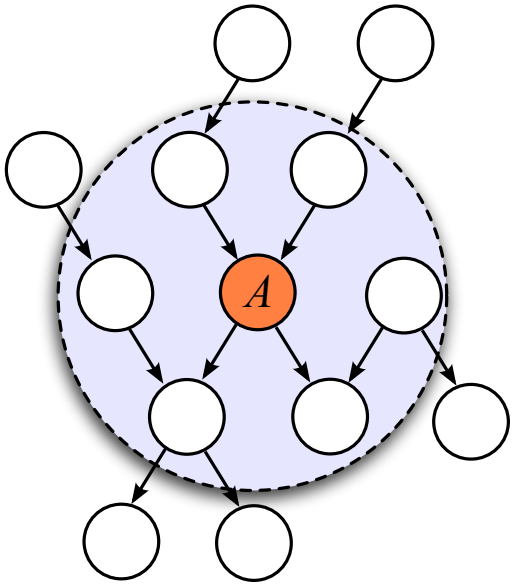
Structure	Path type	A, C independent/dependent?	A, C conditionally independent/dependent given B
Chain	$A \rightarrow B \rightarrow C$	$A \not\perp\!\!\!\perp C$	$A \perp C \mid B$
Other Chain	$A \leftarrow B \leftarrow C$	$A \not\perp\!\!\!\perp C$	$A \perp C \mid B$
Fork (Confounder)	$A \leftarrow B \rightarrow C$	$A \not\perp\!\!\!\perp C$	$A \perp C \mid B$
Collider	$A \rightarrow B \leftarrow C$	$A \perp C$	$A \not\perp\!\!\!\perp C \mid B$

Prediction, GLMs, and Causal Inference

- The goals are usually for GLMs and causal inference models are similar in practice:
  - control for variables to approximate the causal influence of variables of interest
  - if model assumptions are satisfied (linearity, no latent confounding, etc), standard GLMs will give causal parameter estimates
  - in this case, causal inference models *should* give similar parameter estimates to GLM estimates
  - In both cases, we should only control for variables that can *cause* our outcome
    - That is, we should not control for variables that are caused by the outcome (this can lead to collider bias)
    - Some of the time, we will not know which variables are caused by the outcome, a general rule is to not control for variable that are taken or observed after the outcome occurs
- If we do not control for *all* confounding variables, any association between X and Y present through the confounder pathway (red in image below) will likely be included in the parameter quantified by a model
- In *any* inference, we want to quantify the true causal parameter
  - Causal inference models make this easier by relaxing some assumptions



- Causal inference
  - Assumption of linear relationships within the data are likely not true (or at least hard to verify)
  - Causal inference models make it possible to control for confounding variables in a non-linear setting
  - Goal: Estimate a causal parameter of interest (when there are not linear relationships)
    - Causal parameters in our simulations are the parameters in the Bayesian network (DAG)
    - Causal parameter interpretation: if we were to change X, we should be able to use the causal parameter to accurately predict Y
    - Recall: when we did not control for confounding variables, the parameter estimates did not match the DAG
  - Assumptions: no latent confounding (we have all confounding variables included in the data)
- Prediction
  - use all useful information in data to make a prediction
  - we can use causes and effects
  - we also want to use colliders because they also provide information
  - General rule: use Markov Blanket (https://en.wikipedia.org/wiki/Markov_blanket)
  - For prediction, all variables in Markov blanket are useful for prediction, but not causation
    - If we changes causes, we should see changes in outcome
    - If we change variables in collider, we should not
  - Note: Prediction methods try to find all variables in Markov blanket
    - This will depend on signal/noise ratio and model assumptions
  - Given Markov Blanket, all other variables are conditionally independent from outcome
  - If we *know* causal Bayesian network underlying a dataset, we should include all variables in Markov blanket and no others
    - Including variables not in Markov Blanket will be noise variables (given the Markov Blanket)



Markov Blanket Image from Wikipedia

In the image above, the Markov Blanket of A is the set of white variable within the large circle

Evaluating Work Training Programs

- Manpower Demonstration Research Corporation was a federally and privately funded program implemented in the mid-1970s to provide work experience for a period of 6-18 months to individuals who faced economic and social problems prior to enrollment
- Those selected to join the program participated in various types of work such as restaurant and construction
- Pre-treatment information was collected - earnings, education, age, ethnicity, marital status
- All observations here are from men
- See Dehejia, R.H. and Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Re-Evaluating the Evaluation of Training Programs. Journal of the American Statistical Association 94: 1053-1062 (https://www.tandfonline.com/doi/pdf/10.1080/01621459.1999.10473858?casa_token=dsAisSiC-v4AAAAA:Auzr8KHp8-iB9Gy3T5o9hL-usKjKR1rne_TUvZDkHUCcl31OlVk_c0vwikXNTwQVYIKhhgSKqKXRJw)

```
data(lalonde)
dim(lalonde)
```

```
## [1] 614  10
```

```
names(lalonde)
```

```
##  [1] "treat"    "age"      "educ"     "black"    "hispan"   "married"
##  [7] "nodegree" "re74"     "re75"     "re78"
```

```
table(lalonde$treat)
```

```
##
##   0   1
## 429 185
```

```
lalonde$treat <- ifelse(lalonde$treat == 1, TRUE, FALSE)

#age
tapply(lalonde$age, lalonde$treat, mean)
```

```
##    FALSE     TRUE
## 28.03030 25.81622
```

```
t.test(lalonde$age ~ lalonde$treat)
```

```
##
##  Welch Two Sample t-test
##
## data:  lalonde$age by lalonde$treat
## t = 2.9911, df = 510.57, p-value = 0.002914
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7598127 3.6683610
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            28.03030            25.81622
```

```
ggplot(lalonde, aes(x=age, fill=treat)) +  geom_density(alpha=0.25)
```

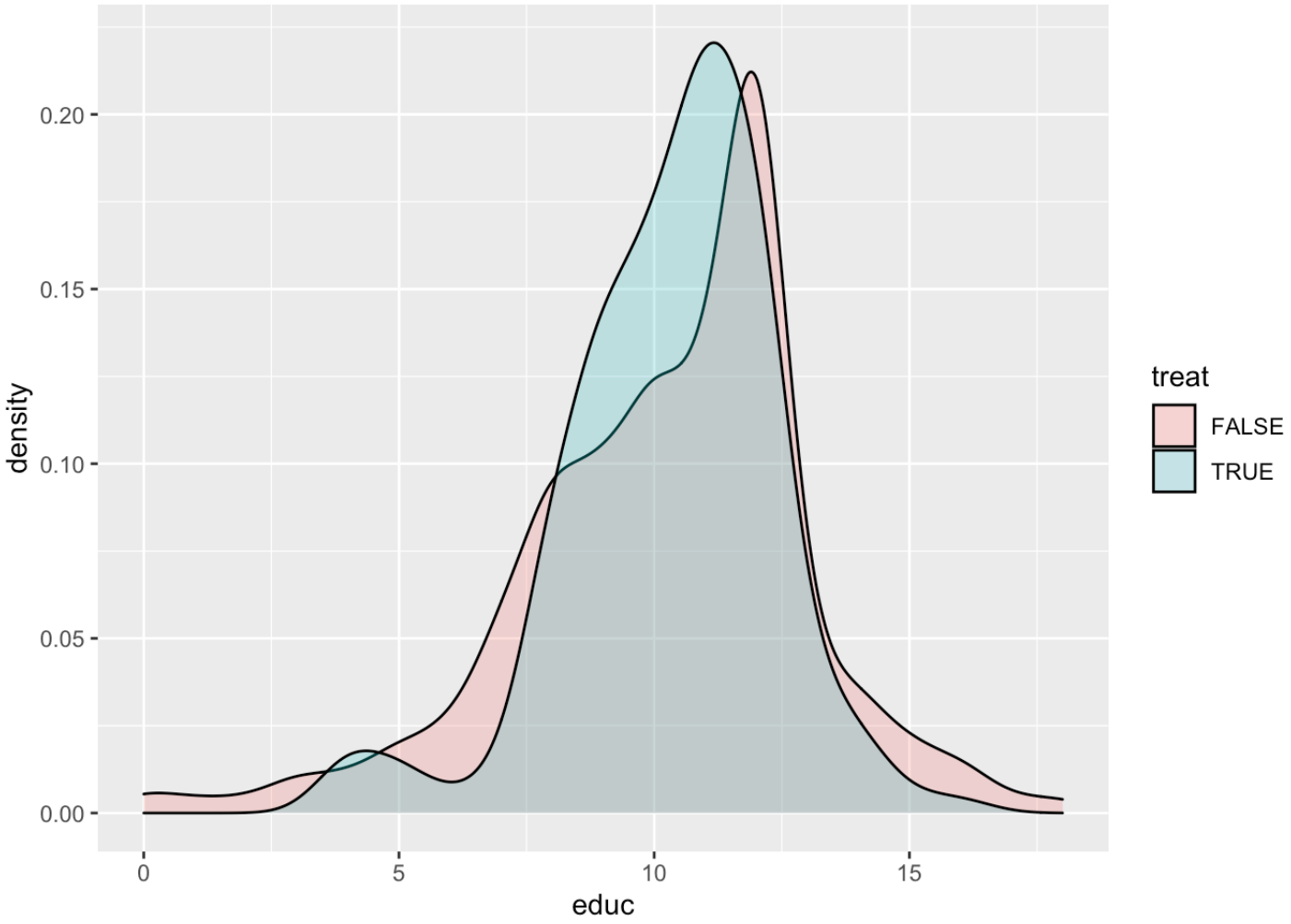

```
#educ
tapply(lalonde$educ, lalonde$treat, mean)
```

```
##    FALSE     TRUE
## 10.23543 10.34595
```

```
t.test(lalonde$educ ~ lalonde$treat)
```

```
##
##  Welch Two Sample t-test
##
## data:  lalonde$educ by lalonde$treat
## t = -0.54676, df = 485.37, p-value = 0.5848
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.5076687  0.2866393
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            10.23543            10.34595
```

```
ggplot(lalonde, aes(x=educ, fill=treat)) +  geom_density(alpha=0.25)
```



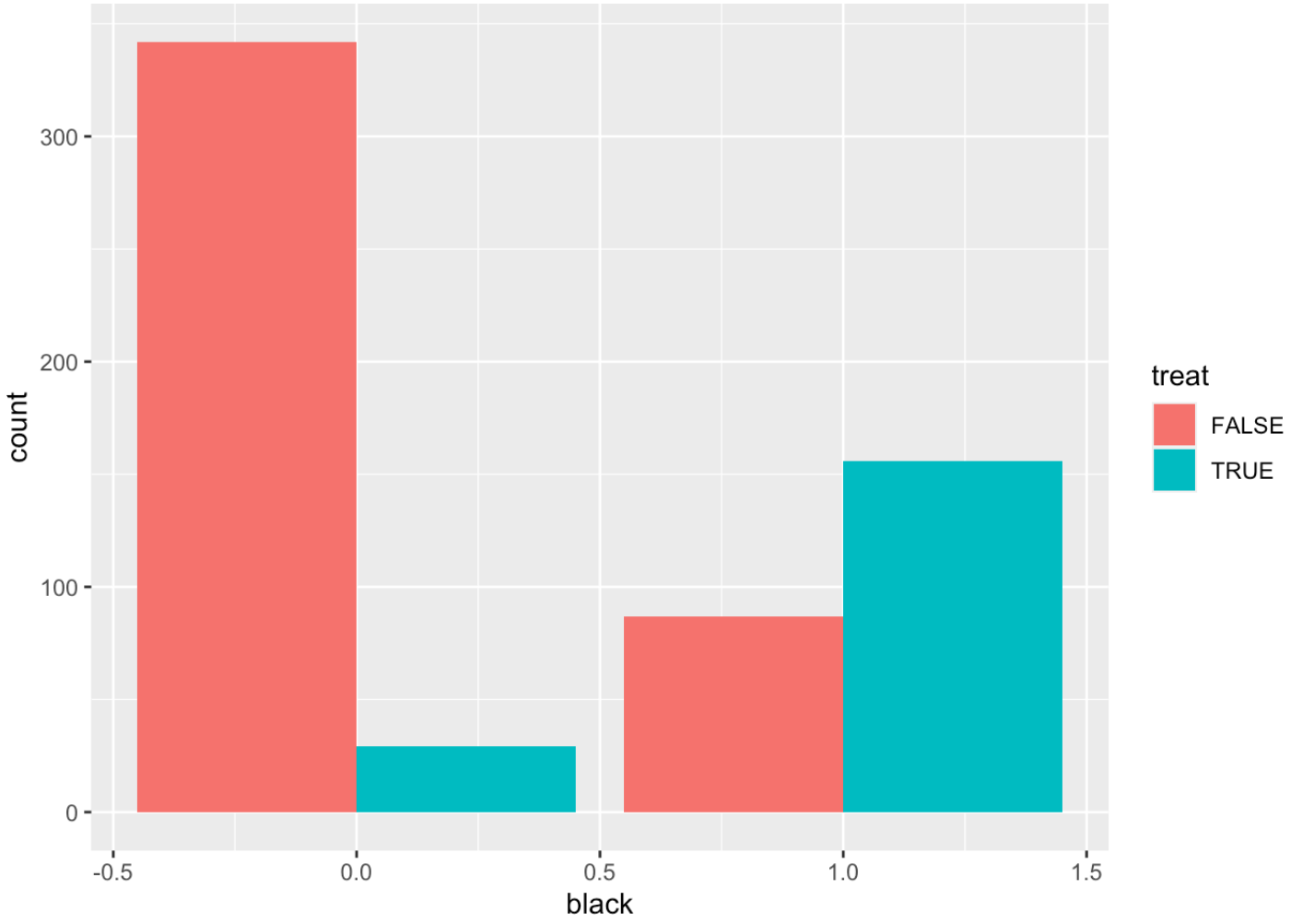

```
#black
table(lalonde$black, lalonde$treat)
```

```
##
##      FALSE TRUE
##    0   342   29
##    1    87  156
```

```
chisq.test(lalonde$black, lalonde$treat)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  lalonde$black and lalonde$treat
## X-squared = 219.04, df = 1, p-value < 2.2e-16
```

```
ggplot(lalonde, aes(x=black, fill=treat)) +  geom_bar(position = 'dodge')
```



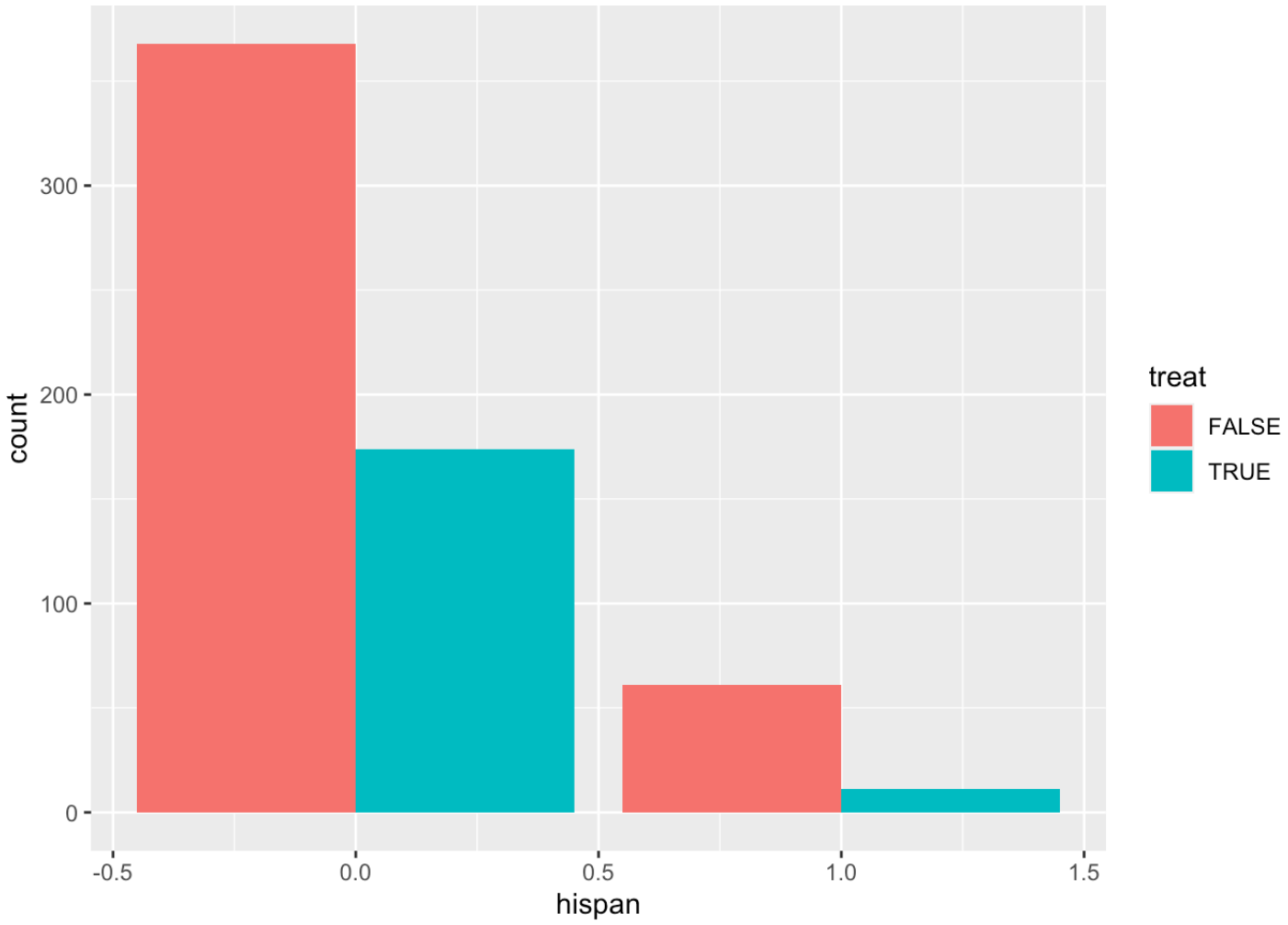
```
#hispan
table(lalonde$hispan, lalonde$treat)
```

```
##
##     FALSE TRUE
##   0   368  174
##   1    61   11
```

```
chisq.test(lalonde$hispan, lalonde$treat)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  lalonde$hispan and lalonde$treat
## X-squared = 7.7664, df = 1, p-value = 0.005323
```

```
ggplot(lalonde, aes(x=hispan, fill=treat)) +  geom_bar(position = 'dodge')
```



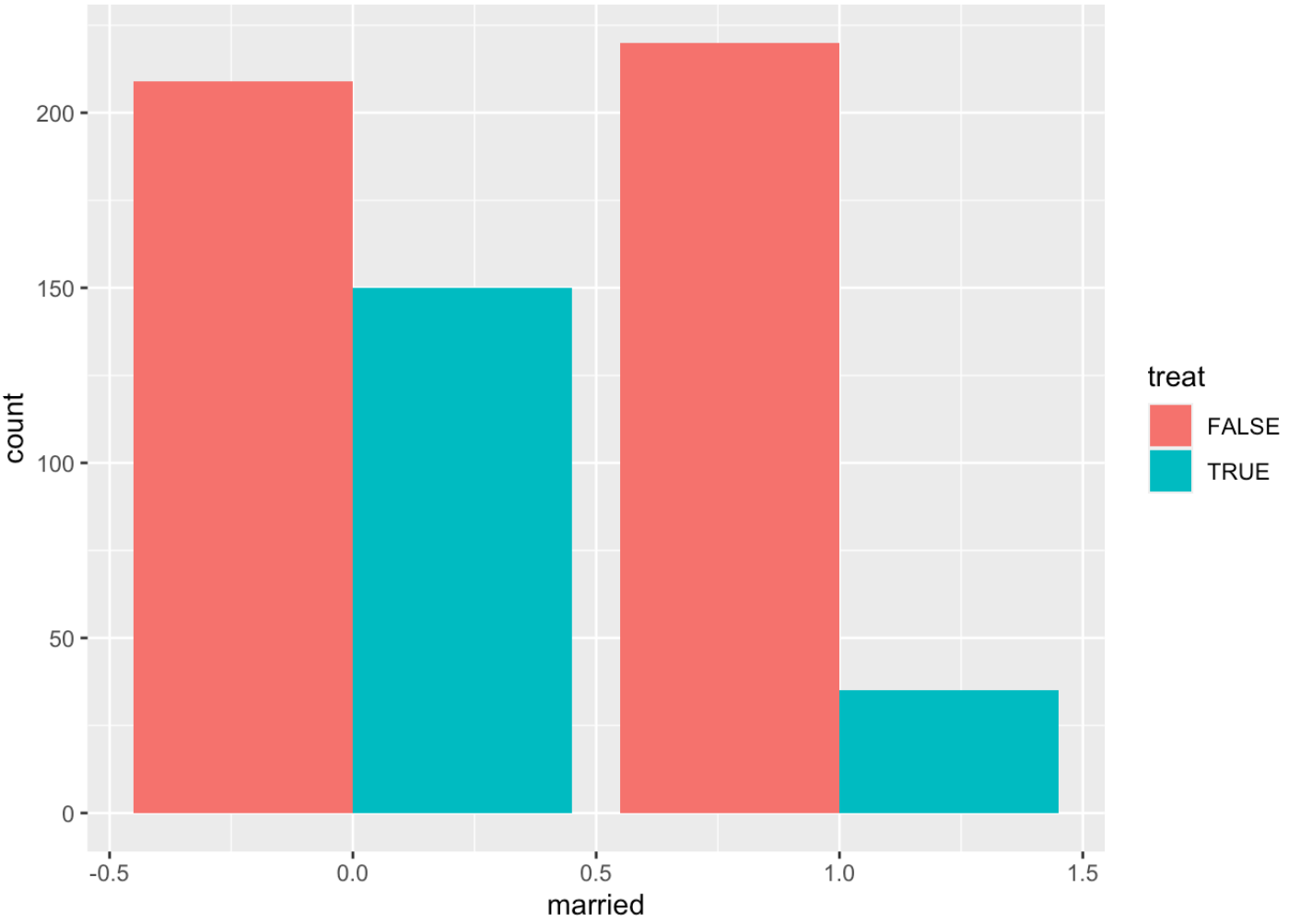

```
#married
table(lalonde$married, lalonde$treat)
```

```
##
##     FALSE TRUE
##   0   209  150
##   1   220   35
```

```
chisq.test(lalonde$married, lalonde$treat)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  lalonde$married and lalonde$treat
## X-squared = 54.428, df = 1, p-value = 1.613e-13
```

```
ggplot(lalonde, aes(x=married, fill=treat)) +  geom_bar(position = 'dodge')
```



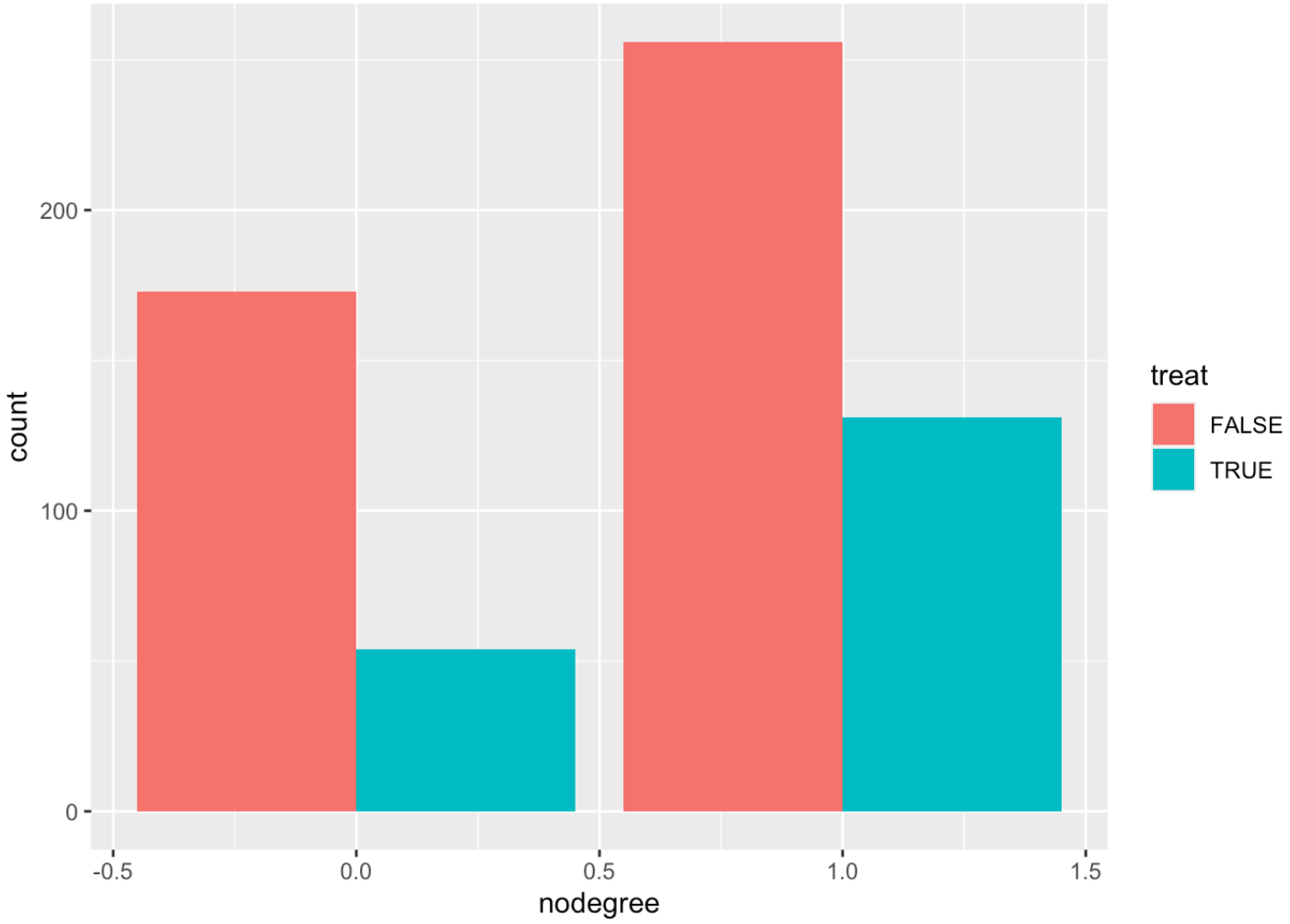

```
#nodegree
table(lalonde$nodegree, lalonde$treat)
```

```
##
##     FALSE TRUE
##   0   173   54
##   1   256  131
```

```
chisq.test(lalonde$nodegree, lalonde$treat)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  lalonde$nodegree and lalonde$treat
## X-squared = 6.4107, df = 1, p-value = 0.01134
```

```
ggplot(lalonde, aes(x=nodegree, fill=treat)) +  geom_bar(position = 'dodge')
```



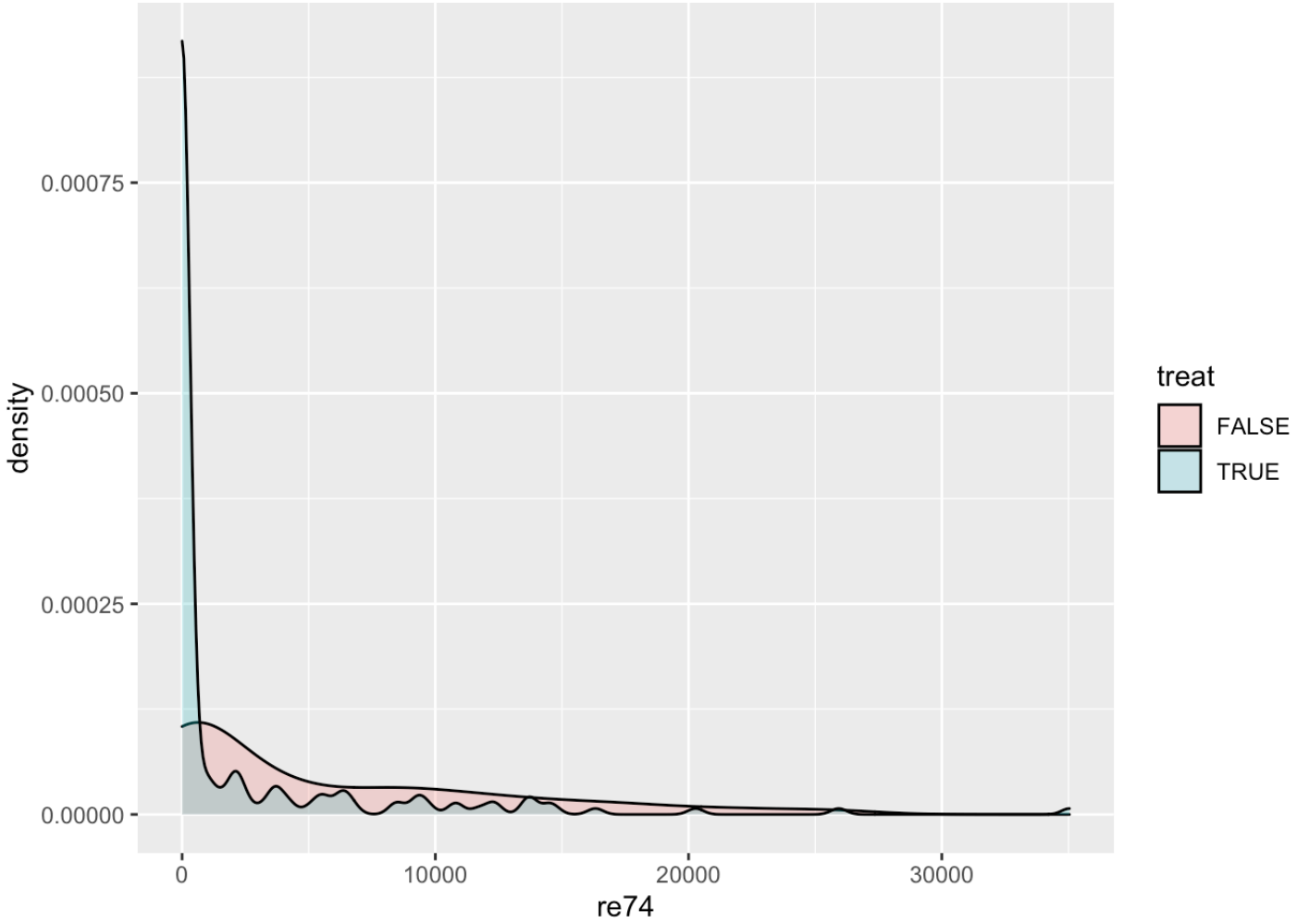

```
#re74
tapply(lalonde$re74, lalonde$treat, mean)
```

```
##    FALSE     TRUE
## 5619.237 2095.574
```

```
t.test(lalonde$re74 ~ lalonde$treat)
```

```
##
##  Welch Two Sample t-test
##
## data:  lalonde$re74 by lalonde$treat
## t = 7.2456, df = 475.99, p-value = 1.748e-12
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2568.067 4479.258
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            5619.237            2095.574
```

```
ggplot(lalonde, aes(x=re74, fill=treat)) +  geom_density(alpha=0.25)
```



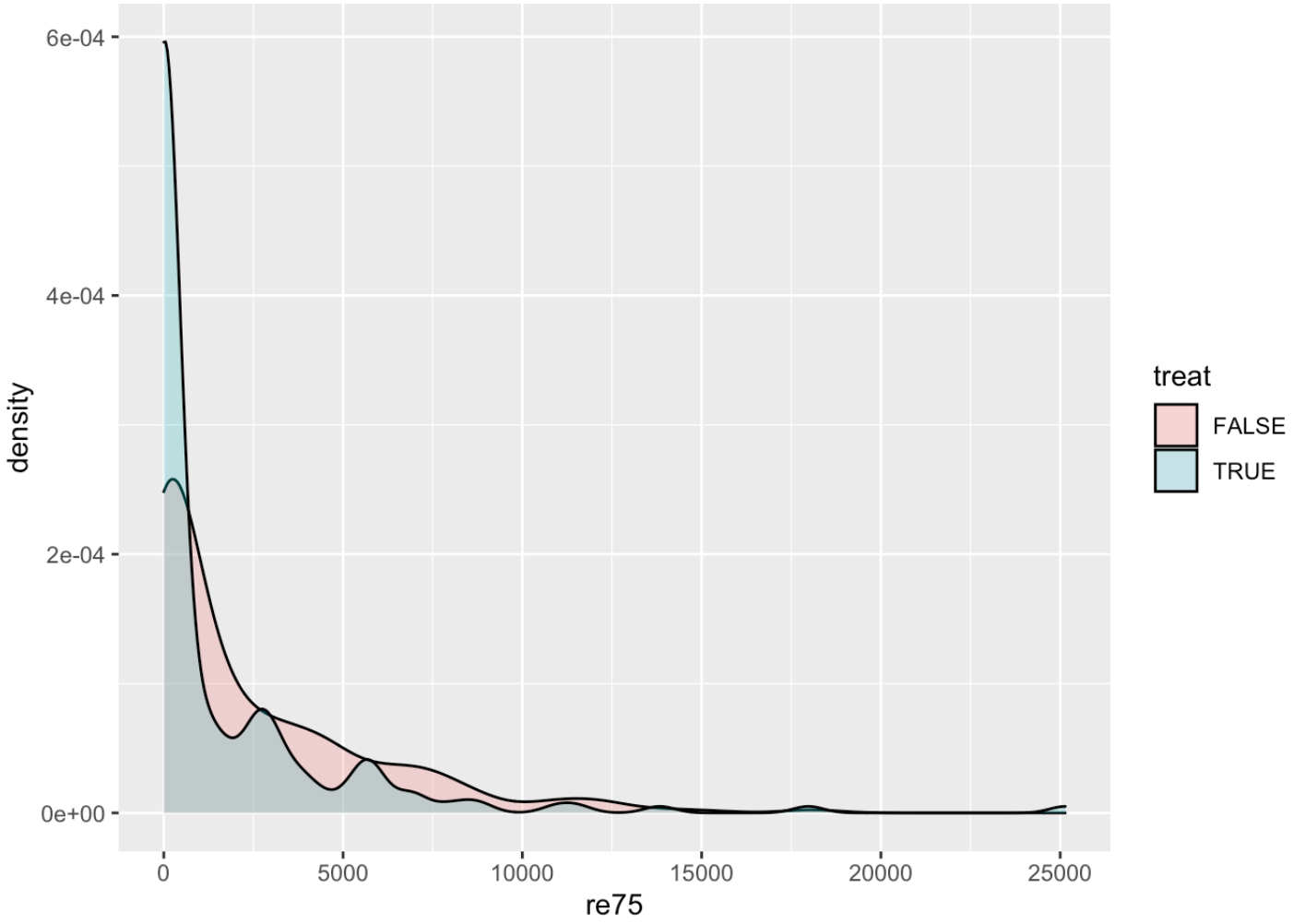

```
#re75
tapply(lalonde$re75, lalonde$treat, mean)
```

```
##    FALSE     TRUE
## 2466.484 1532.055
```

```
t.test(lalonde$re75 ~ lalonde$treat)
```

```
##
##  Welch Two Sample t-test
##
## data:  lalonde$re75 by lalonde$treat
## t = 3.2776, df = 356.22, p-value = 0.00115
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   373.742 1495.116
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            2466.484            1532.055
```

```
ggplot(lalonde, aes(x=re75, fill=treat)) +  geom_density(alpha=0.25)
```



Quantifying Causal Effect with Counterfactuals

- Let Y_i^1 be the outcome under treatment for observation i and let Y_i^0 be the outcome without treatment for observation i.
- Example: Does taking vitamin C prevent sickness?
  - $Y_i^1 = 1$ if i takes vitamin C and stays healthy
  - $Y_i^1 = 0$ if i takes vitamin C and gets sick
  - $Y_i^0 = 1$ if i does not take vitamin C and stays healthy
  - $Y_i^0 = 0$ if i does not take vitamin C and gets sick
- The causal effect for observation i is

$$Y_i^1 - Y_i^0$$

- Unfortunately, it's not possible to any individual's with and without treatment

```
##   A Y Y0 Y1
## 1 0 0  0 NA
## 2 0 0  0 NA
## 3 0 0  0 NA
## 4 0 0  0 NA
## 5 1 0 NA  0
## 6 1 1 NA  1
## 7 1 1 NA  1
## 8 1 1 NA  1
```

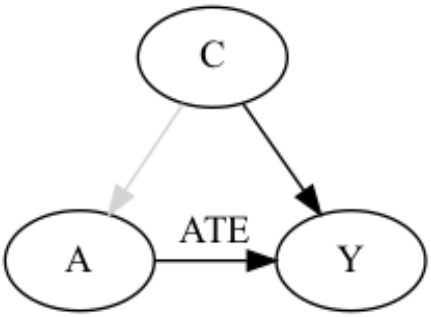
- In a population, the average causal effect (ATE) is

$$\text{ATE} = E\left[Y^1 - Y^0\right] = E[Y^1] - E[Y^0]$$

- Estimated as

$$\widehat{\text{ATE}} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^1 - Y_i^0\right)$$

- Is $E(Y^1|A=1)$ different from $E(Y|A=1)$?
- $Y^1 = Y^{A=1}$ and $Y^0 = Y^{A=0}$ assumes that A is not influenced by any variables, measured or latent



No other variables influence A when we consider Y^1, Y^0

- ATE can be interpreted as the average difference in outcome within the population that is attributed to A
- There may be other reasons Y differs from person to person, like age, severity, etc
- Conditional Average Causal Effect (CATE): if Z is another covariate,

$$\text{CATE}_z = E[Y^1|Z=z] - E[Y^0|Z=z]$$

  is the average causal effect for group $Z = z$.

Randomized Controlled Experiments or Trials

- When does a parameter estimate have a causal interpretation?
- Vaccine trails: a population is randomized to receive a test vaccine or placebo
  - Single or double blind: participants (and sometimes researcher) are not told which group they are

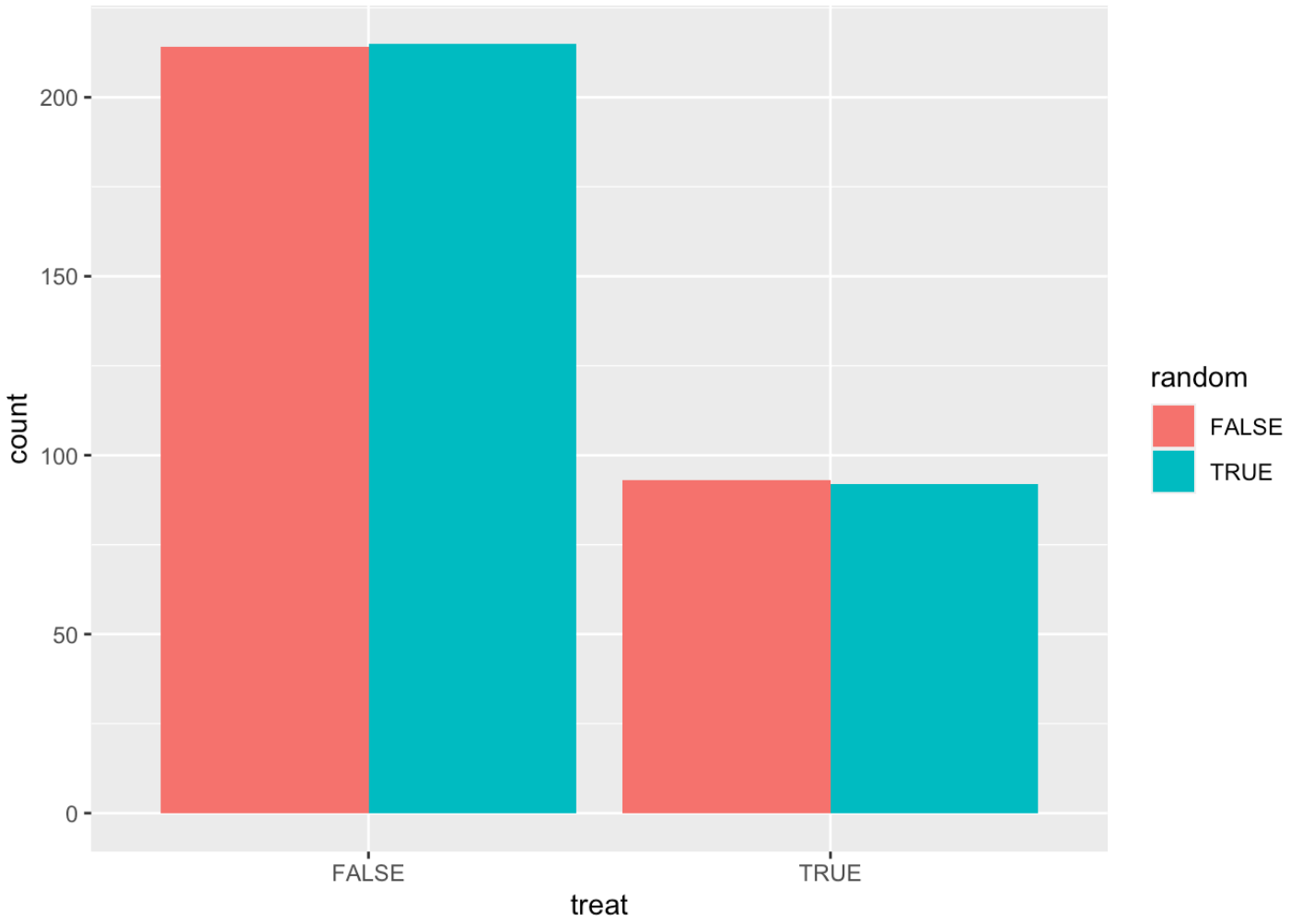
in
- Because of randomization, confounding is not possible and with a large enough sample, the two group will be statistically identical other than vaccine treatment
- Any difference in infection acquisition can be attributed to vaccination.
- Can estimate ATE:

$$\begin{aligned}\widehat{\text{ATE}} &= E[Y^{\text{vaccine}}] - E[Y^{\text{placebo}}] \\ &= \frac{1}{n}\sum_{i=1}^{n} Y_i^{\text{vaccine}} - \frac{1}{m}\sum_{j=1}^{m} Y_j^{\text{placebo}}\end{aligned}$$

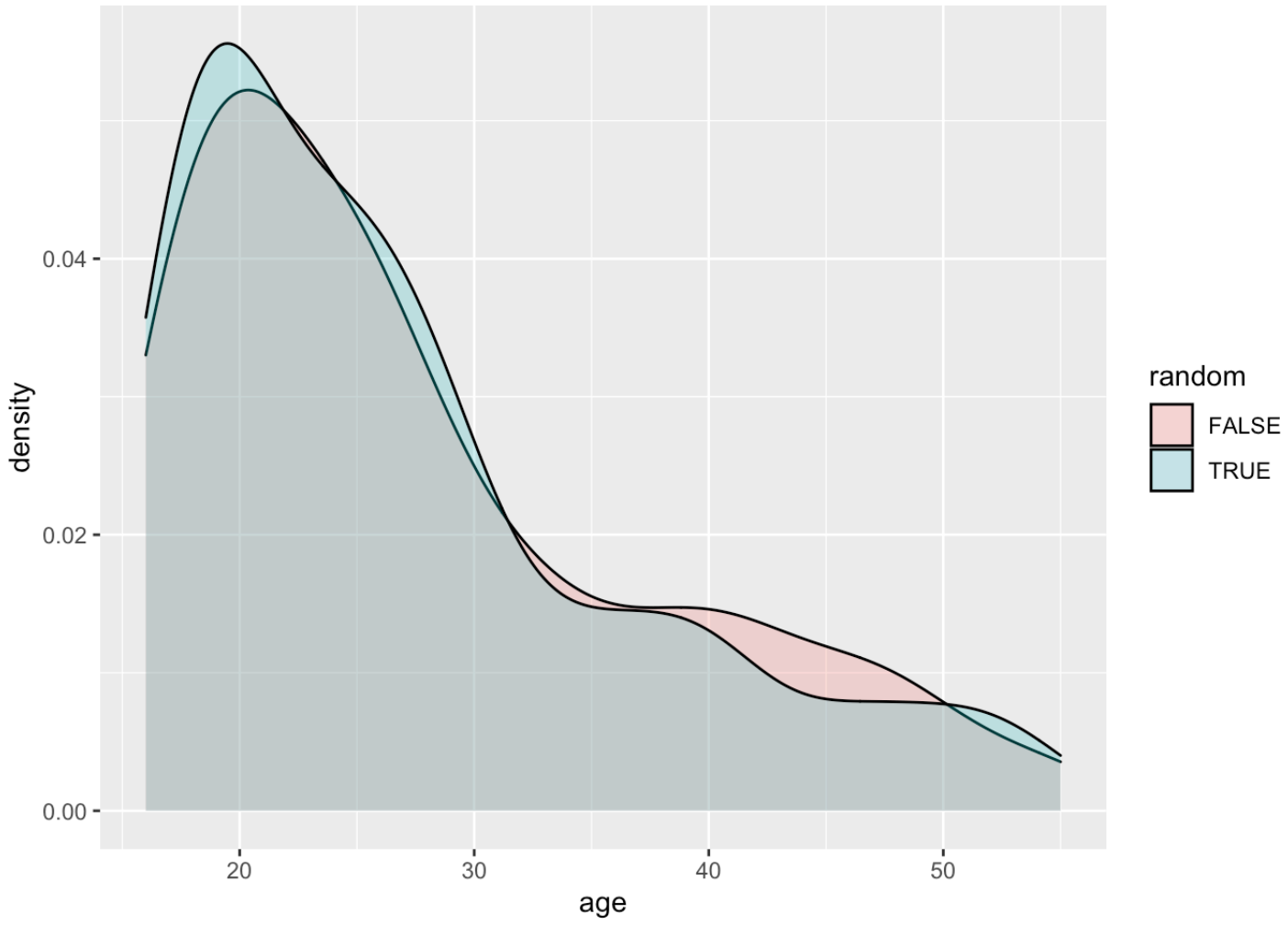
- $H_0 : \widehat{\text{ATE}} = 0, H_1 : \widehat{\text{ATE}} > 0$
- Two sample t-test is sufficient

- Randomization: what would the treatment and control populations look like for the work training data if treatment were randomized?

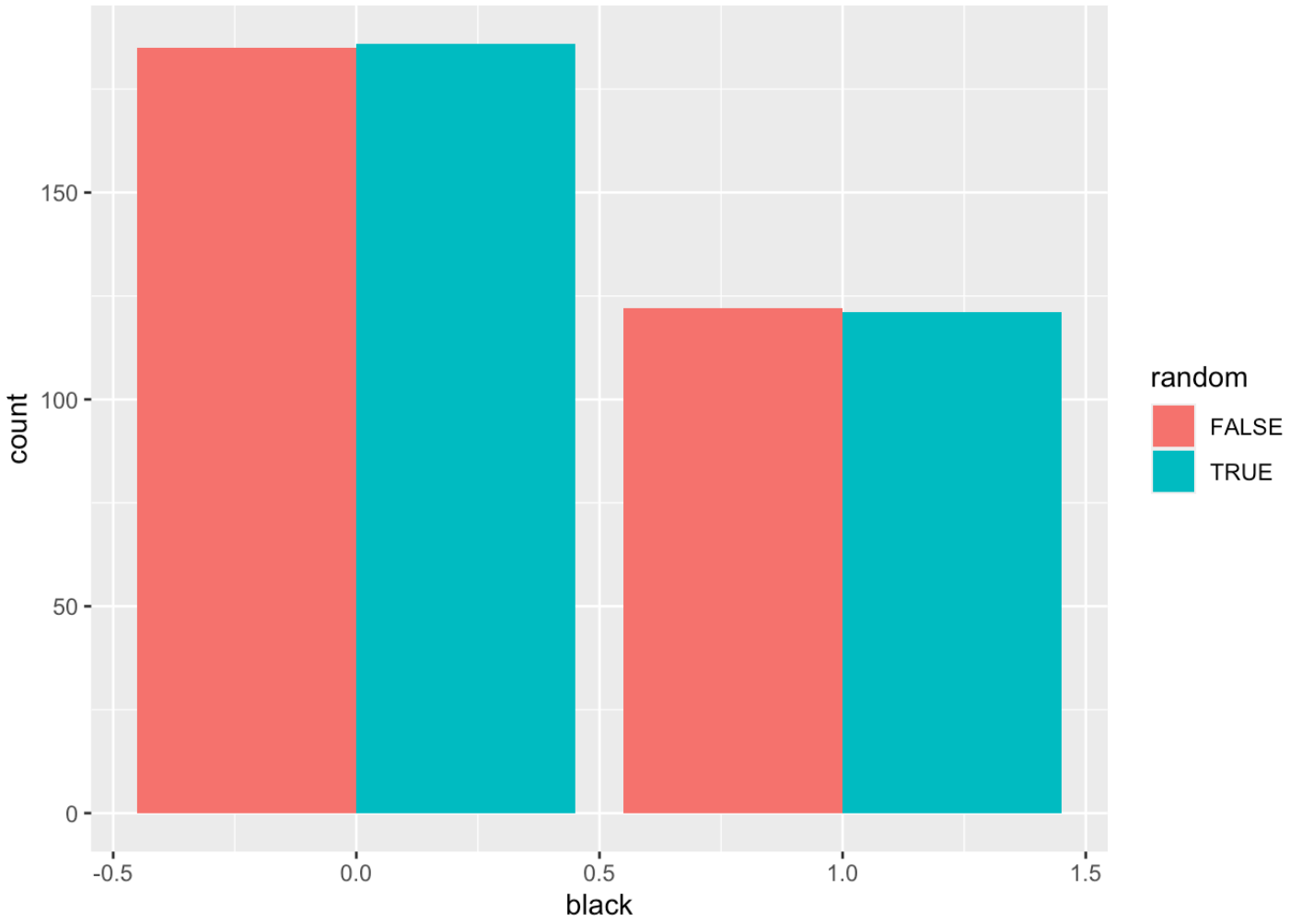
```
set.seed(1234)
lalonde$random <- sample(rep(c(TRUE, FALSE), length.out=dim(lalonde)[1]))
ggplot(lalonde, aes(x=treat, fill=random)) + geom_bar(position = 'dodge')
```



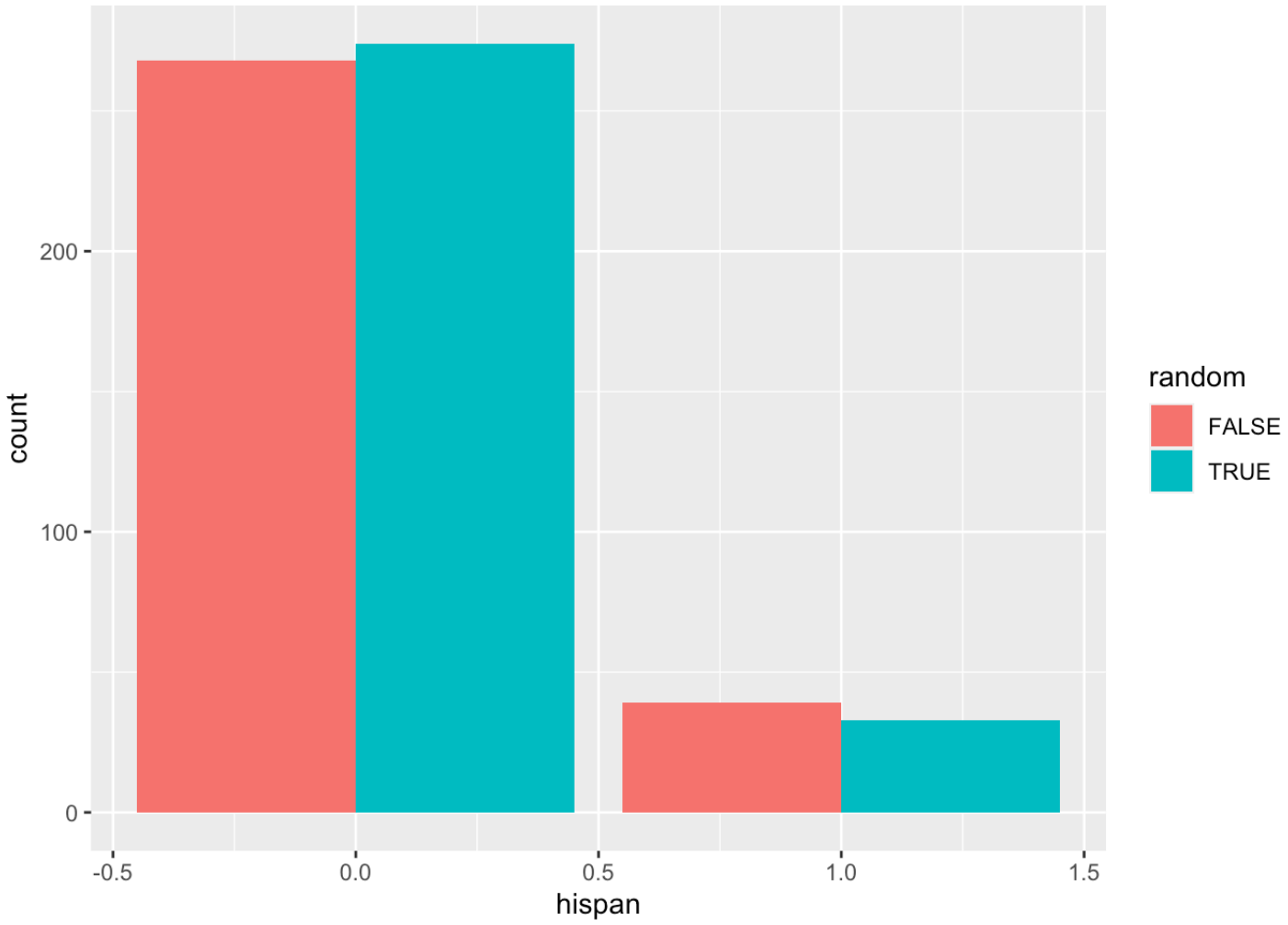

```
ggplot(lalonde, aes(x=age, fill=random)) +  geom_density(alpha=0.25)
```



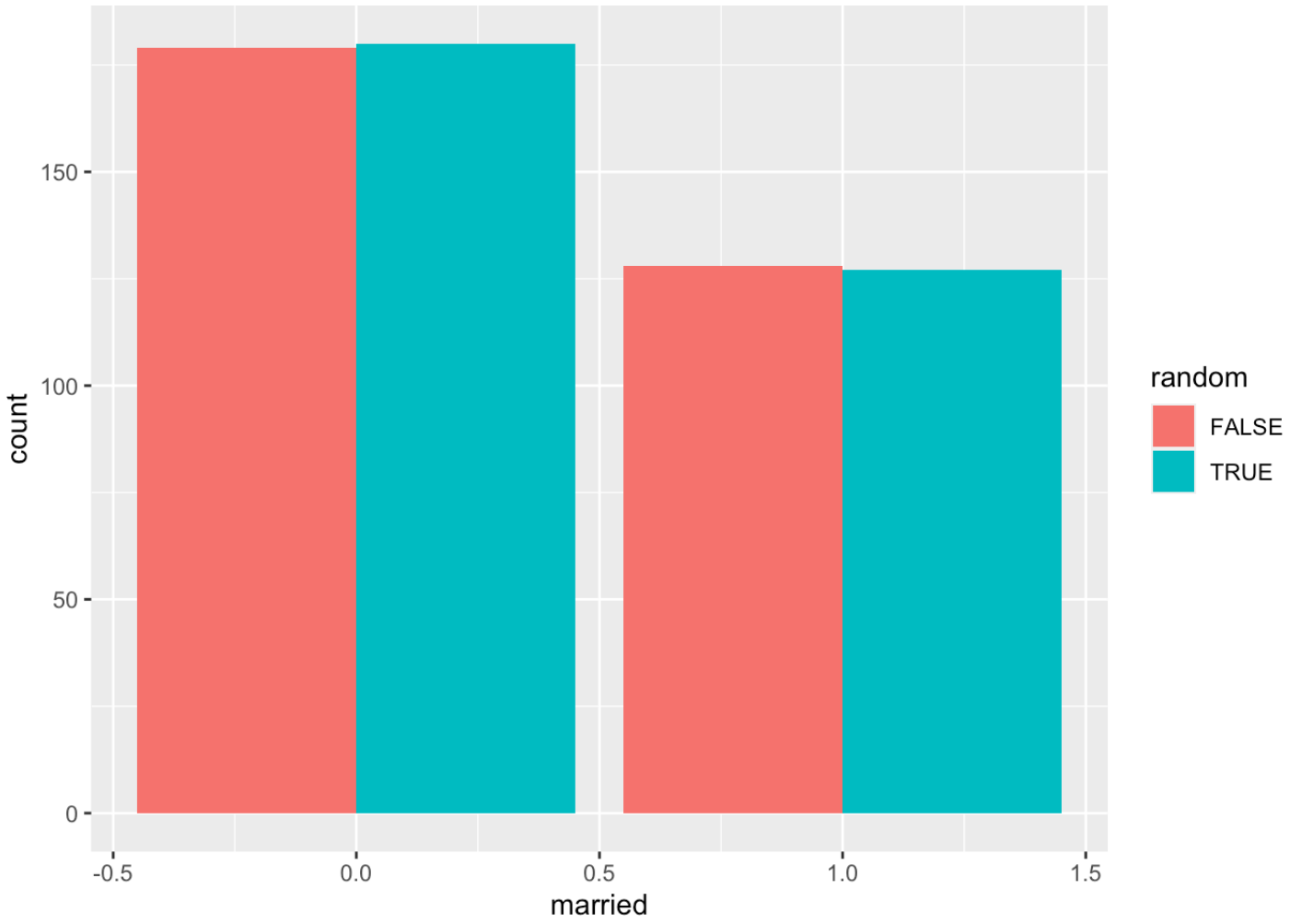

```
ggplot(lalonde, aes(x=black, fill=random)) + geom_bar(position = 'dodge')
```



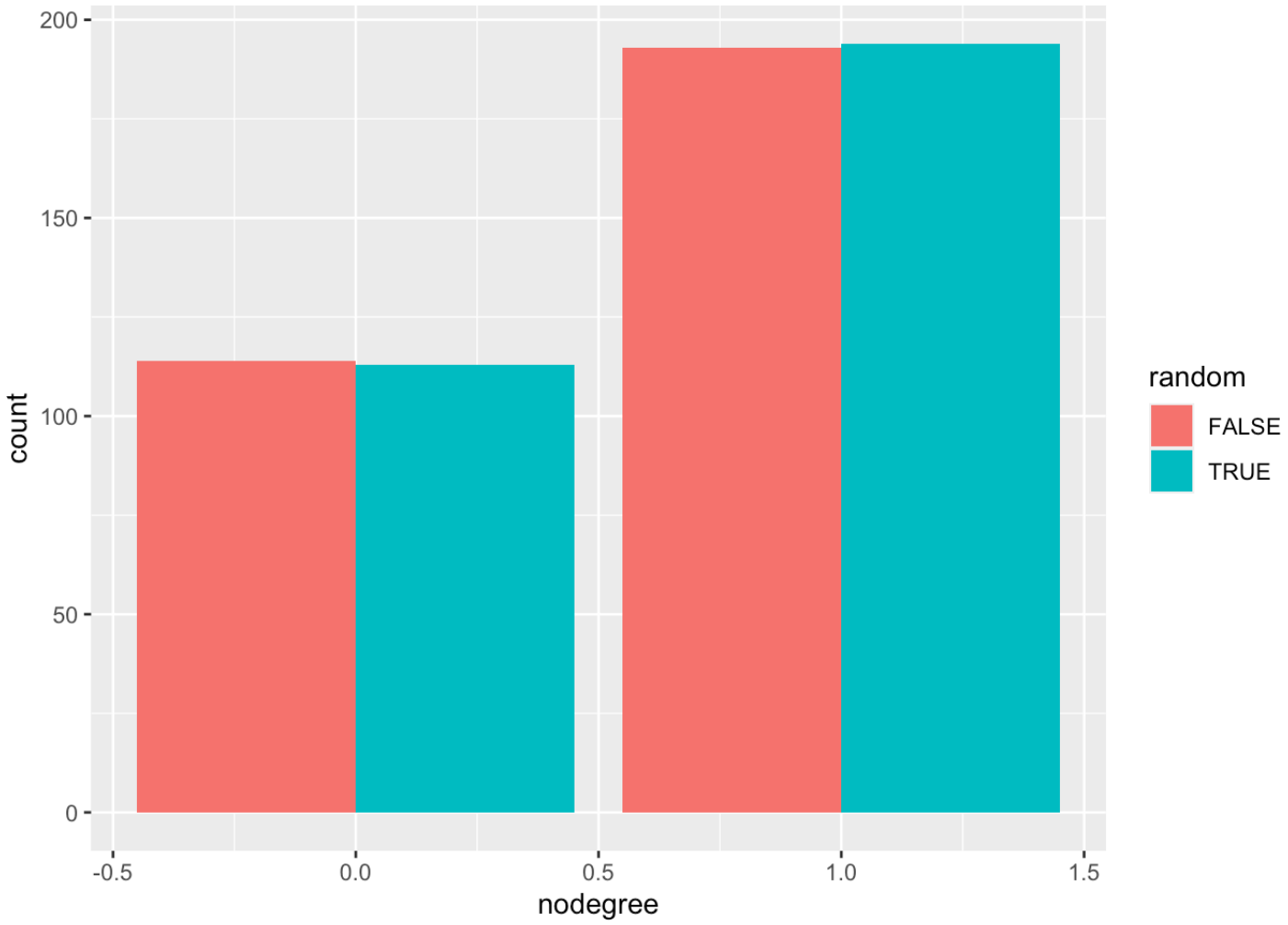

```
ggplot(lalonde, aes(x=hispan, fill=random)) + geom_bar(position = 'dodge')
```



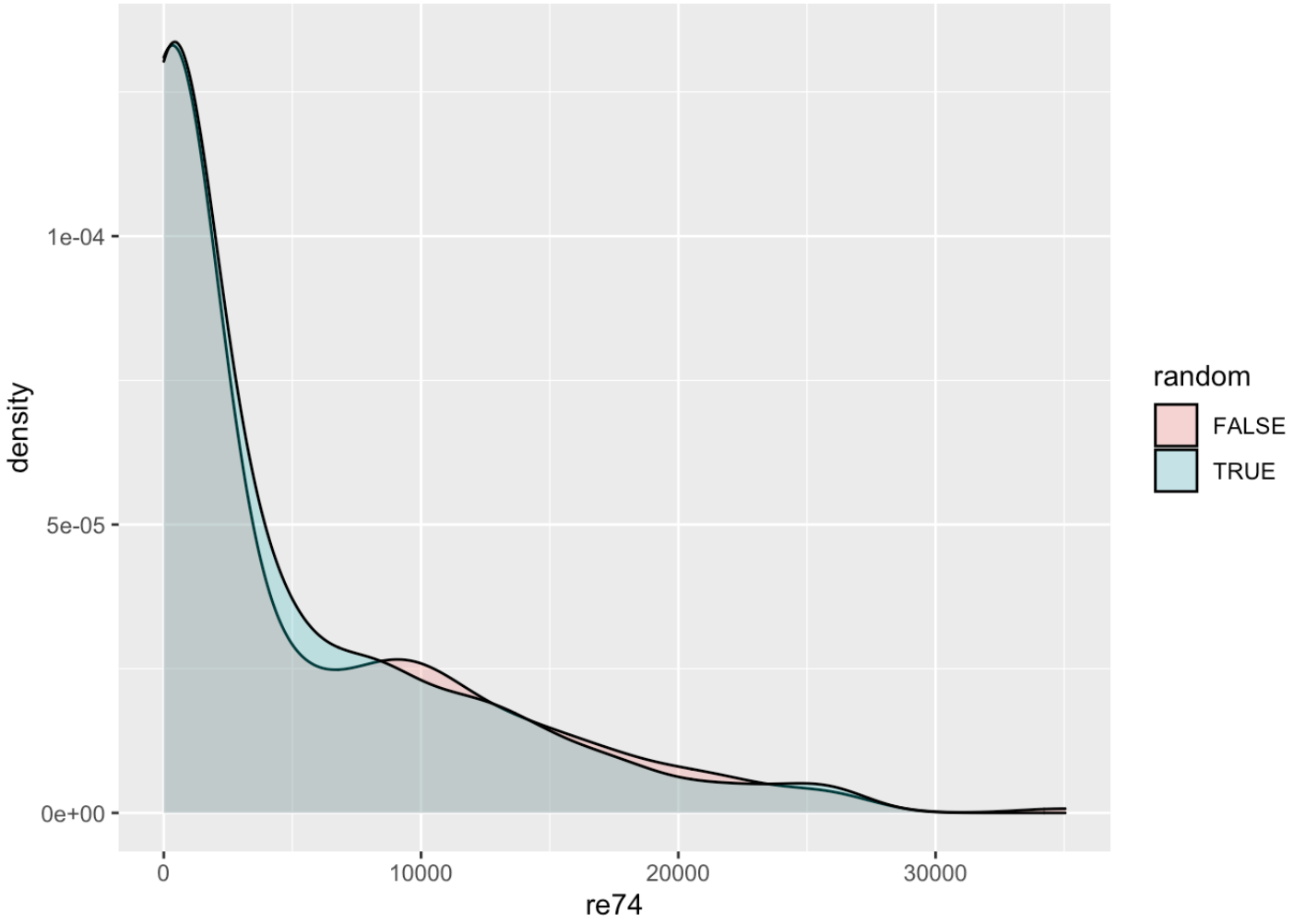

```
ggplot(lalonde, aes(x=married, fill=random)) + geom_bar(position = 'dodge')
```



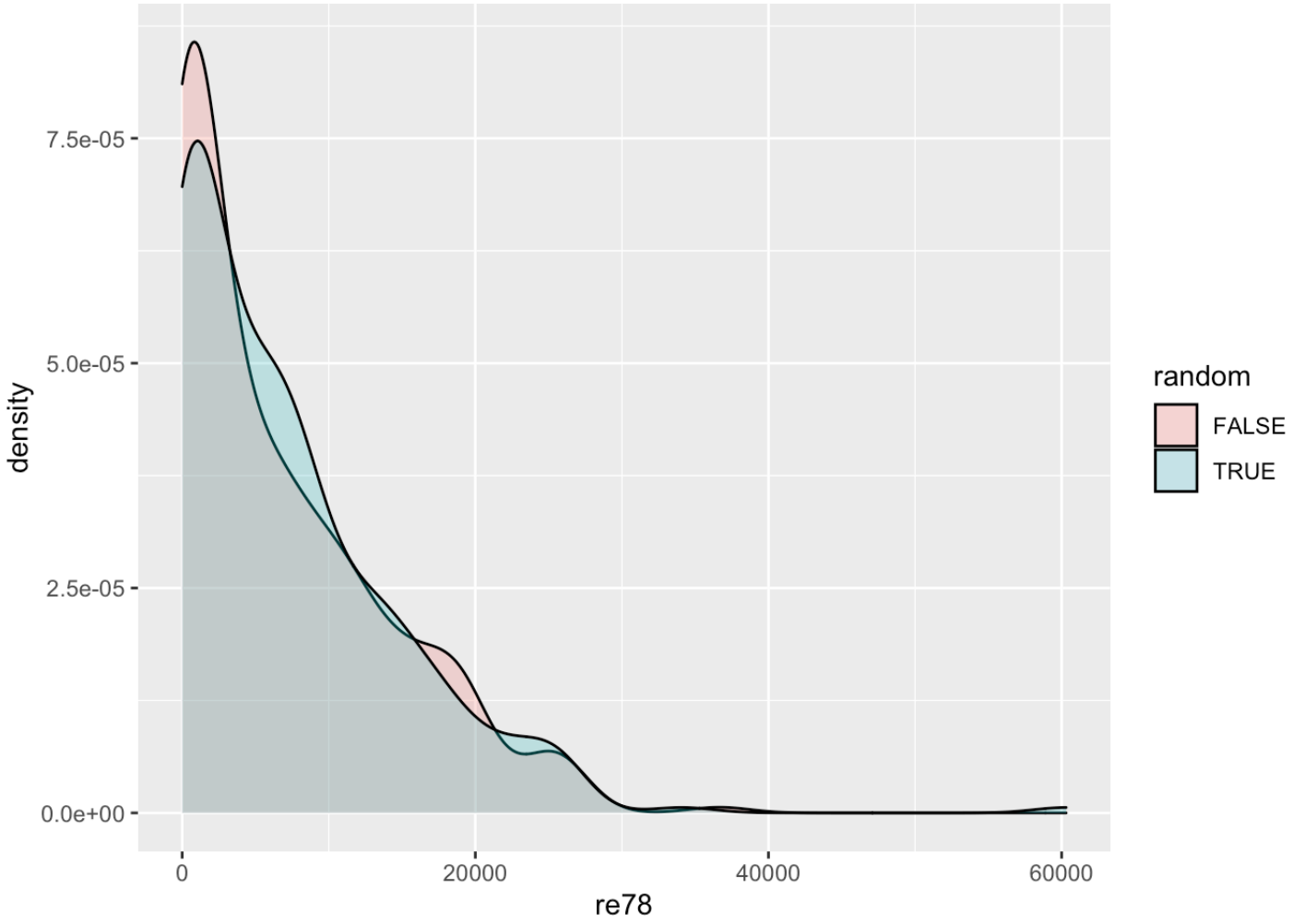

```
ggplot(lalonde, aes(x=nodegree, fill=random)) + geom_bar(position = 'dodge')
```




```
ggplot(lalonde, aes(x=re74, fill=random)) +  geom_density(alpha=0.25)
```




```
ggplot(lalonde, aes(x=re78, fill=random)) +  geom_density(alpha=0.25)
```



- Using natural randomization: In 1973 the Eldfell volcano in Iceland on the island of Heimaey erupted, destroying about 400 homes. The Icelandic government compensated those who lost their homes, many never returned. An economics paper (https://www.nber.org/papers/w22392.pdf) showed that, among people less than 25 years old at the time of the eruption, those who had moved averaged four more years of schooling and earnings $27,000 greater per year than those from families who had kept their home.
    - Because those who lost their homes was naturally randomized, this paper used instrumental variables
    - The treatment was moving away and the outcome was later earnings
    - Losing home -> moving away -> later earnings
    - We are not going to focus on instrumental variables analysis here
    - Used when there is a natural experiement

Inverse Probability Treatment Weighting (treatment not randomized)

- In observational data, the population receiving a treatment is difficult to compare to the other groups

because of possible confounding

- In the first example, the treatment and non-treatment populations had different severity levels, so are hard to compare directly

```
table(df$treatment, df$symptoms)
```

```
##
##                mild severe
##   experimental  136    351
##   placebo       378    135
```

```
chisq.test(df$treatment, df$symptoms)
```

```
##
##    Pearson's Chi-squared test with Yates' continuity correction
##
## data:  df$treatment and df$symptoms
## X-squared = 207.58, df = 1, p-value < 2.2e-16
```

- For comparison, we want the treatment and placebo populations to be generated from the same distribution
- This breaks any association between confounders and treatment

- IPTW uses (possible) confounders to up- or down-weight observations depending on their probability of receiving treatment
  - IPTW are sometimes called propensity scores (PS)
  - These are called propensity scores - they indicate propensity for an observation to be in the the treatment group
  - IPTW makes the populations look similar for the considered covariates
  - Side note: Sometimes people are matched on covariates. What is a possible draw back of matching? What about many covariates
- Why not use regression to control for confounders McCaffrey et all 2013 (https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.5753?casa_token=-_Iuef6qRG0AAAAA:jACp_crwgA9QTkHCNfDi-4lJuMDbCXkHDocwRzvvnky5S6rPDQ-j-Cn8QtXNvK5eCKz0ziJ5cURXTw):
  1. By summarizing all pretreatment variables to a single score, propensity scores are an important dimension reduction tool for evaluating treatment effects. This characteristic of propensity scores is particularly advantageous over standard adjustment methods when there exists a potentially large number of pretreatment covariates.
  2. Propensity score methods derive from a formal model for causal inference, the potential outcomes framework, so that causal questions can be well defined and explicitly specified and not conflated with the modeling approach as they are with traditional regression approaches.
  3. Propensity score methods do not require modeling the mean for the outcome. This can help

avoid bias from misspecification of that model.
4. Propensity score methods avoid extrapolating beyond the observed data unlike parametric regression modeling for outcomes, which extrapolate whenever the treatment and control groups are disparate on pretreatment variables.
5. Propensity score adjustments can be implemented using only the pretreatment covariates and treatment assignments of study participants without any use of the outcomes. This feature of propensity score adjustments is valuable because it eliminates the potential for the choice of model specification for pretreatment variables to be influenced by its impact on the estimated treatment effect.

- Assumptions:
  1. Sufficient overlap: For all i

$$0 < P(A = 1|C = x_i) < 1$$

  2. No unknown confounders: $A \perp Y^t|X$ for $t = 0, 1$
- IPW: For each observation, i, let

$$p_i = P(A = 1|C = x_i) = P(i \text{ gets treatment}|x_i)$$

the associated weight is

$$w_i = \begin{cases} \frac{1}{p_i} & \text{when } A = 1 \\ \frac{1}{1-p_i} & \text{when } A = 0 \end{cases}$$

- Propensity score Theorem

$$(Y^0, Y^1) \perp A|X \Rightarrow (Y^0, Y^1) \perp A|w(X)$$

- Note: This is saying that getting treatment or not is independent of what the response would have been in a counterfactual setting
- Once the weights are estimates, ATE for a binary treatment is estimated as

$$\widehat{\text{ATE}} = \frac{\sum_{i=1}^n A_i Y_i w_i}{\sum_{i=1}^n A_i w_i} - \frac{\sum_{i=1}^n (1 - A_i) Y_i w_i}{\sum_{i=1}^n (1 - A_i) w_i}$$

Estimating Weights

- There is no one standard method for model selection in the context of estimating propensity scores for IPTW for multiple treatments
- It's common to use non-parametric methods for estimating probabilities
- Note that in this context, we care more about the prediction value than the interpretation of parameter estimates
- Let's compare logistic regression and generalized boosted models (GBM) to estimate observation weights

Logistic Regression

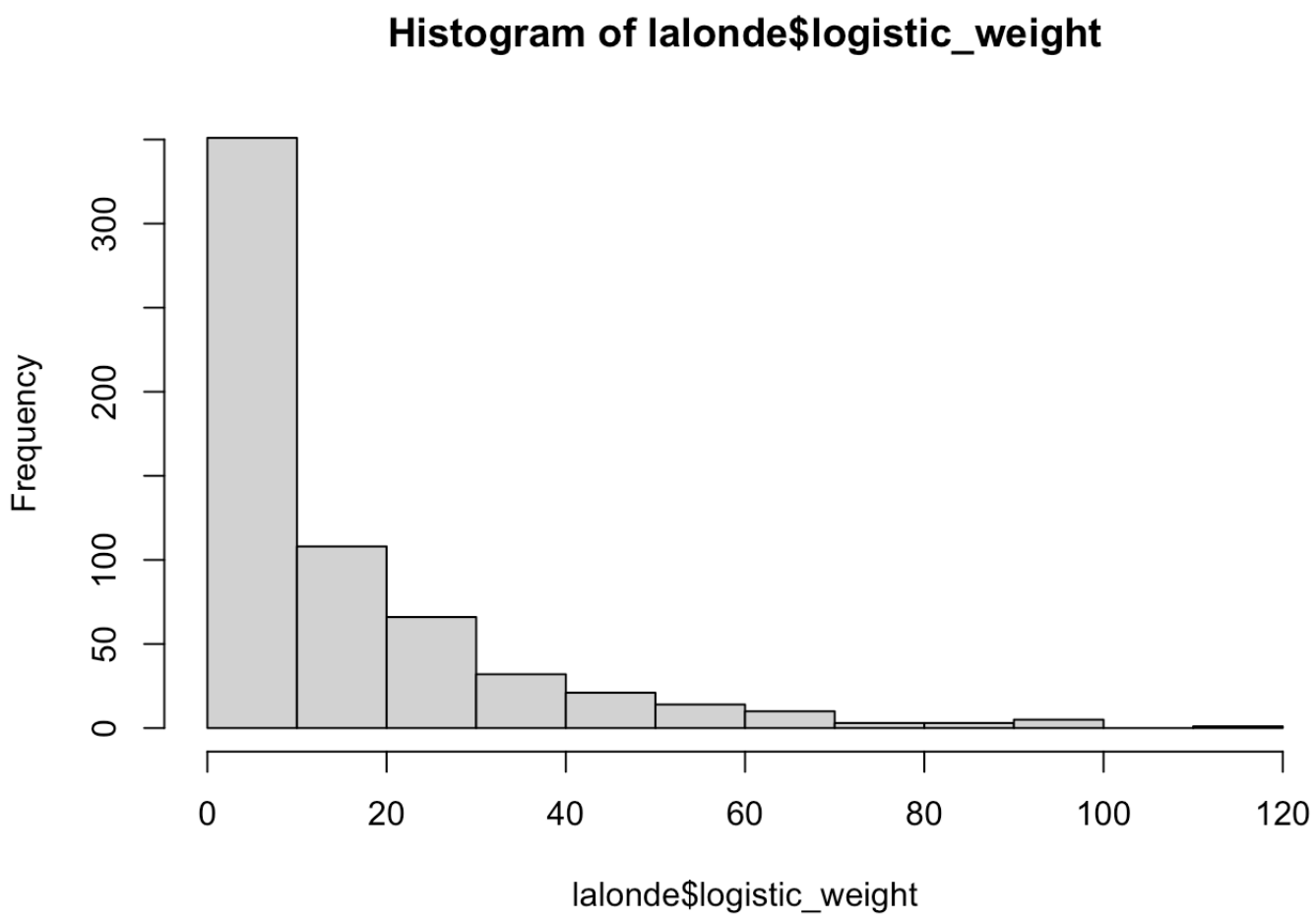
```
prop.mod <- glm(treat ~ age + educ+  black + hispan + married + nodegree + re74 + re7
5, data=lalonde, family=binomial())
summary(prop.mod)
```

```
##
## Call:
## glm(formula = treat ~ age + educ + black + hispan + married +
##     nodegree + re74 + re75, family = binomial(), data = lalonde)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7645  -0.4736  -0.2862   0.7508   2.7169
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.729e+00  1.017e+00  -4.649 3.33e-06 ***
## age          1.578e-02  1.358e-02   1.162  0.24521
## educ         1.613e-01  6.513e-02   2.477  0.01325 *
## black        3.065e+00  2.865e-01  10.699  < 2e-16 ***
## hispan       9.836e-01  4.257e-01   2.311  0.02084 *
## married     -8.321e-01  2.903e-01  -2.866  0.00415 **
## nodegree     7.073e-01  3.377e-01   2.095  0.03620 *
## re74        -7.178e-05  2.875e-05  -2.497  0.01253 *
## re75         5.345e-05  4.635e-05   1.153  0.24884
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 751.49  on 613  degrees of freedom
## Residual deviance: 487.84  on 605  degrees of freedom
## AIC: 505.84
##
## Number of Fisher Scoring iterations: 5
```

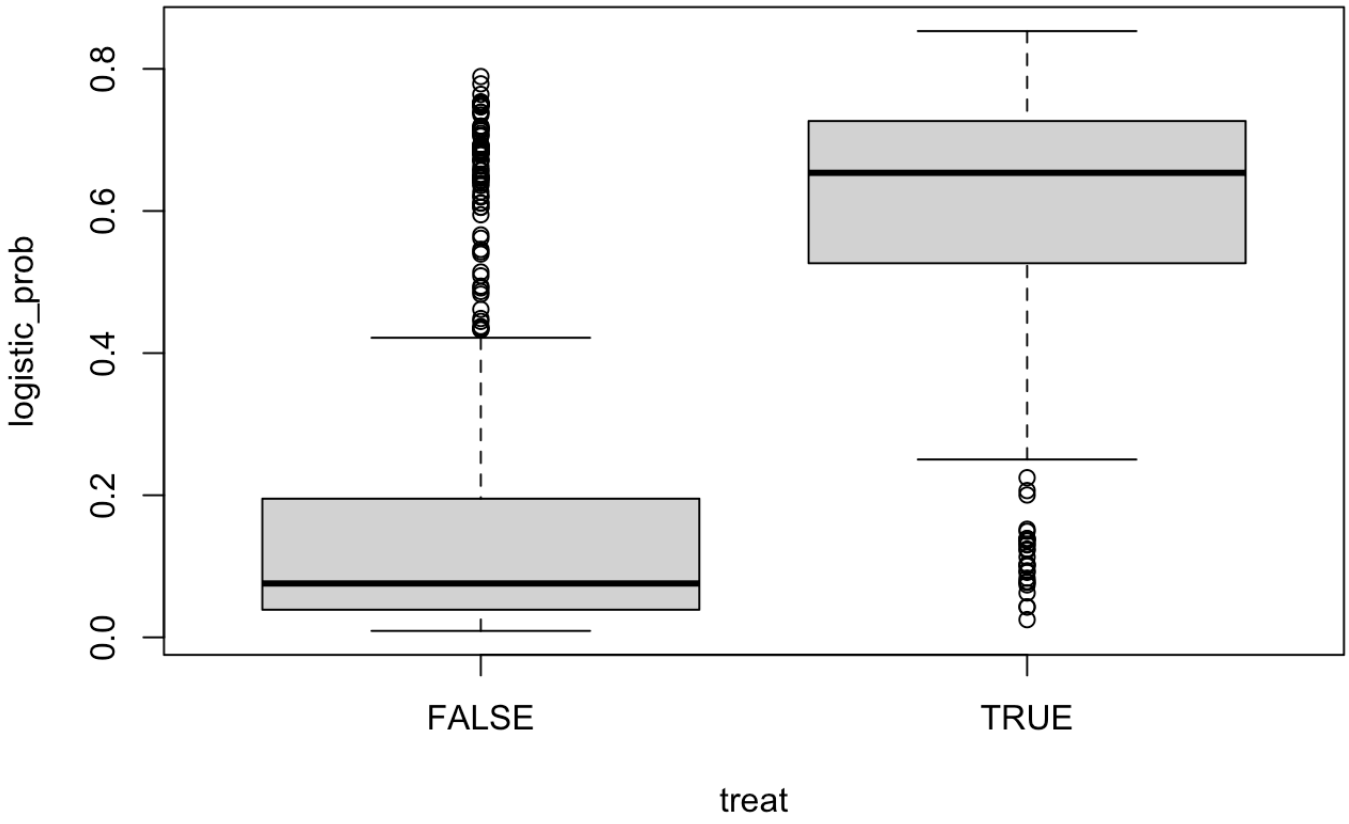
```
treat_prop_logistic <- predict(prop.mod, newdata = lalonde[,-1], type = 'response')
lalonde$logistic_prob <- treat_prop_logistic
lalonde$logistic_weight <- ifelse(lalonde$treat, 1/(1-treat_prop_logistic), 1/treat_p
rop_logistic)
head(lalonde)
```

```
##   treat age educ black hispan married nodegree re74 re75       re78 random
## 1  TRUE  37   11     1      0       1         1    0    0  9930.0460  FALSE
## 2  TRUE  22    9     0      1       0         1    0    0  3595.8940   TRUE
## 3  TRUE  30   12     1      0       0         0    0    0 24909.4500  FALSE
## 4  TRUE  27   11     1      0       0         1    0    0  7506.1460  FALSE
## 5  TRUE  33    8     1      0       0         1    0    0   289.7899   TRUE
## 6  TRUE  22    9     1      0       0         1    0    0  4056.4940  FALSE
##   logistic_prob logistic_weight
## 1     0.6387699        2.768319
## 2     0.2246342        1.289714
## 3     0.6782439        3.107944
## 4     0.7763241        4.470754
## 5     0.7016387        3.351642
## 6     0.6990699        3.323031
```

```
hist(lalonde$logistic_weight)
```



```
boxplot(logistic_prob ~ treat, data=lalonde)
```



- Assessing covariate balance after weighting

$$PSB_k = \frac{|\bar{X}_{k1} - \bar{X}_{k0}|}{\widehat{\sigma}_k}$$

where $\bar{X}_{kt} = \sum_{i=1}^{n} I(A = t) X_{ki} w_i / \sum_{i=1}^{n} I(A = t) w_i$

Boosting

- A problem with logistic regression for estimating weights is that can be challenging to get a good model fit
- Boosting makes this a lot easier and automatic
- We already talked about AdaBoost and a little about general boosting
- For a reference, in AdaBoost
    - each lazy learner is fitted to the data with weights that depend on the last model's performance
    - all of the lazy learners are scored depending on their performance
- For general boosting

- Observations are *not* weighted
- We use learners (models) like regression or trees that are more complex than lazy learner
- Similar to AdaBoost, we must choose the number of sequential learner to use, M
- Each learner tries to predict the *error* of the previous model
- Error: in the literature, the error is the negative gradient of the loss function with respect to the model evaluated at an observation:

$$-\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}$$

- For a continuous outcomes, this is $y_i - f(x_i)$ for continuous outcomes where f is a continuous learner
- For a binary outcomes, this is $I(y_i = 1) - p(x_i)$ where p is a learner like logistic regression model or decision tree
- For categorical outcomes, this is $I(y_i = \text{Category } k) - p_k(x_i)$ where p_k is models the probability of category k
- Learning rate: to avoid over fitting, we reduce the contribution of each learner by $0 < \nu < 1$
- Trees are the most common learner to use for boosting, usually with between 8 and 32 terminal leaves

Gradient Tree Boosting Algorithm from ESL

1. Initialize $f_0(x) = \arg\min_\gamma \sum_{i=1}^n L(y_i, \gamma)$
2. For $m = 1$ to M:
    1. For $i = 1, \ldots, n$ compute the error

$$r_{im} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right]_{f=f_{m-1}}$$

    2. Fit a regression tree to the errors, r_{im} with terminal regions R_{jm} where $j = 1, \ldots, J_m$
    3. For $j = 1, \ldots, J_m$ compute

$$\gamma_{jm} = \arg\min_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

    4. Update $f_m(x) = f_{m-1}(x) + \nu \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$
3. Output $\hat{f}(x) = f_M(x)$

Generalized Boosting for Propensity Scores

- For causal inference, we want to weight the observation to make the treatment and control groups look as if they were randomized
- Using inverse probabilities from boosting model, we want to use as many trees as we need to achieve balance in the covariates

- Ideally, we want the distributions of each covariate in the treatment and control groups to be similar (mean, variance, skew, shape, etc)
- Clearly this is rarely possible, so we might restrict balance to mean and/or standard deviation for example
- Assessing covariate balance with *standardize bias* estimation

$$SB_k = \frac{|\bar{X}_{k1} - \bar{X}_{k0}|}{\hat{\sigma}_k}$$

- Generally, standardized mean differences of less than 0.20are considered small, 0.40 are considered moderate, and 0.60 are considered large
- This cutoff can change within fields and between investigator
- McCaffery et al: $SB > 0.2$ is problematic
- Below the `twang` R package automatically add mores trees until a stopping rule based on balance it met
- `twang` chooses propensity scores based on the boosted model with the best balance
- see twang documentation (https://cran.r-project.org/web/packages/twang/twang.pdf)

```
boosted.mod <- ps(treat ~ age + educ + black + hispan + married + nodegree + re74 + r
e75,
                  data=lalonde,
                  estimand = "ATE",
                  n.trees = 5000,
                  interaction.depth=2,
                  perm.test.iters=0,
                  verbose=FALSE,
                  stop.method = c("es.mean"))
summary(boosted.mod)
```

```
##             n.treat n.ctrl ess.treat ess.ctrl    max.es   mean.es    max.ks
## unw             185    429  185.0000   429.00 1.3085999 0.4432341 0.6404460
## es.mean.ATE     185    429   60.3237   193.34 0.5531452 0.1873810 0.2707165
##             max.ks.p   mean.ks iter
## unw               NA 0.2702451   NA
## es.mean.ATE       NA 0.1168722 2476
```

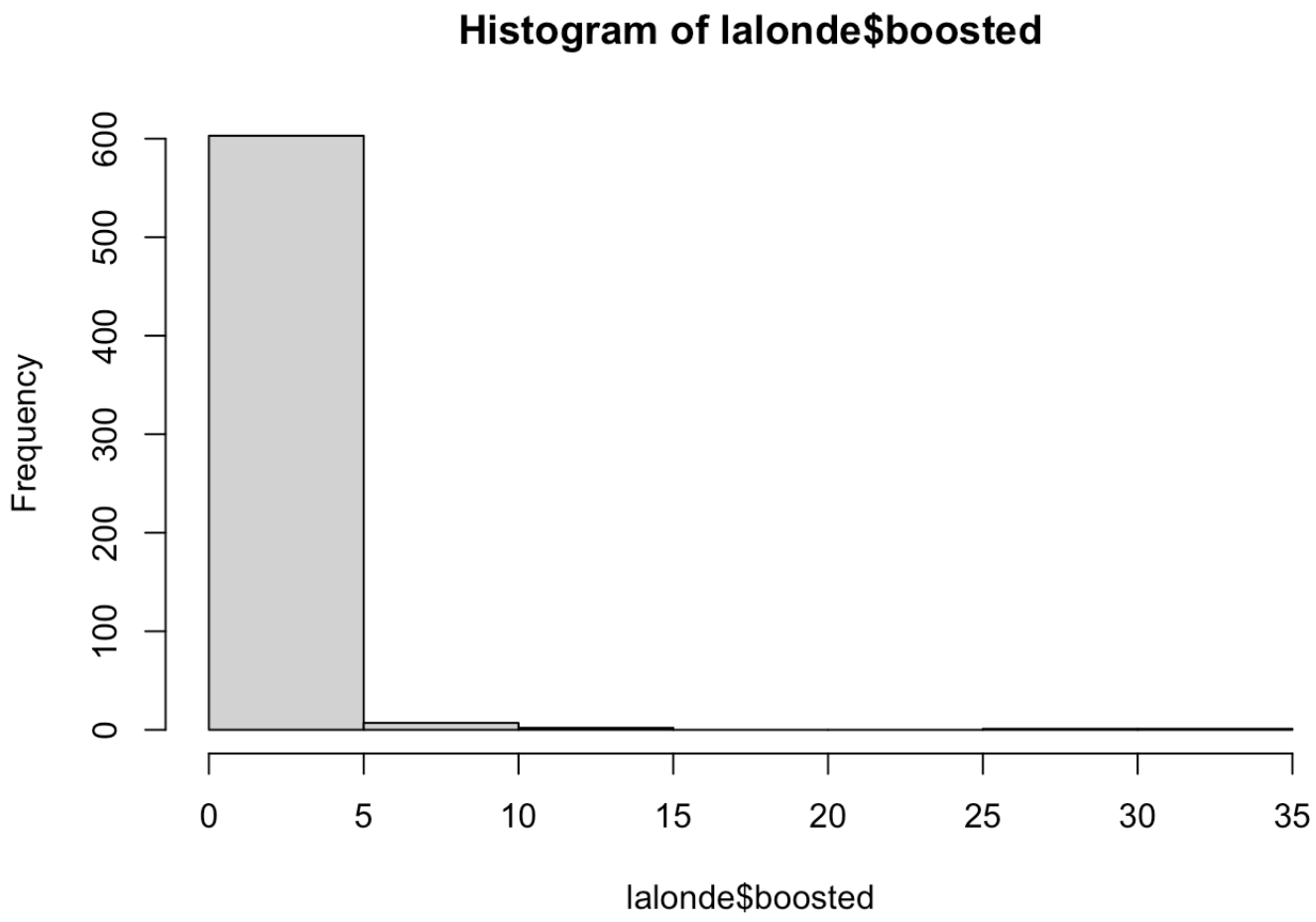
```
summary(boosted.mod$gbm.obj,
        n.trees=boosted.mod$desc$es.mean.ATE$n.trees,
        plot=FALSE)
```

```
##                var     rel.inf
## black       black 56.05496593
## re74         re74 16.55677812
## age           age 16.48874057
## re75         re75  4.21479167
## educ         educ  3.16741532
## married   married  2.97674945
## nodegree nodegree  0.44563787
## hispan     hispan  0.09492107
```

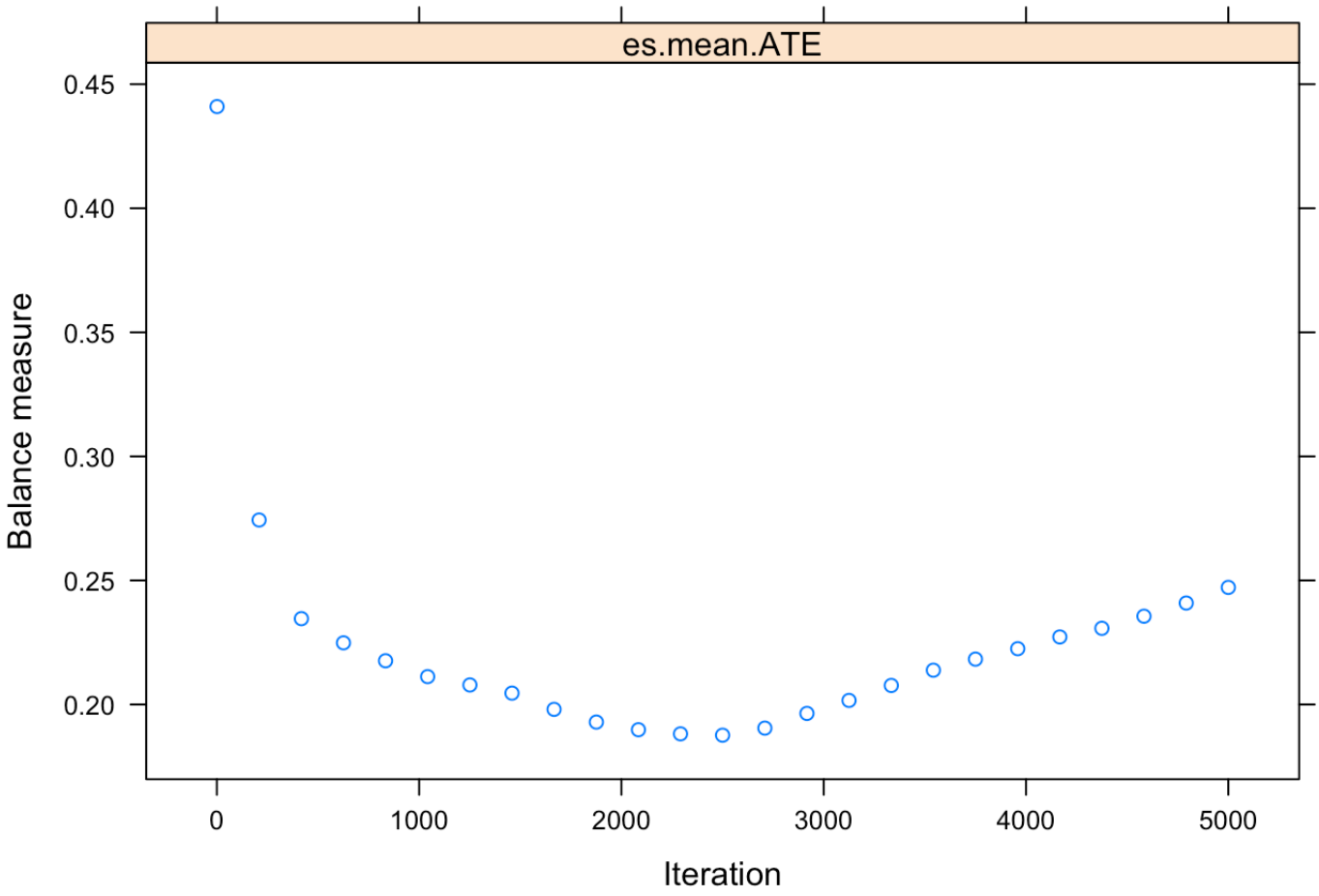
```
lalonde$boosted <- get.weights(boosted.mod)
```

```
## Warning in get.weights(boosted.mod): No stop.method specified.  Using es.mean.ATE
```

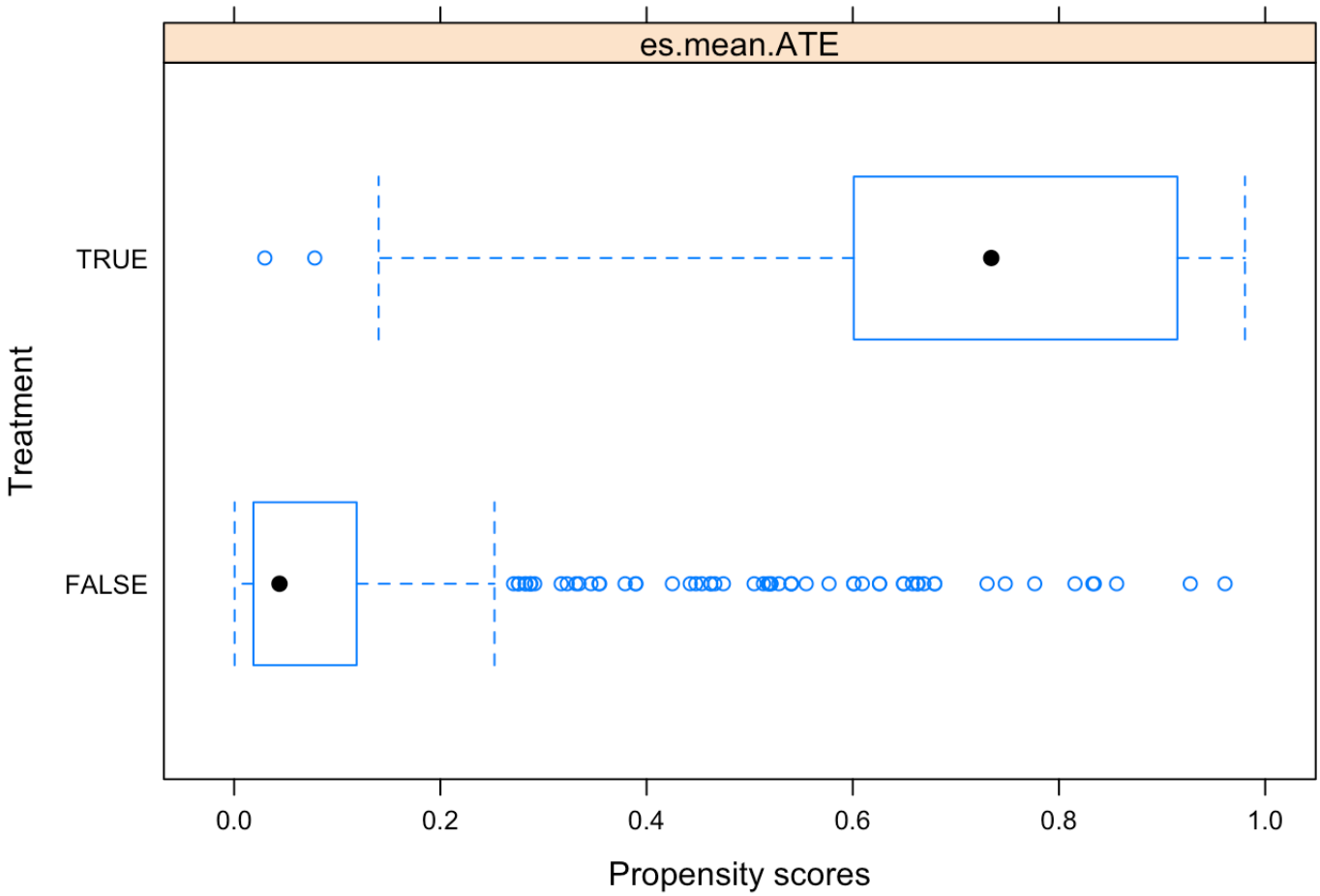
```
hist(lalonde$boosted)
```



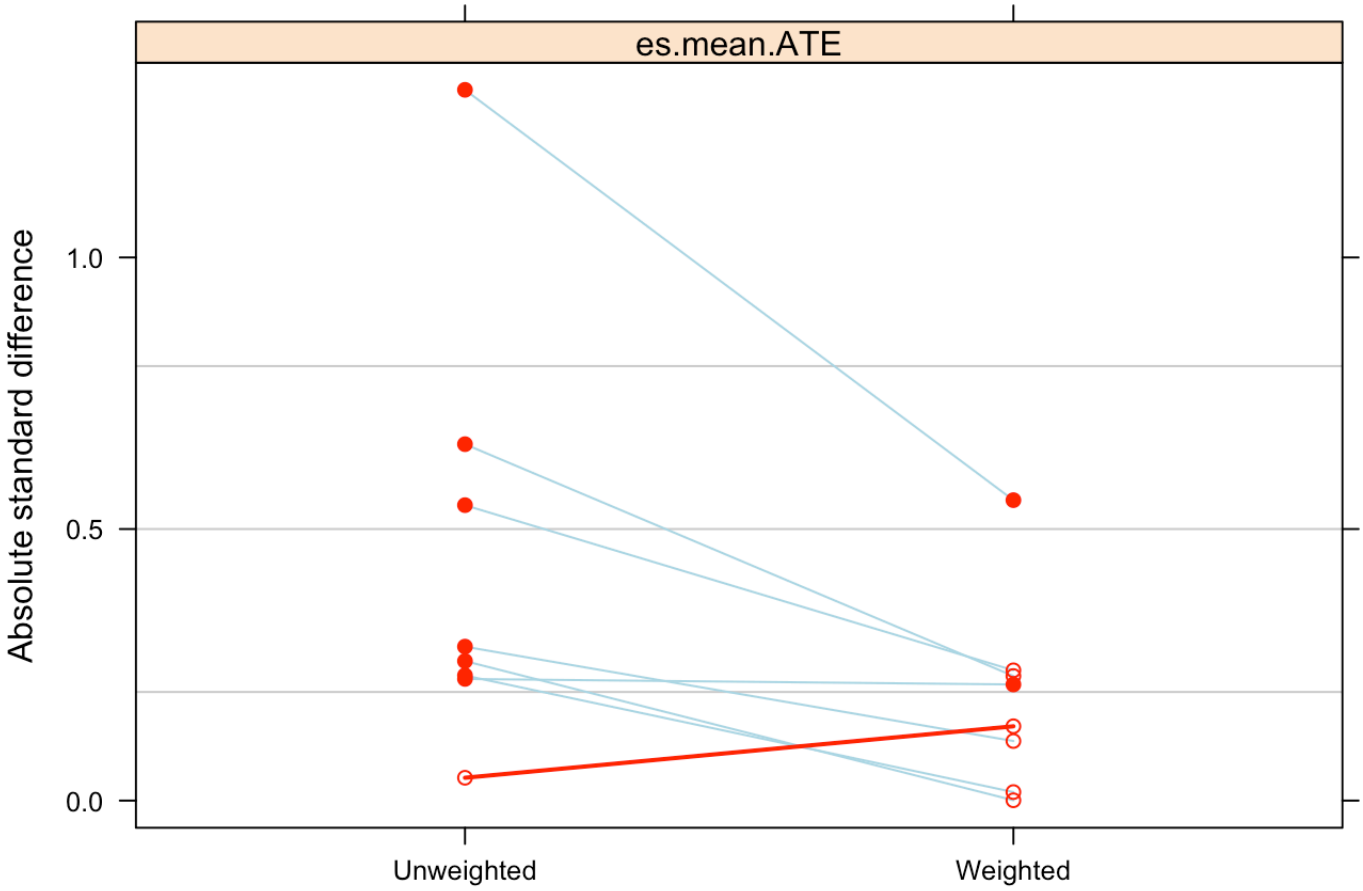
```
plot(boosted.mod)
```




```
plot(boosted.mod, plots=2)
```




```
plot(boosted.mod, plots=3)
```




```
bal.table(boosted.mod)
```

```
## $unw
##             tx.mn    tx.sd    ct.mn    ct.sd std.eff.sz   stat     p    ks
## age        25.816    7.155   28.030   10.787     -0.224 -2.994 0.003 0.158
## educ       10.346    2.011   10.235    2.855      0.042  0.547 0.584 0.111
## black       0.843    0.365    0.203    0.403      1.309 19.371 0.000 0.640
## hispan      0.059    0.237    0.142    0.350     -0.257 -3.413 0.001 0.083
## married     0.189    0.393    0.513    0.500     -0.656 -8.607 0.000 0.324
## nodegree    0.708    0.456    0.597    0.491      0.231  2.716 0.007 0.111
## re74     2095.574 4886.620 5619.237 6788.751     -0.544 -7.254 0.000 0.447
## re75     1532.055 3219.251 2466.484 3291.996     -0.284 -3.282 0.001 0.288
##          ks.pval
## age        0.003
## educ       0.074
## black      0.000
## hispan     0.317
## married    0.000
## nodegree   0.074
## re74       0.000
## re75       0.000
##
## $es.mean.ATE
##             tx.mn    tx.sd    ct.mn    ct.sd std.eff.sz   stat     p    ks
## age        25.366    7.051   27.479   10.057     -0.214 -2.737 0.006 0.135
## educ       10.678    2.035   10.318    2.682      0.137  1.170 0.242 0.090
## black       0.634    0.483    0.364    0.482      0.553  3.087 0.002 0.271
## hispan      0.117    0.323    0.117    0.322      0.001  0.004 0.997 0.000
## married     0.317    0.467    0.430    0.496     -0.229 -1.324 0.186 0.113
## nodegree    0.594    0.492    0.601    0.490     -0.016 -0.095 0.924 0.008
## re74     3033.625 5242.729 4587.500 6420.671     -0.240 -1.877 0.061 0.168
## re75     1784.296 3375.528 2145.837 3190.334     -0.110 -0.940 0.348 0.151
##          ks.pval
## age        0.340
## educ       0.814
## black      0.002
## hispan     1.000
## married    0.555
## nodegree   1.000
## re74       0.133
## re75       0.217
```

Estimating Average Treatment Effect (ATE)

- Want $E[Y^1] - E[Y^0]$
- Treatment population mean estimate for $t = 0, 1$:

$$\widehat{\mu}_t = \frac{\sum_{i=1}^{n} I(T_i = t) Y_i w_i(t)}{\sum_{i=1}^{n} I(T_i = t) w_i(t)}$$

- Estimate $E[Y^1] - E[Y^0]$ as

$$\widehat{\text{ATE}} = \widehat{\mu}_1 - \widehat{\mu}_0$$

- We can use a weighted *t-test* to evaluate $\widehat{\text{ATE}}$
- McCaffery et all suggests using `svyglm` which achieves the same goal

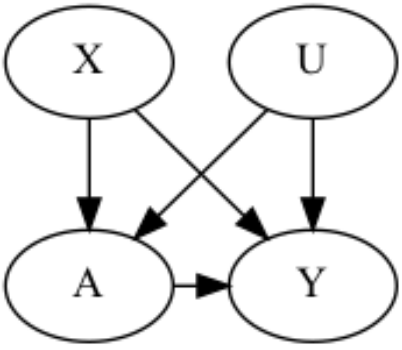
```
library(survey)
design <- svydesign(ids=~1, weights=~boosted, data=lalonde)
glm1 <- svyglm(re78 ~ treat, design=design)
summary(glm1)
```

```
##
## Call:
## svyglm(formula = re78 ~ treat, design = design)
##
## Survey design:
## svydesign(ids = ~1, weights = ~boosted, data = lalonde)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6646.4      388.3  17.116   <2e-16 ***
## treatTRUE     -719.5      865.5  -0.831    0.406
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 49534038)
##
## Number of Fisher Scoring iterations: 2
```

```
summary(lm(re78 ~ treat + age + educ+  black + hispan + married + nodegree + re74 + r
e75, data=lalonde))
```

```
##
## Call:
## lm(formula = re78 ~ treat + age + educ + black + hispan + married +
##     nodegree + re74 + re75, data = lalonde)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13595  -4894  -1662   3929  54570
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.651e+01  2.437e+03   0.027   0.9782
## treatTRUE    1.548e+03  7.813e+02   1.982   0.0480 *
## age          1.298e+01  3.249e+01   0.399   0.6897
## educ         4.039e+02  1.589e+02   2.542   0.0113 *
## black       -1.241e+03  7.688e+02  -1.614   0.1071
## hispan       4.989e+02  9.419e+02   0.530   0.5966
## married      4.066e+02  6.955e+02   0.585   0.5590
## nodegree     2.598e+02  8.474e+02   0.307   0.7593
## re74         2.964e-01  5.827e-02   5.086 4.89e-07 ***
## re75         2.315e-01  1.046e-01   2.213   0.0273 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6948 on 604 degrees of freedom
## Multiple R-squared:  0.1478, Adjusted R-squared:  0.1351
## F-statistic: 11.64 on 9 and 604 DF,  p-value: < 2.2e-16
```

- interpretation: causal vs statistical
- Limitation: Umeasured confounding



Writing tips

- Intro paragraph:
  - Give the big picture in one sentence. What is the general application field?
  - What is the problem you are attempting to answer?
  - How will you answer it?

    - In one sentence, what did you find?
- Give the baseline characteristics of the data after cleaning
    - How many observations with salient brake down
    - A table is a very efficient way to summarize data:

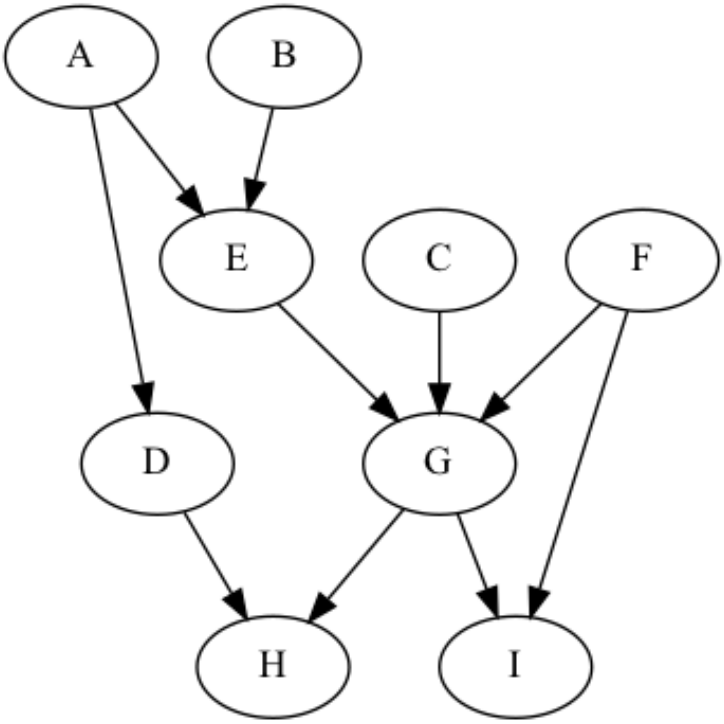
	Total (n)	**Treatment (n1)**	**Control (n2)**	**p-value**
Covariate 1	mean (sd)	mean (sd)	mean (sd)	<0.001
Covariate 2	%	%	%	0.34
Covariate 3	median (q1, q3)	median (q1, q3)	median (q1, q3)	0.02

- Present you most important model (or models)
    - Clearly interpret the parameter of interest and its connection to the larger question that is being asked
    - A lot of this material is technical. You're job is to understand the larger statistical picture and communicate it in an easy to understand way to someone who does not have any background in statistics an who probably does not do math regularly
- Example: how would you describe IPW to a client?
    1. Why do we use IPW in the first place?
    2. What does IPW do?
    3. How does IPW attempt to allow a causal interpretation for ATE?
- Conclusion:
    - Restate what you attempted to answer
    - State any potential limitation
    - Reiterate findings
- Don't
    - try to explain d-separation or Bayesian networks
    - think of a statistician as your audience
- Do
    - rely on intuition about causation
    - use your understanding of causal modeling to answer the question

Graphical Causal Models and D-separation

- Researchers frequently use *directed acyclic graphs* (DAG) to visualize causal relationships between variables in a dataset
    - directed - all edges in graph have direction (all edges are arrows)
    - acyclic - arrow directions do not create a loop
    - DAGs are sometimes called Bayesian networks
- DAG below
    - A, B, C, D, E, F, G, H, I represent variables in a dataset
    - Arrows indicate a direct causal relationship

- A is the *parent* of D
- D is the *child* of A
- B is the *ancestor* of I
- I is the *successor* of B



- Here: we are interested in paths between variables
  - Note: paths can have arrows pointing either direction
  - One path between D and F is

$$D \leftarrow A \rightarrow E \rightarrow G \leftarrow F$$

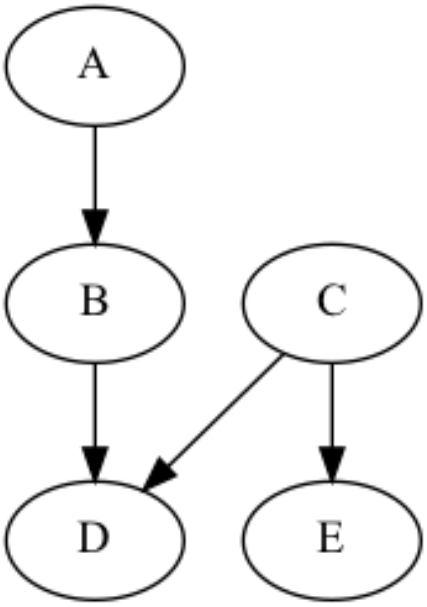
  - Another is

$$D \rightarrow H \leftarrow G \rightarrow I \leftarrow F$$

  - These are of interest if we want to know if D and F are associated or causally linked
- Somewhat technical assumptions: in general assume arrows between nodes indicate a causal relationship
- Text on Probabilistic Graphical Models (https://mitpress.mit.edu/books/probabilistic-graphical-models) for more detail

Factorization and conditional independence

- In causal inference we frequently want to isolate the average causal effect between two variables given some knowledge of the surrounding Bayesian network, i.e. potential confounders or colliders
- In reality we rarely know the Bayesian structure of a dataset, but these concepts help inform causal inference.



- Joint distribution can factored as

$$P(A, B, C, D, E) = P(A)P(B|A)P(C)P(D|B, C)P(E|C)$$

- DAGs give information about conditional independence relationships among variables
- This DAG has one path from A to E: $A \rightarrow B \rightarrow D \leftarrow C \rightarrow E$
- How can we use this DAG to determine statistical associations between variables/nodes?
- With 3 nodes:
  - Chain: $A \rightarrow B \rightarrow C$
  - Other Chain: $A \leftarrow B \leftarrow C$
  - Fork: $A \leftarrow B \rightarrow C$
  - Collider: $A \rightarrow B \leftarrow C$

Structure	**Path type**	A, C **independent/dependent?**	A, C **conditionally independent/dependent given** B
Chain	$A \rightarrow B \rightarrow C$	$A \not\perp\!\!\!\perp C$	$A \perp C \mid B$
Other Chain	$A \leftarrow B \leftarrow C$	$A \not\perp\!\!\!\perp C$	$A \perp C \mid B$
Fork (Confounder)	$A \leftarrow B \rightarrow C$	$A \not\perp\!\!\!\perp C$	$A \perp C \mid B$
Collider	$A \rightarrow B \leftarrow C$	$A \perp C$	$A \not\perp\!\!\!\perp C \mid B$

- Questions:
  1. Are A and B associated?
  2. Are A and D associated?
  3. Are D and E associated?
  4. Are B and C associated?

5. Are A and E associated?

Rule: If a path has no colliders, then yes and we say that the path is open. If there is a collider on the path, then no and we say that the path is blocked

- What out conditioning?
- Questions:

6. Are A and D associated given B?
7. Are D and E associated given C?
8. Are B and C associated given D?
9. Are A and C associated given D?
10. Are A and E associated given D?
11. Are A and E associated given B?
12. Are A and E associated given B, D?

Rule: Conditioning on a collider opens a path. Conditioning on a non-collider blocks a path

- Note: A variable/node and be a collider on one path but not on another.

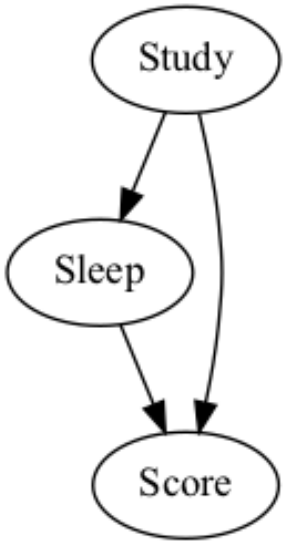
How can we show that if $D \leftarrow C \rightarrow E$ then $D \perp E|C$?

- $D \perp E|C$ if and only if $P(D, E|C) = P(D|C)P(E|C)$

$$P(D, E|C) = \frac{P(D, E, C)}{P(C)} = \frac{P(C)P(D|C)P(E|C)}{P(C)} = P(D|C)P(E|C)$$

What about $A \rightarrow B \rightarrow C$?

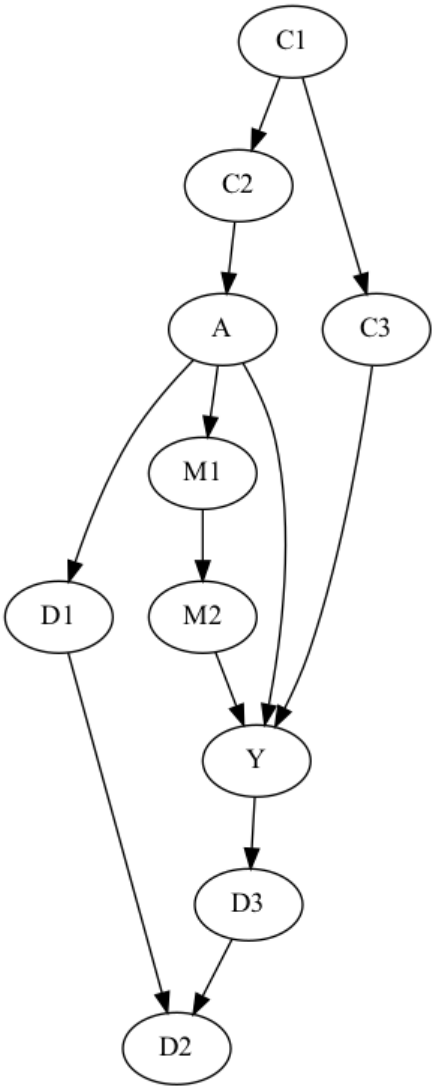
Markov Assumption



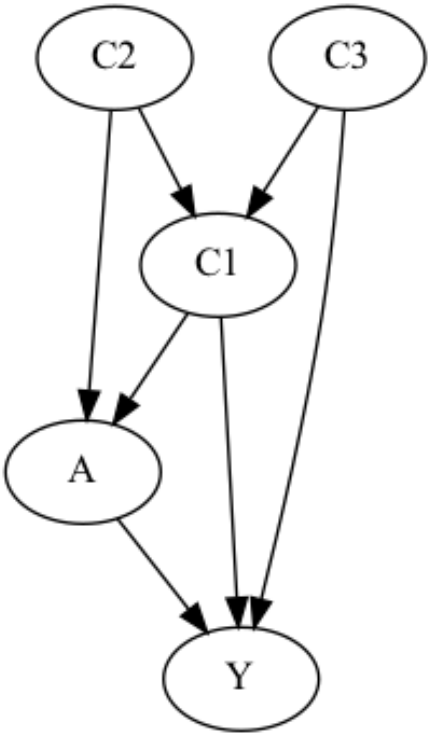
- We are implicitly using the Markov assumption here
- Specifically, we are assuming different paths won't perfectly cancel out
    - I want a good score on my test so I continue studying

- With the additional studying, I sleep less
- With less sleep, my performance isn't as good
- With the extra studying, I know the material better
- Can my lack of sleep and extra studying perfectly cancel out?
- Markov assumption says no

Regression with knowledge of causal structure



- We want to isolate the causal effect of A on Y
- Assume there are no other relevant variable other than what is in the DAG
- Which variables need to be taken into account and which should be avoided?


C2
C3
C1
A
Y