# Regression Adjustment

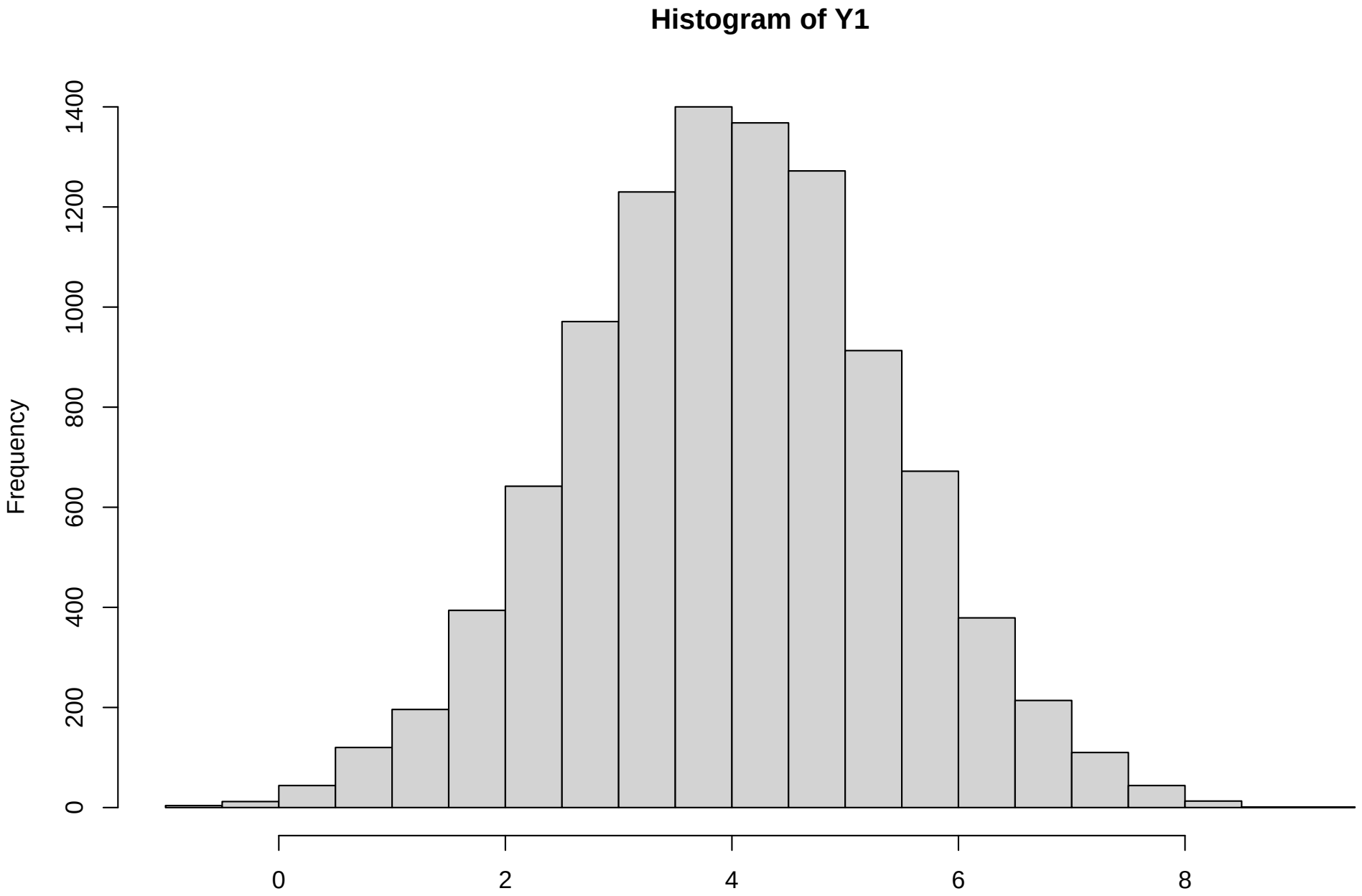
David M. Vock

PubH 7485/8485

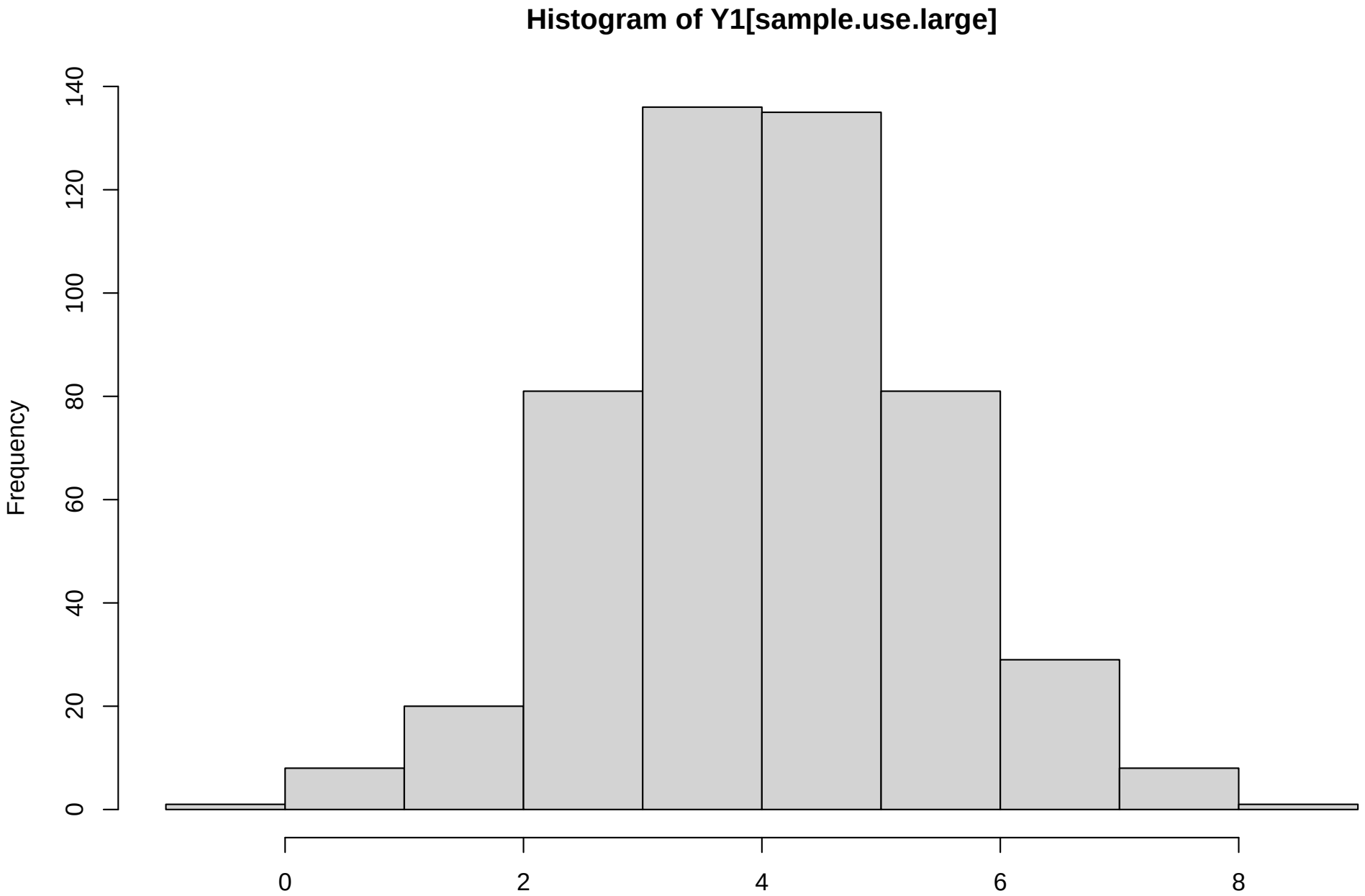
# Observational Studies

- Observational study: Individuals are not assigned to treatment intervention by an experimental design
- Often unethical or impractical to do randomized trial
- In general, $A$ is not independent of $\{Y^1, Y^0\}$
- Thus, the distribution of $Y|A = 1$ or $Y^1|A = 1$ does not equal the distribution of $Y^1$ and similarly the distribution of $Y|A = 0$ does not equal the distribution of $Y^0$
- Why? Confounding
- Heuristically, those who receive treatment may be inherently different than those who do not. Consequently, the associational parameters may reflect such inherent differences as well as any effect of treatment
- In other words, there are common causes of both $Y$ and $A$

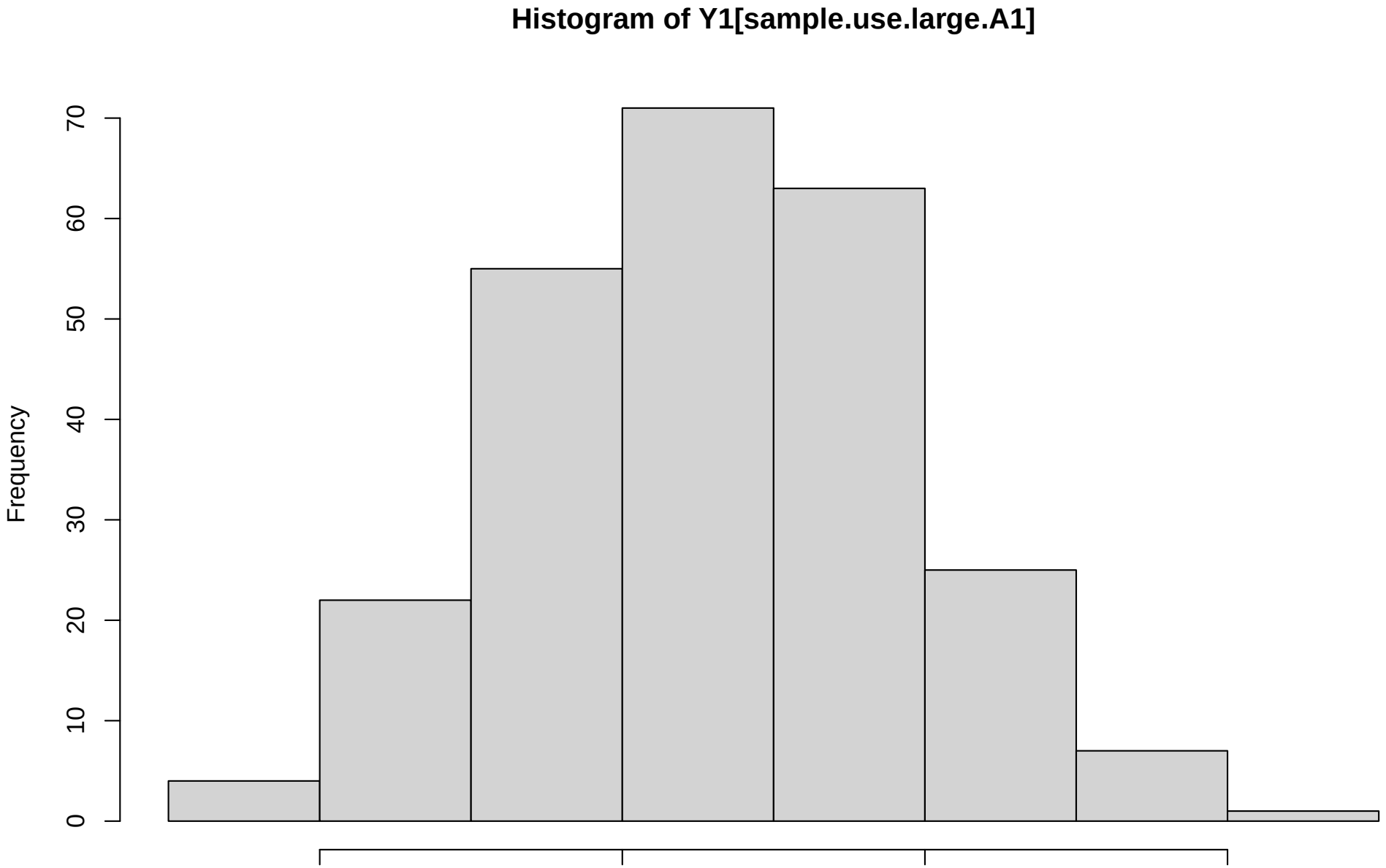
# Original Population



# Identify 200 Participants for Study



# Nonrandom Assignment of Treatment (only observe $Y^1$ on these participants)



# Challenge of Observational Studies

- Key challenge of observational studies - patients receiving treatment 1 may not be prognostically similar to those receiving treatment 0
- In potential outcome notation $\{Y^1, Y^0\}$ are not independent of $A$
- Note that this was the key assumption which allowed us to use sample averages to estimate $E(Y^1)$ and $E(Y^0)$ consistently

# No Unmeasured Confounders

- If we can identify all the pretreatment variables which are believed to explain part of the prognostic variation AND treatment choice, then it may be reasonable to assume that treatment assignment is otherwise random.
- That is $\{Y^1, Y^0\}$ are independent of $A$ conditioned on $X$. In other words within levels of a covariate it is as if we did a randomized study
- Known as the strong ignorability assumption or no unmeasured confounders
- Unidentifiable assumption

# A note on No Unmeasured Confounders

- If the treatment/intervention/action/exposure is selected by natural choice, then at some level the no unmeasured confounders assumption must be true.
- When a physician is deciding which treatment to give a patient, clearly she does not know what the potential outcomes of the patient are.
- Therefore, if all the relevant information that may affect treatment decisions were captured in the data $X$ , then the no unmeasured confounders assumption would be tenable.
- Problem is that if there is some information that affects treatment choice (and outcome) which is not captured in the data $X$ available to the analyst.

# Regression Estimator

- We will argue that under the assumptions of consistency and no unmeasured confounding
  $E\{E(Y|A=1,X) - E(Y|A=0,X)\} = E(Y^1 - Y^0)$
- Note the first expectation is over the distribution of $X$
- This says that I can find the treatment effect within each level of the covariate $X$ and then weight the treatment effect within each level of the covariate by the frequency of each level of $X$

# Regression Estimator

- Assume that $X$ is binary (say young versus old)
- Our assumption is that within each level of $X$ it is as if a randomized trial were done.
- Can find the ATE in young by taking difference in sample average between treatment and control in this stratum (similar for old)
- Overall ATE would be a weighted average of ATE in young and ATE in old

# Regression Modeling

- $X$ may be continuous and involve many different covariates $\rightarrow$ modeling
- If we can develop "good" estimators for the distribution of $Y|A=1,X$ & $Y|A=0,X$ and for the distribution of $X$ then we can get a "good" estimator of $E(Y^1)$ and $E(Y^0)$
- We are really good at specifying models for $Y|A,X$ - these are regression models
- For example, if $Y$ is continuous a natural model might be $E(Y,A,X) = \mu(A,X;\eta) = \eta_0 + \eta_1^T X + \eta_2 A + \eta_3^T AX$ and we could estimate $\eta$ using least squares (call this $\hat{\eta}$)
- We can estimate the distribution of $X$ nonparametrically using the empirical distribution

# Regression Modeling

- Putting it all together

1) $\hat{E}(Y^1) = \frac{1}{n}\sum_{i=1}\mu(1, X_i; \hat{\eta})$
2) $\hat{E}(Y^0) = \frac{1}{n}\sum_{i=1}\mu(0, X_i; \hat{\eta})$
3) $\hat{\delta} = \hat{E}(Y^1) - \hat{E}(Y^0)$

# Regression Modeling

- Putting it all together in words

a) Posit some model for $E(Y, A, X) = \mu(A, X; \eta)$ and get parameter estimates for $\eta$ using least squares or MLEs.
b) Get the predicted value for each subject in the data set assuming that they were treated and untreated. That is, compute $\mu(1, X_i; \hat{\eta})$ and $\mu(0, X_i; \hat{\eta})$ for each $i$
c) Take the average of $\mu(1, X_i; \hat{\eta})$ and $\mu(0, X_i; \hat{\eta})$. This gives an estimate of $\hat{E}(Y^1)$ and $\hat{E}(Y^0)$, respectively.
d) Take their difference

# Regression Modeling

- The preceding approach works regardless of whether the outcome is continuous, binary, etc.
- Note that if we have a linear model for $Y|A, X$ and there is no interaction between treatment and covariates then $\hat{\delta} = \hat{\eta_A}$ (the estimated regression coefficient for the treatment term)
- Note that one cannot "read-off" the ATE if the regression model is not linear (e.g., if we have a binary outcome)
- We have not discussed how to estimate uncertainty (but we will need to)

# Motivating Dataset

- Data from a randomized trial of the effectiveness of various approaches to "getting out the vote"
- Conducted by Gerber and Green in New Haven, CT in 1998
- Tested the efficacy of mailings, door-to-door canvasing, and telephone calls
- Dataset of 10,829 subjects who were in the control groups for mailing and door-to-door canvasing
- Publically available from the Matching package in R

Citations: 1) Gerber, Alan S. and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." American Political Science Review 94: 653663. 2) Gerber, Alan S. and Donald P. Green. 2005. "Correction to Gerber and Green (2000), replication of disputed findings, and reply to Imai (2005)." American Political Science Review 99: 301-313. 3) Imai, Kosuke. 2005. "Do Get-Out-The-Vote Calls Reduce Turnout? The Importance of Statistical Methods for Field Experiments." American Political Science Review 99: 283-300.

# RMarkdown

````
461
462 ## Voing Example: Regression Adjustment Results
463 ```{r, echo = TRUE, cache = TRUE}
464 data_trt <- data_ctr <-imai
465 data_trt$PHN.C1 = 1
466 data_ctr$PHN.C1 = 0
467 pred1 <- predict(m1, newdata = data_trt, type = "response")
468 pred0 <- predict(m1, newdata = data_ctr, type = "response")
469 ATE <- mean(pred1 - pred0)
470 print(ATE, digits = 3)
471 ```
472
473 ## Standard Errors and Statistical Inference
474 - Under the assumptions of consistency and no unmeasured confounding and assuming that the regression model
    for $Y|A, X$ is correctly specified then the proposed estimator is consistent and asymptotically normal (show
    that this is an M-estimator)
475 - Note that if we have a linear model for $Y|A, X$ and there is no interaction between treatment and
    covariates then $\hat{\delta} = \hat{\eta_A}$ (the estimated regression coefficient for the treatment term)
    and the standard error can be read off standard software
476 - Otherwise deriving the sta error is complicated -- use bootstrap
477
478
479 ## Bootstrap
480
````

$\hat{\delta} = \hat{\eta_A}$

**Figure 1:** Example Markdown Code

# Motivating Dataset

- Subjects randomized to receive a phone call reminding them to vote may not have ever answered the phone
- 247 potential voters received and answered the phone
- Those who answer the phone not similar to those who do not

# Key Variables

- VOTED98: Voted in 1998 (outcome)
- PHN.C1: Contact occurred in phntrt1 (treatment)
- PERSONS: Number voters (one or more than one) in household
- WARD: Ward of residence
- AGE: Age of respondent
- MAJORPTY: Democratic or Republican
- VOTE96.1: Voted in 1996
- NEW: New voter

# Differences in Key Variables Between Groups

```
vars <- c("VOTED98F", "PERSONSF", "AGE", "VOTE96.1F",
  "NEWF", "MAJORPTYF",  "WARD")
tabUnmatched <- CreateTableOne(vars = vars, strata = "PHN.C1F"
  data = imai, test = FALSE)
t1 <- print(tabUnmatched, smd = TRUE, showAllLevels = TRUE, va
```

# Differences in Key Variables Between Groups

```
kable(t1)
```

| | level | Not Contacted | Contacted | SMD |
|---|---|---|---|---|
| n | | 10582 | 247 | |
| Voted in 1998 (%) | No | 5881 (55.6) | 87 (35.2) | 0.418 |
| | Yes | 4701 (44.4) | 160 (64.8) | |
| Voters in household (%) | 1 Voter | 5269 (49.8) | 119 (48.2) | 0.032 |
| | 2+ Voters | 5313 (50.2) | 128 (51.8) | |
| Age (years) (mean (SD)) | | 49.43 (18.73) | 58.31 (19.85) | 0.460 |
| Voted in 1996 (%) | No | 4965 (46.9) | 71 (28.7) | 0.382 |
| | Yes | 5617 (53.1) | 176 (71.3) | |
| New voter (%) | Previous Voter | 8452 (79.9) | 219 (88.7) | 0.243 |
| | New Voter | 2130 (20.1) | 28 (11.3) | |
| Party affilation (%) | Republican | 2701 (25.5) | 49 (19.8) | 0.136 |
| | Democrat | 7881 (74.5) | 198 (80.2) | |
| Ward of residence (%) | 2 | 317 ( 3.0) | 3 ( 1.2) | 0.565 |
| | 3 | 273 ( 2.6) | 3 ( 1.2) | |
| | 4 | 234 ( 2.2) | 2 ( 0.8) | |
| | 5 | 200 ( 1.9) | 4 ( 1.6) | |
| | 6 | 435 ( 4.1) | 5 ( 2.0) | |
| | 7 | 337 ( 3.2) | 3 ( 1.2) | |
| | 8 | 360 ( 3.4) | 7 ( 2.8) | |
| | 9 | 387 ( 3.7) | 9 ( 3.6) | |
| | 10 | 452 ( 4.3) | 16 ( 6.5) | |
| | 11 | 451 ( 4.3) | 19 ( 7.7) | |
| | 12 | 364 ( 3.4) | 9 ( 3.6) | |
| | 13 | 435 ( 4.1) | 10 ( 4.0) | |
| | 14 | 383 ( 3.6) | 6 ( 2.4) | |
| | 15 | 329 ( 3.1) | 8 ( 3.2) | |
| | 16 | 240 ( 2.3) | 5 ( 2.0) | |
| | 17 | 500 ( 4.7) | 19 ( 7.7) | |
| | 18 | 578 ( 5.5) | 21 ( 8.5) | |

# Voting Example: Unadjusted Results

```
## [1] "Unadjusted ATE"

## [1] 0.204

## [1] "Standard Error"

##      1
## 0.0308

## [1] "95% CI"

## [1] 0.143 0.264
```

# Unadjusted Results: Key Assumptions

Identifying

1) Consistency
2) No confounding

Modeling

1) None

# Regression Adjustment

- Fit logistic regression model for voting in 1998 with covariates for telephone call, number of voters in household, age, voting in 1996, new voter, party affiliation, and ward
- Include interaction between telephone call and number of voters in household, age, voting in 1996, new voter, party affiliation

# Regression Output

```
m1 <- glm(VOTED98 ~ PHN.C1*(PERSONS + VOTE96.1 + NEW + MAJORPT
round(summary(m1)$coefficients[c(1:7, 36:40), ], digits = 3)
```

```
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.811      0.170 -22.456    0.000
## PHN.C1                0.536      0.845   0.634    0.526
## PERSONS               0.243      0.047   5.211    0.000
## VOTE96.1              2.135      0.062  34.664    0.000
## NEW                   1.255      0.075  16.646    0.000
## MAJORPTY              0.362      0.054   6.745    0.000
## AGE                   0.025      0.001  18.400    0.000
## PHN.C1:PERSONS        0.071      0.324   0.219    0.827
## PHN.C1:VOTE96.1       0.184      0.423   0.434    0.664
## PHN.C1:NEW            0.388      0.582   0.667    0.505
## PHN.C1:MAJORPTY       0.053      0.381   0.139    0.889
## PHN.C1:AGE           -0.007      0.009  -0.843    0.399
```

# Regression Adjustment

- Get predicted value for each individual in the dataset assuming that they are (a) in the treatment group and (b) in the control group
- Take the difference in the mean predicted value to get estimate of ATE

# Voing Example: Regression Adjustment Results

```
data_trt <- data_ctr <-imai
data_trt$PHN.C1 = 1
data_ctr$PHN.C1 = 0
pred1 <- predict(m1, newdata = data_trt, type = "response")
pred0 <- predict(m1, newdata = data_ctr, type = "response")
ATE <- mean(pred1 - pred0)
print(ATE, digits = 3)
```

```
## [1] 0.0965
```

# Standard Errors and Statistical Inference

- Under the assumptions of consistency and no unmeasured confounding and assuming that the regression model for $Y|A, X$ is correctly specified then the proposed estimator is consistent and asymptotically normal (show that this is an M-estimator)
- Note that if we have a linear model for $Y|A, X$ and there is no interaction between treatment and covariates then $\hat{\delta} = \hat{\eta_A}$ (the estimated regression coefficient for the treatment term) and the standard error can be read off standard software
- Otherwise deriving the standard error is complicated – use bootstrap

# Bootstrap for Regression Adjustment

```
set.seed(1101985)
B <- 100
ATE.boot <- NULL
n <- nrow(imai)
for(i in 1:B) {
  imai.boot <- imai[sample(1:n, n, replace = TRUE), ]
  m1.boot <- glm(VOTED98 ~ PHN.C1*(PERSONS + VOTE96.1 +
      NEW + MAJORPTY + AGE) + WARD, data = imai.boot,
    family = "binomial")
  data_trt.boot <- imai.boot
  data_trt.boot$PHN.C1 = 1
  data_ctr.boot <- imai.boot
  data_ctr.boot$PHN.C1 = 0
  pred1.boot <- predict(m1.boot, newdata = data_trt.boot,
    type = "response")
  pred0.boot <- predict(m1.boot, newdata = data_ctr.boot,
    type = "response")
```

# Voting Example: Regression Adjustment Results

```
## [1] "Average Treatment Effect"
## [1] 0.0965
## [1] "Bootstrap SE"
## [1] 0.0335
## [1] "Bootstrap Normal 95% CI"
## [1] 0.0309 0.1621
```

# Regression Adjustment Results: Key Assumptions

Identifying

1) Consistency
2) No Unmeasured confounding

Modeling

1) Outcome model (given all confounders) correctly specified. Note this may involve extrapolation if there is not sufficient overlap in covariates between treatment and control.

# More Flexible Regression Models

- Because we must get outcome model correct (i.e., interpretability not paramount) spurred the use of more flexible, data-adaptive method
- Examples of more flexible regression models

1) BART - Bayesian additive regression trees
2) Random regression forest
3) Support vector regression

- Even though the use of these methods sounds "fancier," still doing regression adjustment to estimate causal effects
- Could we instead try modeling something more straightforward?