Matching for Adjustment and Causal Inference
Class 1: Experiments, Potential Outcomes, and Treatment Effects

Jake Bowers

August 21, 2023

Overview of the class

At the end:

- We estimate the TV ad caused an increase in vaccinations by 5 pct points and that it would be surprising to see this size of effect if the ad had no effect ($p = .001$).
- Recall that we compare pairs of people with the same age, so differences in age cannot explain the difference.
- Any differences within pair in income are smaller than we would see in an equivalent pair-randomized experiment.
- Also, although we did not measure party ID, and surely Republicans and Democrats of the same age are not equally likely to watch the ads, Republicans have to be 5x more likely to watch the ad than Democrats (of the same age) before we would estimate an effect of 0 or $p = .051$.

See also (Rabb et al., 2022).

Overview of the class

- We **infer** about counterfactual causal mechanisms by **estimating** and **testing**: Causal inference (in this class) requires statistical inference.
- Statistical inference requires **distributions** (we cannot know whether an estimator is unbiased without an idea of a distribution, we cannot calculate a p-value without a distribution, etc.)
- Distributions require **thought experiments** about repetition: We imagine that we could (a) re-draw random samples from a the same population or (b) re-assign randomized treatment. (We can also imagine reusing our priors and likelihood.)
- We will use the fact that distributions of certain test statistics is **known for randomized experiments** to (a) reason about whether we have a good non-randomized research design, (b) analyze outcomes, and (c) reason about sensitivity of results fo unobserved confounders.

So: We will start with randomized experiments even though the rest of the course is about non-randomized studies.

Experiments, Potential Outcomes, and Treatment Effects

1. Causal inference in randomized experiments and the idea of only partially observed **potential outcomes**. The idea that we can use what we observe to learn about what we cannot observe.
2. Statistical inference for causal effects in randomized experiments via the Fisher and Neyman approaches (Rosenbaum, 2010, Chap 2), (Gerber and Green, 2012, Chap 1-3): Estimation, Estimators, Tests, Testing.
3. Why are randomized experiments special? Unbiased estimators. The ability to assess bias. The ability to exclude alternative explanations. Tests with known error rates. The ability to assess error rates of tests.
4. What does "controlling for" do in a linear model when we do not have a randomized experiment? How can we make the case that we are "controlling for" enough?

Note: You can download (and contribute to) course materials at https://github.com/bowers-illinois-edu/short-course-causal

Overly Ambitious Plan

- Introduction to Jake
- Introduction to the idea of the course: roadmap
- Introductions to you and a project that you might work on during the week.
- Jake introduces concepts: potential outcomes, treatment effects, unbiased estimation, and encourages questions and answers.
- Coffee Break
- Questions about the lecture
- Exercise 1: Describe your data (and update R).
- Break
- Jake discuss the problem of adjustment in observational studies as compared to experiments
- Open Discussion and/or work on projects

Notation and Concepts for Counterfactual Causal Inference

- Treatment or Intevention $Z_i = 1$ for treatment and $Z_i = 0$ for control for units i. (We mostly assume that all units **could have** $Z_i = 1$ or $Z_i = 0$. That it is not impossible for any unit to have either value.) (Q: What is a unit? Examples of interventions?)
- Each unit has a pair of potential outcomes $(y_{i,Z_i=1}, y_{i,Z_i=0})$ (also written $(y_{i,1}, y_{i,0})$) (given SUTVA).
  - Without the SUTVA assumption, and with 4 units, with two having $Z_i = 1$, unit $i = 1$ would have the following potential outcomes: $(y_{i,1100}, y_{i,1010}, y_{i,1001}, y_{i,0101}, y_{i,0011})$
- Causal Effect under SUTVA when $y_{i,1} \neq y_{i,0}$, $\tau_i = f(y_{i,1}, y_{i,0})$ ex. $\tau_i = y_{i,1} - y_{i,0}$. (Examples of Interfering Units and Not Interfering Units)
- Fundamental Problem of (Counterfactual) Causality We only see one potential outcome $Y_i = Z_i * y_{i,1} + (1 - Z_i)y_{i,0}$ (Examples?)
- Covariates, $\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{nk} \end{bmatrix}$ is a matrix containing background information about the units that might predict $(y_{i,1}, y_{i,0})$ or Z (but that don't predict Z if Z is randomized as in an experiment).

The observation and unobserved causal comparisons

We can learn about unobserved but theorized causal mechanisms by observing the world (Brady, 2008):

- **Persistent association** "We always/mostly see $Y = 1$ when $X = 1$ and $Y = 0$ when $X = 0$."
- **Counterfactual Difference** "If X had not been this value, then Y would not have been that value."
- **Difference after manipulation** "When we change X from one value to another value, then Y changes from one value to another value." (establishes causal priority of X over Y, implied that Y would not have changed.).
- **Difference after operation of a mechanism** "Once upon a time A changed X, and then one day X changed B, and because of that B changed C, and finally C changed Y."

All approaches are useful. This week we are focusing on the counterfactual approach and somewhat on the persistent association.

How to interpret "X causes Y" in counterfactual terms?

The counterfactual approach requires that we can imagine units with and without X (or say, $X = 1$ and $X = 0$).

- We can establish that X causes Y without knowing the mechanism. The mechanism can be complex, and it can involve probability: X causes Y sometimes because of A and sometimes because of B.
- "X causes Y" can mean "With X, probability of Y is higher than would be without X." or "Without X there is no Y." Either is compatible with the counterfactual idea.
- Of course: Correlation is not causation.
- "X causes Y" is a statement about what didn't happen: "If X had not operated, occurred, then Y would not have occurred." (More about the fundamental problem of counterfactual causation later)

Today: Show that a randomized experiment allows us to **learn about** unobserved counterfactual outcomes using statistical inference.

Example: Honey and Colds

Your friend explains a causal mechanism that eating raw honey reduces the duration of colds. What kinds of **alternative** explanations might we come up with for this result? Imagine these were the underlying potential outcomes with x_1 and x_2 representing two of those explanations and that $x_1 \rightarrow y_0$ and that $x_2 \nrightarrow y_0$.

id	x1	x2	y0	y1	tau
1	1	3	6.25	5.25	-1
2	1	8	10.25	9.25	-1
3	2	8	10.00	8.00	-2
4	3	2	12.25	9.25	-3
5	1	6	12.25	11.25	-1
6	0	6	12.00	12.00	0
7	0	7	9.00	9.00	0
8	0	1	5.00	5.00	0
9	0	8	10.00	10.00	0
10	2	7	10.00	8.00	-2

The true, unobserved, average (additive) causal effect is: -1.

Let us run a randomized experiment and see how we do:

An RCT:

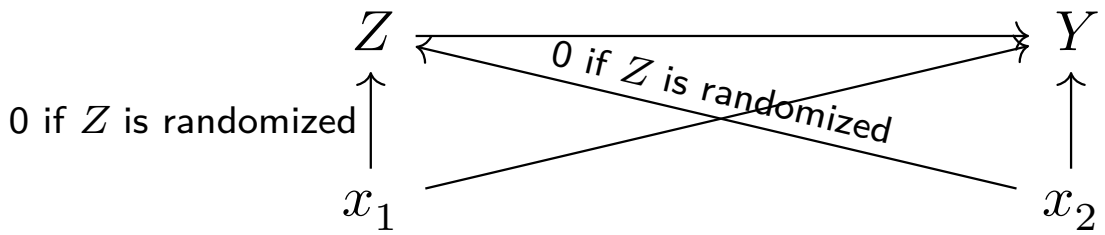
What is happening here? How would we know whether `complete_ra` worked as it should?

```
library(randomizr)
set.seed(12345)
dat$Z <- complete_ra(N = 10, m = 5)
dat$Y <- with(dat, Z * y1 + (1 - Z) * y0)
kable(dat)
```

id	x1	x2	y0	y1	tau	Z	Y
1	1	3	6.25	5.25	-1	0	6.25
2	1	8	10.25	9.25	-1	1	9.25
3	2	8	10.00	8.00	-2	0	10.00
4	3	2	12.25	9.25	-3	0	12.25
5	1	6	12.25	11.25	-1	1	11.25
6	0	6	12.00	12.00	0	1	12.00
7	0	7	9.00	9.00	0	1	9.00
8	0	1	5.00	5.00	0	1	5.00
9	0	8	10.00	10.00	0	0	10.00
10	2	7	10.00	8.00	-2	0	10.00

Assessing randomization I

We expect that the distributions of x_1 and x_2 would be (nearly) the same between the treated and control groups. We write "0" below, but in fact, randomization does not make those relationships exactly 0.



Here, just looking at means:

```
dat %>%
  group_by(Z) %>%
  reframe(mean(x1), mean(x2))
```

```
# A tibble: 2 x 3
      Z `mean(x1)` `mean(x2)`
  <int>      <dbl>      <dbl>
1     0        1.6        5.6
2     1        0.4        5.6
```

Assessing randomization II

```
library(RItools)
bal1 <- balanceTest(Z ~ x1 + x2, data = dat)
bal1$results[, 1:4, ]
```

```
    stat
vars Control Treatment std.diff adj.diff
  x1     1.6       0.4   -1.342     -1.2
  x2     5.6       5.6    0.000      0.0
```

What **should** we expect from an experiment?

...in regards covariate balance? Lets simulate to learn:

```
new_exp <- function(Z) {
  newZ <- sample(Z)
  return(newZ)
}

diff_means <- function(x, Z) {
  mean(x[Z == 1]) - mean(x[Z == 0])
}

all_cov_bal <- replicate(1000, diff_means(x = dat$x1, Z = new_exp(dat$Z)))

summary(all_cov_bal)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.6000 -0.4000  0.0000  0.0232  0.4000  1.6000
```

```
obs_cov_bal <- diff_means(dat$x1, dat$Z)
obs_cov_bal
```

```
[1] -1.2
```

What **should** we expect from an experiment?

So:

1. experiments do not guarantee exact equality of covariates and
2. we can **know** (or closely approximate) what kind of covariate differences a given experimental design would generate.

1. Overview and Review
2. Concepts and Notation for Causal Inference and Statistical Inference
3. Estimation, Estimators, Bias, Consistency, Given Randomization
4. Covariance Adjustment?
5. Summary and Overview
6. Appendix

Estimating the ATE in an RCT:

Here are two proposals for estimating the ATE. How would we know whether either or both of them work well (trick question)? (What do we want estimators to do for us?)

```
est1 <- function(Z, Y) {
  mean(Y[Z == 1]) - mean(Y[Z == 0])
}
est2 <- function(Z, Y) {
  coef(lm(Y ~ Z))[["Z"]]
}

with(dat, est1(Z = Z, Y = Y))
```

```
[1] -0.4
```

```
with(dat, est2(Z = Z, Y = Y))
```

```
[1] -0.4
```

How does randomization help us trust our estimators?

This is a simulation assessing **estimation bias** (and hinting at **consistency**)

```
## The truth:
with(dat, mean(y1 - y0))
```

```
[1] -1
```

```
new_exp <- function(Z) {
  ## This next shuffles Z
  newZ <- sample(Z)
  return(newZ)
}

new_est <- function(newZ, y0, y1, the_est) {
  newY <- newZ * y1 + (1 - newZ) * y0
  result <- the_est(Z = newZ, Y = newY)
}

set.seed(1235)
dist_est1 <- with(dat, replicate(100, new_est(newZ = new_exp(Z), y0 = y0, y1 = y1, the_est = est1
mean(dist_est1)
```

```
[1] -0.806
```

How does randomization help us trust our estimators?

Note: (1) Different simulations give slightly different results and (2) more simulations differ from each other less.

```
set.seed(1235)
dist_est1a <- with(dat, replicate(100, new_est(newZ = new_exp(Z), y0 = y0, y1 = y1, the_est = est
mean(dist_est1a)
```

```
[1] -0.806
```

```
dist_est1b <- with(dat, replicate(100, new_est(newZ = new_exp(Z), y0 = y0, y1 = y1, the_est = est
mean(dist_est1b)
```

```
[1] -1.019
```

```
dist_est2a <- with(dat, replicate(10000, new_est(newZ = new_exp(Z), y0 = y0, y1 = y1, the_est = e
mean(dist_est2a)
```

```
[1] -0.9883
```

```
dist_est2b <- with(dat, replicate(10000, new_est(newZ = new_exp(Z), y0 = y0, y1 = y1, the_est = e
mean(dist_est2b)
```

```
[1] -0.989
```

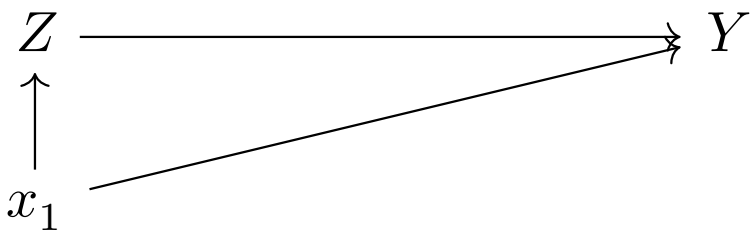
How does randomization help us trust our estimators?

What did randomization provide here?

1. Grounds for repetition (i.e. we **knew** how to repeat the generation of Z),
2. No need to mention `x1` (we could check to see if we should worry about x1).
3. an unbiased estimator.

What about if we didn't know exactly how Z was assigned? I

Imagine that x1 causes Z (here, Z is randomized but x1 changes Z before revealing y0 or y1):




```
new_biased_exp <- function(Z, x1) {
  newZ1 <- sample(Z)
  # newZ <- newZ1*rbinom(10,size=1,prob=(x1+1)/4)
  newZ <- pnorm(x1 + newZ1) > pnorm(median(x1 + newZ1))
  return(as.numeric(newZ))
}
trueATE <- with(dat, mean(y1 - y0))
with(dat, est1(new_biased_exp(Z, x1), Y))
```

```
[1] 0.9
```

What about if we didn't know exactly how Z was assigned? II

```
set.seed(1235)
dist_est_biased <- with(dat, replicate(10000, new_est(newZ = new_biased_exp(Z, x1), y0 = y0, y1 =
summary(trueATE - dist_est_biased)
```

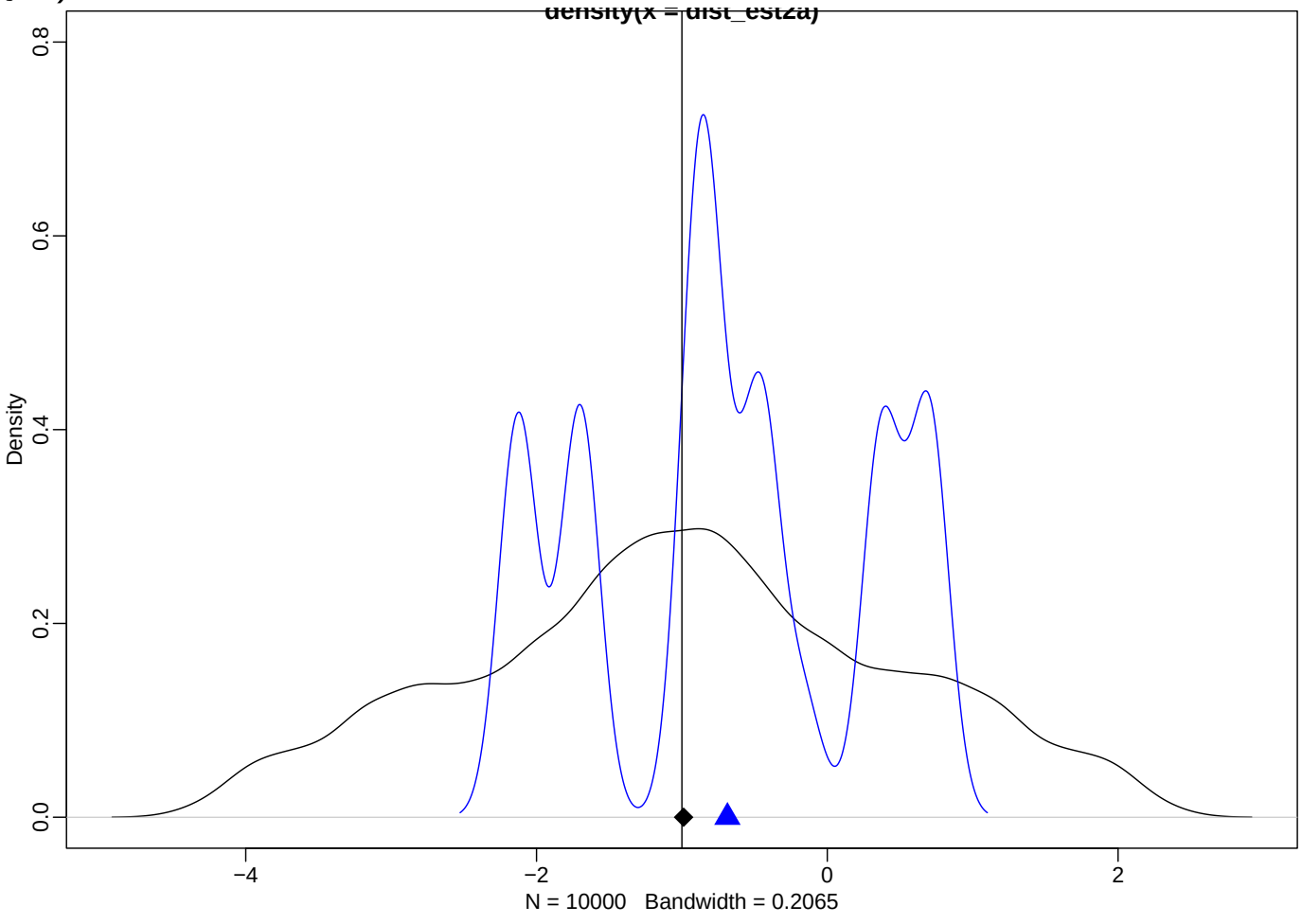
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.700  -1.375  -0.167  -0.313   0.700   1.125
```

```
## And recall our previous distribution of our estimator across randomizations
## This next is unbiased (mean \approx 0)
summary(trueATE - dist_est2a)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -3.300  -1.000  -0.100  -0.012   1.000   3.300
```

What about if we didn't know exactly how Z was assigned?

Imagine that x1 causes Z (here, Z is randomized but x1 changes Z before reveals y0 or y1):



What about if we didn't know how Z was assigned at all?

How would we assess bias if we didn't know that x_1 caused Z?

- We cannot simply shuffle Z. Because we don't know how Z arose.
- Do we know how x1 was generated? If so, we could re-generate x1 and hope that our x1 to Z function is right
- We could repeatedly re-generate Y itself if we **knew** how it was created.
- We could resample the entire dataset if we **knew** how it was sampled.

So: **known randomization allows us to assess bias.**

What does linear regression do in an observational study?

Here is another bit of fake data where we know the true causal effects (the τ_i for each person and the $y_{i,1}, y_{i,0}$, too). In real life we'd only observe Y, $x_1, \ldots, x_4$, and Z.

id	x1	x2	x3	x4	Z	Y	y1	y0	tau
1	11	6	2.633	1	1	9	9	9	0
2	12	2	2.141	0	1	6	6	6	0
3	9	1	3.042	1	1	12	12	12	0
4	8	5	4.058	1	0	19	21	19	2
5	11	1	1.687	0	1	11	11	10	1
6	4	6	3.686	1	0	30	33	30	3

What is the effect of Z on Y?

If we had a dataset, like, say, the number of miles people are willing to travel to get tested by COVID (Y) and whether they downloaded a COVID prevention information kit from a local US municipal government website, (Z), we could estimate the average causal effect of the COVID info kit like so:

```
lm0 <- lm_robust(Y ~ Z, data = dat)
coef(lm0)
```

```
(Intercept)           Z
      14.45       -1.25
```

But how should we interpret this? It looks like the kit causes a reduction in willingness to travel to be tested. This might be true. But we can immediately think of **alternative explanations**:

- Maybe people who download information kits differ from people who don't choose to download such kits in other ways — they might be wealthier, more likely to have a computer (since looking at pdf brochures on an old phone is no fun), be more interested in reading about health, speak English (imagining that the kit is in English), etc..

So, how might we try to set aside, or engage with, those alternative explanations?

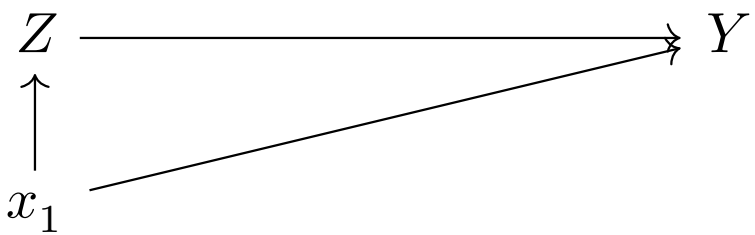
"Controlling for" to remove the effect of x_1 from $\hat{\hat{\tau}}$

A common approach looks like the following — the "effect of Z 'controlling for' x_1".

```
lm1 <- lm_robust(Y ~ Z + x1, data = dat)
coef(lm1)["Z"]
```

```
    Z
1.899
```

Recall that this is the problem — a $Z \to Y$ relationship could easily just reflect the $x_1 \to Z$ and $x_1 \to Y$ relationships and not the $Z \to Y$ relationship.



Today: Let's talk about what "controlling for" means. And then let's ask "How would we know whether we did a good job — did we"control for x_1" **enough**?"
What does "controlling for" mean here? How can we explain it?
How would we know whether we did a good job — did we "control for x_1" **enough**?

First, recall how linear models control or adjust

Notice that the linear model **does not hold constant** x_1. Rather it **removes a linear relationship** – the coefficient of 1.8994 from `lm1` is **the effect of** Z **after removing the linear relationship between** x_1 **and** Y **and between** x_1 **and** Z. (blue is treated)

```
## Adjusting for only x1
lm1 <- lm_robust(Y ~ Z + x1, data = dat)
coef(lm1)["Z"]
```

```
    Z
1.899
```

```
### Notice that this is the same as what follows
lm_Y_x1 <- lm(Y ~ x1, data = dat)
lm_Z_x1 <- lm(Z ~ x1, data = dat)
dat$resid_Y_x1 <- resid(lm_Y_x1)
dat$resid_Z_x1 <- resid(lm_Z_x1)
lm1b <- lm(resid_Y_x1 ~ resid_Z_x1, data = dat)
coef(lm1b)[[2]]
```

```
[1] 1.899
```

```
stopifnot(all.equal(coef(lm1b)[[2]], coef(lm1)[["Z"]]))
```

Linear models use residualization to remove linear relationships

What does **residualization do**? It removes linear relationships. Notice that the coefs on x1 below are 0.

```
lm_resid_Y_x1 <- lm(resid_Y_x1 ~ x1, data = dat)
lm_resid_Z_x1 <- lm(resid_Z_x1 ~ x1, data = dat)
coef(lm_resid_Y_x1)[["x1"]]
```

```
[1] 1.622e-16
```

```
coef(lm_resid_Z_x1)[["x1"]]
```

```
[1] 2.748e-19
```

Linear models use residualization to remove linear relationships

Notice that this works with as many covariates as you'd like:

```
lmYbig <- lm(Y ~ x1 + x2 + x3 + x4, data = dat)
lmZbig <- lm(Z ~ x1 + x2 + x3 + x4, data = dat)
coef(lm(resid(lmYbig) ~ resid(lmZbig)))[2]
```

```
resid(lmZbig)
        1.736
```

```
lmbig <- lm(Y ~ Z + x1 + x2 + x3 + x4, data = dat)
coef(lmbig)[["Z"]]
```

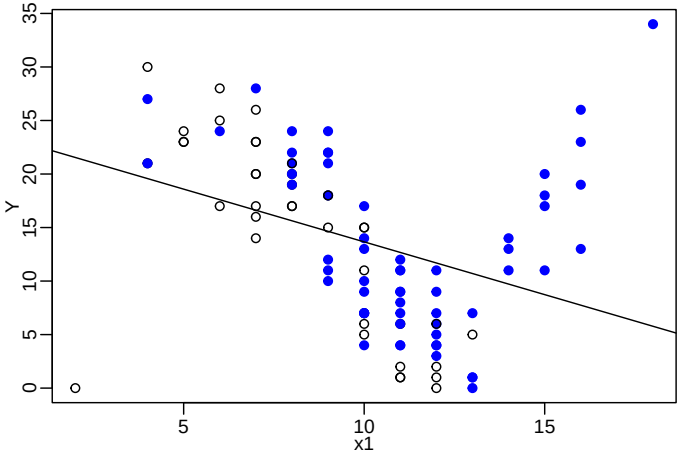
```
[1] 1.736
```

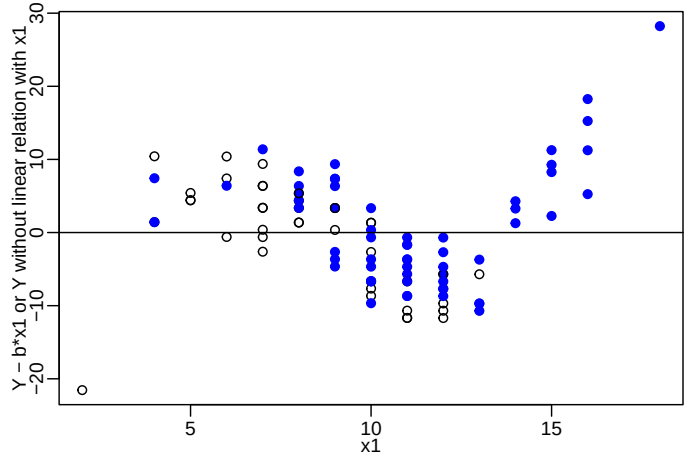
`resid(lmYbig)` is $Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \hat{\beta}_2 x_{i,2} + \hat{\beta}_3 x_{i,3} + \hat{\beta}_4 x_{i,4})$

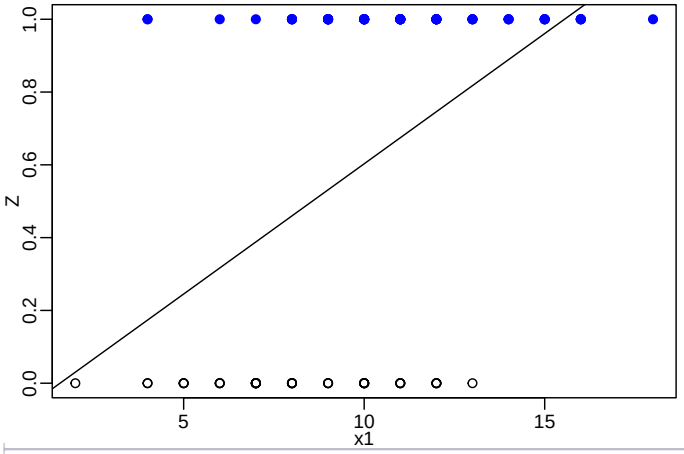
That is, `resid(lmYbig)` is just Y without its linear additive relationship with the covariates (the linear additive relationship that least squares would give you, not the one that, say a quantile regression would give you, or an outlier penalized regression or another target function).

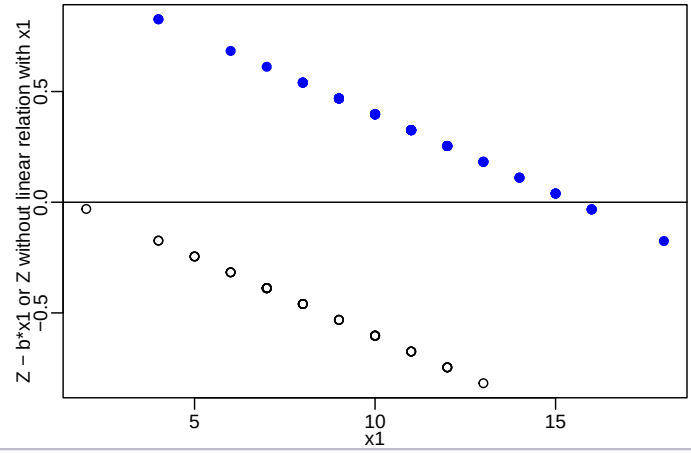
Removing x1 from Y and x1 from Z and then describing the $Z \rightarrow Y$ linear relationship

Notice that the linear model **does not hold constant** x_1. Rather it **removes a linear relationship** – the coefficient of 1.8994 from `lm1` is **the effect of** Z **after removing the linear relationship between** x_1 **and** Y **and between** x_1 **and** Z. (blue is treated)



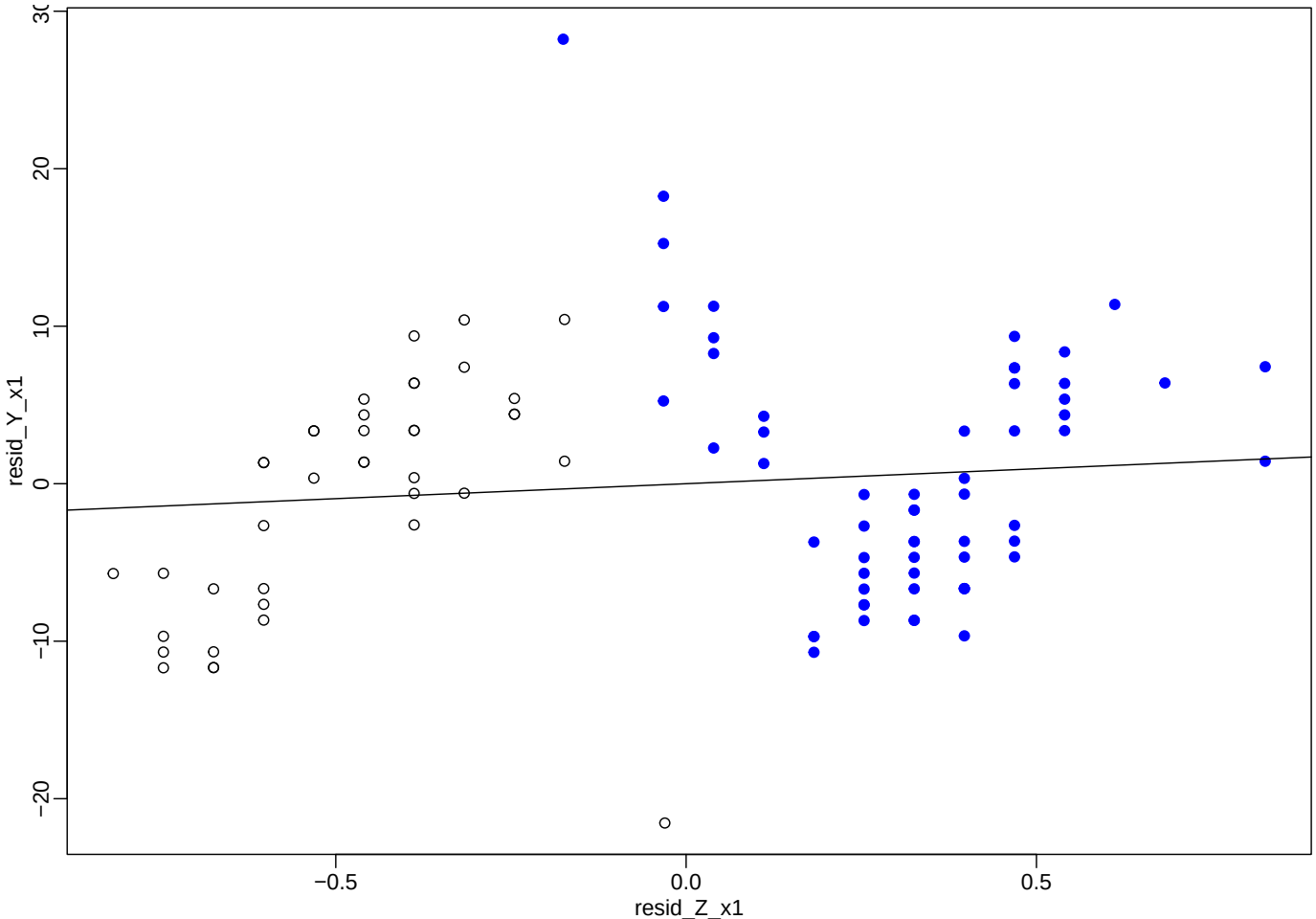






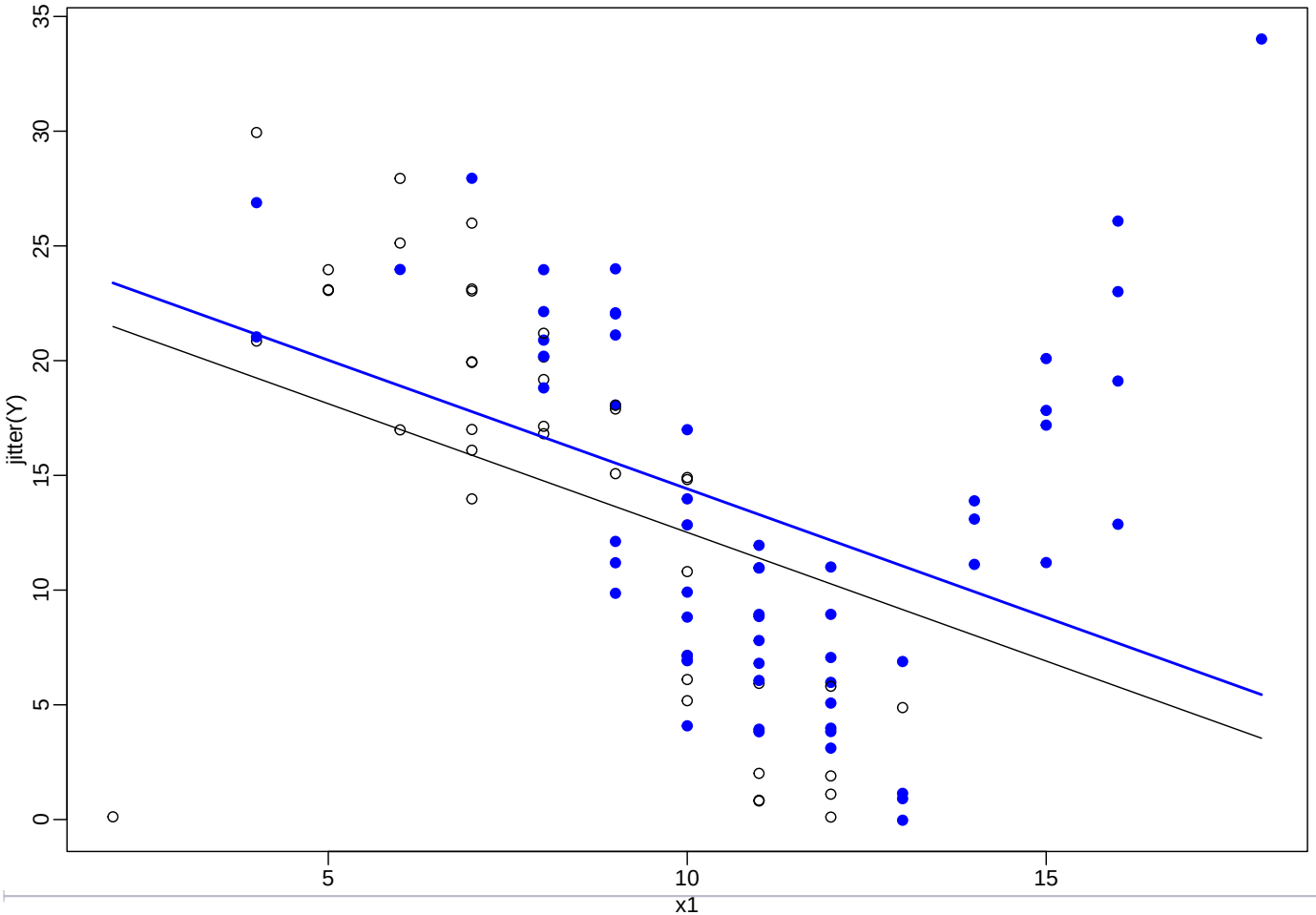
Recall how linear models control or adjust

Notice that the linear model **does not hold constant** x_1. Rather it **removes a linear relationship** – the coefficient of 1.8994 from `lm1` is **the effect of** Z **after removing the linear relationship between** x_1 **and** Y **and between** x_1 **and** Z. (blue=treated, black=control).



Recall how linear models control or adjust

How might this plot help us make decisions about the adequacy of our linear model adjustment strategy? Signs of extrapolation? Non-linearity? (Notice all the blue points with no open circles underneath them — in that range of the plot, our "adjustment" results entirely from the assumption of linearity. We don't know what would happen if we were to observe the open circles: would they follow a straight line?)

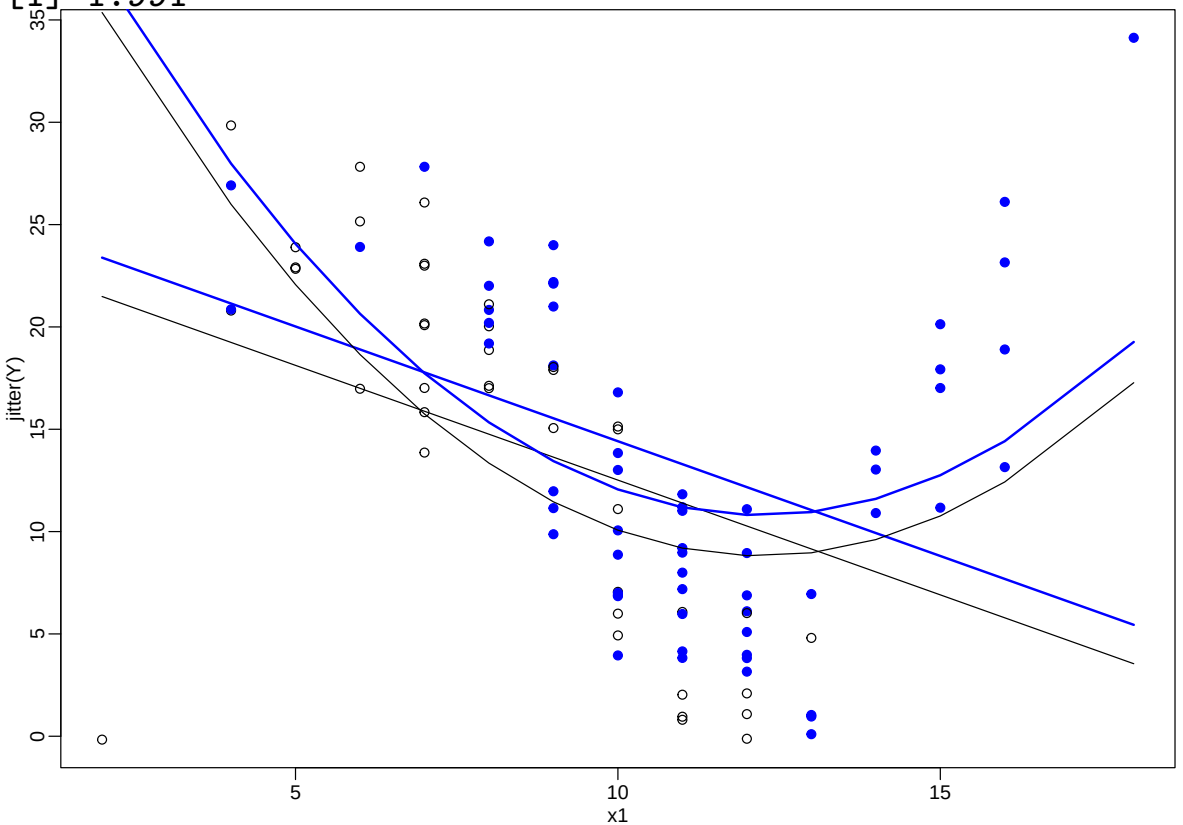


What about improving the model?

Does this help? (not really. Why squared? Why not cubed? Why not cut into pieces? Why not...?)

```
lm2 <- lm(Y ~ Z + x1 + I(x1^2), data = dat)
coef(lm2)[["Z"]]
```

```
[1] 1.991
```



What about when we control for more than one variable?

Is this better? Or worse? (It depends on whether we want to remove additive and linear relationships.)

```
lm3 <- lm(Y ~ Z + x1 + x2 + x3 + x4, data = dat)
coef(lm3)[["Z"]]
```

```
[1] 1.736
```

We could still residualize (removing the multidimensional linear relationship):

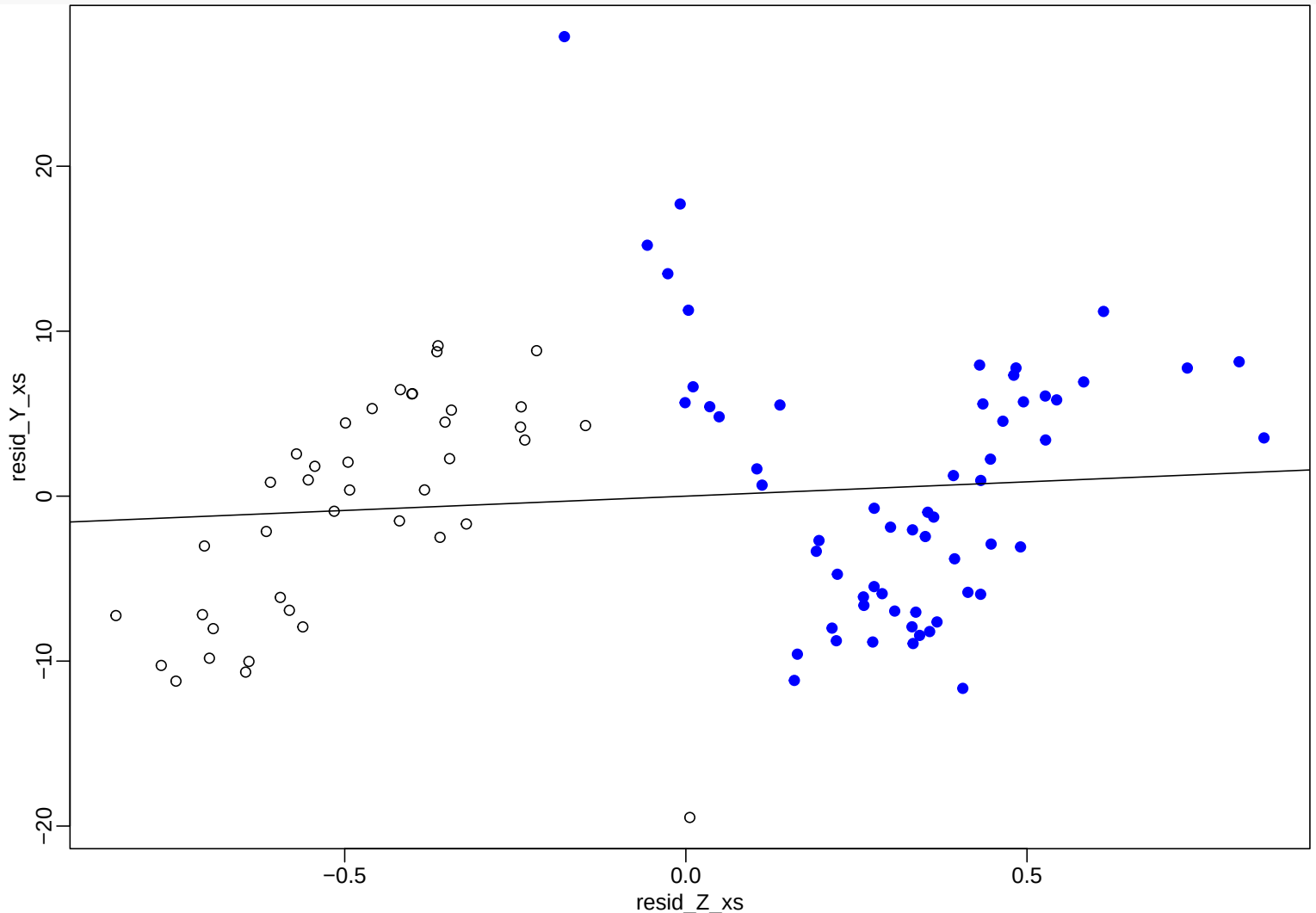
```
dat$resid_Y_xs <- resid(lm(Y ~ x1 + x2 + x3 + x4, data = dat))
dat$resid_Z_xs <- resid(lm(Z ~ x1 + x2 + x3 + x4, data = dat))
lm3_resid <- lm(resid_Y_xs ~ resid_Z_xs, data = dat)
coef(lm3_resid)[[2]]
```

```
[1] 1.736
```

What about when we control for more than one variable?

Is this better? Or worse? Hard to tell. What should we be looking for?

```
with(dat, plot(resid_Z_xs, resid_Y_xs, col = c("black", "blue")[dat$Z + 1], pch = c(1, 19)[dat$Z
abline(lm3_resid)
```



What about when we control for more than one variable? I

Does adding variables help? (Here we can see influential points using the Cook's D statistic. See the code for the different specifications.) Notice that as we add variables, or make the "controlling for" part more complicated, the more single points start to exert more and more influence over the results.

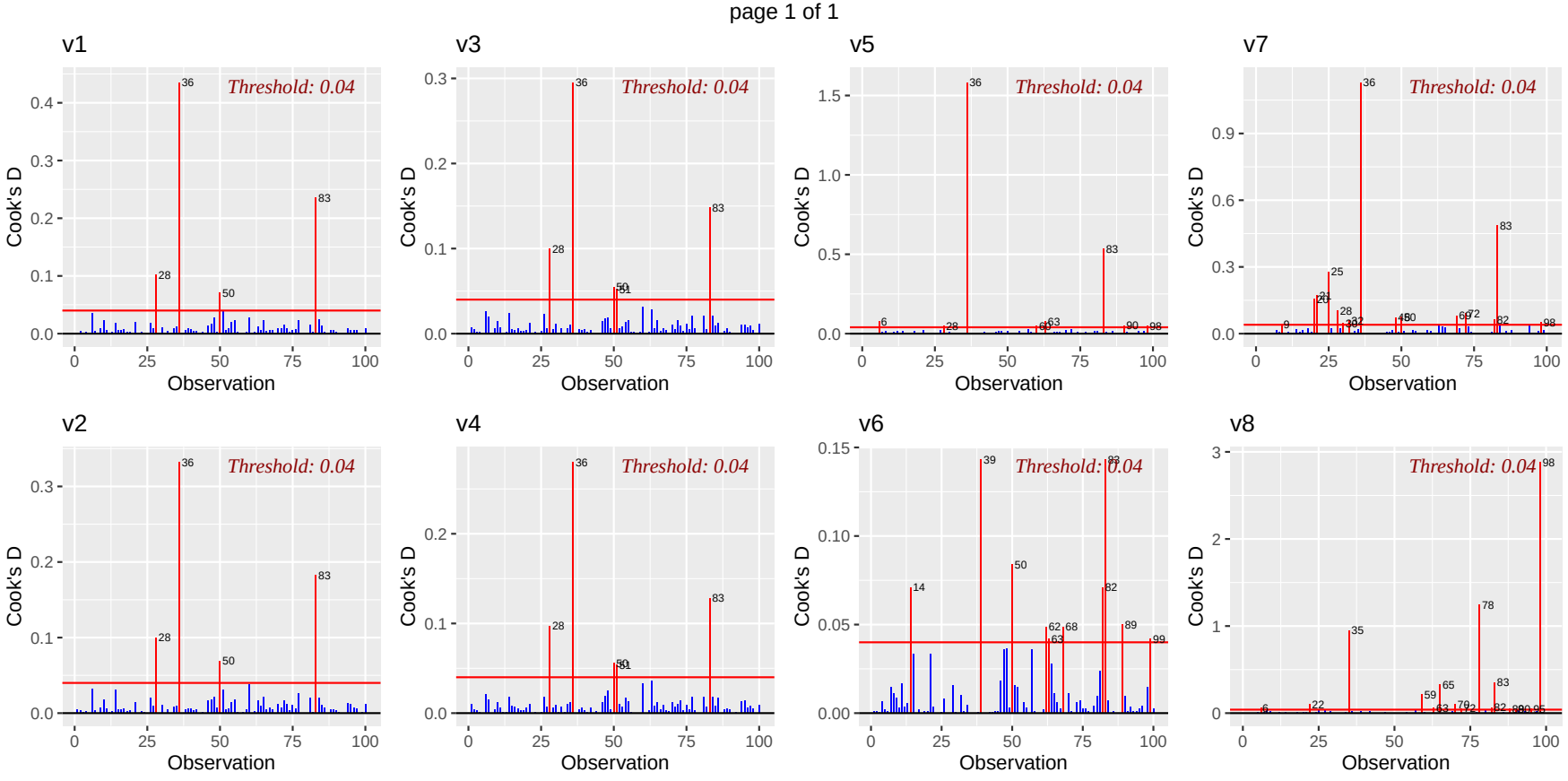
What about when we control for more than one variable? II

```
library(olsrr)
library(splines)
library(gridExtra)
v1 <- ols_plot_cooksd_bar(lm(Y ~ Z + x1, data = dat), print_plot = FALSE)
v2 <- ols_plot_cooksd_bar(lm(Y ~ Z + x1 + x2, data = dat), print_plot = FALSE)
v3 <- ols_plot_cooksd_bar(lm(Y ~ Z + x1 + x2 + x3, data = dat), print_plot = FALSE)
v4 <- ols_plot_cooksd_bar(lm(Y ~ Z + x1 + x2 + x3 + x4, data = dat), print_plot = FALSE)
v5 <- ols_plot_cooksd_bar(lm(Y ~ Z + poly(x1, 3) + poly(x2, 2) + poly(x3, 4) + x4, data = dat), p
v6 <- ols_plot_cooksd_bar(lm(Y ~ Z + I(cut(x1, 3)) * I(cut(x2, 3)) * I(cut(x3, 3)) * x4, data = da
v7 <- ols_plot_cooksd_bar(lm(Y ~ Z * x1 * x2 * x3 * x4, data = dat), print_plot = FALSE)
v8 <- ols_plot_cooksd_bar(lm(Y ~ Z + ns(x1, 3) + ns(x2, 3) * ns(x3, 3) * x4, data = dat), print_pl

plots <- lapply(1:8, function(i) {
  newplot <- get(paste0("v", i))$plot
  return(newplot + ggtitle(paste0("v", i)) + theme(legend.position = "none"))
})

cooksd_plot <- marrangeGrob(plots, nrow = 2, ncol = 4)
ggsave("cooksd.pdf", cooksd_plot, width = 12, height = 6)
```

What about when we control for more than one variable? III



How to choose? Maybe a specification curve?

.{allowframebreaks}

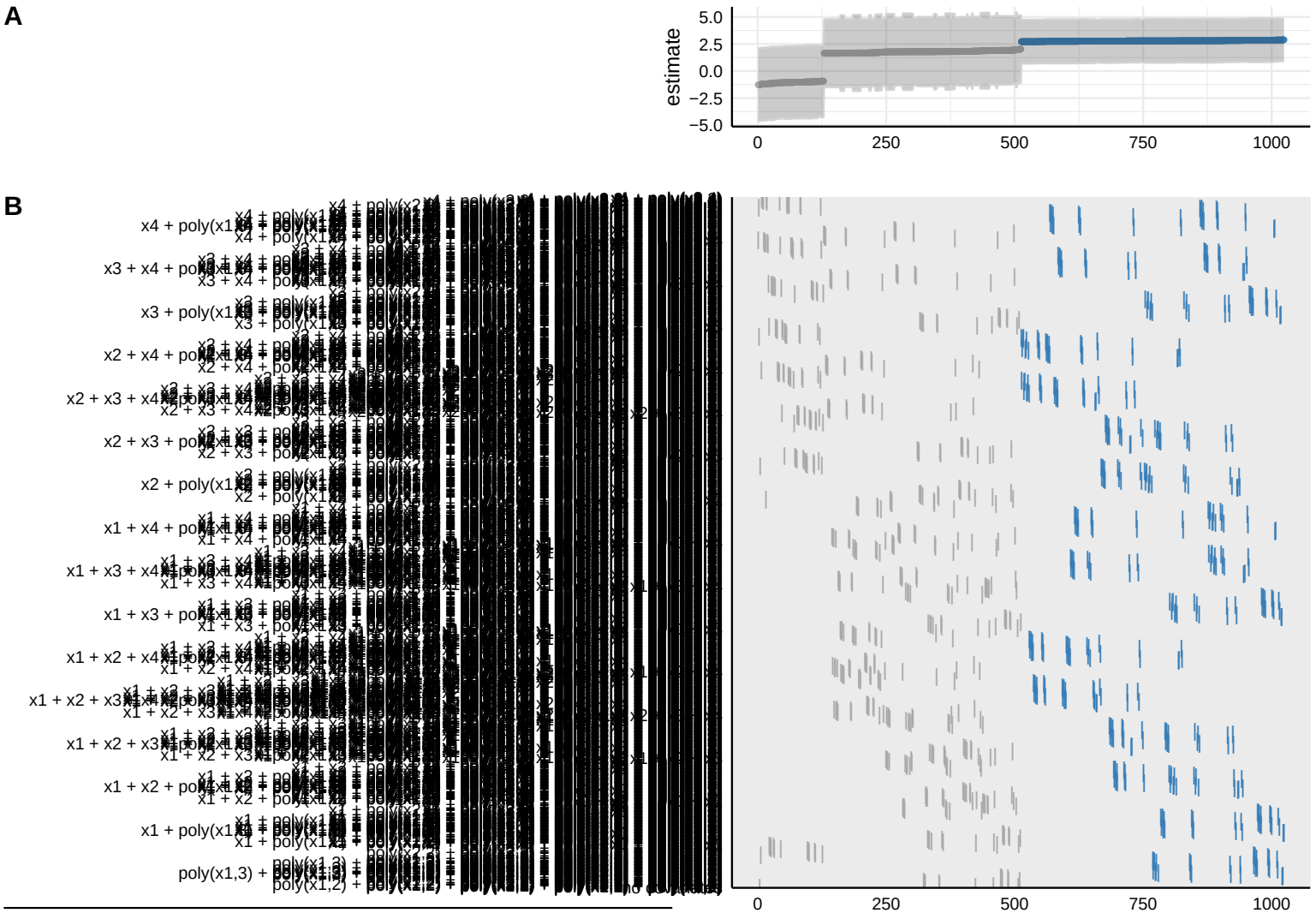
How many choices do we have? Should we try as many choices as possible?[1]

```
library(specr)
## see https://cran.r-project.org/web/packages/specr/vignettes/getting-started.html
## possible covariates:
library(splines)
basecovs <- c("x1", "x2", "x3", "x4")
mf <- model.frame(Y ~ Z + x1 * x2 * x3 * x4 + x1 * poly(x1, 3) + x2 * poly(x2, 2) + x3 * poly(x3,
  I(cut(x1, 3)) * I(cut(x2, 3)) * I(cut(x3, 3)), data = dat)
mm <- model.matrix(mf, data = dat)
thedat <- data.frame(mm[, -1])
thedat$Y <- dat$Y

specr_setup_obj <- specr::setup(
  data = thedat,
  y = c("Y"),
  x = c("Z"),
  model = c("lm"),
  # controls = grep("^x|^poly|^I|^ns",names(thedat),value=TRUE))
  controls = c(
    "x1", "x2", "x3", "x4",
    "poly(x1,3)",
    "poly(x1,2)",
    "poly(x2,2)",
    "poly(x3,2)",
    "poly(x3,3)",
    "poly(x3,4)"
```

How to choose? A specification curve.

How many choices do we have? Should we try as many choices as possible?[2]



[2]see https://masurp.github.io/specr/index.html for more citations

How to choose? Choosing different break-points.

How many choices do we have? Should we try as many choices as possible?

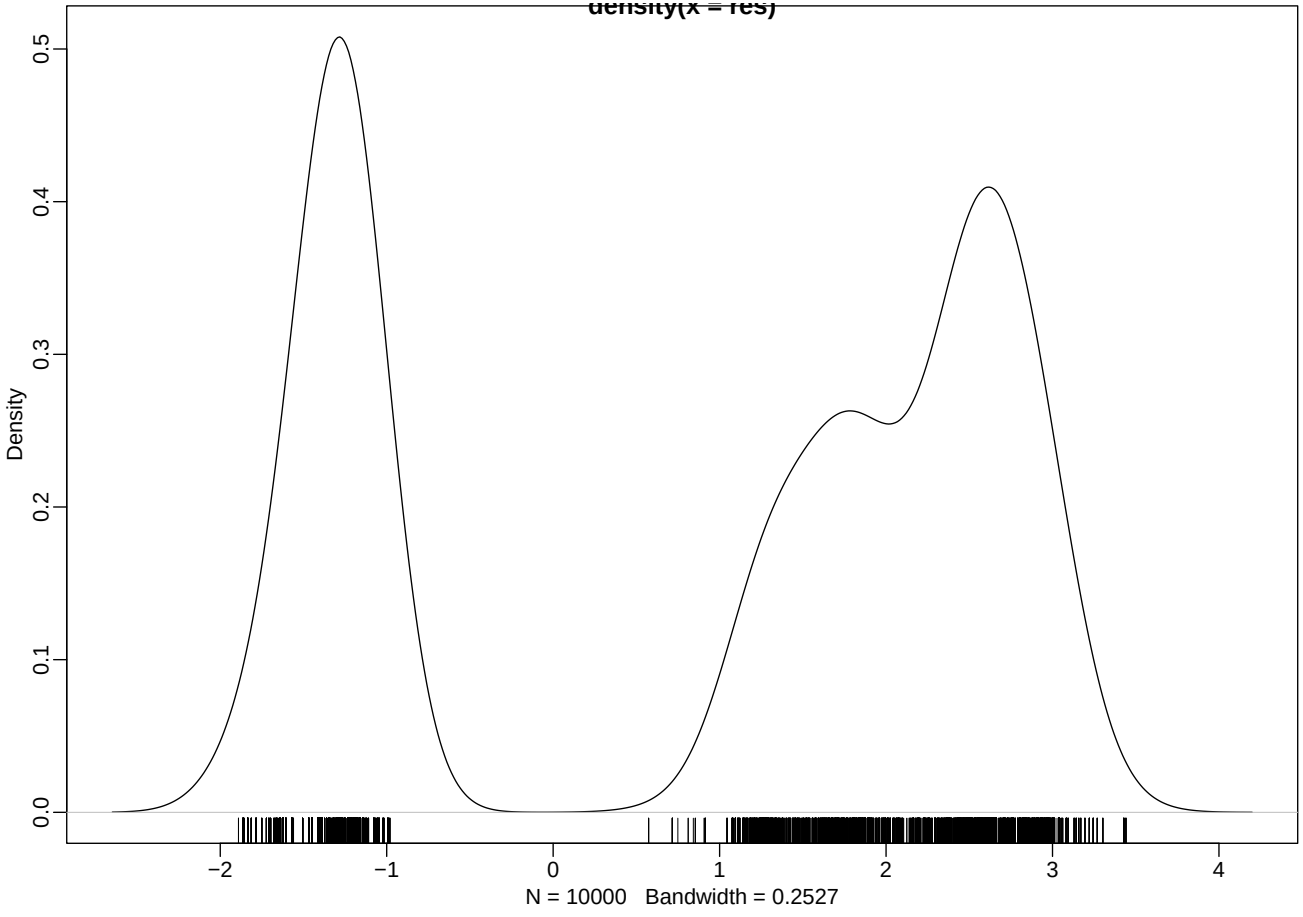
```
lmadjfn <- function() {
  covs <- c("x1", "x2", "x3", "x4")
  ncovs <- sample(1:length(covs), 1)
  somecovs <- sample(covs, size = ncovs)
  ncuts <- round(runif(ncovs, min = 1, max = 8))
  theterms <- ifelse(ncuts == 1, somecovs,
    paste("cut(", somecovs, ",", ncuts, ")", sep = "")
  )
  thefmla <- reformulate(c("Z", theterms), response = "Y")
  thelm <- lm(thefmla, data = dat)
  theate <- coef(thelm)[["Z"]]
  return(theate)
}

set.seed(12345)
res <- replicate(10000, lmadjfn())
summary(res)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -1.890  -1.226   1.635   0.888   2.531   3.441
```

How to choose? Choosing different break-points.

How many choices do we have? Should we try as many choices as possible? Here are the estimates of $Z \to Y$ from 10,000 different ways to "control for" x_1, x_2, x_3, x_4.



Ack. Which one should we choose? It seems like adjusting for covariates using linear models raises many more questions than it answers, and it is not clear how to answer all of those questions!

How about stratification?

Ok. So at this point we are not going to use linear models to adjust for covariates. What to do? What about simplifying the problem? When a person wants to know whether we have "controlled for", say, x_4, I suspect they are really asking for this:

```
lm_x4_0 <- lm(Y ~ Z, data = dat, subset = x4 == 0)
lm_x4_1 <- lm(Y ~ Z, data = dat, subset = x4 == 1)
coef(lm_x4_1)[["Z"]]
```

```
[1] -3.753
```

```
coef(lm_x4_0)[["Z"]]
```

```
[1] 0.5946
```

In this case we can say that we have "held constant" x_4. But what is the **overall estimate** in this case?

Choosing an additive and linear functional form allows us to predict Y for any given x where the differences in predicted Y relate to differences in x in a constant way with respect to the other variables. But this an implication or consequence of the linearity and additivity choice.

Estimate an overall ATE with stratification I

We know how to analyze a block-randomized (or strata-randomized) experiment (see (Gerber and Green, 2012)): each block is a mini-experiment. We estimate the ATE within each block and combine by weighting each block specific estimate.
The block-size weight produces an unbiased estimator in randomized experiments — in an observational study we don't know about the bias since we don't exactly know how to repeat the study. The precision weight (aka the "fixed effects" weights) tends to produce smaller standard errors and confidence intervals but is biased in randomized experiments.

Estimate an overall ATE with stratification II

```r
dat_sets <- dat %>%
  group_by(x4) %>%
  summarize(
    nb = n(),
    ateb = mean(Y[Z == 1]) - mean(Y[Z == 0]),
    prob_trt = mean(Z),
    nbwt = n() / nrow(dat),
    prec_wt = nbwt * prob_trt * (1 - prob_trt),
  )

dat_sets$prec_wt_norm <- with(dat_sets, prec_wt / sum(prec_wt))

print(dat_sets)
```

```
# A tibble: 2 x 7
     x4    nb   ateb prob_trt  nbwt prec_wt prec_wt_norm
  <int> <int>  <dbl>    <dbl> <dbl>   <dbl>        <dbl>
1     0    57  0.595    0.596  0.57   0.137        0.572
2     1    43 -3.75     0.605  0.43   0.103        0.428
```

```r
est_ate1 <- with(dat_sets, sum(ateb * nbwt))
est_ate2 <- with(dat_sets, sum(ateb * prec_wt / (sum(prec_wt))))
```

Estimate an overall ATE with stratification? I

Block- or strata-level weights can also be represented at the individual level — and this allows us to use linear models (least squares) to produce block-weighted estimates of the overall average causal effect after "holding constant" x_4.

Estimate an overall ATE with stratification? II

```r
## Now at the individual level
dat <- dat %>%
  group_by(x4) %>%
  mutate(
    nb = n(),
    mb = sum(Z),
    ateb = mean(Y[Z == 1]) - mean(Y[Z == 0]),
    prob_trt = mean(Z),
    nbwt = (Z / prob_trt) + (1 - Z) / (1 - prob_trt),
    prec_wt = nbwt * prob_trt * (1 - prob_trt)
  ) %>%
  ungroup()

## Two ways to use the block-size weight
est_ate1a <- difference_in_means(Y ~ Z, blocks = x4, data = dat)
est_ate1b <- lm_robust(Y ~ Z, weights = nbwt, data = dat)
est_ate1c <- lm(Y ~ Z, weights = nbwt, data = dat)

## Three other ways to use the precision or harmonic weight
est_ate2a <- lm_robust(Y ~ Z + x4, data = dat)
est_ate2b <- lm_robust(Y ~ Z, fixed_effects = ~x4, data = dat)
est_ate2c <- lm_robust(Y ~ Z, weights = prec_wt, data = dat)

c(est_ate1, coef(est_ate1a)["Z"], coef(est_ate1b)["Z"], coef(est_ate1c)["Z"])
```

Estimate an overall ATE with stratification? III

```
[1] -1.275 -1.275 -1.275 -1.275
```

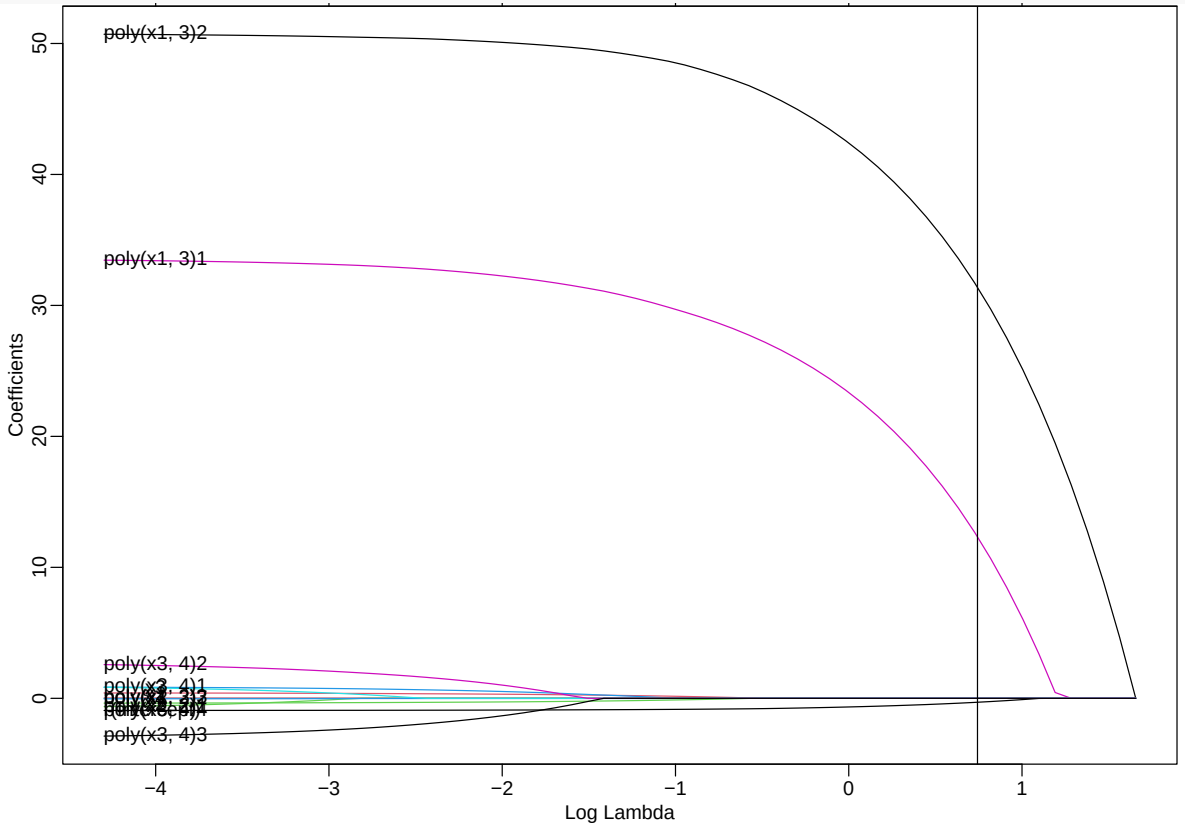
```
c(est_ate2, coef(est_ate2a)[["Z"]], coef(est_ate2b)[["Z"]], coef(est_ate2c)[["Z"]])
```

```
[1] -1.268 -1.268 -1.268 -1.268
```

Finally, what about variable selection? I

We could use a penalized model (like the lasso or adaptive lasso) or some other approach (like random forests) to **automatically choose** a specification.

```
## Here using the mm data with polynomials
library(glmnet)
cv1 <- cv.glmnet(mm[, 3:15], y = dat$Y)
```



Finally, what about variable selection? II

```
14 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)   2.110e+01
x1           -7.434e-01
x2            .
x3            .
x4            .
poly(x1, 3)1 -4.393e-13
poly(x1, 3)2  2.592e+01
poly(x1, 3)3  4.497e+01
poly(x2, 2)1  .
poly(x2, 2)2  .
poly(x3, 4)1  .
poly(x3, 4)2  .
poly(x3, 4)3  .
poly(x3, 4)4  .
```

- Of course, then we have to argue that our **tuning parameter choice** made sense.
- And, again, we have no standard for knowing when we have done **enough**.

1. Overview and Review
2. Concepts and Notation for Causal Inference and Statistical Inference
3. Estimation, Estimators, Bias, Consistency, Given Randomization
4. Covariance Adjustment?
5. Summary and Overview
6. Appendix

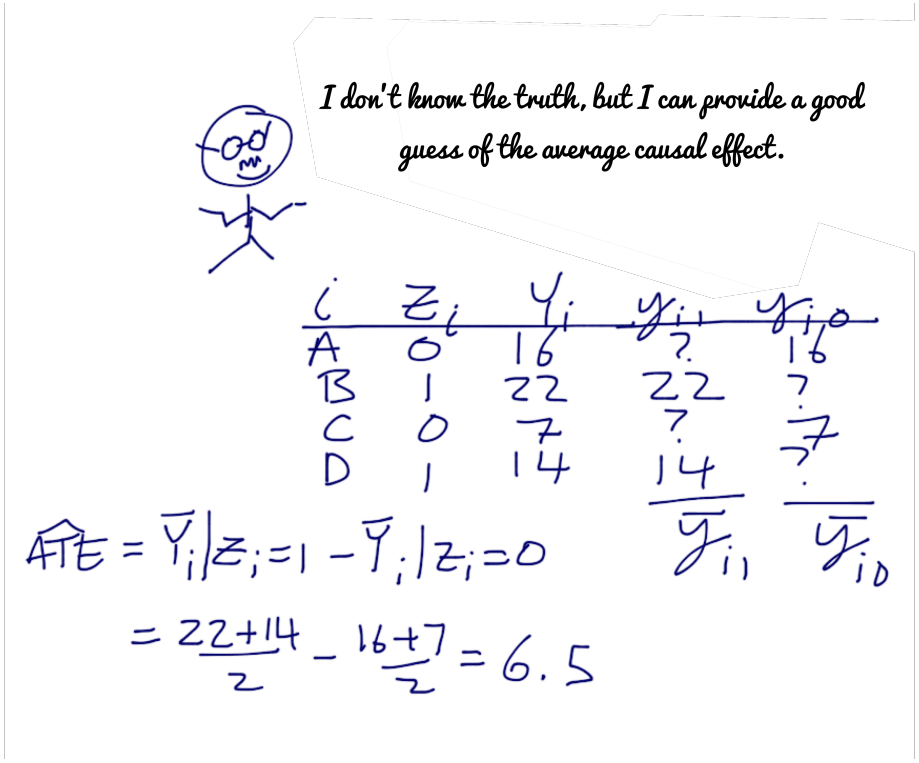
Benefits of Randomized Designs

Randomization makes $y_1, y_0, \mathbf{X} \perp Z$. How to make use of this fact in a randomized experiment?

1. Interpretable comparisons (lack of omitted variable bias, confounding, selection bias)
   - Can I interpret differences in outcome as caused by Z and not X? Is it easy to confuse the effect of Z with the effects of X?
   - How does randomization do this? How does randomization eliminate **alternative explanations**? (Recall that it does not exactly balance X.)
2. Reliable statistical inferences (estimators and tests)
   - The idea of **design-based** versus **model-based** statistical inference (next few slides).

Design Based Approach: Estimate Averages

1. Notice that the observed Y_i are a sample from the (small, finite) population of $(y_{i,1}, y_{i,0})$.
2. Decide to focus on the average, $\bar{\tau}$, because sample averages, $\hat{\bar{\tau}}$ are unbiased and consistent estimators of population averages under random sampling (where no covariate determines the sample inclusion probability or assignment to Z_i).
3. Estimate $\bar{\tau}$ with the observed difference in means.



Design Based Approach: Estimate Averages

I don't know the truth, but I can provide a good guess of the average causal effect.

i	Z_i	Y_i	y_{i1}	y_{i0}
A	0	16	?	16
B	1	22	22	?
C	0	7	?	7
D	1	14	14	?
			$\bar{y}_{i1}$	$\bar{y}_{i0}$

$$ATE = \bar{Y}_i|Z_i=1 - \bar{Y}_i|Z_i=0$$

$$= \frac{22+14}{2} - \frac{16+7}{2} = 6.5$$

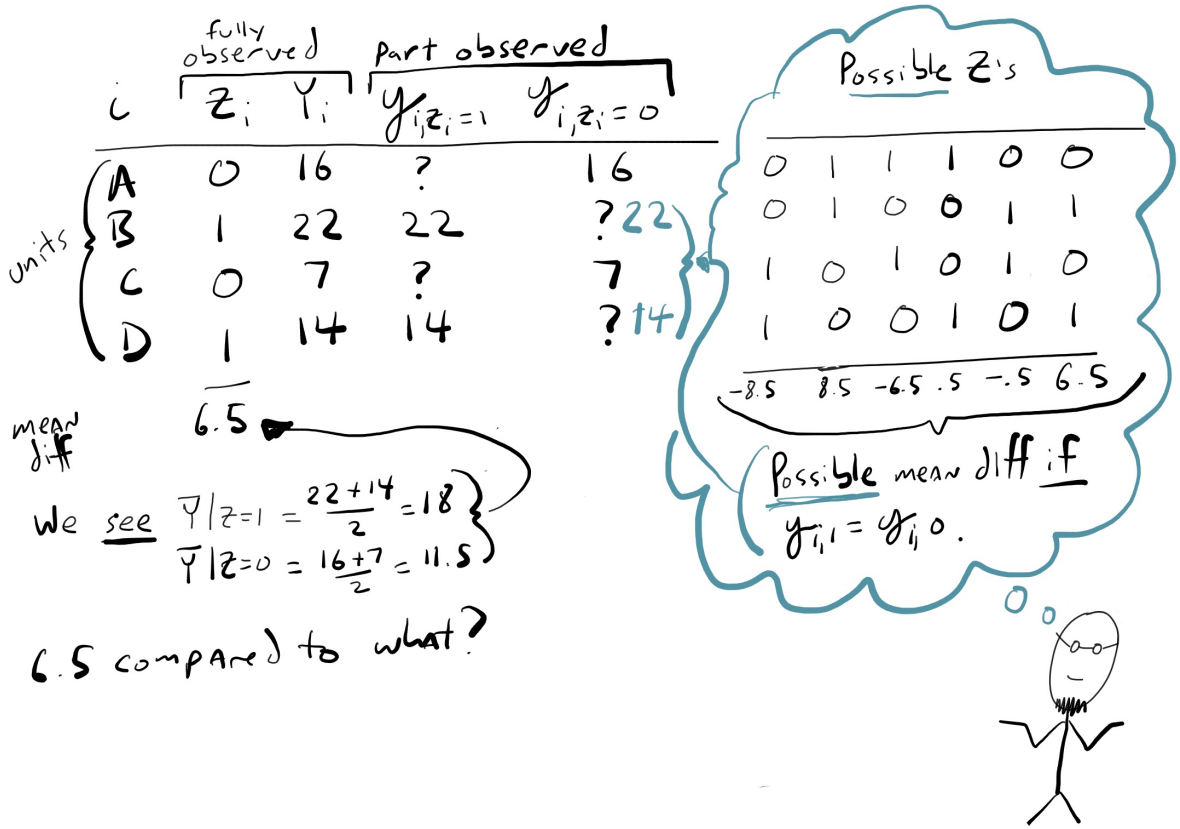
What about when we have not randomized?

We can try to adjust (for example by "controlling for". But adjustment raises new questions: how to adjust? how to assess our adjustment?) After all **Regression is not research design**
Ideas for a simpler way to say "We have held x1 constant?" as a way to discard alternative explanations based on x1?

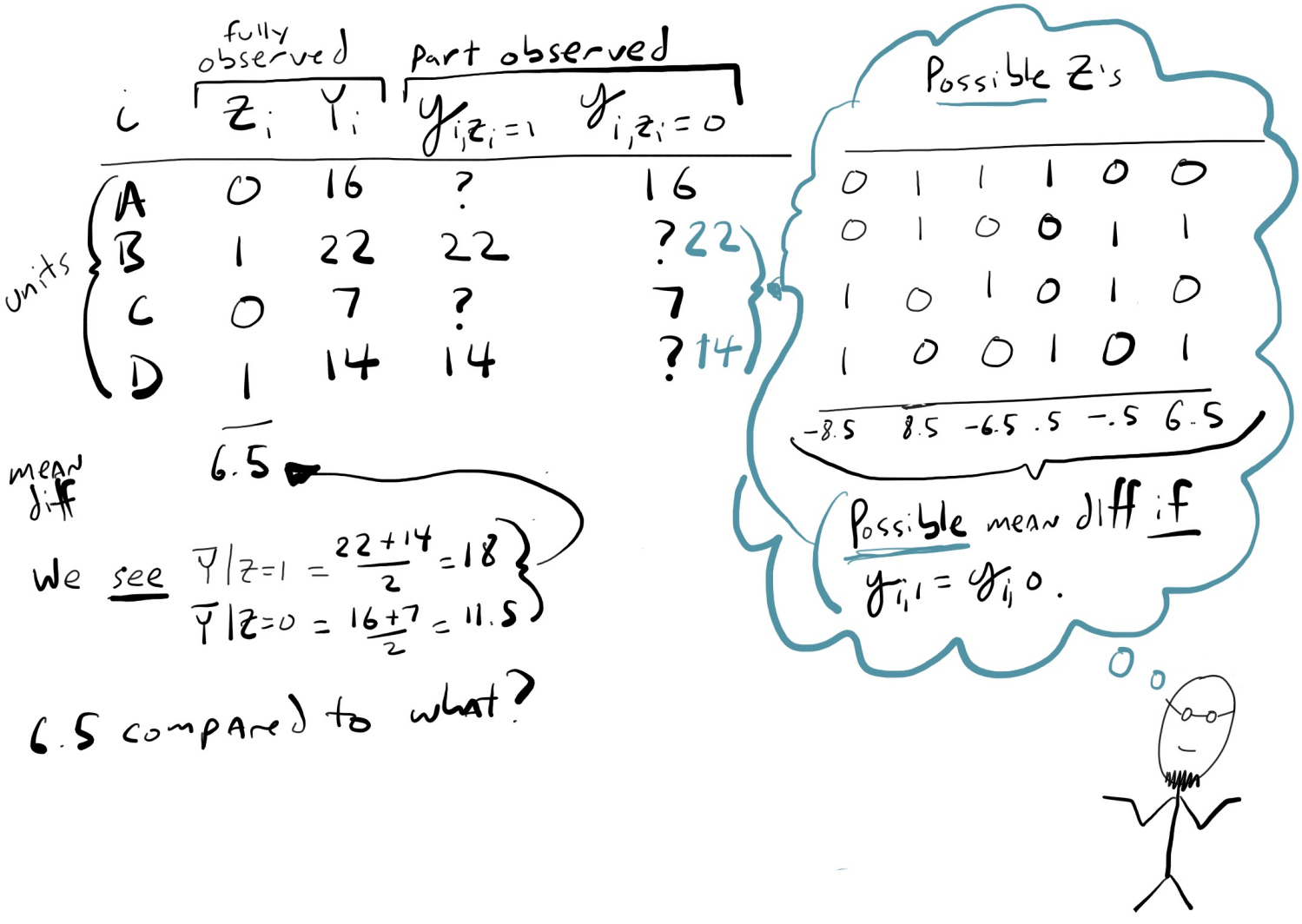
Lingering Questions?

Design Based Approach: Test Hypotheses

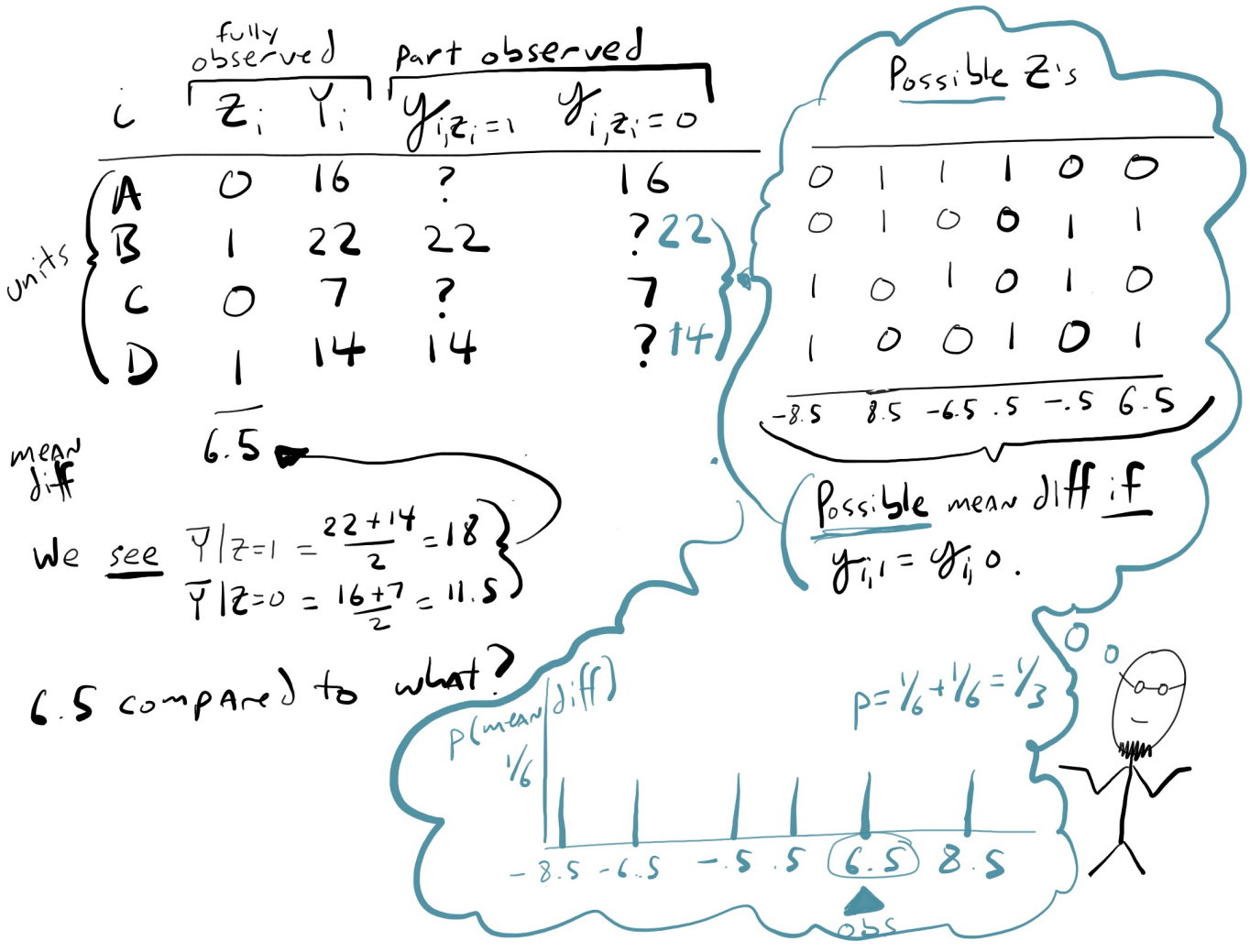
1. Make a guess about τ_i.
2. Then measure surprise or consistency of data with this guess given the design. (Given all of the ways this experiment could have occurred, how many look more extreme than what we observe? Does our observation look typical or rare?)



Design Based Approach: Test Hypotheses



Design Based Approach: Test Hypotheses



Approaches to creating interpretable comparisons:

- Randomized experiments (more precision from reducing heterogeneity in Y)
- Instrumental variables (with randomized Z created D)
- Natural Experiments / Discontinuities (one X creates Z) (includes RDD)
- Difference-in-Differences (reduce bias and increase precision from reducing heterogeneity in Y)
- Semi-/Non-parametric Covariance Adjustment (ex. Matching)
- Parametric covariance adjustment

References


Brady, Henry E. (2008). "Causation and explanation in social science". In: Oxford handbook of political methodology, pp. 217–270.


Gerber, Alan S and Donald P Green (2012). Field Experiments: Design, Analysis, and Interpretation. New York, NY: W.W. Norton.


Rabb, Nathaniel et al. (July 2022). "The influence of social norms varies with "others" groups: Evidence from COVID-19 vaccination intentions". In: Proceedings of the National Academy of Sciences 119.29. DOI: https://doi.org/10.1073/pnas.2118770119.


Rosenbaum, Paul R (2010). Design of Observational Studies. New York, NY: Springer.