Homework 05

Causal Inference

Name

October 15, 2019

Design of an experiment

1. Suppose you are interested in the effect of the presence of vending machines in schools on childhood obesity. What randomized experiment would you want to do (in a perfect world) to evaluate this question?

We draw a sample of shools, then from DAY1, it splited into two parallel universes where one suddently has a vending machine on campus and the other one doesn't. We then travel back and forth from two universes to observe students' behavior and measure their obesity level, then the difference of obesity level of each individual between the two parallel universes are the causal effect of the vending machine in schools posted on childhood obesity. The difference of the average of childhood obesity between the two universes are then the average efficacy of vending machines in schools with respect to childhood obesity.(Okay, this won't happend, but I get the point that randomized control experiment is **hard**)

2. Suppose you are interested in the effect of smoking on lung cancer. What randomized experiment could you plausibly perform (in the real world) to evaluate this effect?

Will it be un-ethical to actually do experiment on human subject that can cause cancer? If this senario can ever be done in the real world, then the **experiment** will be like the following:

- we would probably draw a random sample of smokers, ask them to recall when they started to smoke, the frequency of their somking behavior and other biological and economical or financial measures(as many as possible?).
- Then we will draw from the population of non-smokers, measure their exact same attributes as we measured the smoking subjects.
- We will then wait for years, then record their health status, remeasure previous measures. Maybe after 20-30 years. I record how many of them get lung cancer.
- We compare the averaging effect of length of smoking along with other measures to model a probablity model of one getting a lung cancer.

3. Suppose you are a consultant for a researcher who is interested in investigating the effects of teacher quality on student test scores. Use the strategy of mapping this question to a randomized experiment to help define the question more clearly. Write a memo to the researcher asking for needed clarifications to this study proposal.

The attempt of exploring causal effects of teacher's quality to student test scores is generally very tricky. Before even design an experiment, in order to investigate this, the following needs to be answered.

- How to measure a teacher's quality?
  - Education level?
  - Years as teachers?
- What test scores are measured and how they are measured?
  - Measuring the differences between pre-treatment test scores and post-treatment test scores?
  - If only one measures of test scores can be obtained, then what can we know about the potential test scores of the pre-treatment student groups?
- Does the dynamic between the teacher and the students affect the test scores? How can we measure that?

If all these can be measured and answered. We then can possibly investigate the causal effect of may be one type of teacher quality on students' scores' changes.

Causal effect

The table below describes a hypothetical experiment on 2400 persons. Each row of the table specifies a category of person, as defined by his or her pre-treatment predictor x, treatment indicator T , and potential outcomes y_0, y_1. (For simplicity, we assume unrealistically that all the people in this experiment fit into these eight categories.)

Category	# persons in category	x	T	y_0	y_1
1	300	0	0	4	6
2	300	1	0	4	6
3	500	0	1	4	6
4	500	1	1	4	6
5	200	0	0	10	12
6	200	1	0	10	12
7	200	0	1	10	12
8	200	1	1	10	12

In making the table we are assuming omniscience, so that we know both y_0 and y_1 for all observations. But the (nonomniscient) investigator would only observe x, T, and y^T for each unit. (For example, a person in category 1 would have $x = 0$,$T = 0$,$y = 4$, and a person in category 3 would have $x = 0$,$T = 1$,$y = 6$.)

(a) What is the average treatment effect in this population of 2400 persons? (a.ans) 2 (a.ans) The average treament effect in this case is the average y_1 outcome of treatment group minus the average y_0 outcome of the control group which equals to

(b) Is it plausible to believe that these data came from a randomized experiment? Defend your answer.

(b.ans) If we compare the treament assignment against the pre-treatment predictors:

$x/\#/T$	$T = 0$	$T = 1$
$x = 0$	500	700
$x = 1$	500	700

- Because the marginal distribution of x for different T, as well as the marginal distribution of T for different x are identical.We can assume the data came from a randomized experiment

(c) Another population quantity is the mean of y for those who received the treatment minus the mean of y for those who did not. What is the relation between this quantity and the average treatment effect?

(c.ans) The average y difference between the groups is 3.31 which is larger than the true effect 2.

(d) For these data, is it plausible to believe that treatment assignment is ignorable given sex? Defend your answer.

(ans.d) If sex is refering to x, where $x = 1$ indicates one gender and $x = 0$ indicates another, then yes. because the Treatment group and Control group assignments are indentical.

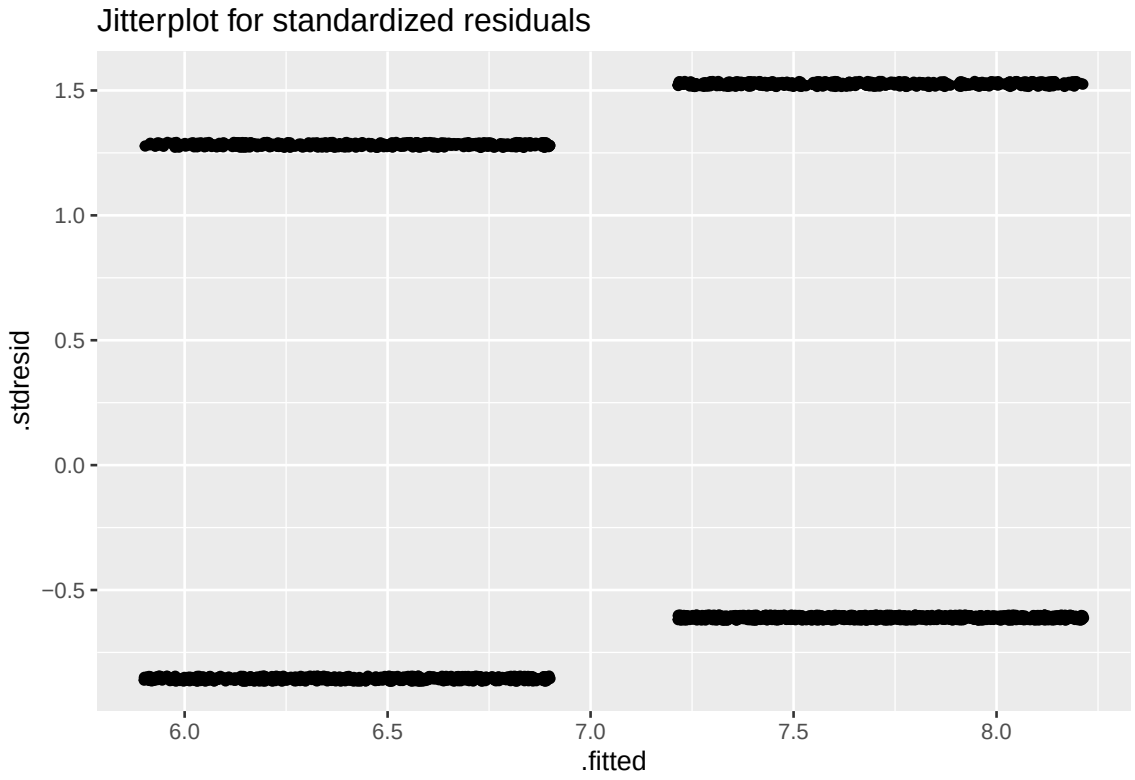
(e) Figure out the estimate and the standard error of the coefficient of T in a regression of y on T and x.

```
numofsubs <- c(300,300,500,500,200,200,200,200)
x = c(0,1,0,1,0,1,0,1)
T1 = c(0,0,1,1,0,0,1,1)
y = c(4,4,6,6,10,10,12,12)
```

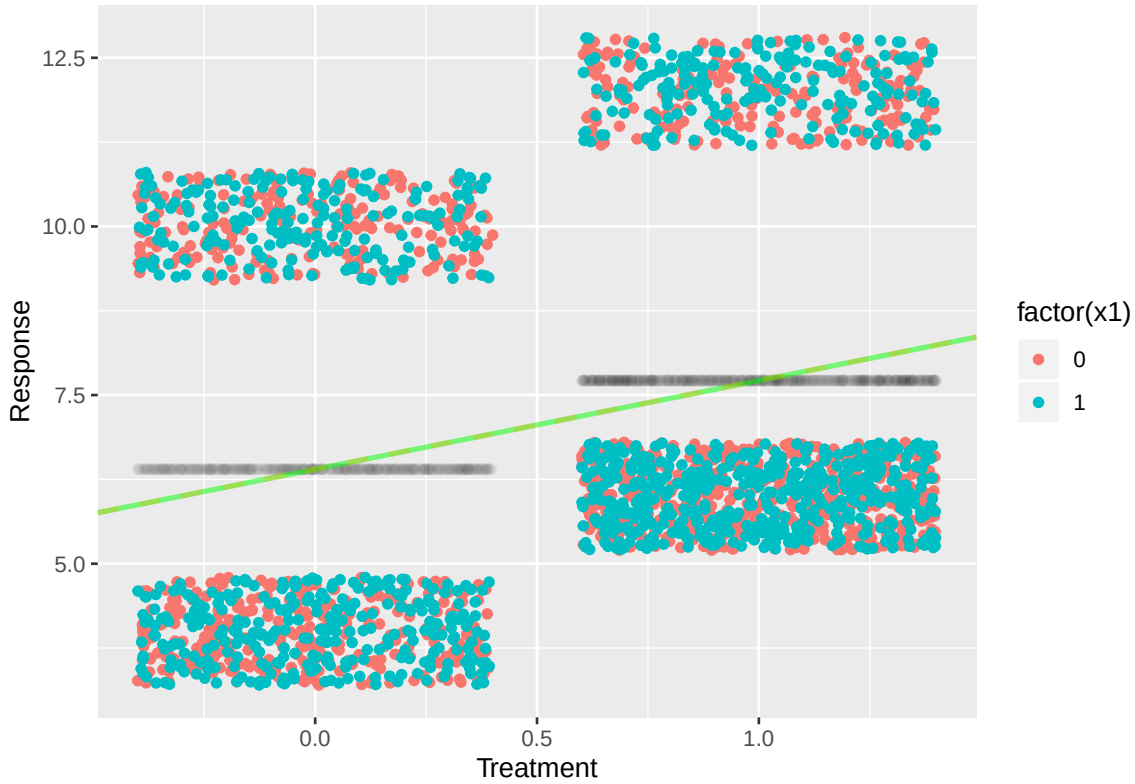
```r
dumb <- function(x,nums){
  a <- list()
  for(i in 1:length(nums)){
    a <- append(a,rep(x[i],nums[i]))
  }
  return(unlist(a))
}
x1 <- dumb(x,numofsubs)
T11 <- dumb(T1,numofsubs)
y1 <- dumb(y,numofsubs)


tab2 <-  data.frame(cbind(x1,T11,y1))

fit.1 <- lm(y1~T11+x1,data = tab2)
summary(fit.1)
```

```
##
## Call:
## lm(formula = y1 ~ T11 + x1, data = tab2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.400 -1.886 -1.714  3.600  4.286
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.400e+00  1.058e-01   60.51   <2e-16 ***
## T11         1.314e+00  1.163e-01   11.30   <2e-16 ***
## x1          2.901e-15  1.147e-01    0.00        1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.81 on 2397 degrees of freedom
## Multiple R-squared:  0.05055,    Adjusted R-squared:  0.04976
## F-statistic: 63.81 on 2 and 2397 DF,  p-value: < 2.2e-16
```

```r
ggplot(fit.1)+aes(x = .fitted, y = .stdresid)+geom_jitter(height = .01,width = .5)+ggtitle("Jitterplot
```




```
ggplot(tab2)+aes(x = T11, y = y1) + geom_jitter(aes(color = factor(x1)))+xlab("Treatment")+ylab("Respons
```



For linear models the coefficient of $T11$ is estimated to be 1.314, and the standard error is estimated as 0.1058.

Consulting

You are consulting for a researcher who has performed a randomized trial where the treatment was a series of 26 weekly therapy sessions, the control was no therapy, and the outcome was self-report of emotional state one year later. However, most people in the treatment group did not attend every therapy session. In fact there was a good deal of variation in the number of therapy sessions actually attended. The researcher is concerned that her results represent "watered down" estimates because of this variation and suggests adding in another predictor to the model: number of therapy sessions attended. What would you advise her?

ans: I think the biggest issue of the experiment is the pre-treatment self-value emotional state of all the participants.

Once we have that. Depending on what kind of questions she tries to answer.

If the question is: does the therapy has a positive influence on peoples emotional state. Then there is no need to add the number of therapy sessions. The variance of sessions attended is just the true nature of how the population will engage with your intervention.

A more challenging question can be in any of the following form: How useful is one session? Does the number of sessions have impact on the usefulness of the intervention? what is the increase in happiness do this intervention cause.

To answer this question, some prerequisits must be satisfied:

1. How to measure emotional status in a continuous value.
2. You really need to have roughly the same pre-treatment sample distribution between different numbers of sessions attended.
3. You need to be able to build a model to predict the potential of attending x number of sessions if given treatment in the control group and then they need to have similar distribution compared to their treated conterparts.

If the conditions listed above are satisfied, we can then make causal inference on the effectiveness of one session of the therapy thus answer those questions.

Gain-score models:

In the discussion of gain-score models in [GH] Section 9.3, we noted that if we include the pre-treatment measure of the outcome in a gain score model, the coefficient on the treatment indicator will be the same as if we had just run a standard regression of the outcome on the treatment indicator and the pre-treatment measure. Show why this is true.

$$g_i = y_i - x_i$$

so if we substitute g_i into the gain score model with x_i we will have:

$$y_i - x_i = \alpha + \theta T + \gamma x_i + error_i$$

$$y_i = \alpha + \theta T + x_i + \gamma x_i + error_i$$

$$y_i = \alpha + \theta T + (1+\gamma) x_i + error_i$$

we rewrite $\beta^\star = 1 + \gamma$,then we yeild:

$$y_i = \alpha + \theta T + \beta^\star x_i + error_i$$

because the model use the same information to fit, so the result of fitting the above model should be the same as fitting

$$y_i = \alpha + \theta T + \beta x_i + error_i$$

linear regression Assume that linear regression is appropriate for the regression of an outcome, y, on treatment indicator, T, and a single confounding covariate, x. Sketch hypothetical data (plotting y versus x, with treated and control units indicated by circles and dots, respectively) and regression lines (for treatment and control group) that represent each of the following situations: a. No treatment effect, b. Constant treatment effect, c. Treatment effect increasing with x.

Hypothetical Study

Consider a study with an outcome, y, a treatment indicator, T , and a single confounding covariate, x. Draw a scatterplot of treatment and control observations that demonstrates each of the following: (a) A scenario where the difference in means estimate would not capture the true treatment effect but a regression of y on x and T would yield the correct estimate. (b) A scenario where a linear regression would yield the wrong estimate but a nonlinear regression would yield the correct estimate.

Messy randomization

The folder `cows` contains data from an agricultural experiment that was conducted on 50 cows to estimate the effect of a feed additive on six outcomes related to the amount of milk fat produced by each cow.

Four diets (treatments) were considered, corresponding to different levels of the additive, and three variables were recorded before treatment assignment: lactation number (seasons of lactation), age, and initial weight of cow.

Cows were initially assigned to treatments completely at random, and then the distributions of the three covariates were checked for balance across the treat- ment groups; several randomizations were tried, and the one that produced the "best"balance with respect to the three covariates was chosen. The treatment assignment is ignorable (because it depends only on fully observed covariates and not on unrecorded variables such as the physical appearances of the cows or the times at which the cows entered the study) but unknown (because the decisions whether to rerandomize are not explained). We shall consider different estimates of the effect of additive on the mean daily milk fat produced. a. Consider the simple regression of mean daily milk fat on the level of additive. Compute the estimated treatment effect and standard error, and explain why this is not a completely appropriate analysis given the randomization used. b. Add more predictors to the model. Explain your choice of which variables to include. Compare your estimated treatment effect to the result from (a). c. Repeat (b), this time considering additive level as a categorical predictor with four letters. Make a plot showing the estimate (and standard error) of the treatment effect at each level, and also showing the inference the model fit in part (b).

sesame

The folder `sesame` contains data from an experiment in which a randomly selected group of children was encouraged to watch the television program Sesame Street and the randomly selected control group was not. a. The goal of the experiment was to estimate the effect on child cognitive devel- opment of watching more Sesame Street. In the experiment, encouragement but not actual watching was randomized. Briefly explain why you think this was done. (Hint: think of practical as well as statistical reasons.) b. Suppose that the investigators instead had decided to test the effectiveness of the program simply by examining how test scores changed from before the intervention to after. What assumption would be required for this to be an appropriate causal inference? Use data on just the control group from this study to examine how realistic this assumption would have been. c. Did encouragement (the variable viewenc in the dataset) lead to an increase in post-test scores for letters (postlet) and numbers (postnumb)? Fit an appropriate model to answer this question. d. We are actually more interested in the effect of watching Sesame Street regu- larly (regular) than in the effect of being encouraged to watch Sesame Street. Fit an appropriate model to answer this question. e. Comment on which of the two previous estimates can plausibly be interpreted causally.