Causal Inference Notes for Students [DRAFT]

Annie Chen

March 2020

This is a collection of common misunderstandings/questions that I encountered while marking your problem sets. I hope that a document initially written for my own understanding can also be of use to you. Here I make a few clarifications (especially regarding questions that require lengthier explanations), share what I consider to be some of the most important takeaways from each question, and expound on the mechanics. Links to additional sources are also provided for those who would like to read more on the topic. The last page contains some useful math definitions and identities. Let me know if you have any questions!

Contents

A Problem set 2

Does failure to reject the null hypothesis of no difference mean there is balance between treated and control groups?

It is important to understand, firstly, why do we not accept the null hypothesis. If the null hypothesis of no difference (typically in a two-sample t-test, $H_0 : \mu_1 - \mu_2 = 0$) is not rejected at some significance level, we cannot discriminate *which* hypothesis is correct. It may be that your test is under-powered because the data doesn't give you enough juice to find a significant difference, for example. So, not finding a statisitically significant difference in your pre-treatment covariates may be *consistent with* balance, but does not prove that the difference between treated and control groups is zero. This is a concept that applies generally in hypothesis testing (in the frequentist tradition).

Searching for covariate balance using this backwards logic has been criticized. Procedures that "invert" the null hypothesis such that researchers test for difference rather than no difference are, in my humble opinion, more sensible. I highly recommend taking a look at the linked paper. It explains equivalence testing using two-one-sided t-tests (TOST).

Further reading: An Equivalence Approach to Balance and Placebo Tests

B Problem set 4

Why does OLS regression and matching yield different results?

In the presence of individual-level heterogeneous treatment effects, least squares regression gives you conditional-variance-weighted causal estimate. This is because OLS regression uses a minimum-variance estimator that imparts greater weight to strata with effects of the lowest expected variance. Additionally, the least squares estimator imposes a functional form on the estimated average causal effect, whereas matching estimators are non-parametric.

Further reading: Morgan & Winship (p.206)

Why do we care about common support? What is the difference between strong and weak overlap?

How do I compare treated group and control group propensity scores after matching?

Once you've successfully matched on your estimated propensity score, you can access the matched values by appending `$mdata` to your matched object, which returns a list including your inputs (`$X, $Y, $Tr`). What you want is to figure out the estimated propensity scores for associated treated and control units. The indices of the treated group and control group can be recovered using `$index.treated` and `$index.control` respectively. For example:

```
ps_model <- glm(...family = binomial("logit"))
ps_match <- Match(..X = ps_model$fitted.values)

#propensity score of treated, likewise for control
ps_model$fitted.values[ps_match$index.treated]
```

```
#or, examine the X input directly from the matched object
ps_match$mdata$X[index.treated, ]
```

Once you have these vectors, it's straightforward to plug them into your `ggplot()`.

Further reading: `Match()` documentation

C Problem set 5

How do I make sense of contour plots for sensitivity analysis?

Part of the confusion stems from the myriad ways to represent the relationship between the treatment, outcome, and covariates. In other words, the sensitivity parameters δ and γ can be reparameterized in different ways – i.e., standardized coefficients, partial R^2 – which affects the scale and interpretability. The following might be more detail than you care to know, so free to skip to the TLDR at the end of this section. While it is easy to get bogged down in the nuances, the basic idea always boils down to a comparison of the relationship between a confounding variable, U, and the outcome, Y, (γ) against the relationship between the confounder and a treatment, D, (δ).

Tired: Sensitivity parameters and their naive interpretation.

Recall from lecture that the bias of an estimated treatment effect in the presence of unobserved confounders can be captured by the product of the sensitivity parameters δ and γ. If both the confounder and treatment are binary, they can be non-parametrically defined as:[1]

$$\delta \equiv P(U = 1|D = 1) - P(U = 1|D = 0)$$

and

$$\gamma \equiv \mathbb{E}[Y|U = 1] - \mathbb{E}[Y|U = 0]$$

This is assuming that the confounder has a constant effect on the outcome between treatment status (U and D do not interact). Notice that, while the intuition for these sensitivity parameters are straightforward ("relationship" between two variables), the precise mathematical definitions differ slightly. So, whereas γ is the *effect* of U on the outcome, δ is the "imbalance" of the confounder with respect to the treatment (and potentially other covariates). Meaning, the difference in U across values of the treatment. In the parametric equivalent, $\hat{\delta}$ is the coefficient from a regression of U on D.

Careful readers will have realized that Imbens' 2003 article begins by estimating the reverse regression (i.e. the propensity score). Again supposing the treatment is binary,[2]

$$P(D|U) = \frac{exp(\alpha + \delta U)}{1 + exp(\alpha + \delta U)}$$

This mirrors what we did in lab (but conditioning on extra covariates):

[1] See the VanderWeele article for how we arrived at these.

[2] Imbens includes some other covariates ($\mathbf{X}$) in his example, which I haven't included for simplicity, but same idea. As another heads-up, he uses different letters, but I alter them here to be consistent with what we did in class.

```
ps_mod <- glm(abd ~ age + fthr.ed + mthr.ed + C.ach, data = child_data,
          family = binomial(link = "logit"))

#all else equal, residency in Ach or not
mod_ach1 <- data.frame(model.matrix(ps_mod)[, 1:4], abd = 1)
mod_ach0 <- data.frame(model.matrix(ps_mod)[, 1:4], abd = 0)

#predict setting C.ach at 1 or 0, conditioned by other covariates
#type="response" tells R to output probabilities, P(D = 1|X)
ps_ach1 <- predict(ps_mod, newdata = mod_ach1, type = "response")
ps_ach0 <- predict(ps_mod, newdata = mod_ach0, type = "response")
```

which was used to calculate $P(D|U = 1) - P(D|U = 0)$, the difference in propensity scores for levels of potential confounders (i.e., $U_{\texttt{C.ach}} \in \{0, 1\}$). Yet, we know that regression coefficients are not symmetric – the effect of `x` $\sim$ `y` is not the same as `y` $\sim$ `x` – and in general, $P(D|U) \neq P(U|D)$. The short answer is that the use of propensity scores in this manner reflects the approach of Rosenbaum and Rubin (1983) who specify the two sensitivity parameters as odds-ratios and a third parameter for $P(U|X)$, the marginal distribution of the confounder (its overall prevalence) – put differently, they are connected via Bayes' Theorem.[3] Suffice it to say, both are legitimate ways of representing the relationship between the U and D, but are interpretable on different scales. It ***is*** important that the sensitivity parameters are ***on the same scale*** when you are plotting the hypothetical curve and the benchmarks.

Wired: Reparameterization as partial R-squared.

Instead of looking at raw regression coefficients, it is common to reparameterize the sensitivity parameters in terms of correlations (as Imbens did). The symmetry of correlations (like R^2) means they are invariant to directionality (i.e., $\rho_{X,Y} = \rho_{Y,X}$). Specifically, the use of partial R^2 is intuitive (*partial* because it is proportion of the original variance that is explained by the confounder – how much does the confounder contribute beyond other covariates in the model). In `R`, you can use `rsq::rsq.partial()`. Similarly, Blattman (2009) uses the difference in R^2 (also ref. the lab slides where we used `lm.beta()`, standardized coefficients) to capture the relative contribution of the confounder. They are, for our purposes, the same thing.[4] For example,

```
library(rsq)

#delta
```

[3]For a formal exposition, see the seminal paper written by Rosenbaum and Rubin (1983). Collorary 4.2 and section *Subclassification on Propensity Scores*, in particular. Their work falls directly from the fact that a propensity score is a *balancing score* and is intimately connected with the use of propensity scores to achieve balance (something we learned from the week on Matching). In Appendix to VanderWeele and Arah (2011), see section *Relation to the Sensitivity Analysis of Rosenbaum and Rubin (1983).*

[4]Technically, the statistics differ in their denominators, but they are all some measure of effect size.

```
delt <- rsq.partial(lm(abd ~ age + C.ach + fthr.ed, data))
partr_age_d <- delt$partial.rsq[1]

#gamma
gamm <- rsq.partial(lm(educ ~ age + C.ach + fthr.ed, data))
partr_age_g <- gamm$partial.rsq[1]
```

These analyses reflect an interest in the magnitude of the relationship between the variables. However, direction matters because it distinguishes between confounders, colliders, mediators. Therefore, strictly speaking, it is the treatment effect of `D` $\sim$ `U` rather than `U` $\sim$ `D` that generates the spuriousness (otherwise, U is a mediator). You might frequently see OVB in regression formulated as `U` $\sim$ `D` because whether some bias exists and the direction of that bias depends only on the numerator $Cov(D, U)$ (and by symmetry, $Cov(U, D)$). A subtle insight is that OVB in regression holds *whether or not* it can be interpreted causally. This speaks to the old "correlation is not causation" adage. Where the magnitude of bias is concerned and is quantified by the effect as opposed to the correlation, the difference between a regression of D on U and the reverse regression is non-trivial!

Inspired: Other Interpretations.

Matt Blackwell's approach involves specifying a "confounding function," $q(D, X)$. Might be of interest to some. See further readings.

TLDR

The curve you plotted are all combinations of δ and γ (on whatever scale you choose to parameterize them) that would reduce the ATE by some magnitude of your choosing. It represents confounding produced by a hypothetical U. To see how this compares to the variables in your current model, perform regressions of the treatment on the covariates (x-axis) and outcome on covariates (y-axis) for each variable separately. Importantly, the parameterization of δ and γ should be consistent between your benchmark points and the hypothetical curve (i.e., extract standardized coefficients in your models). Now that you have a sense for the influence of each variable, you may use the relative position of these observed covariates to reason theoretically (i.e., using your substantive knowledge) about the types of variables likely to be confounding and the plausibility of such confounders. The graph *does not* tell you whether U exists or not, but might be suggestive of how susceptible your model is to misspecification. For instance, you would expect that all included covariates to score very low on the x-axis in experimental research. You can take a look at a full script where the sensitivity parameters are defined as partial R^2s here.

Further reading:

Blackwell, Matthew. 2013. "A Selection Bias Approach to Sensitivity Analysis for Causal Effects." Political Analysis 22:169–182.

Imbens, Guido. 2003. "Sensitivity to Exogeneity Assumptions in Program Evaluation." The American Economic Review. Vol. 93 No. 2.

VanderWeele, Tyler J. and Onyebuchi A. Arah. 2011. "Unmeasured Confounding for General Outcomes, Treatments, and Confounders: Bias Formulas for Sensitivity Analysis" Epidemiology. January ; 22(1): 42–52.

How does randomization inference work in Rosenbaum's sensitivity analysis using Wilcoxon Signed-Rank Test?

Recall that randomization inference works by calculating different test statistics derived from a permutation of randomization schemes. One typical sharp null hypothesis is that there is no effect for all units, $H_0 : Y_i(0) - Y_i(1) = 0 \; \forall i$, *not that there is no average effect across all individual units*. This allows us to assume that treated and untreated units are exchangeable.

One version of Rosenbaum's sensitivity analysis employs the Wilcoxon Signed-Rank Test as the test statistic. Given that we are working with matched data, it is suitable because it tests for the (median) difference between pairs. Concretely, consider a simple example with four observations. Let `Y_i` and `Y_j` be columns in a data frame of matched pairs `data`, which are the outcomes of your matched data where `Y_i` – `Y_j` $\neq 0$.

					G = 1	G = 2	
Y_i	Y_j	Delta	Rank	Sign		max(pi(X))	min(pi(X))
10	5	5	4	+	0.5	0.67	0.33
7	5	2	1	+	0.5	0.67	0.33
3	7	-4	3	-	0.5	0.33	0.67
10	13	-3	2	-	0.5	0.33	0.67

Under conditional ignorability and the sharp null, the probability of assignment, $\pi(X)$, for i and j are equally likely. That is, the probability of treatment is constant *within* matched pairs. Generating different combinations of treatment assignments, and calculating the Wilcoxon signed-rank statistic (W-statistic) for each permutation allows us to form a null distribution. Comparing the observed W-statistic (from your data) to this distribution gives you an exact p-value.

Recall that the W-statistic is equal to the sum of positive ranks.[5] The largest possible value of W is therefore, 10 (if all ranks are positive, their sum is $W = 1+2+3+4 = 10$). Similarly, if no ranks are positive, $W = 0$. Where $\Gamma = 1$, there is equal probability of treatment (0.5) within each pair: $(\frac{1}{2})(\frac{1}{2})(\frac{1}{2})(\frac{1}{2}) = \frac{1}{16}$. The observed W is $4 + 1 = 5$, and there are 6 numbers from the permuted distribution of $W \in 0, ..., 10$ that are ≥ 5. To calculate the p-value, we sum up the probabilities of $P(W \geq 5) = P(W = 5) + ... + P(W = 10)$. Because each W is obtained with equal probability, the p-value is $\frac{1}{16} \times 6 = \frac{6}{16} = 0.375$.[6]

Now, consider the case where the departure from unconfounding is $\Gamma = 2$. The upper bound for the probability of assignment when $\Gamma = 2$ is $max(\pi(X)) = \frac{2}{3}$, and the lower bound is $min(\pi(X)) = \frac{1}{3}$.[7] It gets trickier to calculate the p-values because probabilities of treatment assignment are no longer the same for each value of W. Let's begin with the case where we are looking for the maximum p-value (see table). what is the probability of obtaining $W = 0$? First, we should recognize that assigning the highest probability of treatment, $max(\pi(X))$, to the positively ranked pairs and the lowest, $min(\pi(X))$, to the negatively ranked pairs will produce the maximum p-value for the sharp null. This is

[5] Note that this is slightly different than the Wilcoxon Rank-Sum statistic.

[6] Recall the following: $P(X \cap Y) = P(X) \times P(Y)$, the joint probability of two independent events is the product of their individual probabilities.

[7] Remember: $\frac{1}{1+\Gamma} \leq \pi(X) \leq \frac{\Gamma}{1+\Gamma}$

a "worst" case scenario for finding a significant effect. Intuitively, up-weighting positive ranks means there is a higher probability of seeing large W values. Hence, the probability of observing a W-statistic as great or greater than the one obtained from the data will be larger.

A W-statistic of 0 implies that there are *no pairs* of which the signed rank is positive. Since each pair is negatively signed, there is 0.33 probability of assignment each, and the probability of getting $W = 0$ is $(\frac{1}{3})(\frac{1}{3})(\frac{1}{3})(\frac{1}{3}) = \frac{1}{81}$. Essentially, we are asking: *what is probability of seeing this particular arrangement of signed ranks?* Likewise, $P(W = 1) = (\frac{2}{3})(\frac{1}{3})^3 = \frac{2}{81}$. And so on. Notice that the probability of seeing any particular value of W is equal to the product of the probabilities of assignment exponentiated by the number of positive or negative ranks.

To obtain the minimum p-value, we swap the treatment probabilities such that the probabilities of assignment are now the most in favour of finding an effect. Below is an example (somewhat simplified) of how the upper and lower (exact)[8] p-values for any given Γ can be computed.

```
#upper and lower bound treatment assignment probability
max_pscore <- G/(1 + G)
min_pscore <- 1/(1 + G)

#let permuted_signed_diffs be a 4x16 matrix of signed differences
#vector containing number of -ve and +ve values in each permutation
num_neg <- length(permuted_signed_diffs < 0)
num_pos <- length(permuted_signed_diffs > 0)

#max/min p-values
pmax <- sum((min_pscore^num_neg) * (max_pscore^num_pos))
pmin <- sum((max_pscore^num_neg) * (min_pscore^num_pos))
```

What is an Odds Ratio?

The *odds* of an event is $\frac{p(event)}{p(non-event)}$. The probability of an event *not* occurring is its complement, which may be represented as $1 - p(event)$.[9] An odds ratio is the relative likelihood of something happening under two different conditions/events. The odds ratio for two events, say $event_1$ and $event_0$ is a comparison of their odds:

$$OR = \Big(\frac{p(event_1)}{1 - p(event_1)}\Big) \Big/ \Big(\frac{p(event_0)}{1 - p(event_0)}\Big)$$

An odds ratio of 1 is equal odds of either event. In the context of Rosenbaum's sensitivity analysis, this is $\Gamma = 1$. Taking one example from the problem set, the odds of being abducted (treated) given residency in Acholibur is the probability of being abducted for a child from this village over the probability of not being abducted for a child in the same village. This value is then compared to the odds of abduction for non-Acholibur

[8] In general, you may have many observations, and the number of permutations become unwieldy. In such cases, we can rely on asymptotics.

[9] Alternatively: $p(event^c)$, $p(event')$, $p(\overline{event})$.

children. An odds ratio > 1 means increased chance of treatment under the condition in the numerator.[10] For example, a value of 1.3 can be read as: "the odds of being abducted is 30% higher for a child from Acholibur than a child who is not." It might be helpful to think of it this way:

	Lives in Acholibur (X = 1)	Not in Acholibur (X = 0)
Abducted (D = 1)	a	b
Not Abducted (D = 0)	c	d

Where the odds ratio is:[11]

$$OR = \Big(\frac{a}{c}\Big)\Big/\Big(\frac{b}{d}\Big) = \Big(\frac{p(D=1|X=1)}{1-p(D=1|X=1)}\Big)\Big/\Big(\frac{p(D=1|X=0)}{1-p(D=1|X=0)}\Big)$$

In the homework assignment, your code might look something like this:[12]

```
# i.e., ps_ach1 is p(abd = 1| C.ach = 1)
# and 1 - ps_ach1 is p(abd = 0| C.ach = 1)
(mean(ps_ach1)/(1-mean(ps_ach1)))/(mean(ps_ach0)/(1-mean(ps_ach0)))

# alternatively, you can convert the log odds coefficient
ps_mod <- glm(abd ~ C.ach + age + fthr.ed + mthr.ed,
              data, binomial(logit))

exp(ps_mod$coefficients[2])
```

D Problem set 6

Do overlapping confidence intervals necessarily indicate statistical insignificance? (No.)

One heuristic we frequently use to deduce the statistical significance of a difference in two independent sample means is to look at the separation of errorbars/shaded regions (representing whatever level of confidence). If they do not overlap, we say that the difference is statistically significant at a given confidence level. But, does the reverse also hold? That is, is it true that the difference is not statistically significant when interval estimates overlap? As the subsection title hints, the answer is a firm *no*. This because computing uncertainty for means *separately* is not the same as forming a confidence interval around the difference between these means. Figure XXX illustrates this graphically. In short, you can have a statistically significant result even in the presence of overlapping intervals.

[10] OR < 1 is a reduction in likelihood of that event occurring.

[11] In the denominator, $1 - p(D = 1|X = 1)$ is the same as $p(D = 0|X = 1)$. Also important to note that $1 - p(D = 1|X = 1) \neq p(D = 1|X = 0)$.

[12] You can also exponentiate the coefficient from a logistic regression.

Is the parallel trends assumption testable?

By now, you are probably well-aware that many of the assumptions we make in causal inference relies on counterfactual statements. Why should the parallel trends assumption be an exception? Well, it's not. I think most grasp that divergence between treatment groups in the pre-treatment period is suggestive that parallel trends has been satisfied, but the assumption ultimately requires that there is no differential XXX in the absence of treatment within the period of interest. Explicitly, the assumption is:

$$\mathbb{E}[Y_1(0)|D=1] - \mathbb{E}[Y_0(0)|D=1] = \mathbb{E}[Y_1(0)|D=0] - \mathbb{E}[Y_0(0)|D=0]$$

The subscript here is the time indicator, and D is the treatment group. Read: the difference in potential outcomes under control between the treated group and untreated group is the same. This makes clear that we are really talking about XXX. So, a bit of nuance here.

Are "parallel trends" that move in XX directions more conservative estimates?

This is a misconception.

E Problem set 7

F Some Useful Maths

Some of math in the textbooks can be daunting. Below, I distill some of the most relevant operations that will hopefully minimize the number of times you find yourself asking, "Wait. What happened there?"

Expected Value

If a random variable X is discrete (a.k.a. countable), its expected value is $\mu = \mathbb{E}[X] = \sum_{\forall x} xp(x)$. So, if X takes values $1, 2, ..., 6$, each x is weighted by the probability of obtaining that value like so (supposing equal probability): $\mathbb{E}[X] = (1)(\frac{1}{6}) + ... + (6)(\frac{1}{6})$.

In the continuous case, we must evaluate an integral because the probability that X takes any single value is 0 and we replace addition with integration. The expected value of a continuous variable becomes $\mathbb{E}[X] = \int_{\mathbb{R}} xf(x)dx$. What's going on? We are now weighing by the probability density of x (denoted $f(x)$, the PDF) instead of $p(x)$, and some very small width of x (represented by dx). Here, $xf(x)$ is the integrand and x is the variable of integration.

One trick you will frequently come across when dealing with conditional expectations is the Law of Iterated Expectation, which says $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$. It is also useful to know that an expectation is a linear operator, which is convenient because you can do dope things like $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, as well as, $\mathbb{E}[aX] = a\mathbb{E}[X]$ for some constant a.

Variance, Covariance, Correlation

Variance is actually just a composite of expected values: $\sigma = \mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$, which may or may not aid your understanding. For the purposes of this class, it is more important to grasp that: $Cov(X, Y) = \mathbb{E}[X, Y] - \mathbb{E}[X]\mathbb{E}[Y]$. Moreover, $\rho = Cor(X, Y) = \frac{Cov(X,Y)}{\sigma_{XY}}$ is just the standardized version of $Cov(X, Y)$.

Unlike expectations, variances do not act linearly. Instead, if a is constant, $\mathbb{V}[aX] = a^2\mathbb{V}[X]$.

Combinatorics?