Section 8: Rosenbaum Sensitivity Analysis

Advanced Quantitative Methods (PLSC 504)

Fall 2017

Background

You had some practice with a sensitivity analysis type exercise in Homework 6. This handout is an introduction to the "Rosenbaum Sensitivity Analysis" (see Rosenbaum, 2002 *Observational Studies*, Chapter 4 and Rosenbaum, 2009 *Design of Observational Studies*, Chapter 3) approach as applied to the matched pairs design. This handout will focus specifically on the implementation in Rosenbaum's books, but the notation will be slightly different to be more consistent with the course.

The test statistic we will use is the Wilcoxon signed rank statistic. This test statistic can be viewed as the non-parametric equivalent of the paired t-test. The recipe for the Wilcoxon signed rank statistic is:

1. Calculate the absolute differences in outcomes between units in each pair.
2. Assign ranks from smallest to largest based on the absolute value of the differences.
3. The sum of those ranks *for pairs with positive differences* is the observed test statistic.

Let's consider a simple scenario. Let $s = 1, \ldots, S$ index the matched pairs. Let Y_{si} denote the outcome for unit i in pair s so that $d_s = Y_{si} - Y_{sj}$ is the difference between unit i and j in pair s and $q_s = \text{rank}(|D_s|)$ is the rank for pair s. Let $\text{sng}(a) = 1$ if $a > 0$ and 0 otherwise. Then the test statistic is $T = \sum_s^S \text{sgn}(d_s) \cdot q_s$.

For illustration, consider a toy example with $S = 3$ pairs – 3 units in "treatment" and 3 units in "control".

```
# Treatment units
Y1 <- c(26, 53, 5)

# Matched controls
Y2 <- c(23, 20, 12)
```

First, we calculate the differences between units in each pair:

```
d <- Y1 - Y2
d
```

```
## [1]  3 33 -7
```

Next, we assign ranks:

```
q <- rank(abs(d))
q
```

```
## [1] 1 3 2
```

Finally, we compute the test statistic:

```
sgn <- function(a){
  as.numeric(a > 0)
}
sum(sgn(d)*q)
```

```
## [1] 4
```

Let's make a function for this,

```r
T_Wilcox <- function(d) {
  q <- rank(abs(d))
  T_obs <- sum(sgn(d)*q)
  return(list(T_obs = T_obs, q = q))
}
```

What is the null distribution of this test statistic? Under the null hypothesis of no treatment effect, there are 2^S possible treatment minus control differences. We need some additional notation here to be consistent with the presentation by Rosenbaum.

Let $\mathcal{W}$ be the set of the 2^S possible values of treatment assignment, e.g. $\mathbf{w} = (w_{11}, w_{12}, \ldots, w_{s2})$ of $\mathbf{W}$. Under the matched pairs design, we say that $\mathbf{w} \in \mathbb{W}$ if $w_{s1} = 0$ or $w_{s1} = 1$ such that $w_{s1} + w_{s2} = 1$. In words, one and only one unit in the pair is treated. When we condition on $\mathcal{W}$, we are conditioning on the event $\mathbf{W} \in \mathcal{W}$. This is a subtle but important point.

Some more notation. Let $\mathcal{F} = (Y(1)_{si}, Y(0)_{si}, \mathbf{x}_{si}, \mathbf{u}_{si})$ denote the set of *fixed* values for all subjects. $Y(\cdot)_{si}$ are the potential outcomes for unit i in pair s, $\mathbf{x}_{si}$ are the observed covariates and $\mathbf{u}_{si}$ are the unobserved covariates or "confounders". These quantities are fixed because they do not change when some treatment assignment W_{si} is realized. In a randomized experiment we don't care about $\mathbf{x}_{si}, \mathbf{u}_{si}$ because, by design, they are irrelevant for causal inference.

The "assignment mechanism" for a paired design can be summarized as

$$\Pr(\mathbf{W} = \mathbf{w} \mid \mathcal{F}, \mathcal{W}) = \frac{1}{|\mathcal{W}|} = \frac{1}{2^S} \quad \text{for } \mathbf{w} \in \mathcal{W}$$

Consider the toy example at hand. For the vector of observed responses $\mathbf{Y} = (Y_{11}, Y_{12}, Y_{21}, Y_{22}, Y_{31}, Y_{32}) = (26, 23, 53, 20, 5, 12)$, the realized treatment assignment is $\mathbf{W} = (W_{11}, W_{12}, W_{21}, W_{22}, W_{31}, W_{32}) = (1, 0, 1, 0, 1, 0)$. However, for $S = 3$ there are $|\mathcal{W}| = 2^3 = 8$ possible realizations of $\mathbf{W}$ such that $\mathbf{w} \in \mathcal{W}$. They are enumerated in Table 1,

Table 1: The set $|\mathcal{W}|$ for $S = 3$

Label	W_{11}	W_{21}	W_{31}
1	1	1	1
2	1	1	0
3	1	0	1
4	0	1	1
5	1	0	0
6	0	1	0
7	0	0	1
8	0	0	0

The first row corresponds to the $\mathbf{W}$ we have observed. It's not necessary to write out all 8 elements of this vector since we are imposing the constraint that $W_{s2} = 1 - W_{s1}$. Note that we can write the observed difference for pair s as $d_s = (W_{s1} - W_{s2})(Y_{s1} - Y_{s2})$ or $\mathbf{d} = (d_1, d_2, \ldots, d_2)$ for the S pairs. Under the null hypothesis of no treatment effect we have $d_s = (W_{s1} - W_{s2})(Y(0)_{s1} - Y(0)_{s2})$ so that the randomization just changes the signs of the D_s. Table 2 summarizes the possible differences and the associated value of the Wilcoxon signed rank statistic under the null hypothesis of no treatment effect. The first row corresponds to the condition we have observed in the toy example.

Table 2: Null distribution

Label	**d**	$T(\mathbf{d})$	Probability
1	(3, 33, 7)	4	1/8
2	(3, 33, -7)	2	1/8
3	(-3, -33, 7)	3	1/8
4	(3, -33, 7)	6	1/8
5	(-3, 33, -7)	3	1/8
6	(3, -33, -7)	1	1/8
7	(-3, 33, 7)	5	1/8
8	(-3, -33, -7)	0	1/8

So what is the p-value associated with $T(\mathbf{d}) = 4$? As we can see from Table 2, there are only 3 ways to obtain $T \geq 4$ so a 1-sided p-value is $\Pr(T \geq 4 \mid \mathcal{F}, \mathcal{W}) = \frac{3}{8}$. Because the distribution of Wilcoxin's signed rank statistic is symmetric under the null hypothesis we can multiply this by 2 to get a 2 sided p-value[1]. There is a function in base `R` that will compute these p-values for us,

```
# One sided p-value:
wilcox.test(d, alternative = "g")
```

```
##
##  Wilcoxon signed rank test
##
## data:  d
## V = 4, p-value = 0.375
## alternative hypothesis: true location is greater than 0
```

```
# Two sided p-value
wilcox.test(d)
```

```
##
##  Wilcoxon signed rank test
##
## data:  d
## V = 4, p-value = 0.75
## alternative hypothesis: true location is not equal to 0
```

```
# Alternatively:
wilcox.test(Y1, Y2, paired = TRUE)
```

```
##
##  Wilcoxon signed rank test
##
## data:  Y1 and Y2
## V = 4, p-value = 0.75
## alternative hypothesis: true location shift is not equal to 0
```

Note from Table 2 that the only thing changing under the null are the signs of the d_s. The ranks are fixed because $q_s = |d_s|$ does not change. As we saw from other examples of randomization inference, the sole source of randomness here is the treatment assignment! So the value of the test statistic changes because of the S signs. These S signs are independent. For a true matched pairs experiment we would flip a fair coin and assign one and only unit to treament in each pair. If we are living in this world, we can write the expectation and variance of the null distribution of the test statistic conditional on $\mathcal{F}$ and $\mathcal{W}$,

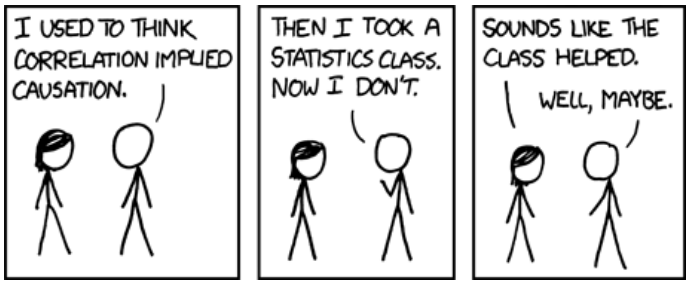
[1] A generic recipe would be to compute both 1 sided p-values, p^U and p^L, and report $\min\left[(2 \cdot \min\{p^U, p^L\}),\ 1\right]$.

$$\mathbb{E}[\ T \mid \mathcal{F}, \mathcal{W}\ ] = (1/2) \sum_{s}^{S} \text{sgn}(|d_s|) \cdot q_s$$

$$\text{Var}(\ T \mid \mathcal{F}, \mathcal{W}\ ) = (1/2)^2 \sum_{s}^{S} [\text{sgn}(|d_s|) \cdot q_s]^2$$

Sensitivity Analysis for Matched Pairs

Sensitivity analysis is a rhetorical device that researchers use to respond to the statement "correlation isn't causation". It was first used as a response to the small but politically powerful group of researchers (including Ronald Fisher) who denied a causal relationship between smoking and lung cancer[2]. The general intuition for the sensitivity analysis is that a researcher who conducts an observational study and wishes to make causal claims responds to critics by formalizing how much "confounding" there would need to be in order to drive the p-value for the obtained result below 0.05.



Source: https://xkcd.com/552/

On one extreme, the result from our observational study is equivalent to a randomized experiment. This is called the "naive model". At the other extreme, we can learn nothing about the causal relationship of interest from the observational study. This is the "association is not causation" model. The truth is definitionally bounded by these extremes.

In Rosenbaum's setup for the matched pairs design we are in a world where we perform 1:1 matching without replacement to form S pairs of units based on their observed covariates. At one end of the spectrum is what I'm going to call the Nihilistic Model. It can be formalized as,

$$\pi_i = \Pr(W_i = 1 \mid Y(1)_i, Y(0)_i, \mathbf{x}_i, \mathbf{u}_i)$$

so that the "assignment mechanism" is

$$\Pr(W_1 = w_1, \ldots, W_N = w_N \mid Y(1)_1, Y(0)_1, \mathbf{x}_1, \mathbf{u}_1, \ldots, Y(1)_N, Y(0)_N, \mathbf{x}_N, \mathbf{u}_N) = \prod_{i=1}^{N} \pi_i^{w_i} (1 - \pi_i)^{1 - w_i}$$

This model is always true, since there is always some unobserved $\mathbf{u}_i$ that makes it true. A sufficiently large amount of "counfounding" can explain away any presumed causal effect.

At the other extreme, the Naive Model says that among units that look similar on the observed covariates, treatment assignment is random. This involves several (familiar) assumptions. First,

[2] see https://academic.oup.com/ije/article/38/5/1175/666926

$$\pi_i = \Pr\left(W_i = 1 \mid Y(1)_i, Y(0)_i, \mathbf{x}_i, \mathbf{u}_i\right) = \Pr(W_i = 1|\mathbf{x}_i)$$

where π_i denotes the (unknown) unit level treatment probability. In words, treatment assignment does not depend on the potential outcomes or unobserved covariates. This is also called "strongly ignorable treatment assignment" or someone might say that the assignment mechanism is "unconfounded".

Second, everyone has a chance of receiving treatment or control, e.g. $0 < \pi_i < 1 \; \forall i$. We might also say that the assignment mechanism is "probabilistic". As we have seen before, an assignment mechanism that is both uncounfounded and probabilistic can be formalized as,

$$\Pr(W_1 = w_1, \ldots, W_N = w_N \mid Y(1)_1, Y(0)_1, \mathbf{x}_1, \mathbf{u}_1, \ldots, Y(1)_N, Y(0)_N, \mathbf{x}_N, \mathbf{u}_N) = \prod_{i=1}^{N} \pi_i^{w_i}(1-\pi_i)^{1-w_i}$$

The Naive Model is, of course, true if we have a randomized controlled trial where units are assigned treatment by the flip of a coin. If this model is true we can simply match on the observed covariates or the propensity score $e(\mathbf{x})$ and $\Pr(W = 1|\mathcal{F}) = \Pr(W = 1|e(\mathbf{x})) = \Pr(W = 1|\mathbf{x})$. As with a randomized experiment, any differences in outcomes due to unobserved "confounders" $\mathbf{u}_i$ are balanced across treatment and control groups under the Naive Model. The Naive Model is, of course, wrong.

Under the Naive Model two units i and j with the same observed covariates $\mathbf{x}_i$, $\mathbf{x}_j$ have the same unit level treatment probabilities: $\pi_i = \Pr(W_i = 1 \mid Y(1)_i, Y(0)_i), \mathbf{x}_i, \mathbf{u}_i) = \Pr(W_j = 1 \mid Y(1)_j, Y(0)_j), \mathbf{x}_j, \mathbf{u}_j)$. The Sensitivity Model paramaterizes the degree to which the Naive Model is false with $\Gamma \geq 1$ such that the "odds of treatment" ratio for two units with the same observed covariates is bounded according to,

$$\frac{1}{\Gamma} \leq \frac{\pi_i/(1-\pi_i)}{\pi_j/(1-\pi_j)} \leq \Gamma$$

If the Naive Model is true then $\pi_i = \pi_k$ and $\Gamma = 1$. If the Nihilistic Model is true then $\Gamma = \infty$. For pairs, the probability that one is in treatment and the other is in control is,

$$\Pr\left(W_i = 1, Z_j = 0 | \;\; Y(1)_i, Y(0)_i), \mathbf{x}_i, \mathbf{u}_i, Y(1)_j, Y(0)_j), \mathbf{x}_j, \mathbf{u}_j, W_i + W_j = 1\right) = \frac{\pi_i}{\pi_i + \pi_j}$$

The Sensitivity Model for S matched pairs can be written as,

$$\frac{1}{1+\Gamma} \leq \frac{\pi_{s1}}{\pi_{s1} + \pi_{s2}} \leq \frac{\Gamma}{1+\Gamma} \text{ for } s = 1, \ldots, S.$$

where $W_{s1} = 1 - W_{s2}$ are mutually independent treatment assignments. Again $\Gamma = 1$ is the Naive Model where the treatment probabilities are all equal to $1/2$ and $\Gamma = \infty$ is the Nihlistic Model. The suggested way to interpret Γ is that larger values correspond to more bias due to failure to account for $\mathbf{u}$. It is difficult to know what values of Γ are "large". A rule of thumb is that 6 is very large. If, for example, $\Gamma = 6$ then two matched units which are identical on observed covariates may differ by a factor of 6 in their (unobserved) "treatment odds" due to differences in their (unobserved) $\mathbf{u}$'s. A sensitivity analysis displays how inferences change as Γ increases.

In a sensitivity analysis for matched pairs, the null distribution for the Wilcoxon signed rank test is unknown but bounded for each fixed Γ. The upper bound T^+ is the sum of S independent random variables that take the value s with probability $p_s^+ = \Gamma/(1+\Gamma)$ and 0 with probability $p_s^- = 1/(1+\Gamma)$. The lower bound T^- is the sum of S independent random variables that take the value s with probability p_s^- or the value 0 with probability p_s^+. The expectations,

$$\mathbb{E}[\; T^+ \mid \mathcal{F}, \mathcal{W} \;] = p_s^+ \sum_{s=1}^{S} \text{sgn}(|d_s|) \cdot q_s$$

and

$$\mathbb{E}[\ T^- \mid \mathcal{F}, \mathcal{W}\ ] = p_s^- \sum_{s=1}^{S} \text{sgn}(|d_s|) \cdot q_s$$

Are equivalent to what we saw in the toy example when $\Gamma = 1$ since this implies $p_s^+ = p_s^- = 1/2$. The variances

$$\text{Var}(T^+ \mid \mathcal{F}, \mathcal{W}) = \text{Var}(T^- \mid \mathcal{F}, \mathcal{W}) = \frac{\Gamma}{(1+\Gamma)^2} \sum_{s=1}^{S} [\text{sgn}(|d_s|) \cdot q_s]^2$$

also reduce to what we saw earlier when $\Gamma = 1$. Although it was easy to do an exact computation in the toy example by hand, this becomes infeasible when S is large. Lucky for us there is an approximation. When $S \to \infty$ the null distribution of $\frac{(T - E[T])}{\sqrt{Var(T)}}$ converges to the standard Normal. Therefore we can use the Normal CDF to get approximate upper and lower bounds on the p-values,

$$1 - \Phi\left(\frac{(T^+ - \mathbb{E}[\ T^+ \mid \mathcal{F}, \mathcal{W}\ ])}{\sqrt{\text{Var}(T^+ \mid \mathcal{F}, \mathcal{W})}}\right)$$

$$1 - \Phi\left(\frac{(T^- - \mathbb{E}[\ T^- \mid \mathcal{F}, \mathcal{W}\ ])}{\sqrt{\text{Var}(T^- \mid \mathcal{F}, \mathcal{W})}}\right)$$

Example: Welding Study

The example used by Rosenbaum in Design of Observational Studies comes from a study of 39 welders who were matched with non-welders on sex (all males), smoking history and age. The authors of the original study wanted to see if the welding fumes that welders were exposed to damaged their DNA. The outcome of interest was the difference in "DNA elution rates" between the welders and non-welders.

Testing the null hypothesis of no treatment effect is similar to what we did with the toy example. Here, however, I'm going to illustrate the Normal approximation approach rather than use an exact test. Let's start by assuming the Naive Model where $\Gamma = 1$.

First, we calculate d_s,

```
load("werfel.RData")

# Calculate the differences in DNA elution rates across s matched pairs.
d <- (werfel$serpc_p - werfel$cerpc_p)
```

We can use our function from before to calculate the observed test statistic,

```
T_obs <- T_Wilcox(d)$T_obs
T_obs
```

```
## [1] 715
```

Then we can calculate the upper and lower bounds on p-values for $Gamma = 1$,

```
G <- 1
Pplus <- G/(1+G)
Pplus
```

```
## [1] 0.5
```

```
Pminus <- 1/(1+G)
Pminus
```

```
## [1] 0.5
```

```
# Expectation of T+
Eplus <- Pplus*sum(q*sgn(abs(d)))
Eplus
```

```
## [1] 39
```

```
# Expectation of T-
Eminus <- Pminus*sum(q*sgn(abs(d)))
Eminus
```

```
## [1] 39
```

```
# Variance
V <- G/(1+G)^2*sum( (sgn(abs(d))*q)^2)
V
```

```
## [1] 45.5
```

```
# The minimum and maximum p-values for a one-sided, upper tailed test
c(1-pnorm((T_obs-Eminus)/sqrt(V)), 1-pnorm((T_obs-Eplus)/sqrt(V)))
```

```
## [1] 0 0
```

O

As expected, when $\Gamma = 1$ the minimum and maximum p-values are equivalent (essentially zero here) and this reduces to the approach we used on the toy example with three matched pairs, the data are the only difference here. As we relax the assumption of random assignment, these p-values will no longer be equivalent. Let's wrap this process into a function

```
WilcoxApprox <- function(d, G, r = 3){

  T_obs <- T_Wilcox(d)$T_obs

  q <- T_Wilcox(d)$q

  Pplus <- G/(1+G)
  Pminus <- 1/(1+G)

  Eplus <- Pplus*sum(sgn(abs(d))*q)
  Eminus <- Pminus*sum(sgn(abs(d))*q)

  V <- (G/(1+G)^2)*sum( (sgn(abs(d))*q)^2 )

  return(list(Gamma = G,
              PMin = round(1-pnorm((T_obs-Eminus)/sqrt(V)), r),
              PMax = round(1-pnorm((T_obs-Eplus)/sqrt(V)), r))
  )

}
```

Let's implement the method for $\Gamma = 1, \ldots, 6$,

```
VF <- Vectorize(WilcoxApprox, "G")
data.frame(t(VF(d = d, G =  1:6)))
```

```
##   Gamma PMin  PMax
## 1     1    0     0
## 2     2    0 0.002
## 3     3    0 0.018
## 4     4    0 0.056
## 5     5    0 0.112
## 6     6    0 0.177
```

This suggests that the result obtained from the study – exposure to welding fumes caused DNA damage – is only "insignificant" under the assumption that there is a "large" amount of "confounding". This reproduces the results reported in Table 3.2 in Design of Observational Studies using approximate rather than exact p-values, but the differences are trivial even with $S = 39$ pairs.

There are many other approaches to sensitivity analysis. We have just covered perhaps the easiest one to present for introductory purposes. Rosenbaum discusses many other approaches in the books. There is a package called `rbounds` that implements some of the different approaches in *Design of Observational Studies*. "Sensitivity to Exogeneity Assumptions in Program Evaluation" by Guido Imbens (2003) outlines a somewhat more general structural equation modelling type approach[3]. A related approach that has become popular in political science is presented in "A Selection Bias Approach to Sensitivity Analysis for Causal Effects" by Matthew Blackwell (2014) with an accompanying `R` package called `causalsens`[4].

[3] https://scholar.harvard.edu/imbens/files/sensitivity_to_exogeneity_assumptions_in_program_evaluation.pdf
[4] http://www.mattblackwell.org/files/papers/sens.pdf