Section 3 : Permutation Inference

Andrew Bertoli

18 September 2013

Roadmap

1. Questions from Last Class
2. Permutation Inference
3. Types of Tests
4. Propensity Score Matching
5. Homework Questions

Questions from Last Class

Multiple or Multivariate Regression?

Most people use multiple regression to describe a model that has one outcome and multiple controls. However, not everyone follows this convention.

Questions from Last Class

A researcher wants to determine if there is discrimination against women in the workforce. He uses the model

$$\text{Salary} = a + b \cdot \text{Education} + c \cdot \text{Experience} + d \cdot \text{Man} + \epsilon$$

Where man is a dummy variable that is coded 1 if the person is a man and 0 if the person is a woman.

After getting the data, he estimates the coefficients as follows.

$$\text{Salary} = \$7,100 + \$1,300 \cdot \text{Edu} + \$2,200 \cdot \text{Exp} + \$700 \cdot \text{Man} + \epsilon$$

To test for discrimination, the researchers checks if the coefficient on Man is statistically significant.

Questions from Last Class

Violations of the assumption that the ϵ are i.i.d. $N(0, \sigma^2)$

According to Freedman, "the error term represents the combined effect of the omitted variables, assuming that

(i) the combined effect of the omitted variables is independent of each variable included in the equation

(ii) the combined effect of the omitted variables is independent across subjects

(iii) the combined effect of the omitted variables has expectation 0."

Permutation Inference

Imagine that we run an experiment. We will have a vector of outcomes

$\mathbf{Y} = (0, 3, -2, ..., 4)$

and a vector for our treatment assignment

$\mathbf{T} = (0, 1, 1, ..., 0)$

We will then compute our test statistic $\hat{\tau} = \bar{T} - \bar{C}$.

Under the sharp null hypothesis of no treatment effect, $E[\hat{\tau}] = 0$. Of course, even if the null is true, we are unlikely to see 0 because $\hat{\tau}$ is a random variable with a variance.

So how do we know if our test statistic is far enough from 0 to reject the null?

Permutation Inference

If the treatment had no effect, the order of 0's and 1's in the vector should have no impact on the outcome.

In other words, if we shuffle the 0's and 1's in $\mathbf{T}$ into a random order and compute a new test statistic, it should be about as far from 0 as our real test statistic $\hat{\tau}$.

So if we compute a large number of new test statistics, our real test statistic should be similar to the new statistics.

However, if the real test statistic is extreme compared to the new test statistics, we should reject the null.

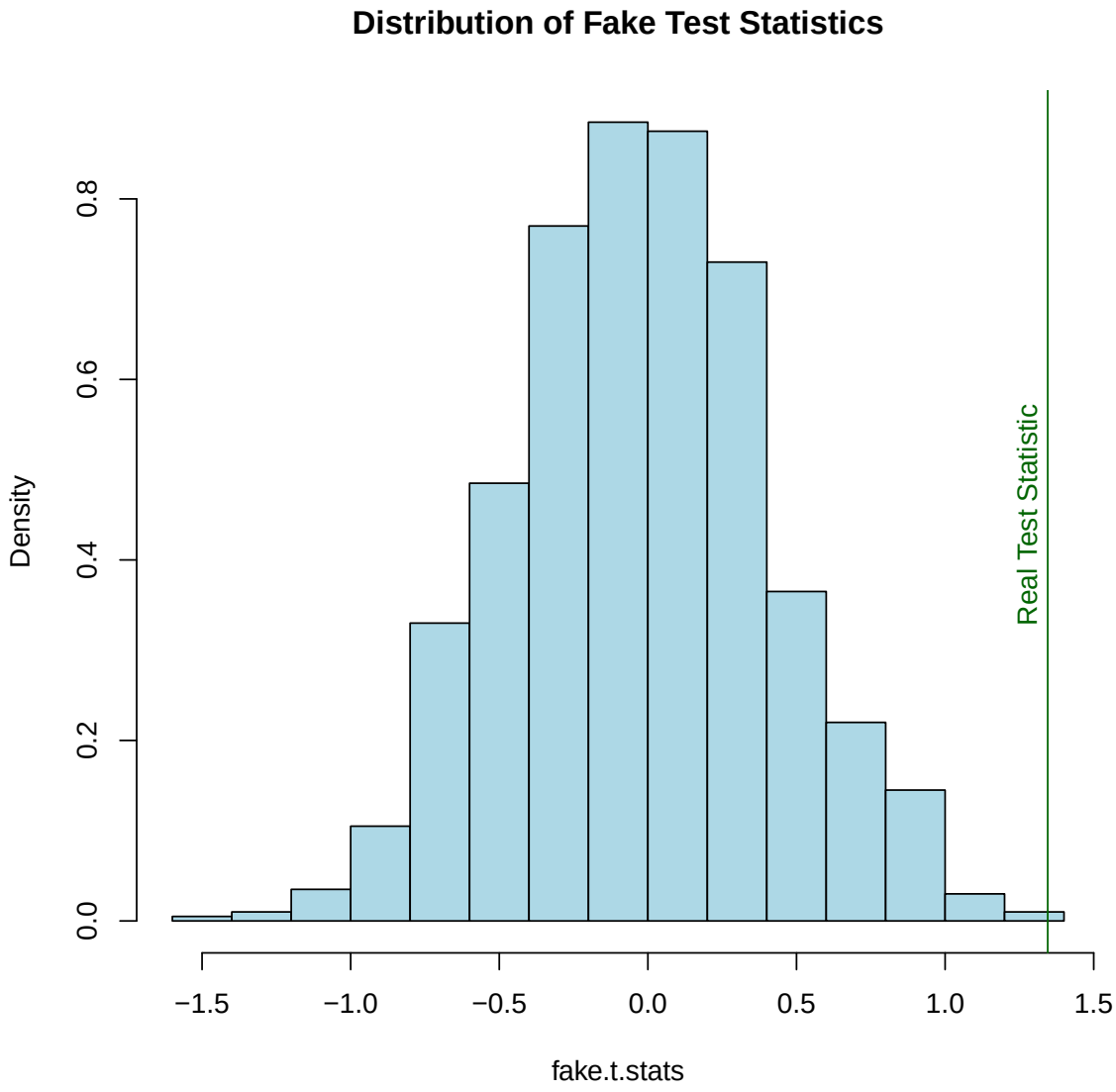
Permutation Inference

```
> t=rnorm(50,1,2)
> c=rnorm(50,0,2)
> real.t.stat=mean(t)-mean(c)
> outcomes=c(t,c)
```

Permutation Inference

```
> fake.t.stats=rep(0,1000)
> for(i in 1:1000){
+ treatmentassignment=sample(c(rep(0,50),rep(1,50)),100,
replace=FALSE)
+ fake.t=outcomes[treatmentassignment==1]
+ fake.c=outcomes[treatmentassignment==0]
+ fake.t.stats[i]=mean(fake.t)-mean(fake.c)
+ }
> pvalue=length(which(abs(fake.t.stats)>=real.t.stat))/
length(fake.t.stats)
> pvalue
[1] 0.002
```

Permutation Inference



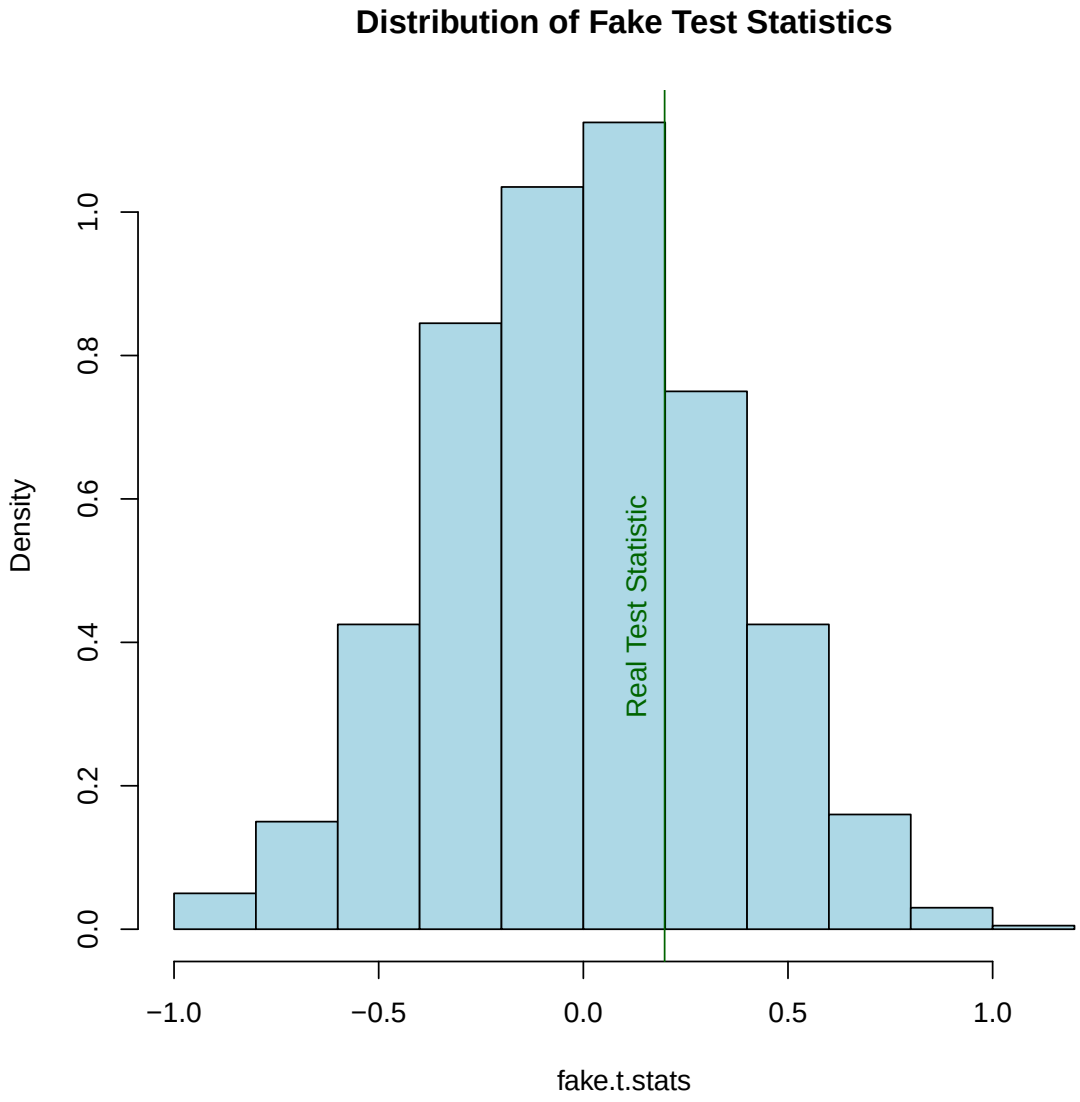
Permutation Inference

```
> t=rnorm(50,0,2) # No Treatment Effect Here
> c=rnorm(50,0,2)
> real.t.stat=mean(t)-mean(c)
> outcomes=c(t,c)
```

Permutation Inference

```
> fake.t.stats=rep(0,1000)
> for(i in 1:1000){
+ treatmentassignment=sample(c(rep(0,50),rep(1,50)),100,
replace=FALSE)
+ fake.t=outcomes[treatmentassignment==1]
+ fake.c=outcomes[treatmentassignment==0]
+ fake.t.stats[i]=mean(fake.t)-mean(fake.c)
+ }
> pvalue=length(which(abs(fake.t.stats)>=real.t.stat))/
length(fake.t.stats)
> pvalue
[1] 0.572
```

Permutation Inference



Permutation Inference

Permutation Inference p-value

$p = 0.014$

t-test p-value

$p = 0.009519$

Types of Tests

Distinguishing between Null Hypotheses

Sharp Null: There is no treatment effect whatsoever.

Tested with Permutation Inference (a.k.a. Randomization Inference)

Regular Null: There is no average treatment effect.

Tested with a t-test.

Types of Tests

Sharp Null Advantages

1. Provides an exact p-value without having to estimate the standard error

2. Does not require any assumptions (including SUTVA, since it is implied under the null)

3. Applicable in more types of experiments (for instance, when we have unusual blocking schemes)

Types of Tests

Sharp Null Disadvantages

1. Less easy to interpret than a test of the average treatment effect.

Rosenbaum: "No effect means no effect. A nonzero effect that varies from one unit to the next and that is hard to fathom or predict is, nonetheless, a non-zero effect. It may not be an immediately useful effect, but it is an effect, perhaps an effect that can someday be understood, tamed, and made useful."

2. Less well known

Types of Tests

Other Notable Tests-Wilcoxon Rank Sum Test

Step 1. Rank the outcomes from smallest to largest

Step 2. Calculate the sum of the ranks of the treated units

Step 3. Let n_t be the number of units in the treatment group, and let n_c be the number of units in the control group. Under the sharp null of no treatment effect, the sum should be distributed $N(\frac{n_t(n_t+n_c+1)}{2}, \frac{n_t n_c(n_t+n_c+1)}{12})$.

Notes: This test is insensitive to outliers. It tends to work better when there are not a lot of ties in the data.

Types of Tests

Other Notable Tests-Wilcoxon Signed Rank Test for Matched Pairs

Step 1. Drop all ties. Let n_r equal the number of remaining pairs.

Step 2. Rank all remaining pairs from smallest difference to largest difference

Step 3. Take the sum of the ranks of the pairs where the treated unit had the higher outcome. Then subtract the sum of the units where the control unit had the higher outcome. The test statistic W is the absolute value of this difference.

Step 4. To get the z-score, we use the formula $z = \frac{W-0.5}{\sigma_W}$, where $\sigma_W = \frac{n_r(n_r+1)(2n_r+1)}{6}$ (use a two-sided test here).

Notes: This tests differs slightly from the one presented in Rosenbaum, where he simply sums the ranks of the treated units. The procedure described here is more commonly used. Again, this test is insensitive to outliers and tends to work better when there are not a lot of ties in the data.

Types of Tests

R Code

```
> library(stats)

> wilcox.test(t,c)

Wilcoxon rank sum test with continuity correction
data: t and c
W = 1608, p-value = 0.01372
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(t, c, paired=TRUE)

Wilcoxon signed rank test with continuity correction
data: t and c
V = 947, p-value = 0.002856
alternative hypothesis: true location shift is not equal to 0
```

Propensity Score Matching

Question: Say we have two units. For each unit, we flip a coin and give treatment if the coin flip is heads. The coin happens to be biased and lands on heads 90% of the time. Imagine that exactly one of the two units got heads. What is the probability that it is Unit 1 vs. Unit 2.

Answer: There is a 50% for each unit, since they each had a 90% chance of getting heads.

Propensity Score Matching

Now imagine that we are doing an observational study with n units.

Pretend that we know the real the propensity score (the probability of treatment) for each unit i.

$\pi_i = P(T_i = 1)$

Then for each treated unit, we could find a control unit with the same propensity score.

It would be random which of the two units got treated, since they had the same probability of being assigned to treatment.

Propensity Score Matching

Problem

This method works if you know the true propensity score. But in most situations, the true propensity score is unknown, and researchers try to estimate it from the covariates.

Propensity Score Matching

Example

Question: Does eating fast food every day cause heart disease?

Imagine that the probability that a person eats fast food every day is determined by their age, gender, stress level, and whether their parents eat fast food. We observe everything but stress levels.

Assume the following relationships hold.

$$\pi_i = \frac{1}{300} \cdot Age_i + 0.1 \cdot Gender_i + 0.2 \cdot Parents.Eaters_i + 0.3 \cdot Stress_i$$

$$\text{P(H. Disease)} = 0.3 \cdot Treat + \frac{1}{400} \cdot Age + 0.05 \cdot Gender + 0.1 \cdot Stress$$

Propensity Score Matching

We would have a data frame that looked like this

	Age	Gender	Parents.Eaters	Treat	Heart.Disease
1	90	1	0	1	0
2	90	1	1	1	1
3	39	1	0	0	0
4	71	1	1	0	0
5	41	0	0	0	0
6	52	0	1	0	0
7	65	1	0	0	1
8	80	1	1	0	1
9	64	1	0	0	0
10	55	1	1	1	1
11	42	0	0	0	0
12	35	1	1	1	1
13	38	0	0	1	0
14	39	0	1	0	0
15	81	1	0	0	0
16	58	0	1	1	0

Propensity Score Matching

Next, we want to estimate the probability of being treated based on the covariates that we have data for. We use the code

```
> pscore=glm(Treat  Age + Gender + Parents.Eaters, family=
binomial(link=logit),data=data)$fitted.values
```

Propensity Score Matching

Now for each treated unit, we will find the closest control unit to use as a match. We will do our matching with replacement.

```
> t=data[data$Treat==1,]
> c=data[data$Treat==0,]
> Controls=rep(0,length(t$pscore))
> for(i in 1:length(t$pscore)){
+ Controls[i]=which.min(abs(c$pscore-t$pscore[i]))
+ }
```

Propensity Score Matching

Lastly, we can do a t-test using the treated units and the controls that we matched them to.

>t.test(t$Heart.Disease,c$Heart.Disease[Controls],paired=TRUE)

t = 3.1311, df = 44, p-value = 0.003024

We can also verify that the treated units and matched controls are balanced on all the covariates.

>t.test(tAge,cAge[Controls],paired=TRUE)

t = 0.341, df = 44, p-value = 0.7347

>t.test(t$Gender,c$Gender[Controls],paired=TRUE)

t = 0, df = 44, p-value = 1

>t.test(t$Parents.Eaters,c$Parents.Eaters[Controls],paired=TRUE)

t = -0.4958, df = 44, p-value = 0.6225

Propensity Score Matching

Question 1: What parameter were we estimating in this example?

Question 2: How does this example relate to Jas's discussion about reweighting observations in an experiment that has different sized treatment and control groups.