Microeconometrics (Causal Inference)
Week 4 - Introduction to causality

Joshua D. Merfeld
KDI School of Public Policy and Management

2023-09-18

What are we doing today?

- This week we will discuss causation
- For example, what prevents us from saying that something is a causal effect?
- Then:
  - How do RCTs help?
  - Propensity scores

What are we doing today?

or. . .

or. . .

Correlation is not (necessarily) causation

Let me start with an example

- I used to live in Atlanta
  - There was a large hospital in downtown: Grady

- Grady was a trauma center
  - It was the only one in the immediate area

Let me start with an example

- Something I once heard: Mortality rates at Grady are so much higher than at other hospitals. Is it a bad hospital?

- What do you think? Is Grady necessarily a worse hospital?

- Let's introduce one of the most common ways to think about causality (in economics): the potential outcomes framework
  - This framework is also known as the Rubin Causal Model

- The potential outcomes framework is a way to think about causality
  - It is not the only way
  - It is not necessarily the best way
  - But it's what we're going to use

- Suppose someone is in a car accident
  - They could be taken to Grady or another hospital
- Grady is the treatment:
  - $D = 1$ if they go to Grady
  - $D = 0$ if they go to another hospital
- We can think of the outcome of interest as the person's health
  - We can think of this as a *potential outcome*
  - We can think of the person's health if they go to Grady as Y_1
  - We can think of the person's health if they go to another hospital as Y_0

- Let's write this all out, imagining different possible people, indexed by i:

$$\text{Potential outcomes} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} \tag{1}$$

- Y_{1i} is the outcome for person i if they go to Grady
- Y_{0i} is the outcome for *the same person at the same time* if they go to another hospital

$$\text{Potential outcomes} = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases} \tag{2}$$

- Y_{1i} is the outcome for person i if they go to Grady
- Y_{0i} is the outcome for *the same person at the same time* if they go to another hospital

- What's the problem we have?

What's the problem?

We never observe the same person at the same time in two different states of nature (i.e. going to Grady AND going to another hospital at the same time)

Let's just compare those who go to Grady with those who don't?

- Actual causal effect:

$$\mathbb{E}(Y_{1i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0)$$

- Comparing across individuals:

$$\mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0)$$

- Note that this is *not* the same thing. We are comparing different people.

- We can break down this comparison into two separate terms:

$$\mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0) =$$
$$[\mathbb{E}(Y_{1i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 1)] \quad (3)$$
$$+ [\mathbb{E}(Y_{0i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0)] \quad (4)$$

Let's just compare those who go to Grady with those who don't?

$$\mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0) =$$
$$[\mathbb{E}(Y_{1i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 1)] \quad (3)$$
$$+ [\mathbb{E}(Y_{0i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0)] \quad (4)$$

- Line 3 is the causal effect we're interested in:

$$\mathbb{E}(Y_{1i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 1) = \mathbb{E}(Y_{1i} - Y_{0i}|D_i = 1)$$

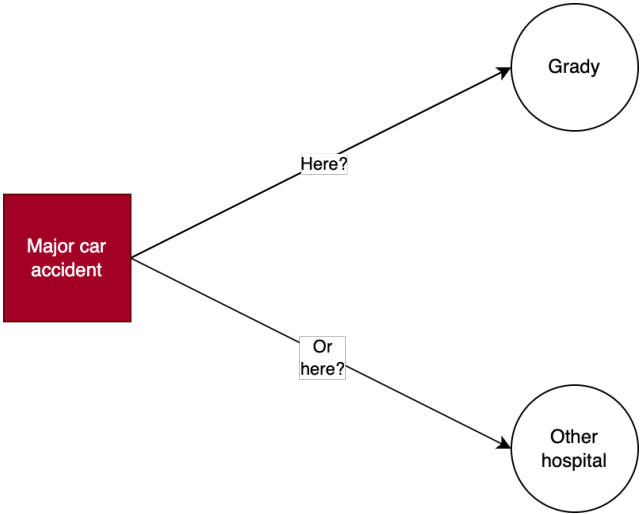
  - This is the *treatment effect of going to Grady on those who went to Grady*
  - Their actual outcome minus their counterfactual outcome

Let's just compare those who go to Grady with those who don't?

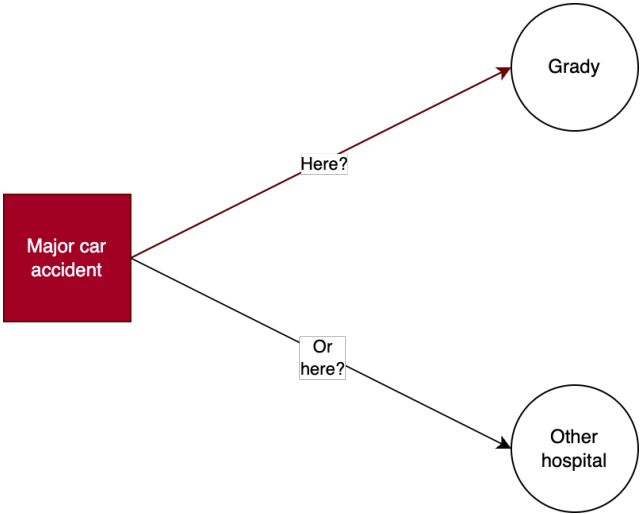
$$\mathbb{E}(Y_i|D_i = 1) - \mathbb{E}(Y_i|D_i = 0) =$$
$$[\mathbb{E}(Y_{1i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 1)] \quad (3)$$
$$+ [\mathbb{E}(Y_{0i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0)] \quad (4)$$

- Line 4 is the real problem. This is *selection bias*.
  - This is the difference between the potential outcomes for those who went to Grady and those who didn't.
  - The question: What might be the *difference* between people who go to Grady and people who go to another hospital?
  - If you ever hear an economist say "selection", this is what they're talking about: the systematic differences between the two groups.

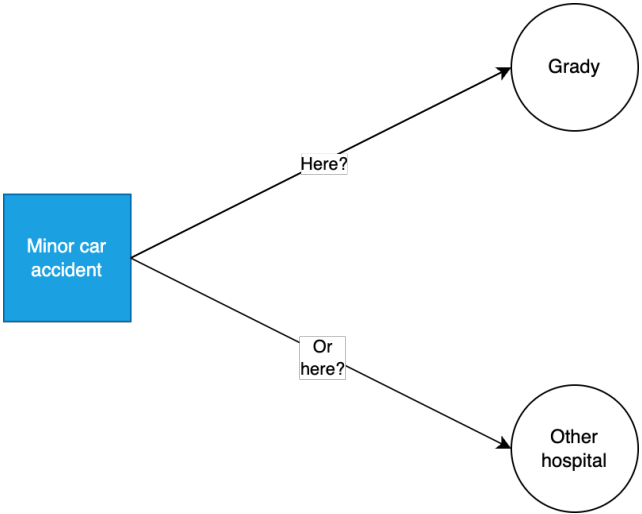
Who goes to Grady?



Who goes to Grady?



Who goes to Grady?



Who goes to Grady?

- So if people with much worse injuries are going to Grady, then we might expect that the people who go to Grady have worse outcomes...

- Even if Grady is better at treating bad injuries!

So what to do?

- The rest of this course is about ways to address this problem
- Let's start with the "gold standard": randomized controlled trials (RCTs)

- So how do RCTs help?
- RCTs can't help with the fact that we can never observe the same unit in both treatment states at the same time
- Instead, RCTs rely on groups

- Let's go back to our diabetes/retinopathy example from last week
- Why wouldn't we want to simply compare individuals who received a new laser eye treatment for retinopathy to people who don't?

- The key is random assignment
- Random assignment means that, *on average*, those who receive treatment and those who do not are the same!
- Our selection bias equation is:

$$\mathbb{E}(Y_{0i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0) \tag{5}$$

- Guess what this equals when we have random assignment?

- Mathematically, random assignment means that potential outcomes are *independent* of treatment assignment:

$$\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp D_i \tag{6}$$

- This allows us to do the following:

$$\begin{aligned} \text{selection} &= \mathbb{E}(Y_{0i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 0) && (7) \\ &= \mathbb{E}(Y_{0i}|D_i = 1) - \mathbb{E}(Y_{0i}|D_i = 1) && (8) \end{aligned}$$

- The last line is zero! In expectation, there is no selection bias when treatment is randomized.

- We're going to use a new dataset to illustrate the analysis of RCTs.
  - We need to download the package from GitHub!

```
install.packages("remotes") # install package remotes
remotes::install_github("higgi13425/medicaldata") # download datasets from github repo
# repo here: https://higgi13425.github.io/medicaldata/
library(medicaldata) # load datasets
```

Licorice Gargle

Dataset Introduction

Abstract

This study enrolled 236 adult patients undergoing elective thoracic surgery requiring a double-lumen endotracheal tube. Gender, physical status, BMI, age, Mallampati score, smoking status, preoperative pain, surgery size, intervention and the outcomes (cough, sore throat and pain swallowing at various time points) are provided. The dataset is cleaned and complete (missing outcomes for 2 patients). There are no outliers or data problems. These are data from a study by Ruetzler et al. "A Randomized, Double-Blind Comparison of Licorice Versus Sugar-Water Gargle for Prevention of Postoperative Sore Throat and Postextubation Coughing". *Anesth Analg* 2013; 117: 614 - 21.

```
colnames(licorice_gargle)
```

```
##  [1] "preOp_gender"           "preOp_asa"              "preOp_calcBMI"
##  [4] "preOp_age"              "preOp_mallampati"       "preOp_smoking"
##  [7] "preOp_pain"             "treat"                  "intraOp_surgerySize"
## [10] "extubation_cough"       "pacu30min_cough"        "pacu30min_throatPain"
## [13] "pacu30min_swallowPain"  "pacu90min_cough"        "pacu90min_throatPain"
## [16] "postOp4hour_cough"      "postOp4hour_throatPain" "pod1am_cough"
## [19] "pod1am_throatPain"
```

- Analyzing RCTs with simple randomization is easy!

- We can simply compare the average outcome for those who received treatment to the average outcome for those who did not:

$$y_i = \beta_0 + \beta_1 D_i + \epsilon_i,$$

where y_i is the outcome of interest and D_i is a dummy variable for treatment.

Effect of licorice gargle on sore throat

```
licresults <- feols(pacu30min_throatPain ~ treat, data = licorice_gargle)
summary(licresults)
```

```
## OLS estimation, Dep. Var.: pacu30min_throatPain
## Observations: 233
## Standard-errors: IID
##              Estimate Std. Error  t value   Pr(>|t|)
## (Intercept)  1.025862   0.110666  9.26988  < 2.2e-16 ***
## treat       -0.752358   0.156171 -4.81753 2.6317e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.18678   Adj. R2: 0.087364
```

```
etable(licresults, licresults,
       vcov = list("iid", "HC1"),
       digits = 3,
       headers = c("sore throat", "sore throat"),
       depvar = FALSE)
```

```
##                     licresults      licresults.1
##                    sore throat       sore throat
##
## Constant       1.03*** (0.111)   1.03*** (0.144)
## treat        -0.752*** (0.156) -0.752*** (0.157)
## ------------ ----------------- -----------------
## S.E. type                  IID Heteroskeda.-rob.
## Observations               233               233
## R2                     0.09130           0.09130
## Adj. R2                0.08736           0.08736
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Make it a pretty table

```
etable(licresults, licresults,
       vcov = list("iid", "HC1"),
       digits = 3,
       se.below = TRUE, # add SEs BELOW coefficients (the norm)
       depvar = FALSE)
```

```
##              licresu.. licresu...1
## Constant       1.03***     1.03***
##               (0.111)     (0.144)
## treat         -0.752***   -0.752***
##               (0.156)     (0.157)
## ------------ ---------   ---------
## S.E. type           IID   Het.-rob.
## Observations        233         233
## R2              0.09130     0.09130
## Adj. R2         0.08736     0.08736
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Even nicer

```
table <- etable(licresults, licresults,
          vcov = list("iid", "HC1"),
          digits = 3, digits.stats = 3,
          se.below = TRUE, # add SEs BELOW coefficients (the norm)
          depvar = FALSE) # don't list dep var at top
table <- table[-c(5:6, 9:11), ]
colnames(table) <- c("", "(1)", "(2)")
kable(table,
          caption = "Effect of licorice gargle on sore throat",
          booktabs = TRUE, align = "lcc", linesep = "", escape = FALSE, row.names = FALSE) %>%
          footnote("* p<0.1 ** p<0.05 *** p<0.01", general_title = "", footnote_as_chunk = TRUE) %>%
          row_spec(4, hline_after = TRUE) %>%
          kable_classic_2()
```

Table 1: Effect of licorice gargle on sore throat

	(1)	(2)
Constant	1.03***	1.03***
	(0.111)	(0.144)
treat	-0.752***	-0.752***
	(0.156)	(0.157)
Observations	233	233
R2	0.091	0.091

* p<0.1 ** p<0.05 *** p<0.01

Controlling for variables

- We often have baseline values of the outcome of interest
  - For example, in an intervention about income, we have a baseline measure of income
- It is quite common to use baseline values as controls in regressions
  - Why? It helps improve power!
- ANCOVA, but it's really just simple OLS regression

Some complications in analyzing RCTs

- What if we have multiple treatment groups?
- What if randomization is at an aggregate level?
- What if we have multiple outcomes?
- What if randomization depends on other variables?

Multiple treatment groups

- Suppose we have three groups:
  - $D_i = 0$ if control
  - $D_i = 1$ if treatment 1
  - $D_i = 2$ if treatment 2

- This is pretty common
  - Sometimes there are multiple treatments
  - Sometimes treatments are layered on top of one another

TABLE 1—INSURANCE INNOVATIONS

	Individual (P1)	Individual High (P2)	Household (P3)	Group (P4)
Eligibility	Individual	Individual	Household	Household
Add. requirement				50% uptake in the group
Coverage limit (pp)	15,000	30,000	15,000	15,000
Premium (pp)	100	150	100	100
Premium discounts (pp)	0–30	0–30	0–30	0–30

Notes: Numbers are in PKR; US$1 $\approx$ PKR 101; PKR 15,000 $\approx$ US$148 (in February 2015); pp $=$ per person. Individual eligibility: client allowed to insure any number and any combination of dependents. Household eligibility: client has to insure either all or none of the dependents. Premium discounts: discount vouchers of PKR 0, 10, 20, or 30 (pp) were randomized with equal probability at the household level.

Note that they randomize *villages*, not households

TABLE 2—TREATMENT ALLOCATION

	Control	Awareness	P1	P2	P3	P4	Total (policies)	Total
Villages	86	82	82	84	82	86	334	502
Groups	283	230	268	266	252	264	1,050	1,563
HHs	1,154	1,026	1,022	1,083	1,058	1,120	4,283	6,463
HHs attending	0	822	856	870	830	877	3,433	4,255
Dependents (dep.)	4,183	3,539	3,560	3,920	3,797	4,085	15,362	23,084
Attending dep.	0	2,798	2,981	3,209	2,938	3,156	12,284	15,082

Notes: Both the "Control" and "Awareness" groups are not used in this paper, except for the discussion of expected costs and moral hazard in online Appendix D.

- We can still use the same equation:

$$y_i = \beta_0 + \sum_{k=1}^{K} \beta_k D_k + \epsilon_i,$$

  where y_i is the outcome of interest and k indexes the different treatment groups.

- In the simplest case, we can simply compare the average outcome for each treatment group to the average outcome for the control group.

- We can also compare the average outcome for each treatment group to the average outcome for each other treatment group.
  - Note that we use an F-test for this: we are testing whether some of the β_k are equal to *one another*, which is a coefficient restriction.

Aggregate randomization

- We've talked about individual randomization
  - But in this example, randomization is at the village level
- So how do we take this into account?

Aggregate randomization

- We've talked about individual randomization
  - But in this example, randomization is at the village level
- So how do we take this into account?
- Cluster standard errors at the level of randomization!
  - In this case, cluster standard errors at the village level
  - If you randomize schools, you would cluster at the school level
  - Etc.

Aggregate randomization: clustered standard errors

TABLE 4—INSURANCE UPTAKE AND ENFORCEMENT OF ELIGIBILITY

	Individual (P1)		Household (P3)		Group (P4)	
	Dependents	HH	Dependents	HH	Dependents	HH
PKR 100	0.166	0.410	0.182	0.258	0.167	0.265
	(0.025)	(0.048)	(0.031)	(0.040)	(0.034)	(0.043)
	[577]	[166]	[660]	[186]	[648]	[189]
PKR 90	0.303	0.651	0.420	0.472	0.269	0.332
	(0.026)	(0.037)	(0.042)	(0.040)	(0.039)	(0.041)
	[792]	[235]	[752]	[214]	[897]	[247]
PKR 80	0.341	0.740	0.484	0.510	0.427	0.477
	(0.026)	(0.032)	(0.053)	(0.048)	(0.046)	(0.044)
	[785]	[227]	[741]	[208]	[850]	[239]
PKR 70	0.385	0.776	0.708	0.739	0.656	0.683
	(0.033)	(0.031)	(0.048)	(0.040)	(0.055)	(0.050)
	[827]	[228]	[784]	[222]	[761]	[202]
Observations	2,981	856	2,937	830	3,156	877
F-stat (P1 versus P*i*)			3.84	65.58	8.68	79.84

Notes: Standard errors in parentheses are clustered at the level of the village. Number of dependents (households) are in square brackets. *F*-stat is the *F*-statistics of a joint hypothesis test for equality of take-up rates between policy P1 and policies P3 and P4, respectively.

Note the use of the F-test here

TABLE 4—INSURANCE UPTAKE AND ENFORCEMENT OF ELIGIBILITY

	Individual (P1)		Household (P3)		Group (P4)	
	Dependents	HH	Dependents	HH	Dependents	HH
PKR 100	0.166	0.410	0.182	0.258	0.167	0.265
	(0.025)	(0.048)	(0.031)	(0.040)	(0.034)	(0.043)
	[577]	[166]	[660]	[186]	[648]	[189]
PKR 90	0.303	0.651	0.420	0.472	0.269	0.332
	(0.026)	(0.037)	(0.042)	(0.040)	(0.039)	(0.041)
	[792]	[235]	[752]	[214]	[897]	[247]
PKR 80	0.341	0.740	0.484	0.510	0.427	0.477
	(0.026)	(0.032)	(0.053)	(0.048)	(0.046)	(0.044)
	[785]	[227]	[741]	[208]	[850]	[239]
PKR 70	0.385	0.776	0.708	0.739	0.656	0.683
	(0.033)	(0.031)	(0.048)	(0.040)	(0.055)	(0.050)
	[827]	[228]	[784]	[222]	[761]	[202]
Observations	2,981	856	2,937	830	3,156	877
F-stat (P1 versus P*i*)			3.84	65.58	8.68	79.84

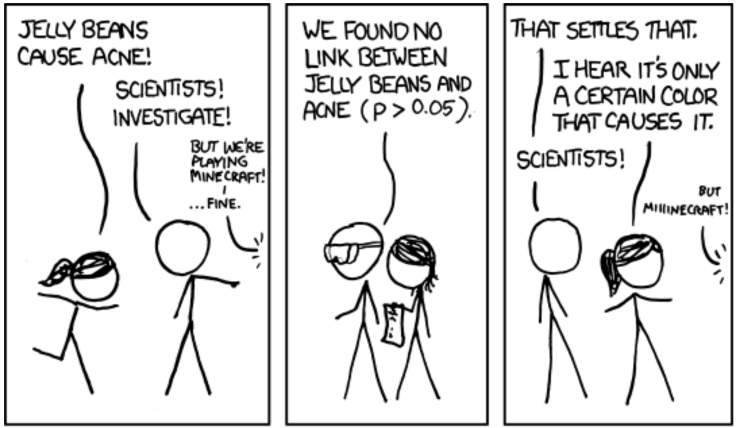
Notes: Standard errors in parentheses are clustered at the level of the village. Number of dependents (households) are in square brackets. *F*-stat is the *F*-statistics of a joint hypothesis test for equality of take-up rates between policy P1 and policies P3 and P4, respectively.

Complication two: multiple outcomes

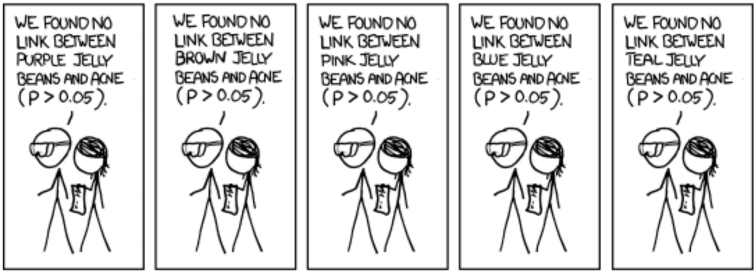
- Suppose we have many outcomes
  - For argument's sake, let's say we have 10 outcomes

- A single p-value can be misleading
  - If we have 10 outcomes and two treatments, we expect one of them to be significant at the 5% level by chance alone!
  - The more outcomes you have, the more likely you are to find significant effects, *even if there are none*!

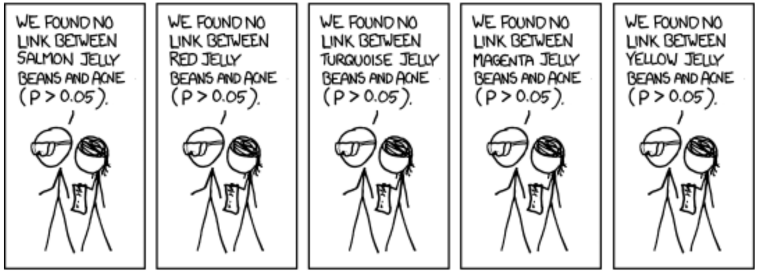
Relevant xkcd (882)



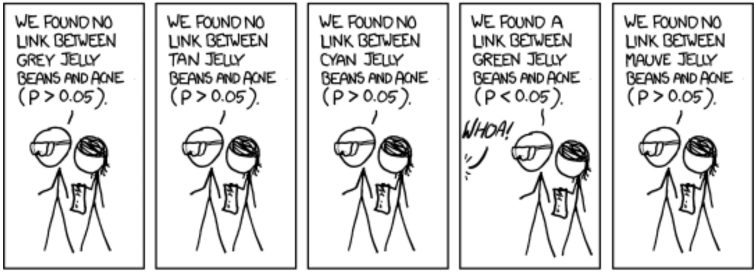
Relevant xkcd (882)



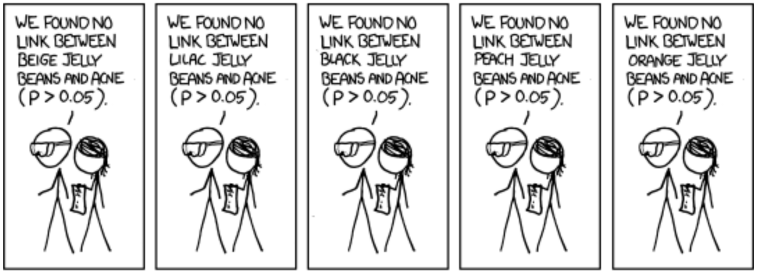
Relevant xkcd (882)



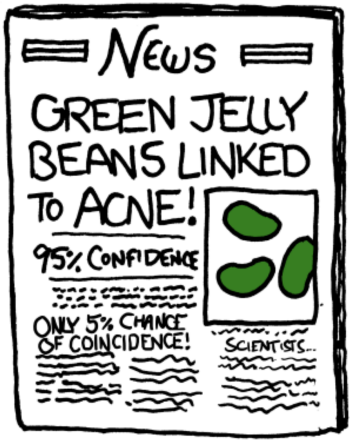
Relevant xkcd (882)



Relevant xkcd (882)



Relevant xkcd (882)



- Suppose you have one outcome
  - Suppose the null hypothesis is true
  - What is the probability of finding a significant effect at the 5% level? What is the probability of not finding a significant effect?

- Suppose you have one outcome
  - Suppose the null hypothesis is true
  - What is the probability of finding a significant effect at the 5% level? What is the probability of not finding a significant effect?
  - 5% and 95%, respectively

- Suppose you have 20 outcomes
  - Suppose all null hypotheses are true
  - What is the probability of finding NO significant effects, assuming independence across outcomes?

The problem, mathematically

- Suppose you have one outcome
  - Suppose the null hypothesis is true
  - What is the probability of finding a significant effect at the 5% level? What is the probability of not finding a significant effect?
  - 5% and 95%, respectively

- Suppose you have 20 outcomes
  - Suppose all null hypotheses are true
  - What is the probability of finding NO significant effects, assuming independence across outcomes?
  - $0.95^{20} = 0.358$

- Bonferroni correction
  - This is the simplest correction
  - If you have 10 outcomes, you multiply your p-value by 10
  - If you have 20 outcomes, you multiply your p-value by 20 (cap p-values at 1.000)

- This is very rare nowadays in applied microeconometrics
  - It's too conservative

Correcting for multiple outcomes

- Sharpened q-values (Benjamini, Krieger, and Yekutieli, 2006; Anderson, 2008)
  - R package `qvalue`
    - Just requires a vector of p-values
  - An advantage is their comparison to p-values
  - A disadvantage is that it does not take into account correlations across outcomes

- This is related to the False Discovery Rate (FDR)
  - Read more here: https://blogs.worldbank.org/impactevaluations/overview-multiple-hypothesis-testing-commands-stata

Table 4. Effects on preteen IQ scores

			Female					Male					Gender difference
Outcome	Age	Project	Effect	CM	Naive p value	FDR q value	n	Effect	CM	Naive p value	FDR q value	n	t statistic
IQ	5	ABC	4.94	96.76	.176	.304	48	10.19	90.81	.005	.082	47	−1.05
			(3.58)					(3.52)					
IQ	6.5	ABC	5.13	92.96	.134	.271	46	7.18	92.10	.053	.517	45	−.41
			(3.35)					(3.65)					
IQ	12	ABC	8.35	87.35	.004	.048	52	3.21	90.48	.294	1.000	49	1.24
			(2.75)					(3.10)					
IQ	5	Perry	12.67	81.65	.004	.048	39	10.61	84.79	.001	.049	54	.40
			(4.30)					(2.84)					
IQ	6	Perry	3.75	87.16	.241	.318	48	5.66	85.82	.037	.451	72	−.46
			(3.21)					(2.68)					
IQ	10	Perry	4.96	81.79	.173	.304	43	−2.33	86.03	.372	1.000	71	1.70
			(3.45)					(2.56)					
IQ	5	ETP	13.55	87.60	.015	.077	30	4.43	87.18	.232	1.000	34	1.28
			(6.09)					(3.75)					
IQ	7	ETP	8.61	89.89	.118	.271	29	4.11	92.89	.344	1.000	30	.57
			(6.69)					(4.25)					
IQ	10	ETP	9.79	81.56	.067	.216	29	−3.17	88.33	.511	1.000	27	1.68
			(5.73)					(5.15)					

NOTE: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. The p and q values are computed as described in Section 3; t statistics test the difference between female and male treatment effects.

- List et al. (2016) is a good example of how to deal with multiple outcomes
  - Their procedure allows for p-values to be correlated
    - This is about familywise error rates (FWER)
  - Much more conservative as you add outcomes, because it is about avoiding *any* false positives (type one errors)

- Unfortunately, I haven't found an R package for this yet, and we aren't going to do this by hand
  - In Stata: `ssc install mhtexp`

- For this class, you can use q-values

- Final complication: what if randomization depends on other variables?

- Stratification
  - People in different "strata" have differential probability of being treated
  - Example:
    - Probability of male-owned firm being selected is 0.5
    - Probability of female-owned firm being selected is 0.75
    - In this case, we can say that selection is stratified by gender

- How do we deal with stratification?
  - By including dummy variables for the strata!

$$y_i = \beta_0 + \beta_1 D_i + \sum_{s=2}^{S} I(G_s) + \epsilon_i,$$

where s is the stratum number, $I(G_s)$ is an indicator for being in stratum s, and S is the number of strata. (Note that we don't include a dummy for the first stratum.)

- Let's talk about propensity scores!
- Propensity scores are a way to deal with selection bias, provided some assumptions are met
- This section draws heavily from Elizabeth Stuart's slides: http://www.preventionresearch.org/wp-content/uploads/2011/07/SPR-Propensity-pc-workshop-slides.pdf
  - We will use one of her packages: MatchIt (`install.packages("MatchIt")`), note the capitalization!
  - https://kosukeimai.github.io/MatchIt/index.html

Let's first talk about matching

- The idea behind matching is to approach the experimental idea:
  - The treatment group and the control group are approximately the same

- The catch: we can only do this for *observables*
  - While RCTs also match on unobservables, we can't do that here

Consider the following example

- We are interested in the effect of gender on wages
  - `matchingdata.csv`, data from South Africa's LFS

	male	female
meanwage	2.619	2.240
agemployment	0.237	0.124
african	0.747	0.720
coloured	0.159	0.180
married	0.625	0.437
age	37.214	37.306
noschooling	0.108	0.099
secondary	0.239	0.241
hhchildsuppgrant	0.062	0.094
urban	0.548	0.630

Exact matching

- One option is to match exactly on observables
  - For example, we could match on age, education, and marital status

- This is straightforward with just a couple variables

Exact matching

```
match <- matchit(female ~ married + secondary, data = df, exact = c("married", "secondary"), replace = TRUE)
summary(match)
```

```
##
## Call:
## matchit(formula = female ~ married + secondary, data = df, exact = c("married",
##     "secondary"), replace = TRUE)
##
## Summary of Balance for All Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance         0.4431        0.4081          0.3799     1.0528    0.0938
## married          0.4371        0.6254         -0.3796          .    0.1883
## secondary        0.2406        0.2390          0.0037          .    0.0016
##           eCDF Max
## distance    0.1883
## married     0.1883
## secondary   0.0016
##
## Summary of Balance for Matched Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance         0.4431        0.4431               0     0.9999         0
## married          0.4371        0.4371               0          .         0
## secondary        0.2406        0.2406               0          .         0
##           eCDF Max Std. Pair Dist.
## distance         0               0
## married          0               0
## secondary        0               0
##
## Sample Sizes:
##                Control Treated
## All           20495.     15019
## Matched (ESS)  5201.43   15019
```

- This becomes problematic when we have *many* covariates
  - It is often impossible to find exact matches across all covariates
- Enter: the propensity score
  - The propensity score is the probability of being treated, conditional on covariates
  - The key: we can match on the propensity score, rather than on the covariates themselves
- Rosenbaum and Rubin (1983), Dehejia and Wahba (2002)

Propensity scores

```
match <- matchit(female ~ married + secondary + age + hhchildsuppgrant + no_schooling + groupafrican + groupcoloured + agemp,
                 data = df, replace = TRUE, method = "nearest")
```

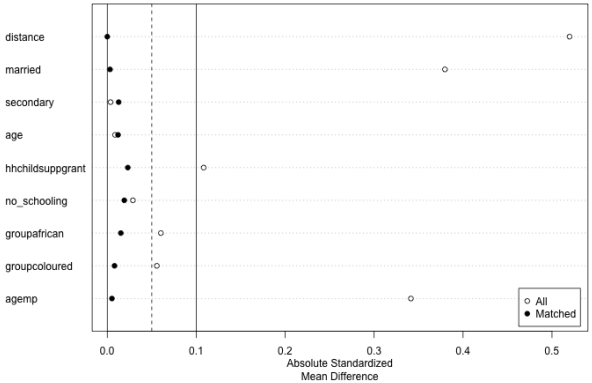
Propensity scores - All observations

```
##                   Means Treated Means Control Std. Mean Diff.
## distance             0.46066320     0.3952330     0.519724170
## married              0.43711299     0.6254208    -0.379630364
## secondary            0.24056195     0.2389851     0.003689156
## age                 37.30574605    37.2141986     0.008727805
## hhchildsuppgrant     0.09374792     0.0621615     0.108366496
## no_schooling         0.09940742     0.1080263    -0.028805791
## groupafrican         0.71975498     0.7468163    -0.060254249
## groupcoloured        0.18023836     0.1588192     0.055723018
## agemp                0.12437579     0.2370334    -0.341376649
```

Propensity scores - Matched observations

```
##                  Means Treated Means Control Std. Mean Diff.
## distance           0.46066320    0.46066152    1.333037e-05
## married            0.43711299    0.43558160    3.087304e-03
## secondary          0.24056195    0.23510220    1.277360e-02
## age               37.30574605   37.17870697    1.211145e-02
## hhchildsuppgrant   0.09374792    0.08702310    2.307146e-02
## no_schooling       0.09940742    0.09368134    1.913744e-02
## groupafrican       0.71975498    0.71289700    1.526986e-02
## groupcoloured      0.18023836    0.18336773   -8.141219e-03
## agemp              0.12437579    0.12264465    5.245725e-03
```

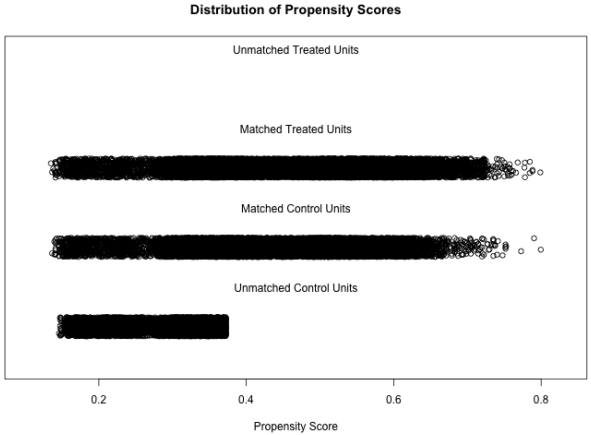
Plotting


distance
married
secondary
age
hhchildsuppgrant
no_schooling
groupafrican
groupcoloured
agemp
All
Matched
0.0
0.1
0.2
0.3
0.4
0.5
Absolute Standardized
Mean Difference

Using the propensity scores

- You can then use the propensity score in different ways
- Examples:
  - Can match treatment units to control units based on propensity scores
  - Can control for the propensity score in a regression
  - Can match and then use something like differences-in-differences
- We won't go into detail, as this is just something I want you to *know* about

Checking common support



In-class practice

- Let's now do some in-class practice
- Goals:
  - Practice regression
  - Practice creating tables
- Data: deserrannoetal.dta (on GitHub)