Precept 10: Causal identification and estimation

Soc 500: Applied Social Statistics

Ian Lundberg

Princeton University

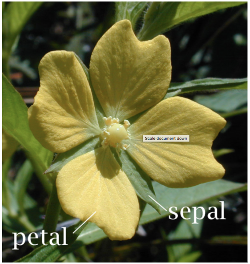
December 1, 2016

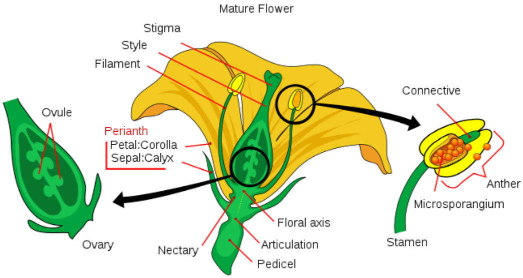
Learning Objectives

1. More data manipulation with `dplyr`
2. Practice evaluating causal identification strategies in published papers

R.A. Fisher's Irises

R.A. Fisher published a paper in 1936 using data on the length and width of the **petals** and **sepals** of a sample of irises. The data is now a common R example dataset.





(Flower above is not an iris)

Three species of irises
Photos from Wikipedia

Setosa

Virginica

Versicolor

Explore and load the iris data

See where this comes from:
https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/iris.html

```
> head(iris)
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

Iris goal 1

Goal: Plot the distribution of sepal lengths and petal lengths in the data, by species

Steps:

1. **select** the variables of interests

Iris goal 1

Goal: Plot the distribution of sepal lengths and petal lengths in the data, by species

Steps:

1. **select** the variables of interests
2. **melt** the data so that the lengths are stored in a single column

Iris goal 1

Goal: Plot the distribution of sepal lengths and petal lengths in the data, by species

Steps:

1. **select** the variables of interests
2. **melt** the data so that the lengths are stored in a single column
3. **ggplot** the density plot, using **fill** to distinguish the petal vs. sepal

Iris goal 1

Goal: Plot the distribution of sepal lengths and petal lengths in the data, by species

Steps:

1. **select** the variables of interests
2. **melt** the data so that the lengths are stored in a single column
3. **ggplot** the density plot, using **fill** to distinguish the petal vs. sepal
4. **facet_wrap** to separate by species

Start with the data frame

Code	**Output**
iris	<pre> Sepal.Length Sepal.Width Petal.Length Petal.Width Species 1 5.1 3.5 1.4 0.2 setosa 2 4.9 3.0 1.4 0.2 setosa 3 4.7 3.2 1.3 0.2 setosa 4 4.6 3.1 1.5 0.2 setosa 5 5.0 3.6 1.4 0.2 setosa 6 5.4 3.9 1.7 0.4 setosa</pre>

Select the variables of interest

Reference:
https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html

Code	Output
`iris %>% select(Sepal.Length, Sepal.Width, Species)`	see below

Code

```
iris %>%
  select(Sepal.Length,
         Sepal.Width,
         Species)
```

Output

```
  Sepal.Length Sepal.Width Species
1          5.1         3.5  setosa
2          4.9         3.0  setosa
3          4.7         3.2  setosa
4          4.6         3.1  setosa
5          5.0         3.6  setosa
6          5.4         3.9  setosa
```

Melt so that the variable to be plotted is in one column

Reference: `http://seananderson.ca/2013/10/19/reshape.html`

Code	Output

```
iris %>%
  select(Sepal.Length,
         Sepal.Width,
         Species) %>%
  melt(id.vars = "Species")
```

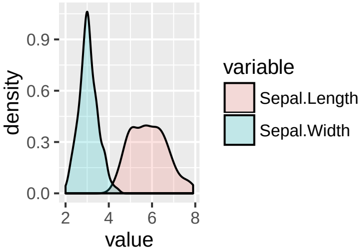
```
  Species     variable value
1  setosa Sepal.Length   5.1
2  setosa Sepal.Length   4.9
3  setosa Sepal.Length   4.7
4  setosa Sepal.Length   4.6
5  setosa Sepal.Length   5.0
6  setosa Sepal.Length   5.4
```

Plot the density

Reference: `http://docs.ggplot2.org/current/`

Code	**Output**

```
iris %>%
  select(Sepal.Length,
         Sepal.Width,
         Species) %>%
  melt(id.vars = "Species") %>%
  ggplot(aes(x = value,
             fill = variable)) +
  geom_density(alpha = .2)
```

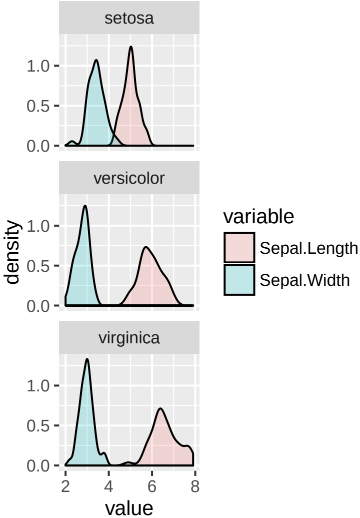


Facet by species

Reference: `http://docs.ggplot2.org/current/`

Code **Output**

```
iris %>%
  select(Sepal.Length,
         Sepal.Width,
         Species) %>%
  melt(id.vars = "Species") %>%
  ggplot(aes(x = value,
             fill = variable)) +
  geom_density(alpha = .2) +
  facet_wrap(~Species,
             ncol = 1)
```



Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

	Tepal	
Species	Petal	Sepal
Setosa	Mean(L - W)	Mean(L - W)
Versicolor	Mean(L - W)	Mean(L - W)
Virginica	Mean(L - W)	Mean(L - W)

Table: Structure of the goal data table

Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

Steps:

1. **mutate** the iris data to add an id number to the rows

Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

Steps:

1. **mutate** the iris data to add an id number to the rows
2. **melt** the data so that the lengths are stored in a single column

Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

Steps:

1. **mutate** the iris data to add an id number to the rows
2. **melt** the data so that the lengths are stored in a single column
3. **separate** the petal/sepal and length/width into two variables

Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

Steps:

1. **mutate** the iris data to add an id number to the rows
2. **melt** the data so that the lengths are stored in a single column
3. **separate** the petal/sepal and length/width into two variables
4. **spread** the data wider so that length and width are each variables

Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

Steps:

1. **mutate** the iris data to add an id number to the rows
2. **melt** the data so that the lengths are stored in a single column
3. **separate** the petal/sepal and length/width into two variables
4. **spread** the data wider so that length and width are each variables
5. **mutate** the data to calculate the difference in length and width

Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

Steps:

1. **mutate** the iris data to add an id number to the rows
2. **melt** the data so that the lengths are stored in a single column
3. **separate** the petal/sepal and length/width into two variables
4. **spread** the data wider so that length and width are each variables
5. **mutate** the data to calculate the difference in length and width
6. **group_by** species and type

Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

Steps:

1. **mutate** the iris data to add an id number to the rows
2. **melt** the data so that the lengths are stored in a single column
3. **separate** the petal/sepal and length/width into two variables
4. **spread** the data wider so that length and width are each variables
5. **mutate** the data to calculate the difference in length and width
6. **group_by** species and type
7. **summarize** to calculate the mean difference within each group

Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

Steps:

1. **mutate** the iris data to add an id number to the rows
2. **melt** the data so that the lengths are stored in a single column
3. **separate** the petal/sepal and length/width into two variables
4. **spread** the data wider so that length and width are each variables
5. **mutate** the data to calculate the difference in length and width
6. **group_by** species and type
7. **summarize** to calculate the mean difference within each group
8. **select** the variables for our table

Iris goal 2

Goal: Compare length vs. width, within strata defined by

- species AND
- petal/sepal

Steps:

1. **mutate** the iris data to add an id number to the rows
2. **melt** the data so that the lengths are stored in a single column
3. **separate** the petal/sepal and length/width into two variables
4. **spread** the data wider so that length and width are each variables
5. **mutate** the data to calculate the difference in length and width
6. **group_by** species and type
7. **summarize** to calculate the mean difference within each group
8. **select** the variables for our table
9. **spread** them out to make a nice table

Making a table
Start with a data frame

Code	Output
`iris`	<pre> Sepal.Length Sepal.Width Petal.Length Petal.Width Species 1 5.1 3.5 1.4 0.2 setosa 2 4.9 3.0 1.4 0.2 setosa 3 4.7 3.2 1.3 0.2 setosa 4 4.6 3.1 1.5 0.2 setosa 5 5.0 3.6 1.4 0.2 setosa 6 5.4 3.9 1.7 0.4 setosa</pre>

Making a table

Add id number. Reference:
https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html

Code	Output
`iris %>% mutate(idnum = 1:nrow(iris))`	(see below)

```
iris %>%
  mutate(idnum = 1:nrow(iris))
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species idnum
1          5.1         3.5          1.4         0.2  setosa     1
2          4.9         3.0          1.4         0.2  setosa     2
3          4.7         3.2          1.3         0.2  setosa     3
4          4.6         3.1          1.5         0.2  setosa     4
5          5.0         3.6          1.4         0.2  setosa     5
6          5.4         3.9          1.7         0.4  setosa     6
```

Making a table

Melt to make the data long. Reference:
http://seananderson.ca/2013/10/19/reshape.html

Code

```
iris %>%
  mutate(idnum = 1:nrow(iris)) %>%
  melt(id = c("idnum","Species"))
```

Output

```
  idnum Species     variable value
1     1  setosa Sepal.Length   5.1
2     2  setosa Sepal.Length   4.9
3     3  setosa Sepal.Length   4.7
4     4  setosa Sepal.Length   4.6
5     5  setosa Sepal.Length   5.0
6     6  setosa Sepal.Length   5.4
```

Making a table

Separate type and measure. Reference:
https://blog.rstudio.org/2014/07/22/introducing-tidyr/

Code	Output

<pre>iris %>%
  mutate(idnum = 1:nrow(iris)) %>%
  melt(id = c("idnum","Species")) %>%
  separate(col = "variable",
           into = c("Type","Measure"),
           sep = "\\.")</pre>
	<pre> idnum Species Type Measure value
	1 1 setosa Sepal Length 5.1
	2 2 setosa Sepal Length 4.9
	3 3 setosa Sepal Length 4.7
	4 4 setosa Sepal Length 4.6
	5 5 setosa Sepal Length 5.0
	6 6 setosa Sepal Length 5.4</pre>

Making a table

Spread to make wide. Reference:
https://blog.rstudio.org/2014/07/22/introducing-tidyr/

Code

```
iris %>%
  mutate(idnum = 1:nrow(iris)) %>%
  melt(id = c("idnum","Species")) %>%
  separate(col = "variable",
           into = c("Type","Measure"),
           sep = "\.") %>%
  spread(key = Measure, value = value)
```

Output

```
  idnum Species  Type Length Width
1     1  setosa Petal    1.4   0.2
2     1  setosa Sepal    5.1   3.5
3     2  setosa Petal    1.4   0.2
4     2  setosa Sepal    4.9   3.0
5     3  setosa Petal    1.3   0.2
6     3  setosa Sepal    4.7   3.2
```

Making a table

Mutate to make a difference variable. Reference:
https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html

Code	Output

Code

```
iris %>%
  mutate(idnum = 1:nrow(iris)) %>%
  melt(id = c("idnum","Species")) %>%
  separate(col = "variable",
           into = c("Type","Measure"),
           sep = "\.") %>%
  spread(key = Measure, value = value) %>%
  mutate(difference = Length - Width)
```

Output

```
  idnum Species  Type Length Width difference
1     1  setosa Petal    1.4   0.2        1.2
2     1  setosa Sepal    5.1   3.5        1.6
3     2  setosa Petal    1.4   0.2        1.2
4     2  setosa Sepal    4.9   3.0        1.9
5     3  setosa Petal    1.3   0.2        1.1
6     3  setosa Sepal    4.7   3.2        1.5
```

Making a table

Group by species and type, summarize to get mean, sd, and n in each group. Reference: https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html

Code

```
iris %>%
  mutate(idnum = 1:nrow(iris)) %>%
  melt(id = c("idnum","Species")) %>%
  separate(col = "variable",
           into = c("Type","Measure"),
           sep = "\.") %>%
  spread(key = Measure, value = value) %>%
  mutate(difference = Length - Width) %>%
  group_by(Species,Type) %>%
  summarize(mean_difference = mean(difference),
            sd_difference = sd(difference),
            count = n())
```

Output

```
Source: local data frame [6 x 5]
Groups: Species [3]

     Species  Type mean_difference sd_difference count
      (fctr) (chr)           (dbl)         (dbl) (int)
1     setosa Petal           1.216     0.1706650    50
2     setosa Sepal           1.578     0.2636401    50
3 versicolor Petal           2.934     0.3372215    50
4 versicolor Sepal           3.166     0.4410609    50
5  virginica Petal           3.526     0.5313863    50
6  virginica Sepal           3.614     0.5664101    50
```

Making a table

Select the variables of interest. Reference:
`https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html`

Code

```
iris %>%
  mutate(idnum = 1:nrow(iris)) %>%
  melt(id = c("idnum","Species")) %>%
  separate(col = "variable",
           into = c("Type","Measure"),
           sep = "\.") %>%
  spread(key = Measure, value = value) %>%
  mutate(difference = Length - Width) %>%
  group_by(Species,Type) %>%
  summarize(mean_difference = mean(difference),
            sd_difference = sd(difference),
            count = n()) %>%
  select(mean_difference, Species, Type)
```

Output

	mean_difference	Species	Type
	(dbl)	(fctr)	(chr)
1	1.216	setosa	Petal
2	1.578	setosa	Sepal
3	2.934	versicolor	Petal
4	3.166	versicolor	Sepal
5	3.526	virginica	Petal
6	3.614	virginica	Sepal

Making a table

spread to make the data wide for a pretty table. Reference:
https://blog.rstudio.org/2014/07/22/introducing-tidyr/

Code	**Output**

```
iris %>%
  mutate(idnum = 1:nrow(iris)) %>%
  melt(id = c("idnum","Species")) %>%
  separate(col = "variable",
           into = c("Type","Measure"),
           sep = "\.") %>%
  spread(key = Measure, value = value) %>%
  mutate(difference = Length - Width) %>%
  group_by(Species,Type) %>%
  summarize(mean_difference = mean(difference),
            sd_difference = sd(difference),
            count = n()) %>%
  select(mean_difference, Species, Type) %>%
  spread(Type, mean_difference)
```

```
Source: local data frame [3 x 3]
Groups: Species [3]

     Species Petal Sepal
      (fctr) (dbl) (dbl)
1     setosa 1.216 1.578
2 versicolor 2.934 3.166
3  virginica 3.526 3.614
```

Making a figure
Going back a few steps

Code	**Output**

```
iris %>%
  mutate(idnum = 1:nrow(iris)) %>%
  melt(id = c("idnum","Species")) %>%
  separate(col = "variable",
           into = c("Type","Measure"),
           sep = "\.") %>%
  spread(key = Measure, value = value) %>%
  mutate(difference = Length - Width) %>%
  group_by(Species,Type) %>%
  summarize(mean_difference = mean(difference),
            sd_difference = sd(difference),
            count = n())
```

```
Source: local data frame [6 x 5]
Groups: Species [3]

     Species  Type mean_difference sd_difference count
      (fctr) (chr)           (dbl)         (dbl) (int)
1     setosa Petal           1.216     0.1706650    50
2     setosa Sepal           1.578     0.2636401    50
3 versicolor Petal           2.934     0.3372215    50
4 versicolor Sepal           3.166     0.4410609    50
5  virginica Petal           3.526     0.5313863    50
6  virginica Sepal           3.614     0.5664101    50
```

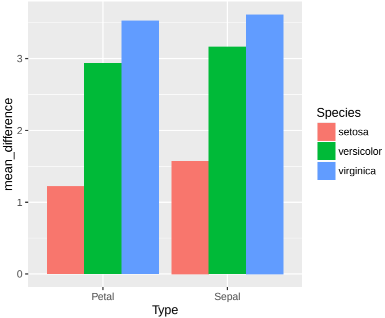
Making a figure

Using geom_bar. Reference: `http://docs.ggplot2.org/current/`

Code

```
iris %>%
  mutate(idnum = 1:nrow(iris)) %>%
  melt(id = c("idnum","Species")) %>%
  separate(col = "variable",
           into = c("Type","Measure"),
           sep = "\.") %>%
  spread(key = Measure, value = value) %>%
  mutate(difference = Length - Width) %>%
  group_by(Species,Type) %>%
  summarize(mean_difference = mean(difference),
            sd_difference = sd(difference),
            count = n()) %>%
  ggplot(aes(x = Type,
             fill = Species)) +
  geom_bar(stat = "identity", position = "dodge")
```

Output

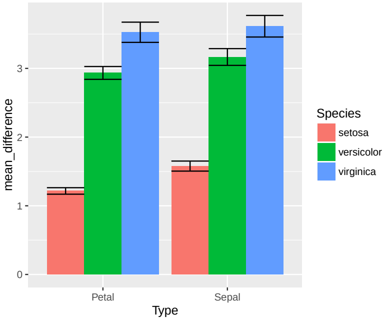


Making a figure

Adding error bars. Reference: `http://docs.ggplot2.org/current/`

Code	Output

```
iris %>%
  mutate(idnum = 1:nrow(iris)) %>%
  melt(id = c("idnum","Species")) %>%
  separate(col = "variable",
           into = c("Type","Measure"),
           sep = "\.") %>%
  spread(key = Measure, value = value) %>%
  mutate(difference = Length - Width) %>%
  group_by(Species,Type) %>%
  summarize(mean_difference = mean(difference),
            sd_difference = sd(difference),
            count = n()) %>%
  ggplot(aes(x = Type,
             y = mean_difference,
             ymin = mean_difference -
               qnorm(.975) * sd_difference /
               sqrt(count),
             ymax = mean_difference +
               qnorm(.975) * sd_difference /
               sqrt(count),
             fill = Species)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_errorbar(position = "dodge")
```

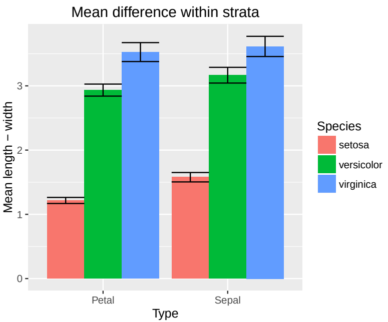


Making a figure

Add titles. We're finished! Reference: `http://docs.ggplot2.org/current/`

Code	**Output**

```
iris %>%
  mutate(idnum = 1:nrow(iris)) %>%
  melt(id = c("idnum","Species")) %>%
  separate(col = "variable",
           into = c("Type","Measure"),
           sep = "\.") %>%
  spread(key = Measure, value = value) %>%
  mutate(difference = Length - Width) %>%
  group_by(Species,Type) %>%
  summarize(mean_difference = mean(difference),
            sd_difference = sd(difference),
            count = n()) %>%
  ggplot(aes(x = Type,
             y = mean_difference,
             ymin = mean_difference -
               qnorm(.975) * sd_difference /
               sqrt(count),
             ymax = mean_difference +
               qnorm(.975) * sd_difference /
               sqrt(count),
             fill = Species)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_errorbar(position = "dodge") +
  ylab("Mean length - width") +
  ggtitle("Mean difference within strata")
```



Slight shift of focus

Slight shift of focus

- We've been doing lots of Applied Social **Statistics**.

Slight shift of focus

- We've been doing lots of Applied Social **Statistics**.
- Let's do some **Applied Social** Statistics!

Causal inference examples

We will walk through examples of causal social science papers that assume observed confounding. For each paper, we will:

- Draw the DAG

Causal inference examples

We will walk through examples of causal social science papers that assume observed confounding. For each paper, we will:

- Draw the DAG
- Define the potential outcomes: $Y_i(0)$, $Y_i(1)$

Causal inference examples

We will walk through examples of causal social science papers that assume observed confounding. For each paper, we will:

- Draw the DAG
- Define the potential outcomes: $Y_i(0), Y_i(1)$
- Discuss potential violations of the identifying assumptions.

Causal inference examples

We will walk through examples of causal social science papers that assume observed confounding. For each paper, we will:

- Draw the DAG
- Define the potential outcomes: $Y_i(0)$, $Y_i(1)$
- Discuss potential violations of the identifying assumptions.
- Conclude: **Do we buy it?**

Example 1: An ethnographic experiment

Duneier, Mitchell. 2001. *Sidewalk.* New York: Farrar, Straus, and Giroux.

- Ethnographic study of book vendors in Greenwich Village in NYC.

Example 1: An ethnographic experiment

Duneier, Mitchell. 2001. *Sidewalk.* New York: Farrar, Straus, and Giroux.

- Ethnographic study of book vendors in Greenwich Village in NYC.
- Duneier noticed that black vendors were pushed around by police officers.

Example 1: An ethnographic experiment

Duneier, Mitchell. 2001. *Sidewalk.* New York: Farrar, Straus, and Giroux.

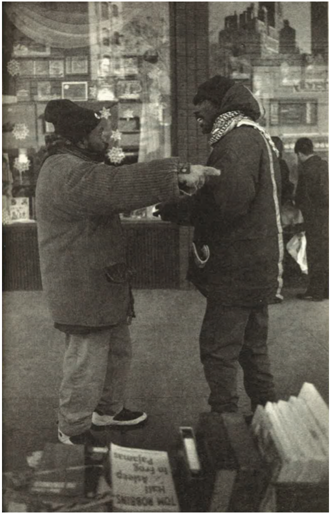
- Ethnographic study of book vendors in Greenwich Village in NYC.
- Duneier noticed that black vendors were pushed around by police officers.
- **Question:** Does a vendors race and legal knowledge affect how the police treat him?

Example 1: An ethnographic experiment

Duneier, Mitchell. 2001. *Sidewalk.* New York: Farrar, Straus, and Giroux.

- Ethnographic study of book vendors in Greenwich Village in NYC.
- Duneier noticed that black vendors were pushed around by police officers.
- **Question:** Does a vendors race and legal knowledge affect how the police treat him?
- **Approach:** A creative small-scale experiment.

NYC street vendors



Duneier 2001: The treated situation

Selections from p. 266-272

When the Law "Means Nothing" to the Police

Two days later, on Christmas afternoon, I saw Ishmael again. When I arrived, Hakim was standing on the corner. Ishmael had set up his table in his usual spot on the corner of Sixth Avenue and Eighth Street. Ten minutes later, Officer X (as I'll call him) approached and said something to the effect of: Ishmael, you have to break down, guy.[15]

I'm not breaking down, man, he responded.

Ishmael clearly was not showing the kind of deference the men on the block normally observe. I took out my tape recorder and turned it on, though neither Ishmael nor the officer saw me do so.

"You have to break down," the officer insisted.

"But I'm not. Because there's no such thing as a law telling me that. I'm not gonna break down, man. If I can't work, what the hell you working for?"

"Step over here for a second. Ishmael . . ."

Duneier 2001: Stating confounders

Selections from p. 266-272

If this was a test designed to find out whether an upper-middle-class white person would be treated differently from an unhoused, poor black vendor, I thought to myself, then it was not a good one. To begin with, the officer had just closed Ishmael down. The odds were very small that a black police officer who had to enforce the law against black vendors every day would let himself be seen as one who would allow a white man to stay in the same spot. Furthermore, he might notice the microphone sticking out of my pocket, and this would probably affect what he'd say to me.

I had been standing at the table for about ten minutes when I saw the officer and his beat partner walking toward me.

As I waited, approximately ten black vendors, including Hakim and Ishmael, stood by, offering their support.

"It's showtime!" yelled Ishmael.

Duneier 2001: Control

Selections from p. 266-272

"My man. There's no selling here today. Break it down."

"Excuse me," I said.

"No selling here today. Break it down."

I took a copy of the municipal law out of my pocket. "I'm exercising my right under Local Law 33 of 1982, and Local Law 45 of 1993, to sell written matter."

Potential outcomes

- Units of analysis are interactions with police

Potential outcomes

- Units of analysis are interactions with police
- Sample size is 2, but 2 high-quality observations

Potential outcomes

- Units of analysis are interactions with police
- Sample size is 2, but 2 high-quality observations
- **Treatment:** Vendor is a black male Greenwich Village bookseller.

Potential outcomes

- Units of analysis are interactions with police
- Sample size is 2, but 2 high-quality observations
- **Treatment:** Vendor is a black male Greenwich Village bookseller.
- **Control:** Vendor is Mitch Duneier who explicitly defends his rights

Example 2: Occupational attainment model

Blau, Peter Michael, and Otis Dudley Duncan. 1967. *The American Occupational Structure.* New York: Wiley.

- **Research question:** How does family background affect the educational and occupational attainment of the next generation?
- **Method:** Linear structural equation models, which were the precursor to DAGs

Example 2: Blau-Duncan (1967) status attainment model

170 THE PROCESS OF STRATIFICATION

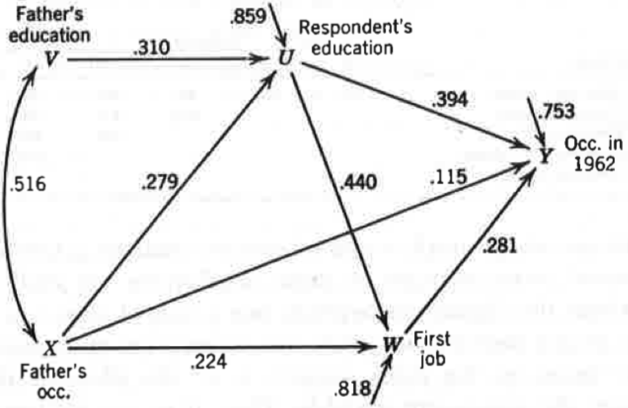


Figure 5.1. Path coefficients in basic model of the process of stratification.

What to condition on for the effect of..

1. first job on occ. in 1962?

What to condition on for the effect of..

1. first job on occ. in 1962?
2. respondent's education on first job?

What to condition on for the effect of..

1. first job on occ. in 1962?
2. respondent's education on first job?
3. respondent's education on occ. in 1962?

What to condition on for the effect of..

1. first job on occ. in 1962?
2. respondent's education on first job?
3. respondent's education on occ. in 1962?
4. father's occupation/education on son's occupation in 1962?

What to condition on for the effect of..

1. first job on occ. in 1962?
2. respondent's education on first job?
3. respondent's education on occ. in 1962?
4. father's occupation/education on son's occupation in 1962?
5. If we condition on respondent's education and first job, will father's education be associated with son's occupation in 1962?

What to condition on for the effect of..

1. first job on occ. in 1962?
2. respondent's education on first job?
3. respondent's education on occ. in 1962?
4. father's occupation/education on son's occupation in 1962?
5. If we condition on respondent's education and first job, will father's education be associated with son's occupation in 1962?

What to condition on for the effect of..

1. first job on occ. in 1962?
2. respondent's education on first job?
3. respondent's education on occ. in 1962?
4. father's occupation/education on son's occupation in 1962?
5. If we condition on respondent's education and first job, will father's education be associated with son's occupation in 1962?

Answers:

1. Respondent's education, father's occupation
2. father's occupation is sufficient
3. father's occupation is sufficient
4. No conditioning needed! But I doubt the DAG holds.
5. No. But only because the DAG assumes the unobserved influences are uncorrelated!

Blau-Duncan assumptions

PATH COEFFICIENTS

Whether a path diagram, or the causal scheme it represents, is adequate depends on both theoretical and empirical considerations. At a minimum, before constructing the diagram we must know, or be willing to assume, a causal ordering of the observed variables (hence the lengthy discussion of this matter earlier in this chapter). This information is external or *a priori* with respect to the data, which merely describe associations or correlations. Moreover, the causal scheme must be complete, in the sense that all causes are accounted for. Here, as in most problems involving analysis of observational data, we achieve a formal completeness of the scheme by representing unmeasured causes as a residual factor, presumed to be uncorrelated with the remaining factors lying behind the variable in question. If

Side note - incredible pre-analysis plan (p. 18)

By the time the data were actually collected the investigators had developed the first of two major sets of specifications for tabulations. It should be mentioned here that at no time have we had access to the original survey documents or to the computer tapes on which individual records are stored. This information is confidential and not available to private research workers. Consequently it was necessary for us to provide detailed outlines of the statistical tables we desired for analysis without inspecting the "raw" data, and to provide these, moreover, some 9 to 12 months ahead of the time when we might expect their delivery. This lead time was required for programming the computer runs that would produce the tables. Evidently this circumstance precluded our following the common strategy of looking at a few marginal totals before running some two-way tables and deciding on interesting three-way or higher-order tabulations after having studied the two-way tables. We had to state in advance just which tables were wanted, out of the virtually unlimited number that conceivably might have been produced, and to be prepared to make the best of what we got. Cost factors, of course, put strict limits on how many tables we could request. We had to imagine in advance most of the analysis we would want to make, before having any advance indications of what any of the tables would look like.

Example 3: Bringing in aspirations

Sewell, William H., Archibald O. Haller, and Alejandro Portes. 1969. "The Educational and Early Occupational Attainment Process." *American Sociological Review* 34 (1): 82?92. doi:10.2307/2092789.

- Challenged Blau and Duncan

Example 3: Bringing in aspirations

Sewell, William H., Archibald O. Haller, and Alejandro Portes. 1969. "The Educational and Early Occupational Attainment Process." *American Sociological Review* 34 (1): 82?92. doi:10.2307/2092789.

- Challenged Blau and Duncan
- Argued that **aspirations** of children were an important pathway linking parental and child attainment

Example 3: Bringing in aspirations

Sewell, William H., Archibald O. Haller, and Alejandro Portes. 1969. "The Educational and Early Occupational Attainment Process." *American Sociological Review* 34 (1): 82?92. doi:10.2307/2092789.

- Challenged Blau and Duncan
- Argued that **aspirations** of children were an important pathway linking parental and child attainment
- Became known as the Wisconsin model of status attainment

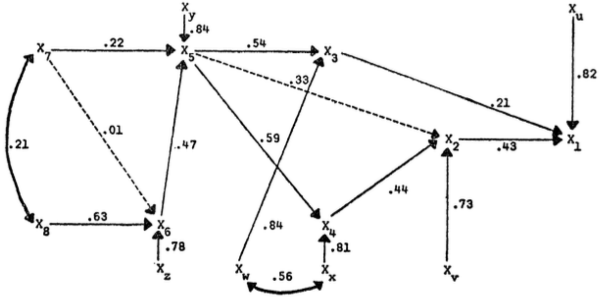
Example 3: Wisconsin model of status attainment

Sewell, Haller, and Portes (1969), *ASR*

OCCUPATIONAL ATTAINMENT 85

DIAGRAM 1

PATH COEFFICIENTS OF ANTECEDENTS OF EDUCATIONAL AND OCCUPATIONAL ATTAINMENT LEVELS



X_1 - Occupational Attainment

X_2 - Educational Attainment

X_3 - Level of Occupational Aspiration

X_4 - Level of Educational Aspiration

X_5 - Significant Others' Influence

X_6 - Academic Performance

X_7 - Socioeconomic Status

X_8 - Mental Ability

Wisconsin model: What to condition on to identify the effect of...

1. X_2 on X_1?

Wisconsin model: What to condition on to identify the effect of...

1. X_2 on X_1?
2. X_5 on X_2?

Wisconsin model: What to condition on to identify the effect of...

1. X_2 on X_1?
2. X_5 on X_2?

Wisconsin model: What to condition on to identify the effect of...

1. X_2 on X_1?
2. X_5 on X_2?

Answers:

1. X_5 or X_3
2. No conditioning needed!

Example 4: Heterogeneous effects of college

Brand, Jennie E., and Yu Xie. 2010. “Who Benefits Most from College? Evidence for Negative Selection in Heterogeneous Economic Returns to Higher Education.” *American Sociological Review.*

Brand and Xie (2010)

- Research question:

Brand and Xie (2010)

- Research question:
  - **Does college affect earnings?**

Brand and Xie (2010)

- Research question:
  - **Does college affect earnings?**
  - Is the effect moderated by social origin?

Brand and Xie (2010)

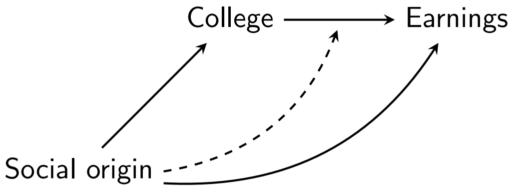
- Research question:
  - **Does college affect earnings?**
  - Is the effect moderated by social origin?
- Identification strategy: Selection on observables

Brand and Xie (2010)

- Research question:
    - **Does college affect earnings?**
    - Is the effect moderated by social origin?
- Identification strategy: Selection on observables

Brand and Xie (2010)

- Research question:
    - **Does college affect earnings?**
    - Is the effect moderated by social origin?
- Identification strategy: Selection on observables



Theoretically: Why heterogeneous effects?

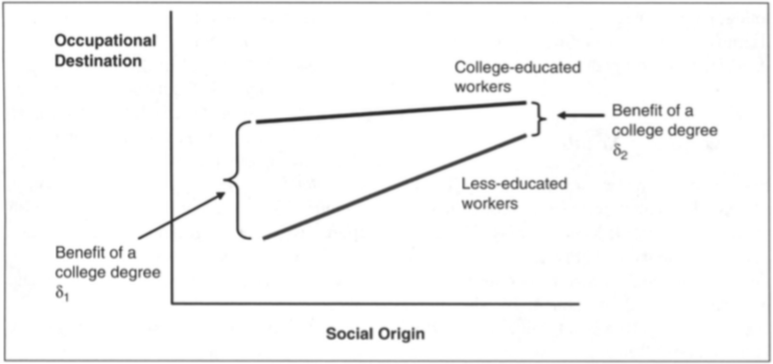


Figure 1. Hypothetical Model: Origin, Education, and Destination

Theoretically: Why heterogeneous effects?

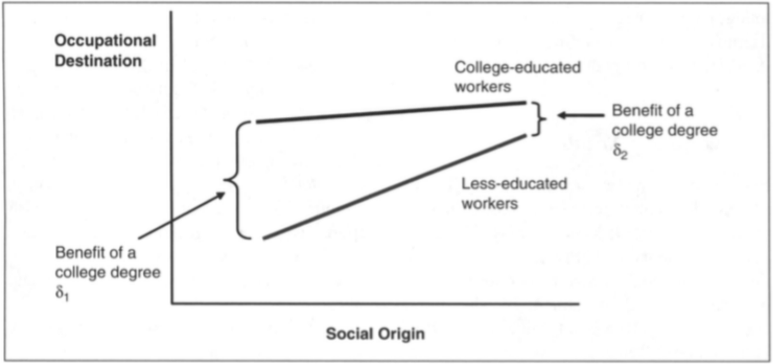


Figure 1. Hypothetical Model: Origin, Education, and Destination

Question: Can we write the potential outcomes here?

Ignorability

What is the assumption of ignorability here?

Ignorability

What is the assumption of ignorability here?

Ignorability

What is the assumption of ignorability here?

To infer causality with observational data, it is necessary to introduce unverifiable assumptions. In this research, we first introduce the ignorability assumption:

$$E(y^0|\boldsymbol{X}, d = 1) = E(y^0|\boldsymbol{X}, d = 0) \qquad (6a)$$

and

$$E(y^1|\boldsymbol{X}, d = 0) = E(y^1|\boldsymbol{X}, d = 1). \qquad (6b)$$

Equation 6a assumes that the average earnings of college-educated workers, had they not completed college, would be the same as the average earnings of non-college-educated workers, conditional on observed covariates. Likewise, Equation 6b assumes that the average earnings of non-college-educated workers, had they completed college, would be the same as the average earnings of college-educated workers, conditional on observed covariates.

Conditioning set: Measuring "social origin"

Table 1. Descriptive Statistics of Precollege Covariates

	NLSY Means				WLS Means			
	Men (N = 1,265)		Women (N = 1,209)		Men (N = 3,690)		Women (N = 4,215)	
Variables	Non-College Graduate	College Graduate	Non-College Graduate	College Graduate	Non-College Graduate	College Graduate	Non-College Graduate	College Graduate
Race								
Black	.18	.07	.15	.07				
Hispanic	.07	.03	.07	.03				
Social Background								
Parents' income	17870	26538	18174	25991	5605	8123	5622	9262
Mother's education	11.26	13.32	11.18	13.37	10.15	11.56	9.94	12.02
Father's education	11.23	14.39	11.16	14.14	9.10	11.37	9.21	11.79
Intact family (0–1)	.72	.83	.67	.85	.90	.92	.90	.92
Number of siblings	3.29	2.34	3.40	2.45	3.45	2.61	3.51	2.40
Rural residence (0–1)	.25	.19	.24	.21	.22	.12	.20	.16
Urban res./prox. to college	.77	.78	.75	.80	.42	.50	.50	.53
Jewish (0–1)	.00	.03	.00	.04	.00	.02	.00	.03
Ability and Academics								
Class rank					35.76	65.49	53.78	79.51
Mental ability (IQ)	−.09	.69	−.04	.64	97.03	111.75	98.67	112.00
College-prep (0–1)	.23	.59	.23	.49	.54	.91	.46	.89
Social-Psychological								
Teachers' encouragement					.35	.75	.36	.77
Parents' encouragement					.47	.91	.39	.90
Friends' college plans	.42	.79	.48	.81	.22	.66	.30	.76
Weighted Sample Proportion	.76	.24	.77	.23	.69	.31	.82	.18

Note: Parents' income is measured as total net family income in 1979 dollars in the NLSY and in 1957 dollars in the WLS. Urban residency/proximity to college indicates whether a respondent lived in an SMSA in the NLSY and whether a respondent's high school was within 15 miles of a college or university in the WLS. Mental ability is measured with a scale of standardized residuals of the ASVAB in the NLSY and with the Henmon-Nelson IQ test in the WLS. College-prep indicates whether a student was enrolled in a college-preparatory curriculum in the NLSY or whether a student completed the requirements for UW-Madison in the WLS.

Conditioning set: Measuring "social origin"

Table 1. Descriptive Statistics of Precollege Covariates

	NLSY Means				WLS Means			
	Men (N = 1,265)		Women (N = 1,209)		Men (N = 3,690)		Women (N = 4,215)	
Variables	Non-College Graduate	College Graduate	Non-College Graduate	College Graduate	Non-College Graduate	College Graduate	Non-College Graduate	College Graduate
Race								
Black	.18	.07	.15	.07				
Hispanic	.07	.03	.07	.03				
Social Background								
Parents' income	17870	26538	18174	25991	5605	8123	5622	9262
Mother's education	11.26	13.32	11.18	13.37	10.15	11.56	9.94	12.02
Father's education	11.23	14.39	11.16	14.14	9.10	11.37	9.21	11.79
Intact family (0–1)	.72	.83	.67	.85	.90	.92	.90	.92
Number of siblings	3.29	2.34	3.40	2.45	3.45	2.61	3.51	2.40
Rural residence (0–1)	.25	.19	.24	.21	.22	.12	.20	.16
Urban res./prox. to college	.77	.78	.75	.80	.42	.50	.50	.53
Jewish (0–1)	.00	.03	.00	.04	.00	.02	.00	.03
Ability and Academics								
Class rank					35.76	65.49	53.78	79.51
Mental ability (IQ)	−.09	.69	−.04	.64	97.03	111.75	98.67	112.00
College-prep (0–1)	.23	.59	.23	.49	.54	.91	.46	.89
Social-Psychological								
Teachers' encouragement					.35	.75	.36	.77
Parents' encouragement					.47	.91	.39	.90
Friends' college plans	.42	.79	.48	.81	.22	.66	.30	.76
Weighted Sample Proportion	.76	.24	.77	.23	.69	.31	.82	.18

Note: Parents' income is measured as total net family income in 1979 dollars in the NLSY and in 1957 dollars in the WLS. Urban residency/proximity to college indicates whether a respondent lived in an SMSA in the NLSY and whether a respondent's high school was within 15 miles of a college or university in the WLS. Mental ability is measured with a scale of standardized residuals of the ASVAB in the NLSY and with the Henmon-Nelson IQ test in the WLS. College-prep indicates whether a student was enrolled in a college-preparatory curriculum in the NLSY or whether a student completed the requirements for UW-Madison in the WLS.

Question: How might the identifying assumptions be violated? Can we write it in terms of DAGs? Potential outcomes?

Example 5: Neighborhoods

Wodtke, Geoffrey T., David J. Harding, and Felix Elwert. 2011. "Neighborhood Effects in Temporal Perspective: The Impact of Long-Term Exposure to Concentrated Disadvantage on High School Graduation." *American Sociological Review* 76(5):713-736.

- **Research question:** How does long-term exposure to disadvantaged neighborhoods affect one's probability of high school graduation?

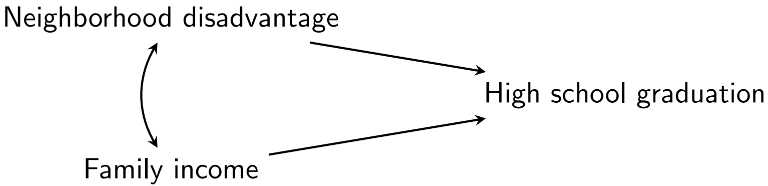
Example 5: Neighborhoods

Wodtke, Geoffrey T., David J. Harding, and Felix Elwert. 2011. "Neighborhood Effects in Temporal Perspective: The Impact of Long-Term Exposure to Concentrated Disadvantage on High School Graduation." *American Sociological Review* 76(5):713-736.

- **Research question:** How does long-term exposure to disadvantaged neighborhoods affect one's probability of high school graduation?
- **Problem:** Family income and neighborhood disadvantage affect each other through childhood

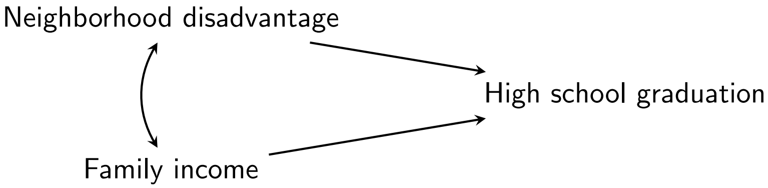
Wodtke, Harding, and Elwert 2011

We might want to have a bidirectional arrow linking neighborhood disadvantage and family income.



Wodtke, Harding, and Elwert 2011

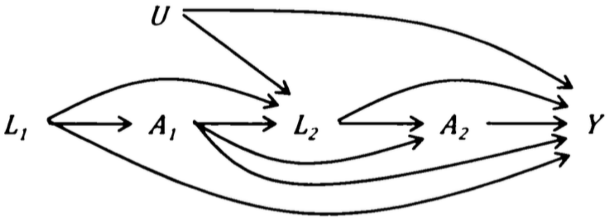
We might want to have a bidirectional arrow linking neighborhood disadvantage and family income.



Can we write sequentially to avoid the bi-directional edge?

Neighborhood Effects in Temporal Perspective

Wodtke, Harding, and Elwert 2011 *ASR*



- L = Family income
- A = Neighborhood disadvantage
- Y = High school graduation
- Subscripts = time

What do you condition on to identify:

1. The effect of A_2 on Y?

What do you condition on to identify:

1. The effect of A_2 on Y?
2. The effect of A_1 on Y?

What do you condition on to identify:

1. The effect of A_2 on Y?
2. The effect of A_1 on Y?

What do you condition on to identify:

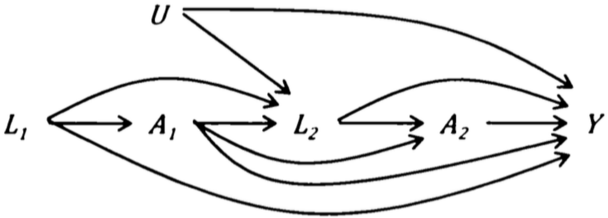
1. The effect of A_2 on Y?
2. The effect of A_1 on Y?

Answers:

1. $\{L_2, A_1\}$
2. $\{L_1\}$

Neighborhood Effects in Temporal Perspective

Wodtke, Harding, and Elwert 2011 *ASR*



Key point: We cannot just condition on family income (L) since part of it is caused by neighborhood disadvantage (A). Nor can we not condition on it. What to do?

A challenging identification problem!

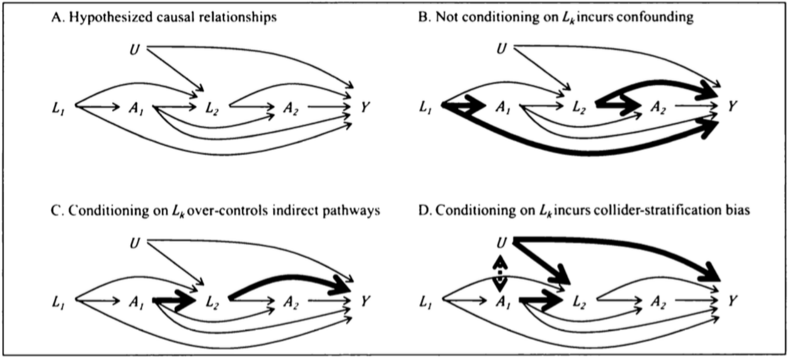


Figure 1. Causal Graphs for Exposure to Disadvantaged Neighborhoods with Two Waves of Follow-up

Note: A_k = neighborhood context, L_k = observed time-varying confounders, U = unobserved factors, Y = outcome.

Result from Wodtke, Harding, and Elwert 2011

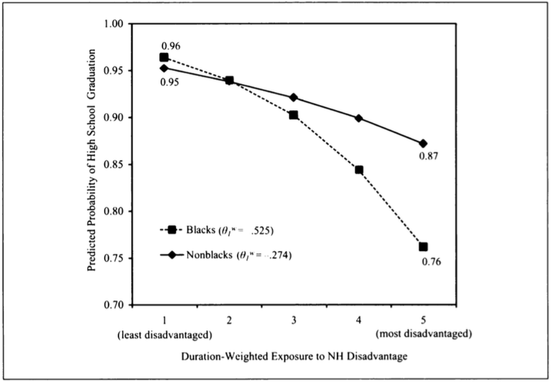


Figure 3. Predicted Probability of High School Graduation by Neighborhood Exposure History
Note: NH = Neighborhood

Example 6. Wealth and college attainment

Conley, Dalton. 2001. "Capital for College: Parental Assets and Postsecondary Schooling." *Sociology of Education*.

- **Research question:** Does family wealth affect educational attainment?

Conditioning set
Conley 2001, *Sociology of Education*

Model	Total Years of Schooling (Ages 19–30)
Respondents' Characteristics	
Black	.320** (.104)
Latino	-.113 (.354)
Other	.866 (.548)
Female	.372*** (.086)
Age (1992)	.095*** (.015)
Number of siblings	-.107*** (.027)
Parental Characteristics	
Age of head of household (1984)	.017* (.008)
Proportion of years female head (1980-84)	-.001 (.162)
Education of head of household (1984)	.167*** (.021)
Proportion of years head of household unemployed (1980-84)	-.587* (.296)
Occupational prestige of head of household (1980-84)	.017*** (.004)
Natural logarithm of income (1980-84, constant dollars)	.017 (.114)
Natural logarithm of net worth (1984)	.172*** (.033)
Constant	5.311 (1.056)

Can we draw the DAG? What assumptions are made?

Tying causal inference to big theories

Conley 2001, *Sociology of Education*

DISCUSSION

Parkin (1979:47–48) argued that "in modern capitalist society the two main exclusionary devices by which the bourgeoisie constructs and maintains itself as a class are, first, those surrounding the institutions of property; and second, academic or professional qualifications and credentials." This article has shown that these two "exclusionary devices" are not independent of each other, since parents may use wealth—that is, property—to finance their children's educational and professional credentials, thereby solidifying their class position on the human capital dimension. In other words, nonhuman capital (property) and human capital are linked across generations. The analysis presented here demonstrated the impact of parental wealth on the educational outcomes of young adults, specifically in the transition to postsecondary schooling.

Example 7: Divorce and child development

The Causal Effects of Father Absence

Sara McLanahan,[1] Laura Tach,[2] and Daniel Schneider[3]

[1]Office of Population Research, Princeton University, Princeton, New Jersey 08544; email: mclanaha@princeton.edu

[2]Department of Policy Analysis and Management, Cornell University, Ithaca, New York 14853; email: lauratach@cornell.edu

[3]Department of Sociology and Robert Wood Johnson Scholars in Health Policy Research Program, University of California, Berkeley, California 94720; email: djschneider@berkeley.edu

Annual Review of Sociology piece summarizes many causal research designs (it's a good overview). We will focus on one.

Example 7: Divorce and child development

Cherlin, Andrew J., Frank F. Furstenberg, Jr., P. Lindsay Chase-Linsdale, Kathleen E. Kiernan, Philip K. Robins, Donna Ruane Morrison and Julien O. Teitler. "Longitudinal Studies of Effects of Divorce on Children in Great Britain and the United States." *Science* 252:1386-1389.

Cherlin et al. 1991

- **Research question:** Is divorce bad for kids?

Cherlin et al. 1991

- **Research question:** Is divorce bad for kids?
- **Controls:** Social class, race, mother employed outside the home in 1976, outcome measured in 1976

Cherlin et al. 1991

- **Research question:** Is divorce bad for kids?
- **Controls:** Social class, race, mother employed outside the home in 1976, outcome measured in 1976
- **Treatment:** Parental divorce in 1976-1981

Cherlin et al. 1991

- **Research question:** Is divorce bad for kids?
- **Controls:** Social class, race, mother employed outside the home in 1976, outcome measured in 1976
- **Treatment:** Parental divorce in 1976-1981
- **Outcome:** Behavior problems in 1981

Cherlin et al. 1991

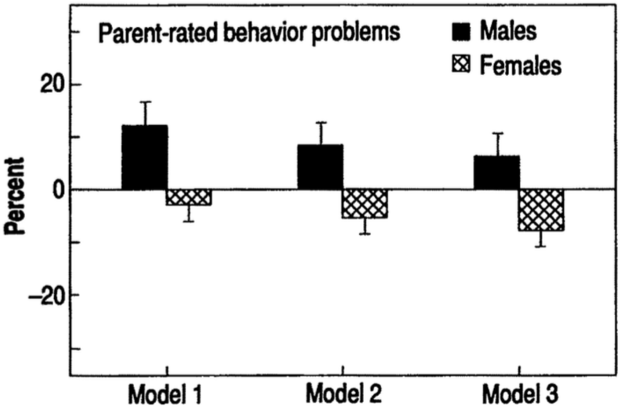
- **Research question:** Is divorce bad for kids?
- **Controls:** Social class, race, mother employed outside the home in 1976, outcome measured in 1976
- **Treatment:** Parental divorce in 1976-1981
- **Outcome:** Behavior problems in 1981

Cherlin et al. 1991

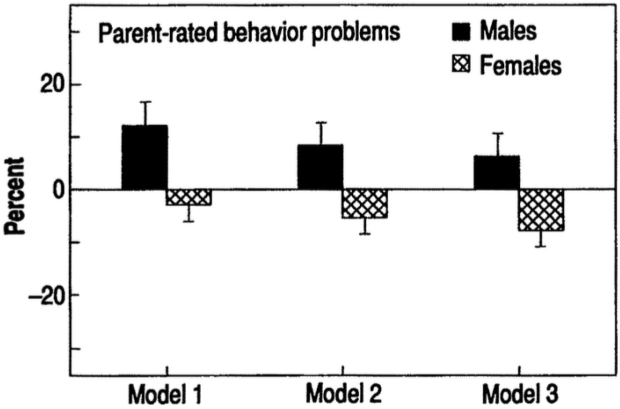
- **Research question:** Is divorce bad for kids?
- **Controls:** Social class, race, mother employed outside the home in 1976, outcome measured in 1976
- **Treatment:** Parental divorce in 1976-1981
- **Outcome:** Behavior problems in 1981

Can we draw the DAG? Write the potential outcomes? Critique the paper?

A 1991-era way of showing results



A 1991-era way of showing results



How could this figure be improved?

A 1991-era way of showing results

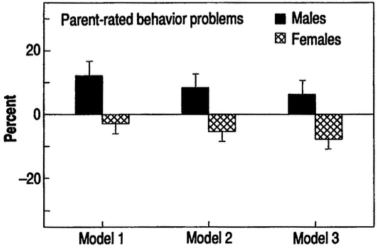


Fig. 2. Effects of a parental divorce between 1976 and 1981 on the behavior problems of children in 1981, when the children were ages 11 to 16, based on a behavior problems scale score as reported by parents from the U.S. National Survey of Children (estimates are restricted to children living with two married parents in 1976). The height of the boxes shows the percentage by which the score of children whose parents divorced between 1976 and 1981 was greater (or less) than the score of children whose parents remained married. Three estimates of the effects of divorce are shown: model 1 controls only for social class, race, and whether the mother was employed outside the home in 1976; model 2 controls additionally for the child's score on the behavior problems scale in 1976, as reported by parents, before anyone's parents were divorced; and model 3 adds further controls for the parents' score on a nine-item marital conflict scale in 1976. Error bars represent one standard error.

Example 8: Heterogeneous treatment effects

Hill, Jennifer. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217-240.

- This is a paper that **looks very hard**

Example 8: Heterogeneous treatment effects

Hill, Jennifer. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217-240.

- This is a paper that **looks very hard**
- There are lots of equations

Example 8: Heterogeneous treatment effects

Hill, Jennifer. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217-240.

- This is a paper that **looks very hard**
- There are lots of equations
- BUT it's really just a fancy version of the **imputation estimator** Brandon showed on Wednesday!

Example 8: Heterogeneous treatment effects

Hill, Jennifer. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1):217-240.

- This is a paper that **looks very hard**
- There are lots of equations
- BUT it's really just a fancy version of the **imputation estimator** Brandon showed on Wednesday!
- You already know what you need to understand the key concepts!

Substantive question
Hill (2011)

Do home visits and child care promote child cognitive development?

- **Sample:** Low birth weight, premature infants in 1985

Substantive question

Hill (2011)

Do home visits and child care promote child cognitive development?

- **Sample:** Low birth weight, premature infants in 1985
- **Treatment:** Randomly chosen treated infants received home visits and child care

Substantive question

Hill (2011)

Do home visits and child care promote child cognitive development?

- **Sample:** Low birth weight, premature infants in 1985
- **Treatment:** Randomly chosen treated infants received home visits and child care
- **Outcome:** Cognitive test scores

Substantive question

Hill (2011)

Do home visits and child care promote child cognitive development?

- **Sample:** Low birth weight, premature infants in 1985
- **Treatment:** Randomly chosen treated infants received home visits and child care
- **Outcome:** Cognitive test scores
- **Pretreatment covariates:**

Substantive question

Hill (2011)

Do home visits and child care promote child cognitive development?

- **Sample:** Low birth weight, premature infants in 1985
- **Treatment:** Randomly chosen treated infants received home visits and child care
- **Outcome:** Cognitive test scores
- **Pretreatment covariates:**
  - Infant characteristics: birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index, sex, twin status

Substantive question

Hill (2011)

Do home visits and child care promote child cognitive development?

- **Sample:** Low birth weight, premature infants in 1985
- **Treatment:** Randomly chosen treated infants received home visits and child care
- **Outcome:** Cognitive test scores
- **Pretreatment covariates:**
    - Infant characteristics: birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index, sex, twin status
    - Mother variables in pregnancy: smoked cigarettes, drank alcohol, took drugs

Substantive question

Hill (2011)

Do home visits and child care promote child cognitive development?

- **Sample:** Low birth weight, premature infants in 1985
- **Treatment:** Randomly chosen treated infants received home visits and child care
- **Outcome:** Cognitive test scores
- **Pretreatment covariates:**
  - Infant characteristics: birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index, sex, twin status
  - Mother variables in pregnancy: smoked cigarettes, drank alcohol, took drugs
  - Mother variables at birth: age, marital status, educational attainment, whether she worked during pregnancy, whether she received prenatal care

Substantive question

Hill (2011)

Do home visits and child care promote child cognitive development?

- **Sample:** Low birth weight, premature infants in 1985
- **Treatment:** Randomly chosen treated infants received home visits and child care
- **Outcome:** Cognitive test scores
- **Pretreatment covariates:**
  - Infant characteristics: birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health index, sex, twin status
  - Mother variables in pregnancy: smoked cigarettes, drank alcohol, took drugs
  - Mother variables at birth: age, marital status, educational attainment, whether she worked during pregnancy, whether she received prenatal care
  - A few residential location variables

Heterogeneous effects in terms of potential outcomes

Recall potential outcomes:

- Potential outcome under control: $Y_i(0) = f(X_i)$

All are functions of pre-treatment covariates.

Heterogeneous effects in terms of potential outcomes

Recall potential outcomes:

- Potential outcome under control: $Y_i(0) = f(X_i)$
- Potential outcome under control: $Y_i(1) = g(X_i)$

All are functions of pre-treatment covariates.

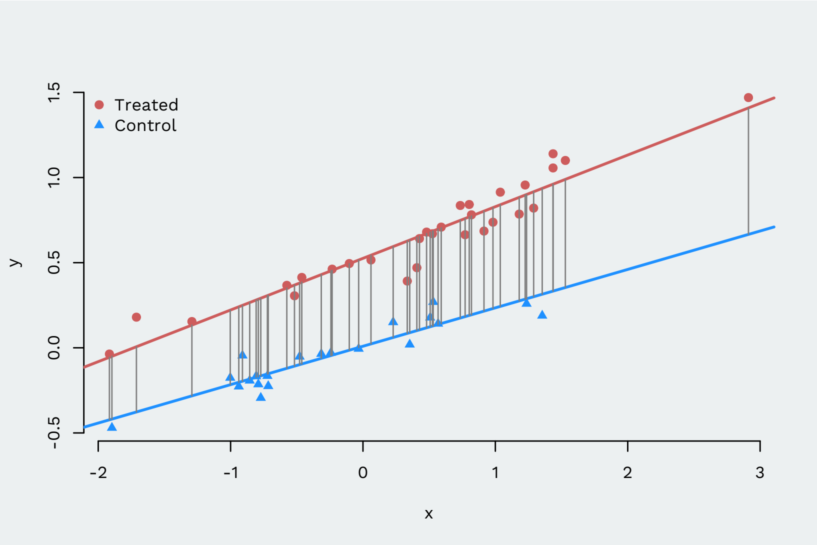
Heterogeneous effects in terms of potential outcomes

Recall potential outcomes:

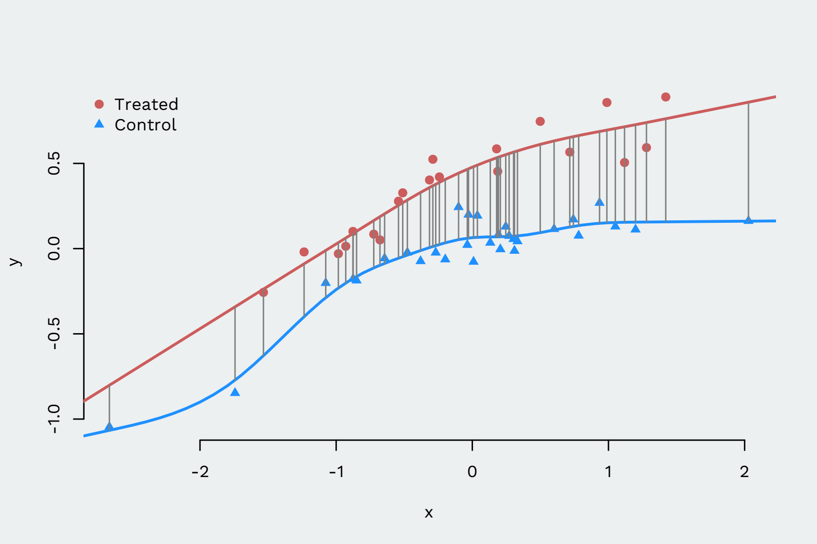
- Potential outcome under control: $Y_i(0) = f(X_i)$
- Potential outcome under control: $Y_i(1) = g(X_i)$
- The treatment effect is $\tau_i = g(X_i) - f(X_i) = h(X_i)$

All are functions of pre-treatment covariates.

Imputation approach from lecture



Imputation approach from lecture



Visualizing heterogeneous effects

nature of the algorithm, which conditions on the X values in the sample, a natural set of estimands are the conditional average treatment effect (CATE)

$$\frac{1}{n}\sum_{i=1}^{n} E(Y_i(1) \mid X_i) - E(Y(0) \mid X_i) = \frac{1}{n}\sum_{i=1}^{n} f(1, x_i) - f(0, x_i),$$

and the conditional average treatment effect for the treated (CATT)

$$\frac{1}{n_t}\sum_{i:Z_i=1} E(Y_i(1) \mid X_i) - E(Y(0) \mid X_i) = \frac{1}{n_t}\sum_{i:Z_i=1} f(1, x_i) - f(0, x_i).$$

Defining causal effects with covariate-based heterogeneity

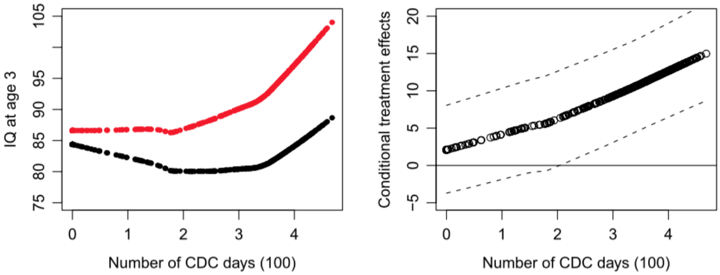


Figure 6. Left panel displays plot of BART-predicted 3-year IQ test scores against CDC participation (in hundreds of days) for children in the treatment group (upper line). The lower line shows predicted scores for the same children if they had not attended any CDC days. Lines were smoothed using lowess. The right panel displays a smoothed function of the treatment effect estimates at each level of CDC participation (conditional on having that level of participation in the treatment group). Dashed lines represent 95% uncertainty bounds. A color version of this figure is available in the electronic version of this article.

Example 9: Contagion in social networks

Christakis, Nicholas A., and James H. Fowler. 2007. "The Spread of Obesity in a Large Social Network over 32 Years." *New England Journal of Medicine* 357(4):370-379.

Also a related book, which is a good read.

Christakis and Fowler 2007

- **Research question:** Does having obese friends cause you to become obese?

Christakis and Fowler 2007

- **Research question:** Does having obese friends cause you to become obese?
- **Sample:** Framingham Heart Study, 1971-2003, following 12,067 people

Christakis and Fowler 2007

- **Research question:** Does having obese friends cause you to become obese?
- **Sample:** Framingham Heart Study, 1971-2003, following 12,067 people
- **Measured confounders:** Ego's age, sex, education

Christakis and Fowler 2007

- **Research question:** Does having obese friends cause you to become obese?
- **Sample:** Framingham Heart Study, 1971-2003, following 12,067 people
- **Measured confounders:** Ego's age, sex, education
- **Lagged dependent variable:** obesity at $t - 1$ (control for U)

Christakis and Fowler 2007

- **Research question:** Does having obese friends cause you to become obese?
- **Sample:** Framingham Heart Study, 1971-2003, following 12,067 people
- **Measured confounders:** Ego's age, sex, education
- **Lagged dependent variable:** obesity at $t-1$ (control for U)
- **Lagged predictor:** Alter's weight at $t-1$ (control for homophily)

Christakis and Fowler 2007

- **Research question:** Does having obese friends cause you to become obese?
- **Sample:** Framingham Heart Study, 1971-2003, following 12,067 people
- **Measured confounders:** Ego's age, sex, education
- **Lagged dependent variable:** obesity at $t-1$ (control for U)
- **Lagged predictor:** Alter's weight at $t-1$ (control for homophily)
- **Treatment:** Alter's obesity at $t+1$

Christakis and Fowler 2007

- **Research question:** Does having obese friends cause you to become obese?
- **Sample:** Framingham Heart Study, 1971-2003, following 12,067 people
- **Measured confounders:** Ego's age, sex, education
- **Lagged dependent variable:** obesity at $t-1$ (control for U)
- **Lagged predictor:** Alter's weight at $t-1$ (control for homophily)
- **Treatment:** Alter's obesity at $t+1$
- **Outcome:** Ego's obesity at $t+1$

Christakis and Fowler 2007: Conclusion

"A person's chances of becoming obese increased by 57% (95% confidence interval [CI], 6 to 123) if he or she had a friend who became obese in a given interval." (quoted from abstract)

Regression weighting from lecture

Slide by Brandon Stewart

$$W_i = \frac{\sigma_d^2(X_i)}{E[\sigma_d^2(X_i)]}$$

- Why does OLS weight like this?
- OLS is a **minimum-variance estimator** $\rightsquigarrow$ more weight to more precise within-strata estimates.
- Within-strata estimates are most precise when the treatment is evenly spread and thus has the highest variance.
- If D_i is binary, then we know the conditional variance will be:

$$\sigma_d^2(x) = P(D_i = 1 \mid X_i = x)[1 - P(D_i = 1 \mid X_i = x)]$$

- Maximum variance with $P[D_i = 1 | X_i = x] = 1/2$.

OLS weighting example
Slide by Brandon Stewart

- Binary covariate:

$$\begin{aligned} \mathbb{P}[X_i = 1] &= 0.75 & \mathbb{P}[X_i = 0] &= 0.25 \\ \mathbb{P}[D_i = 1 | X_i = 1] &= 0.9 & \mathbb{P}[D_i = 1 | X_i = 0] &= 0.5 \\ \sigma_d^2(1) &= 0.09 & \sigma_d^2(0) &= 0.25 \\ \tau(1) &= 1 & \tau(0) &= -1 \end{aligned}$$

- Implies the ATE is $\tau = 0.5$
- Average conditional variance: $\mathbb{E}[\sigma_d^2(X_i)] = 0.13$
- $\rightsquigarrow$ weights for $X_i = 1$ are: $0.09/0.13 = 0.692$, for $X_i = 0$: $0.25/0.13 = 1.92$.

$$\begin{aligned} \tau_R &= \mathbb{E}[\tau(X_i) W_i] \\ &= \tau(1) W(1) \mathbb{P}[X_i = 1] + \tau(0) W(0) \mathbb{P}[X_i = 0] \\ &= 1 \times 0.692 \times 0.75 + -1 \times 1.92 \times 0.25 \\ &= 0.039 \end{aligned}$$