Advanced Quantitative Techniques

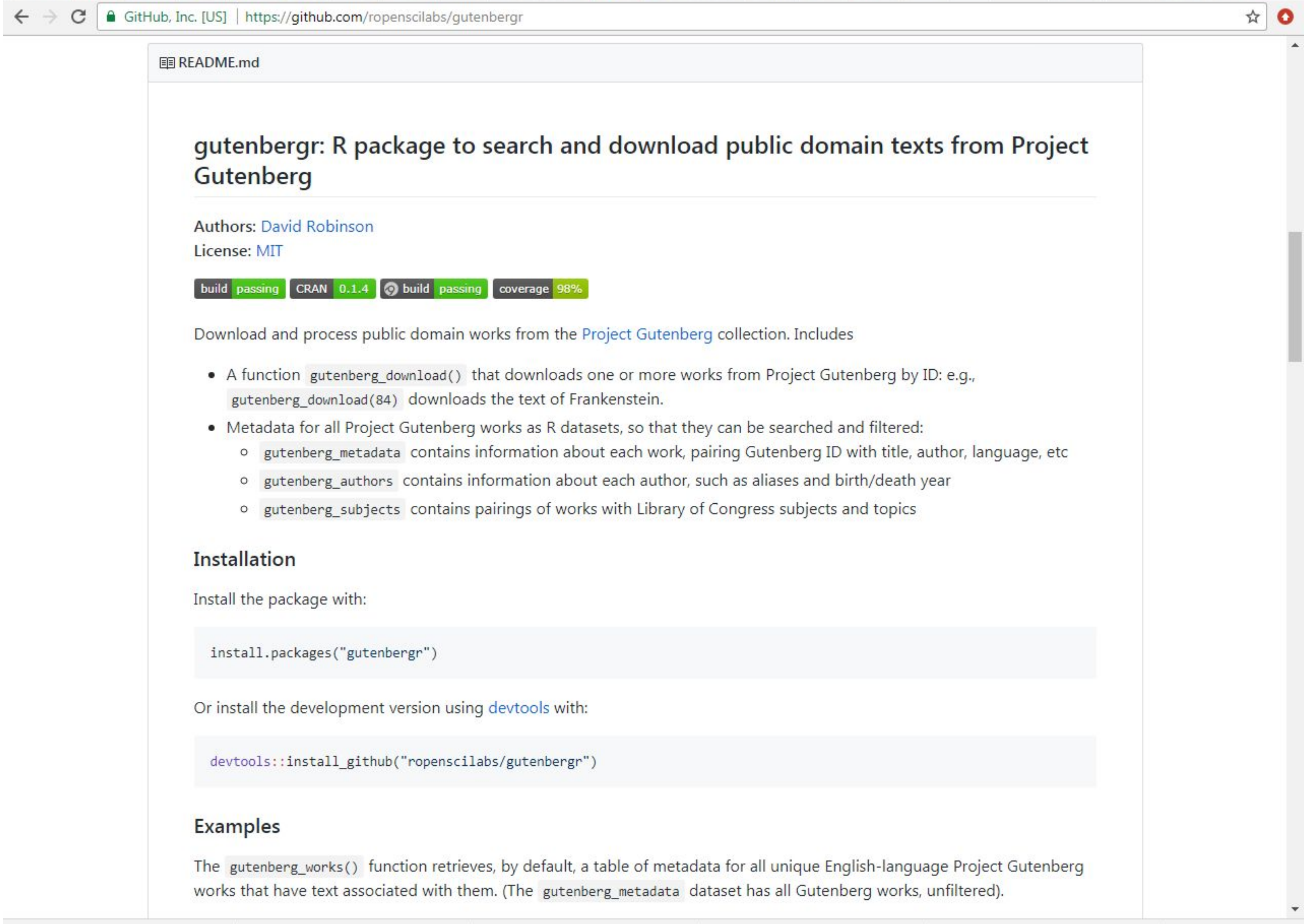
(Class 8)

Gregory M. Eirich

QMSS

Where to get text data?

Literature



GitHub, Inc. [US] | https://github.com/ropenscilabs/gutenbergr

gutenbergr: R package to search and download public domain texts from Project Gutenberg

Authors: David Robinson
License: MIT

build passing
CRAN 0.1.4
build passing
coverage 98%

Download and process public domain works from the Project Gutenberg collection. Includes

- A function `gutenberg_download()` that downloads one or more works from Project Gutenberg by ID: e.g., `gutenberg_download(84)` downloads the text of Frankenstein.
- Metadata for all Project Gutenberg works as R datasets, so that they can be searched and filtered:
  - `gutenberg_metadata` contains information about each work, pairing Gutenberg ID with title, author, language, etc
  - `gutenberg_authors` contains information about each author, such as aliases and birth/death year
  - `gutenberg_subjects` contains pairings of works with Library of Congress subjects and topics

Installation

Install the package with:

```
install.packages("gutenbergr")
```

Or install the development version using devtools with:

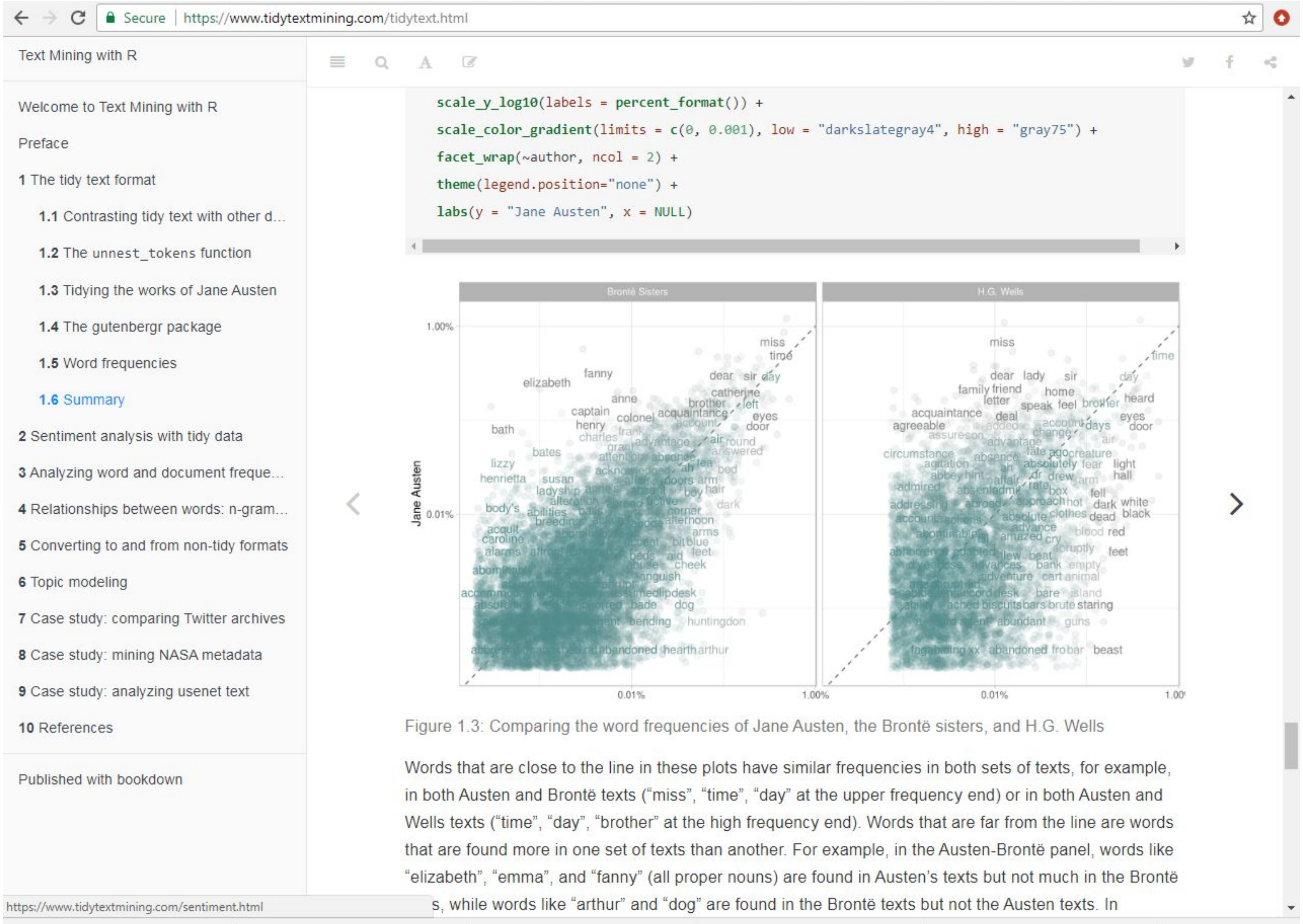
```
devtools::install_github("ropenscilabs/gutenbergr")
```

Examples

The `gutenberg_works()` function retrieves, by default, a table of metadata for all unique English-language Project Gutenberg works that have text associated with them. (The `gutenberg_metadata` dataset has all Gutenberg works, unfiltered).

*

Literature - example ...



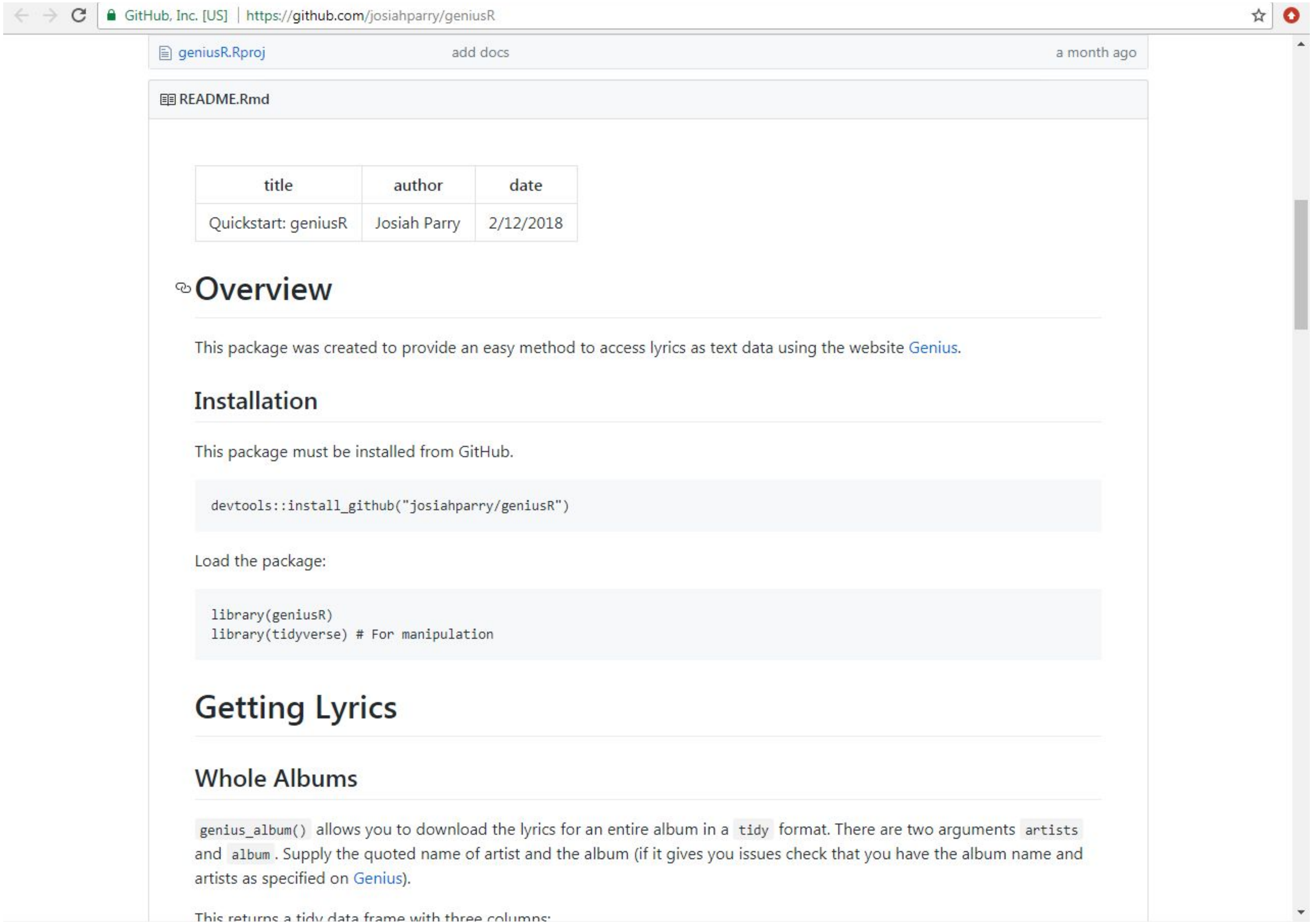
```
scale_y_log10(labels = percent_format()) +
scale_color_gradient(limits = c(0, 0.001), low = "darkslategray4", high = "gray75") +
facet_wrap(~author, ncol = 2) +
theme(legend.position="none") +
labs(y = "Jane Austen", x = NULL)
```

Figure 1.3: Comparing the word frequencies of Jane Austen, the Brontë sisters, and H.G. Wells

Words that are close to the line in these plots have similar frequencies in both sets of texts, for example, in both Austen and Brontë texts (“miss”, “time”, “day” at the upper frequency end) or in both Austen and Wells texts (“time”, “day”, “brother” at the high frequency end). Words that are far from the line are words that are found more in one set of texts than another. For example, in the Austen-Brontë panel, words like “elizabeth”, “emma”, and “fanny” (all proper nouns) are found in Austen's texts but not much in the Brontë s, while words like “arthur” and “dog” are found in the Brontë texts but not the Austen texts. In

*

Music lyrics



geniusR.Rproj add docs a month ago

README.Rmd

title	author	date
Quickstart: geniusR	Josiah Parry	2/12/2018

Overview

This package was created to provide an easy method to access lyrics as text data using the website Genius.

Installation

This package must be installed from GitHub.

```
devtools::install_github("josiahparry/geniusR")
```

Load the package:

```
library(geniusR)
library(tidyverse) # For manipulation
```

Getting Lyrics

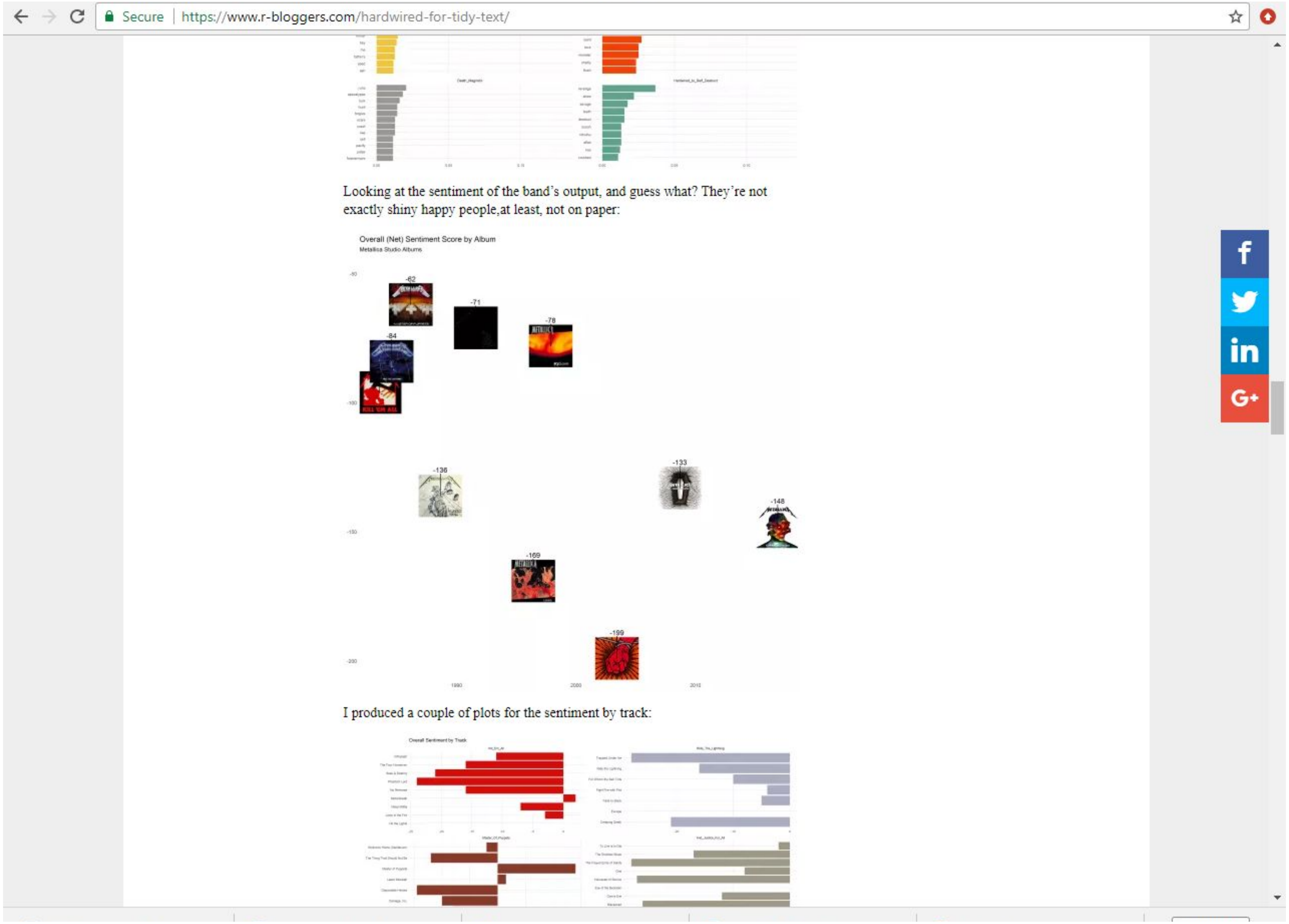
Whole Albums

`genius_album()` allows you to download the lyrics for an entire album in a `tidy` format. There are two arguments `artists` and `album`. Supply the quoted name of artist and the album (if it gives you issues check that you have the album name and artists as specified on Genius).

This returns a tidy data frame with three columns:

*

Music lyrics example ("hardwired … for tidytext")

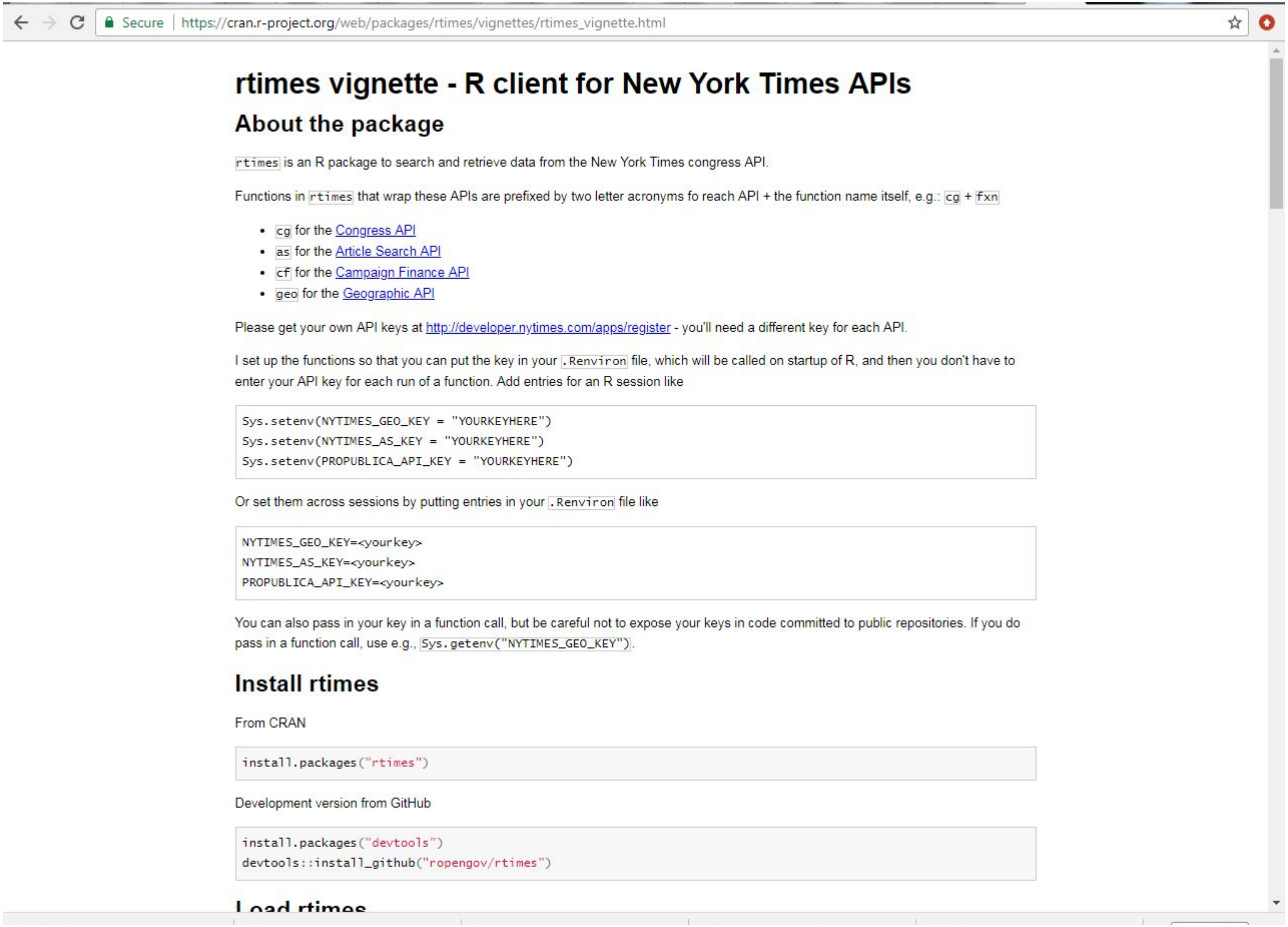


Looking at the sentiment of the band's output, and guess what? They're not exactly shiny happy people,at least, not on paper:

I produced a couple of plots for the sentiment by track:

*

News



https://cran.r-project.org/web/packages/rtimes/vignettes/rtimes_vignette.html

rtimes vignette - R client for New York Times APIs

About the package

`rtimes` is an R package to search and retrieve data from the New York Times congress API.

Functions in `rtimes` that wrap these APIs are prefixed by two letter acronyms fo reach API + the function name itself, e.g.: `cg` + `fxn`

- `cg` for the Congress API
- `as` for the Article Search API
- `cf` for the Campaign Finance API
- `geo` for the Geographic API

Please get your own API keys at http://developer.nytimes.com/apps/register - you'll need a different key for each API.

I set up the functions so that you can put the key in your `.Renviron` file, which will be called on startup of R, and then you don't have to enter your API key for each run of a function. Add entries for an R session like

```
Sys.setenv(NYTIMES_GEO_KEY = "YOURKEYHERE")
Sys.setenv(NYTIMES_AS_KEY = "YOURKEYHERE")
Sys.setenv(PROPUBLICA_API_KEY = "YOURKEYHERE")
```

Or set them across sessions by putting entries in your `.Renviron` file like

```
NYTIMES_GEO_KEY=<yourkey>
NYTIMES_AS_KEY=<yourkey>
PROPUBLICA_API_KEY=<yourkey>
```

You can also pass in your key in a function call, but be careful not to expose your keys in code committed to public repositories. If you do pass in a function call, use e.g., `Sys.getenv("NYTIMES_GEO_KEY")`.

Install rtimes

From CRAN

```
install.packages("rtimes")
```

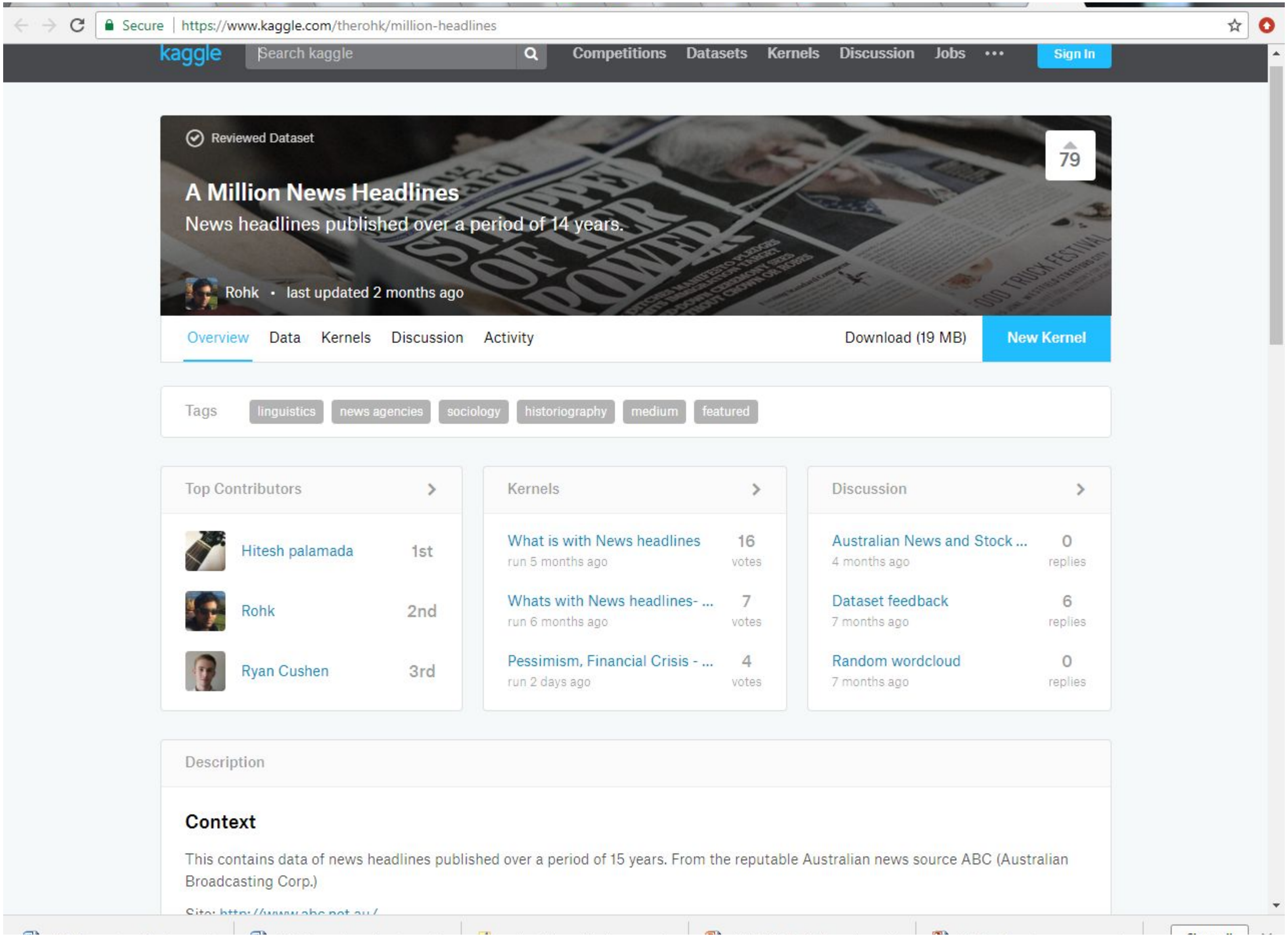
Development version from GitHub

```
install.packages("devtools")
devtools::install_github("ropengov/rtimes")
```

Load rtimes

*

News - another source...



A Million News Headlines

News headlines published over a period of 14 years.

Rohk • last updated 2 months ago

Tags: linguistics, news agencies, sociology, historiography, medium, featured

Top Contributors

Hitesh palamada	1st
Rohk	2nd
Ryan Cushen	3rd

Kernels

What is with News headlines run 5 months ago	16 votes
Whats with News headlines- ... run 6 months ago	7 votes
Pessimism, Financial Crisis - ... run 2 days ago	4 votes

Discussion

Australian News and Stock ... 4 months ago	0 replies
Dataset feedback 7 months ago	6 replies
Random wordcloud 7 months ago	0 replies

Description

Context

This contains data of news headlines published over a period of 15 years. From the reputable Australian news source ABC (Australian Broadcasting Corp.)

Site: http://www.abc.net.au/

*

1. Instrumental variables (IVs)
2. Two Stage Least Squares (2SLS)

1. Instrumental variables (IVs)

Our old nemesis: Omitted variables

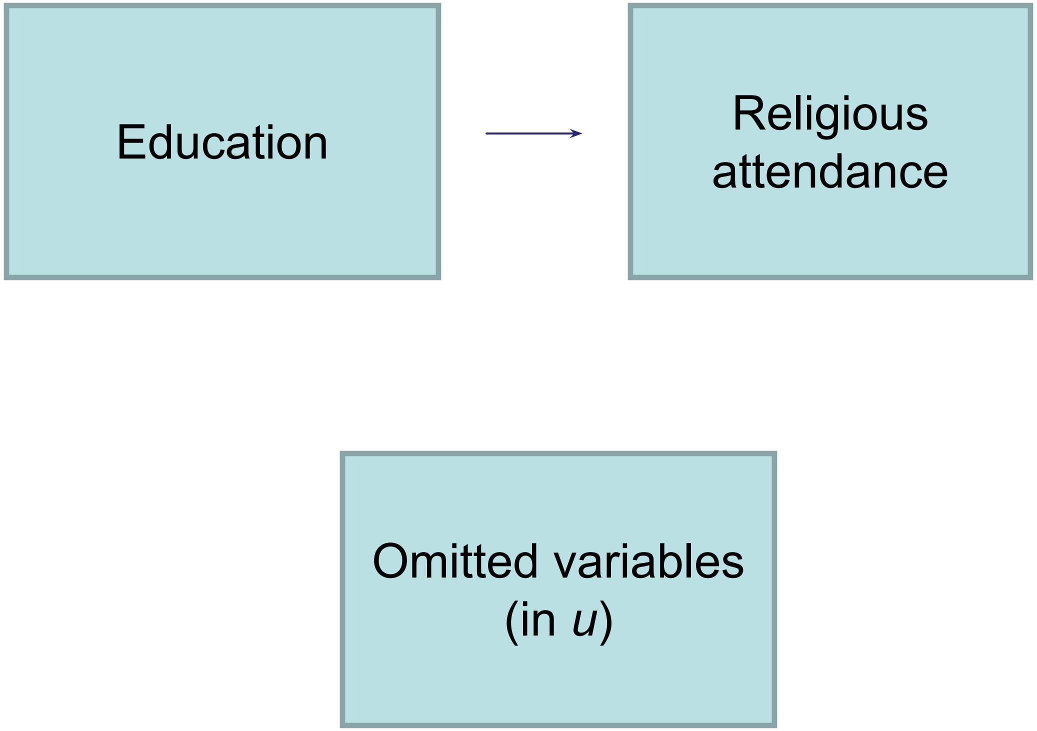
- We have our equation:

$$Attend = \beta_0 + \beta_1 educ + u$$

- But we think that education is endogenous to attendance, not exogenous:

$$cov(educ, u) \neq 0$$

Why would we think that education is endogenous to religious attendance?

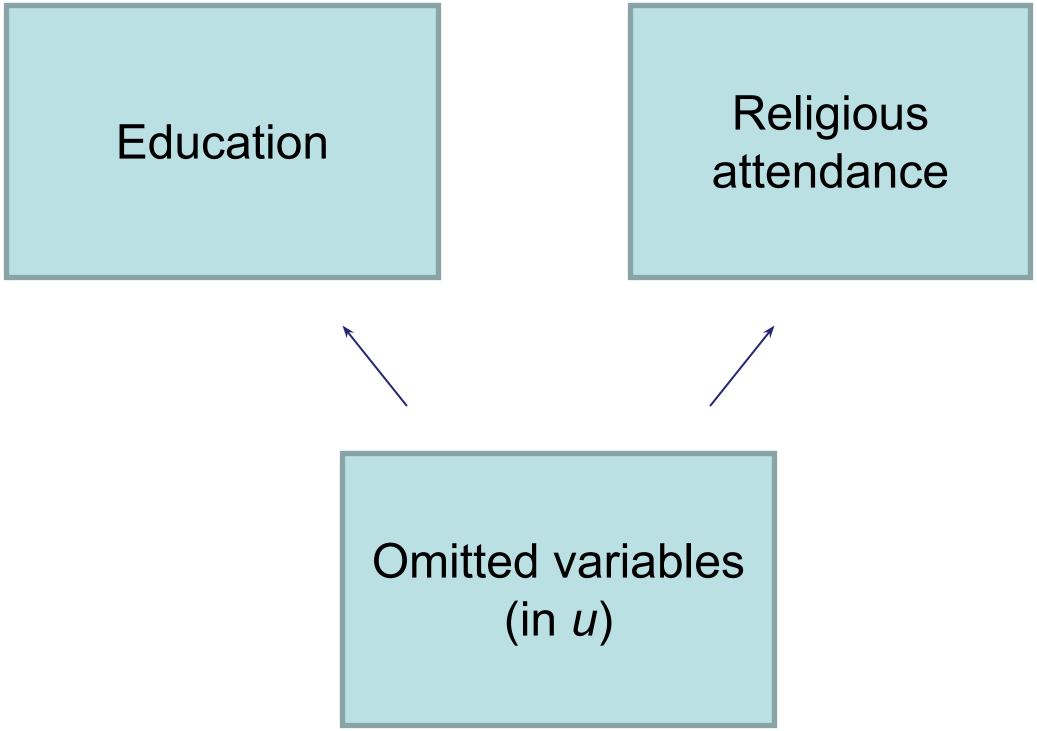


Education

Religious attendance

Omitted variables (in *u*)

Why would we think that education is endogenous to religious attendance?



Education

Religious attendance

Omitted variables (in *u*)

What can we do?

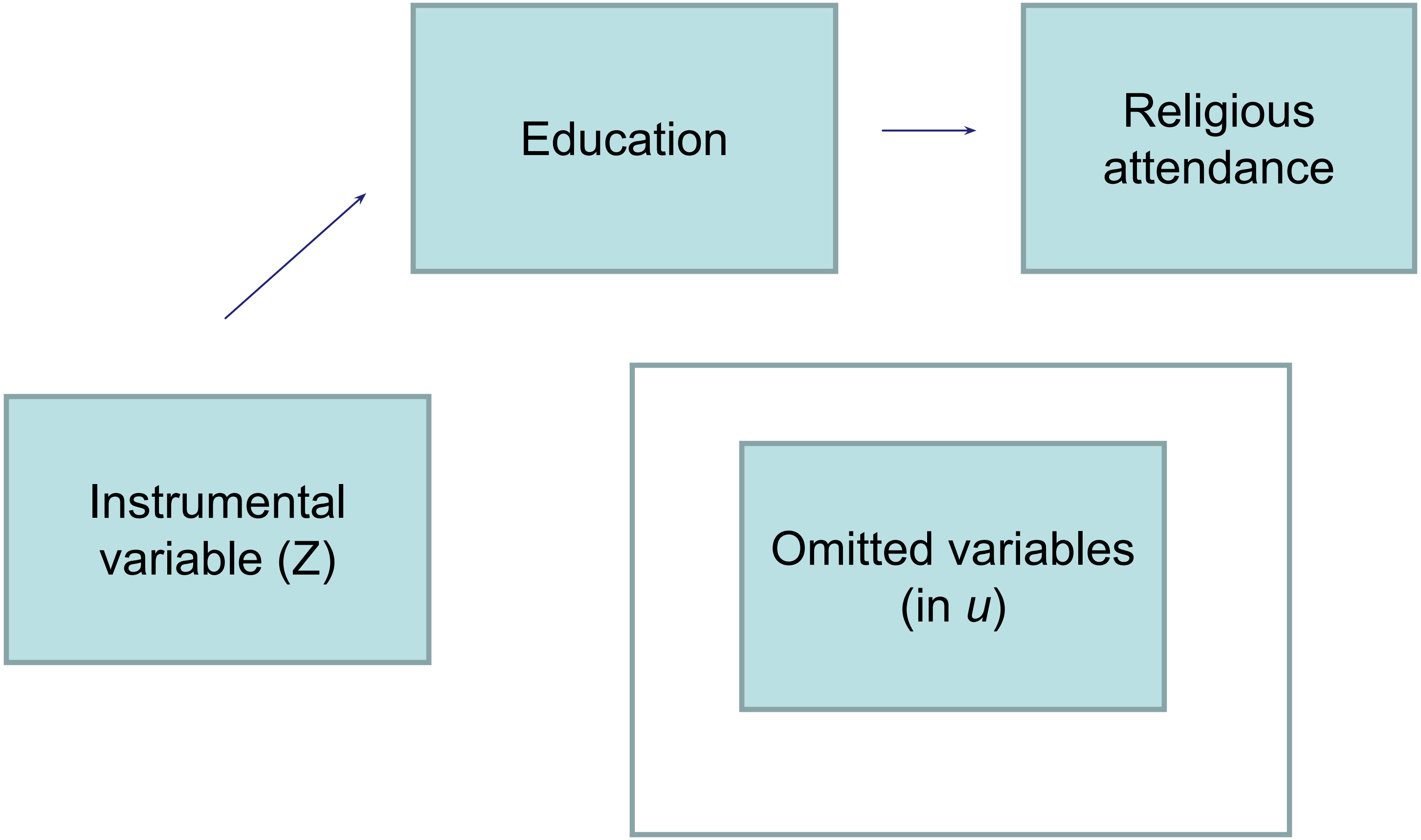
1. Nothing. Just run OLS and explain the nature of the likely bias.
2. Find a proxy variable.
3. Find a suitable instrument for our X variable of interest.

Find a suitable instrument. How?

Find an observable variable (*Z*) that is:

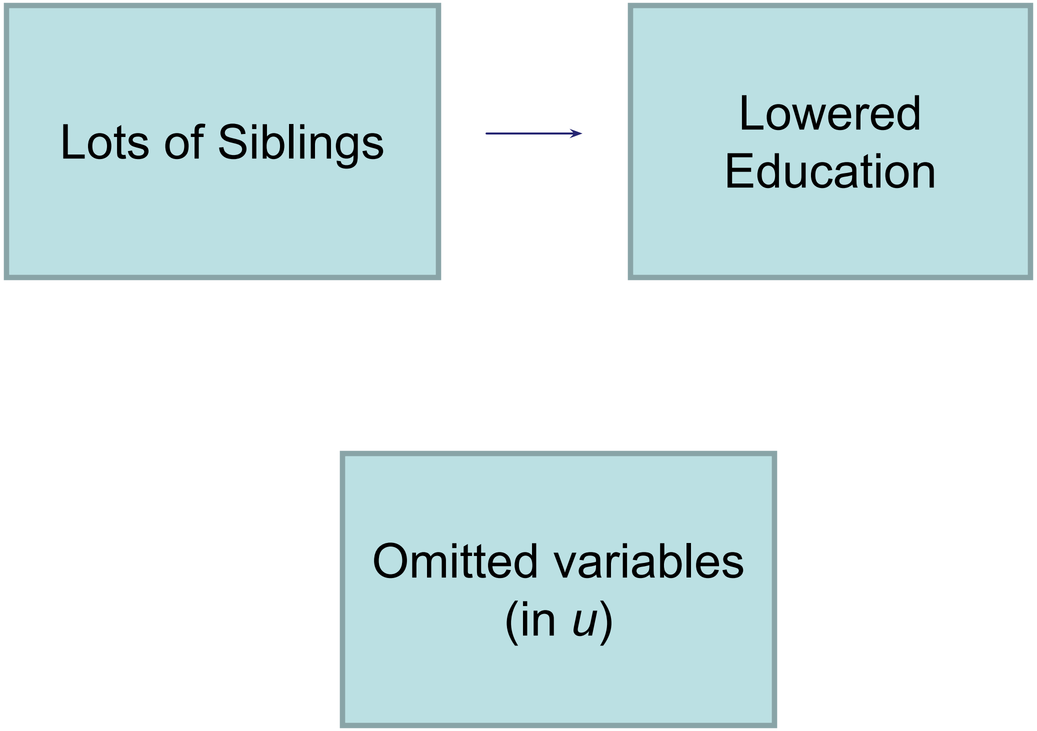
- Correlated with our X, but …
- Not correlated with *u*

What would be a good instrument for education to predict religious attendance?



Some examples of famous IVs

Example 1: Sibling dilution of resources effects – Theory 1

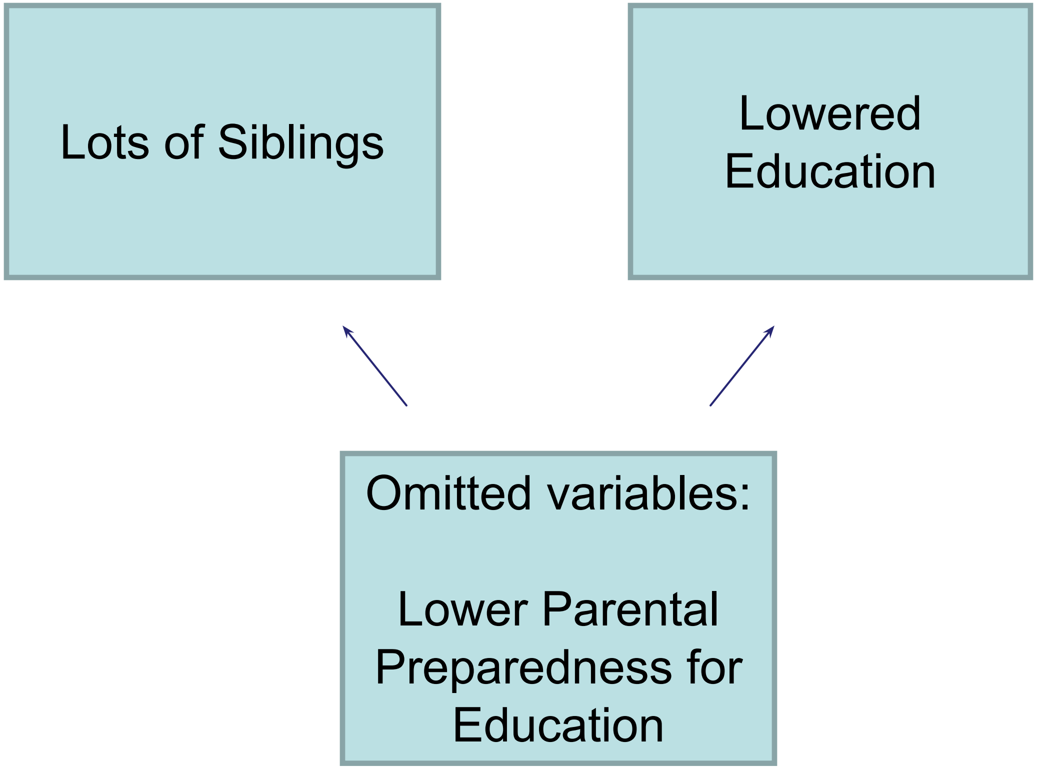


Lots of Siblings

Lowered Education

Omitted variables (in u)

Example 1: Sibling dilution of resources effects – Theory 2



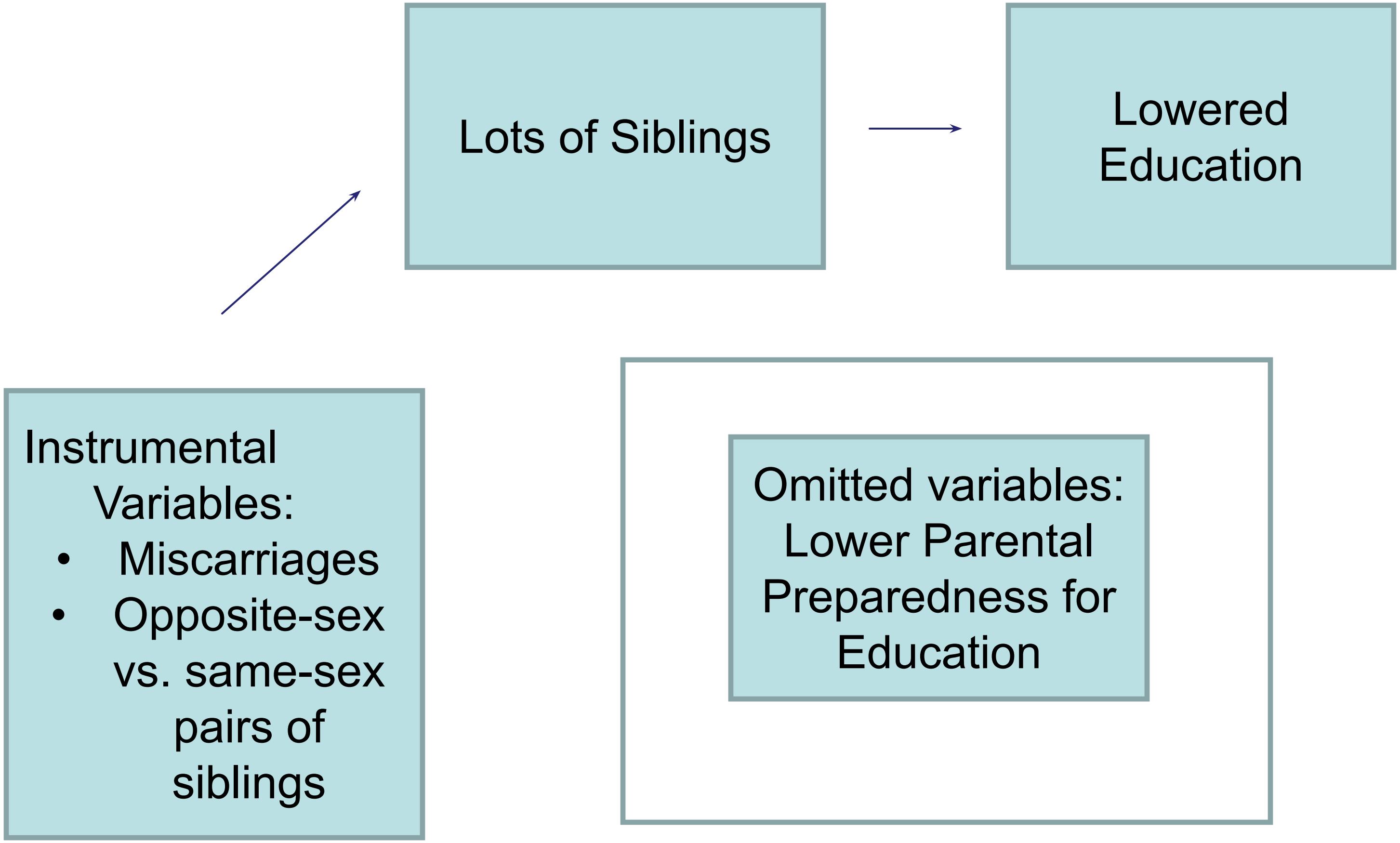
Lots of Siblings

Lowered Education

Omitted variables:

Lower Parental Preparedness for Education

Example 1: Sibling dilution of resources effects – Theory 3



Lots of Siblings

Lowered Education

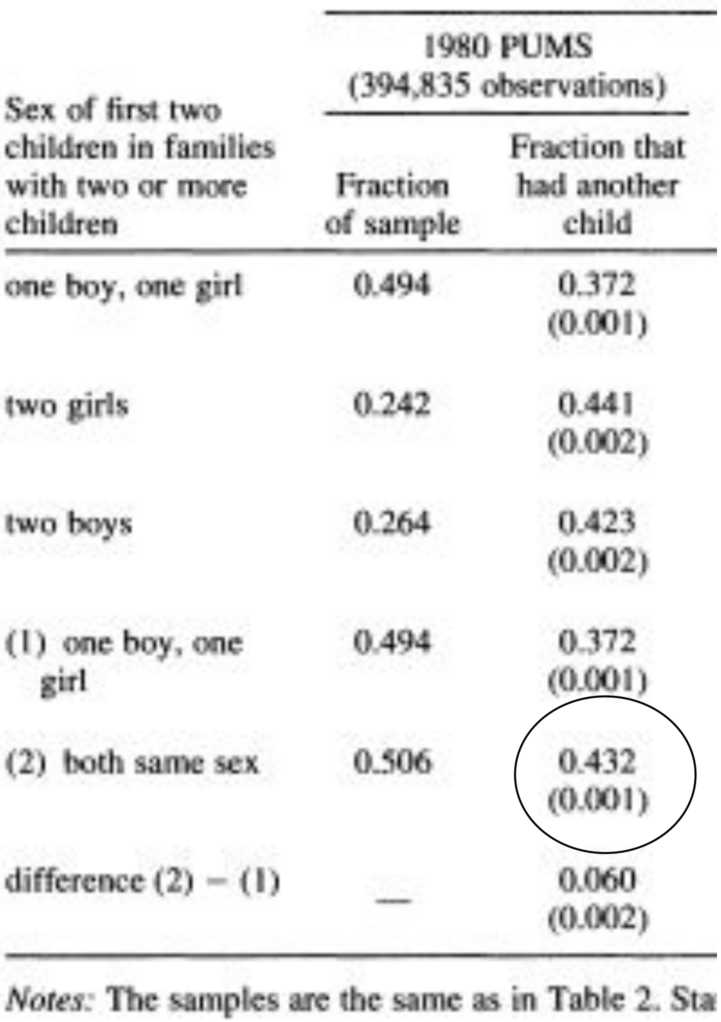
Instrumental Variables:

- Miscarriages
- Opposite-sex vs. same-sex pairs of siblings

Omitted variables: Lower Parental Preparedness for Education

Child sex composition as IV

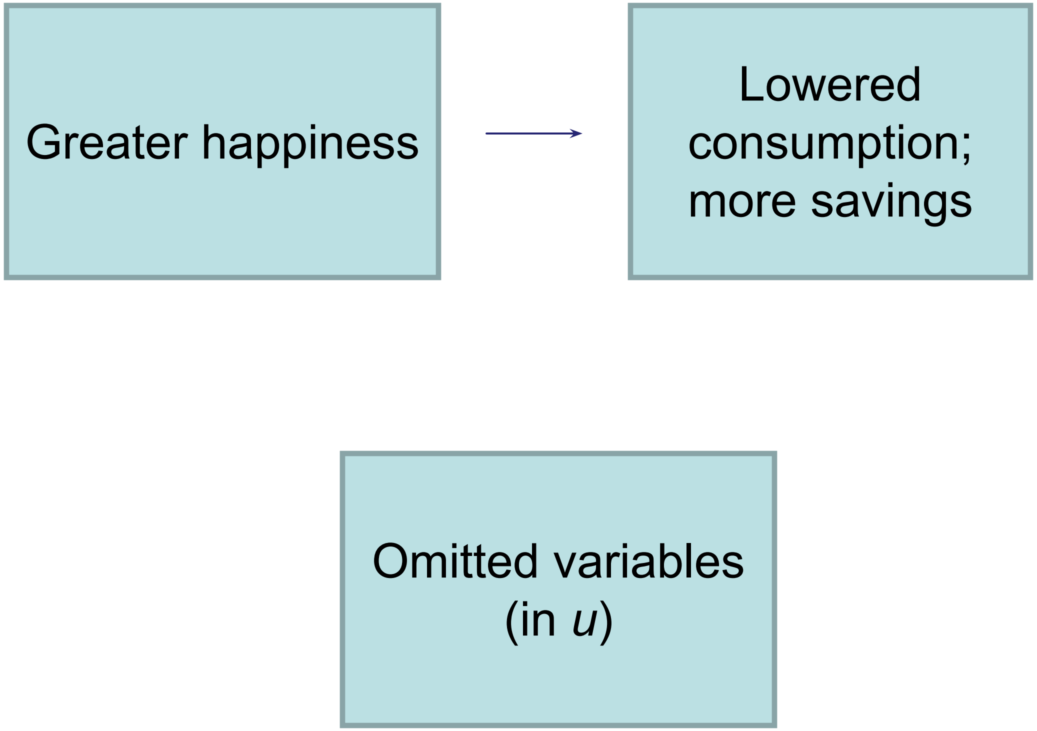
Angrist, Joshua D., and William N. Evans. *Children and their parents' labor supply: Evidence from exogenous variation in family size*. No. w5778. National bureau of economic research, 1996.



Sex of first two children in families with two or more children	1980 PUMS (394,835 observations)	
	Fraction of sample	Fraction that had another child
one boy, one girl	0.494	0.372 (0.001)
two girls	0.242	0.441 (0.002)
two boys	0.264	0.423 (0.002)
(1) one boy, one girl	0.494	0.372 (0.001)
(2) both same sex	0.506	0.432 (0.001)
difference (2) − (1)	—	0.060 (0.002)

Notes: The samples are the same as in Table 2. Sta

Example 2: Effects of happiness on consumption and saving – Theory 1

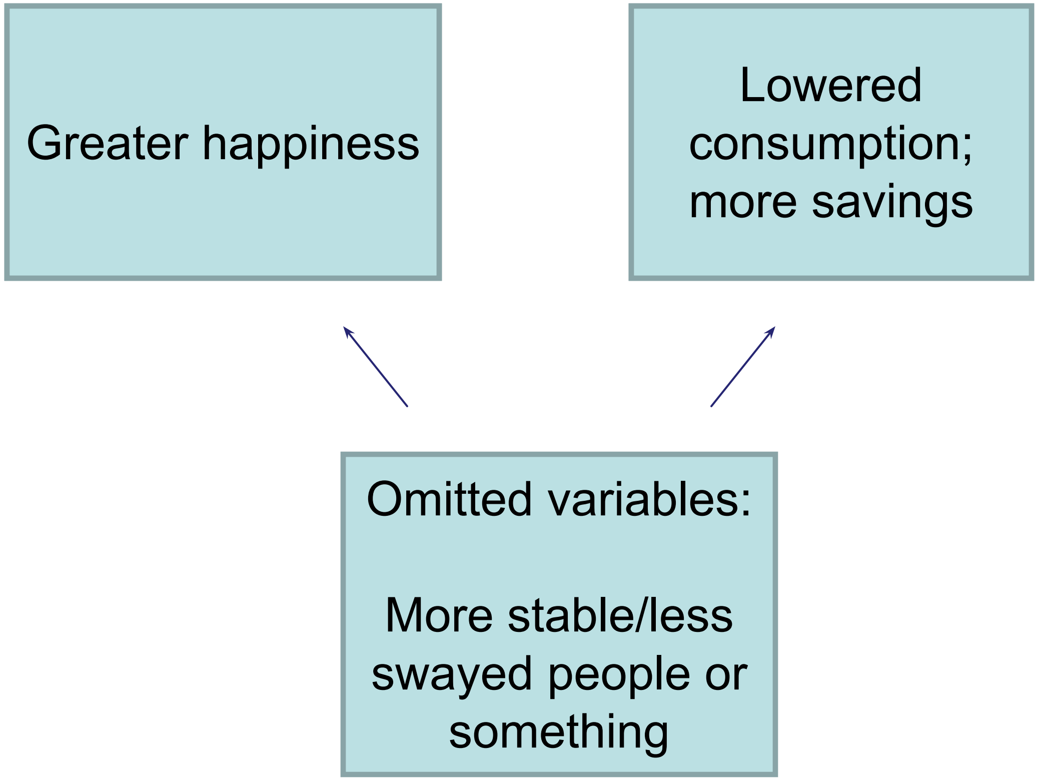


Greater happiness

Lowered consumption; more savings

Omitted variables (in *u*)

Example 2: Effects of happiness on consumption and saving – Theory 2

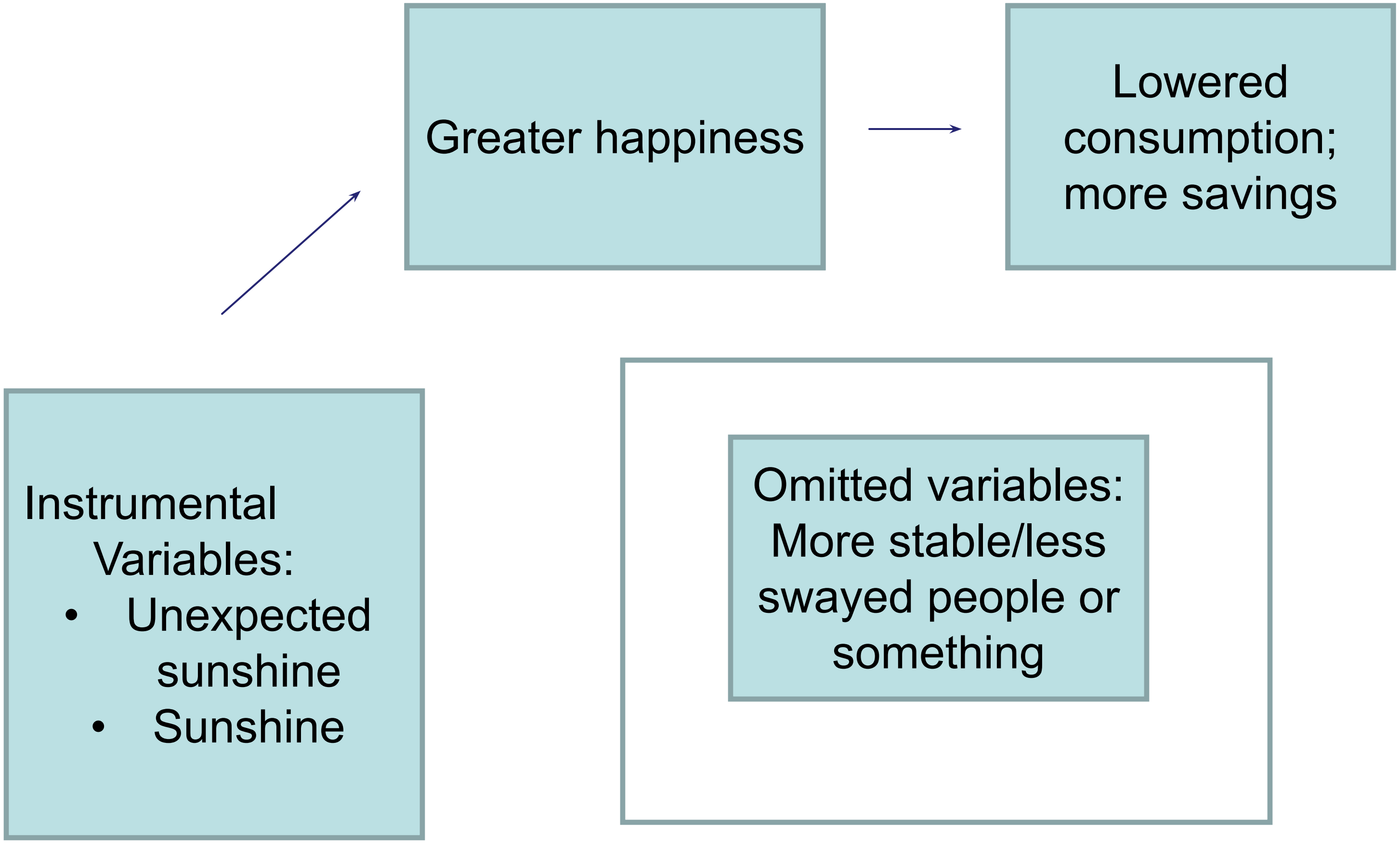

Greater happiness

Lowered consumption; more savings

Omitted variables:

More stable/less swayed people or something

Example 2: Effects of happiness on consumption and saving – Theory 3



Greater happiness

Lowered consumption; more savings

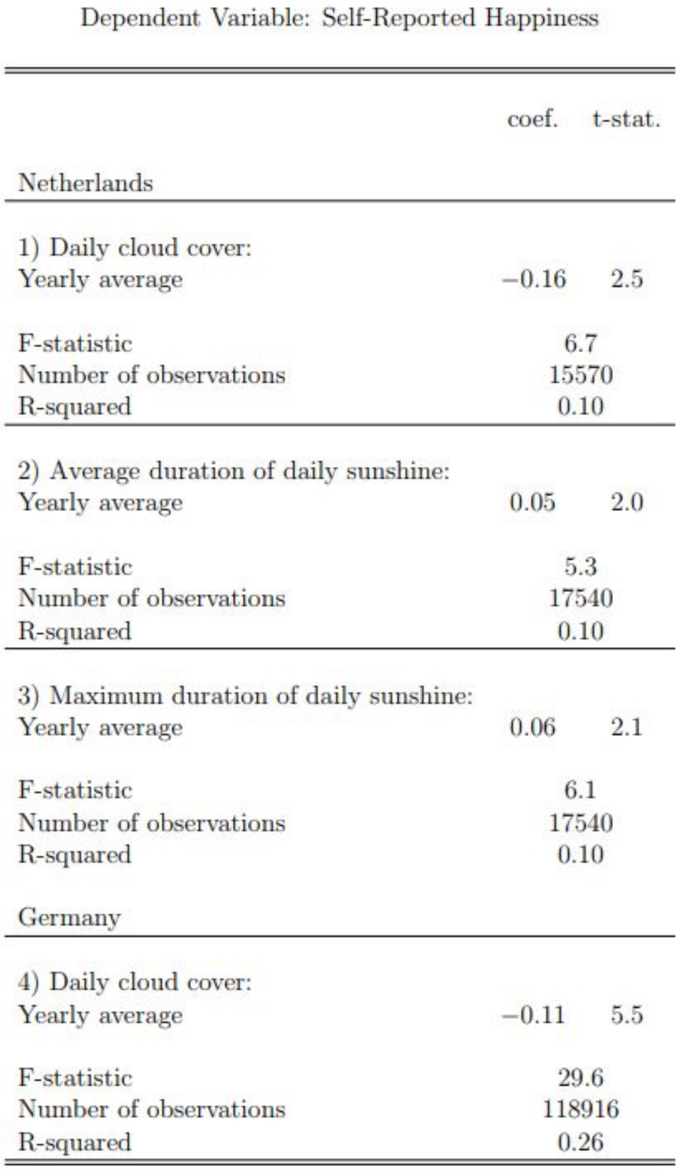
Instrumental Variables:

- Unexpected sunshine
- Sunshine

Omitted variables: More stable/less swayed people or something

Unexpected weather as IV

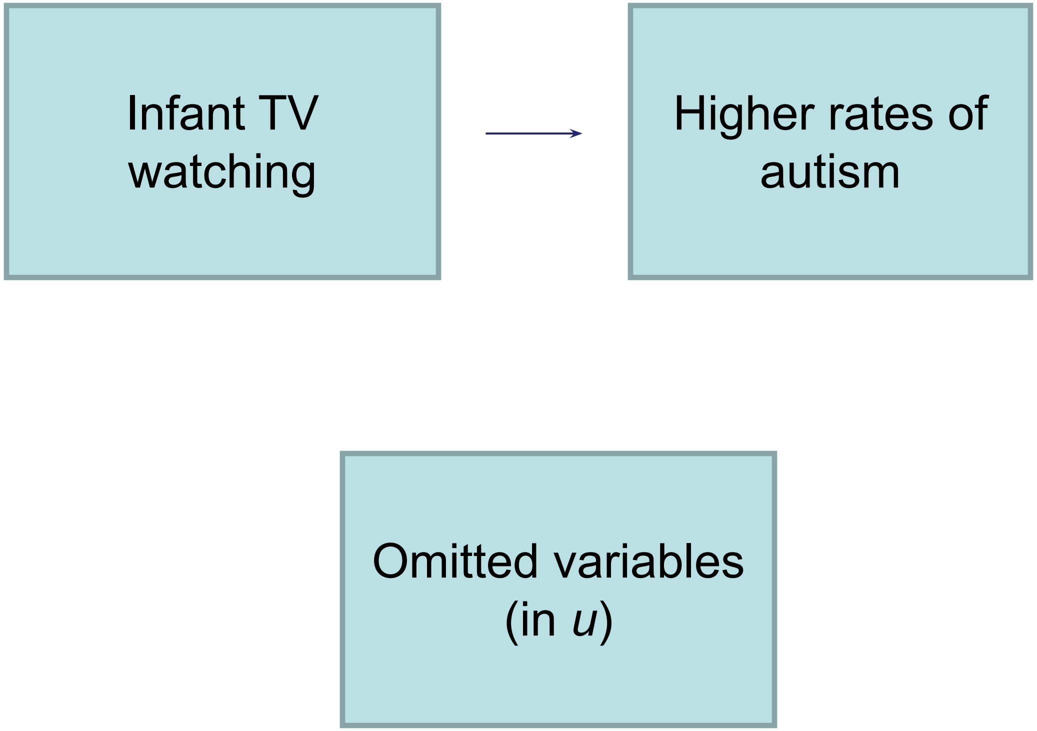
Guven, Cahit. "Reversing the question: Does happiness affect consumption and savings behavior?." *Journal of Economic Psychology* 33.4 (2012): 701-717.



Dependent Variable: Self-Reported Happiness

	coef.	t-stat.
Netherlands		
1) Daily cloud cover:		
Yearly average	−0.16	2.5
F-statistic	6.7	
Number of observations	15570	
R-squared	0.10	
2) Average duration of daily sunshine:		
Yearly average	0.05	2.0
F-statistic	5.3	
Number of observations	17540	
R-squared	0.10	
3) Maximum duration of daily sunshine:		
Yearly average	0.06	2.1
F-statistic	6.1	
Number of observations	17540	
R-squared	0.10	
Germany		
4) Daily cloud cover:		
Yearly average	−0.11	5.5
F-statistic	29.6	
Number of observations	118916	
R-squared	0.26	

Example 3: Infant TV leads to autism – Theory 1

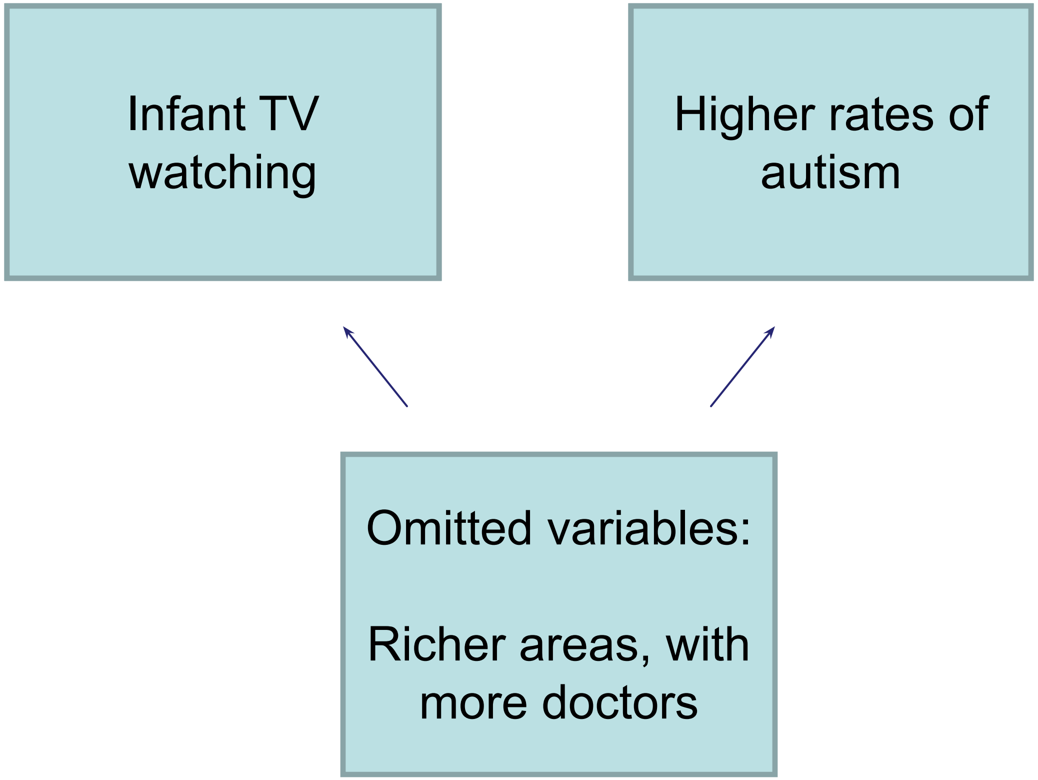


Infant TV watching

Higher rates of autism

Omitted variables (in *u*)

Example 3: Infant TV leads to autism – Theory 2



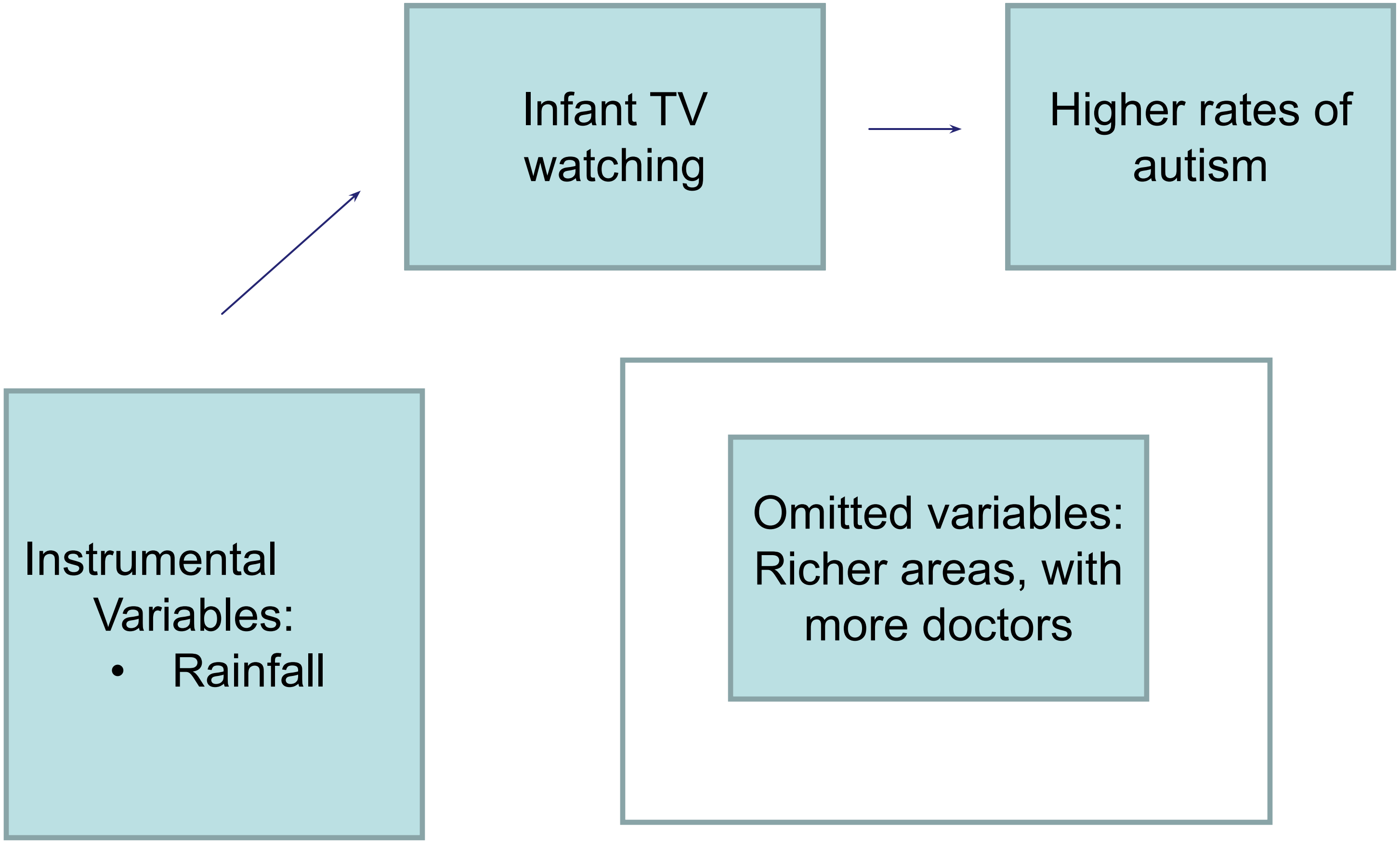
Infant TV watching

Higher rates of autism

Omitted variables:

Richer areas, with more doctors

Example 3: Infant TV leads to autism – Theory 3



Infant TV watching

Higher rates of autism

Instrumental Variables:

- Rainfall

Omitted variables: Richer areas, with more doctors

Precipitation as IV

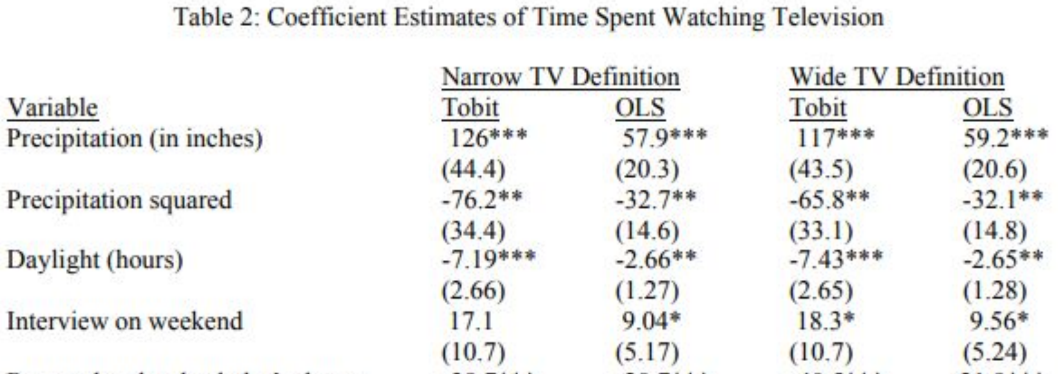
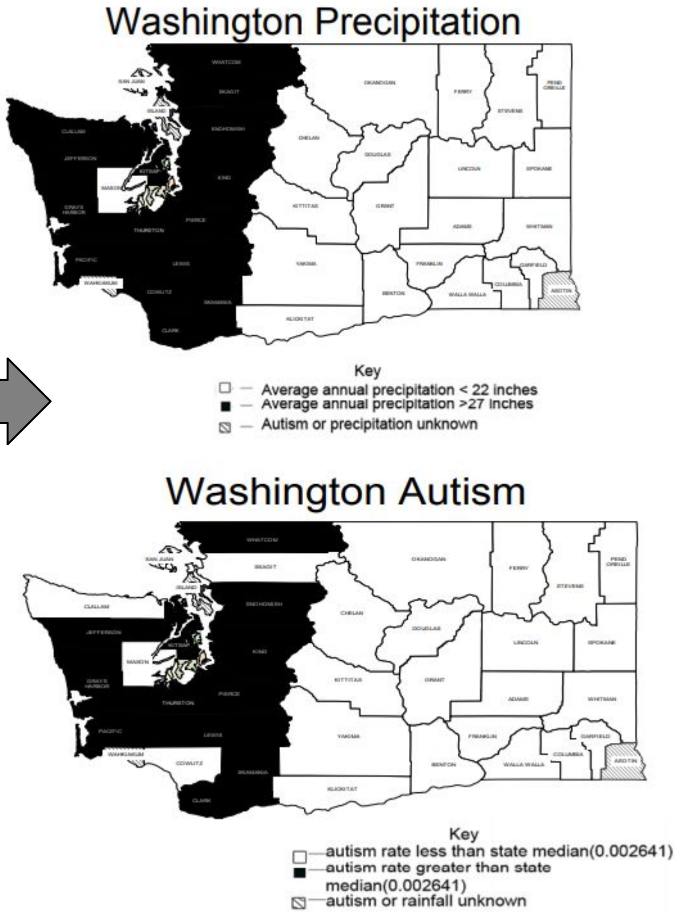


Table 2: Coefficient Estimates of Time Spent Watching Television

	Narrow TV Definition		Wide TV Definition	
Variable	Tobit	OLS	Tobit	OLS
Precipitation (in inches)	126***	57.9***	117***	59.2***
	(44.4)	(20.3)	(43.5)	(20.6)
Precipitation squared	-76.2**	-32.7**	-65.8**	-32.1**
	(34.4)	(14.6)	(33.1)	(14.8)
Daylight (hours)	-7.19***	-2.66**	-7.43***	-2.65**
	(2.66)	(1.27)	(2.65)	(1.28)
Interview on weekend	17.1	9.04*	18.3*	9.56*
	(10.7)	(5.17)	(10.7)	(5.24)

Figure 3

Precipitation (1990-2001) and Autism Rates (2005) of Washington Counties


Washington Precipitation
Key
□ — Average annual precipitation < 22 inches
■ — Average annual precipitation >27 inches
▨ — Autism or precipitation unknown
Washington Autism
Key
□ —autism rate less than state median(0.002641)
■ —autism rate greater than state median(0.002641)
▨ —autism or rainfall unknown

Notes: autism rates are for children between the ages of six and 18. Precipitation data are measured from July 1, 1990 through June 30, 2001.

Waldman, Michael, Sean Nicholson, and Nodir Adilov. *Does television cause autism?*. No. w12632. National Bureau of Economic Research, 2006.

Ok. Find a good instrument for X.

What makes a good instrument?

1. *Z* correlated with *X*

- This we can test “instrument’s relevance” by running a regression of Z on X
- They ought to be strongly correlated (high R-sq)

2. *Z* uncorrelated with *u*

- Cannot test this assumption directly
- Can only ask: Why would our variable Z *not* be correlated with u?
- We usually need some random (exogenous) variation in the amount of *Z* given to people (e.g., acts of God – birth dates, weather – or random processes like getting a boy or a girl baby, or being assigned to one side of a building or another, etc.)

In our case, we need …

… something that has no direct effect on a person’s religious attendance; is not correlated with a person’s underlying personality/temperament/values; but is correlated with a person’s education:

- Quarter of the year born in
- Distance of 2- and 4-year college to family home
- Parental education level
- No. of siblings

Why would quarter of the year be a good or bad IV?

1. Something that has no direct effect on attendance. Okay here.
2. Something that is not correlated with a person’s underlying personality/temperament. When you’re born makes no difference.
3. Something that is correlated with a person’s education. Due to age cut-offs, some people get more mandatory education than others (more true in the past).

Quarter year of birth as IV

Angrist, Joshua D., and Alan B. Keueger. "Does compulsory school attendance affect schooling and earnings?." The Quarterly Journal of Economics 106.4 (1991): 979-1014.

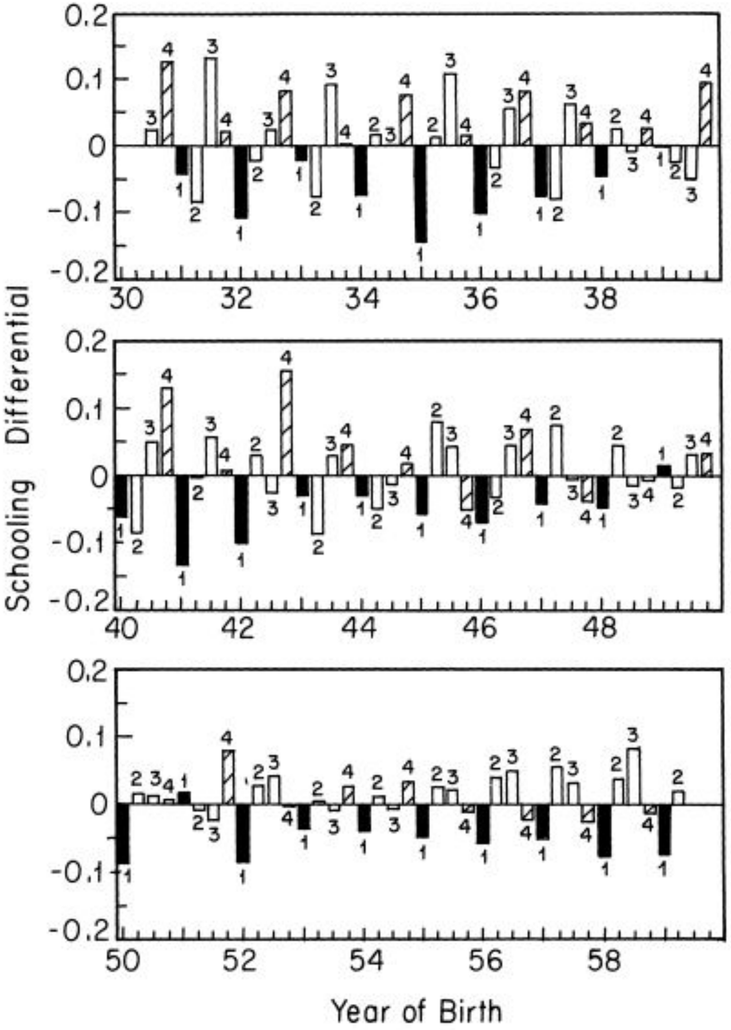
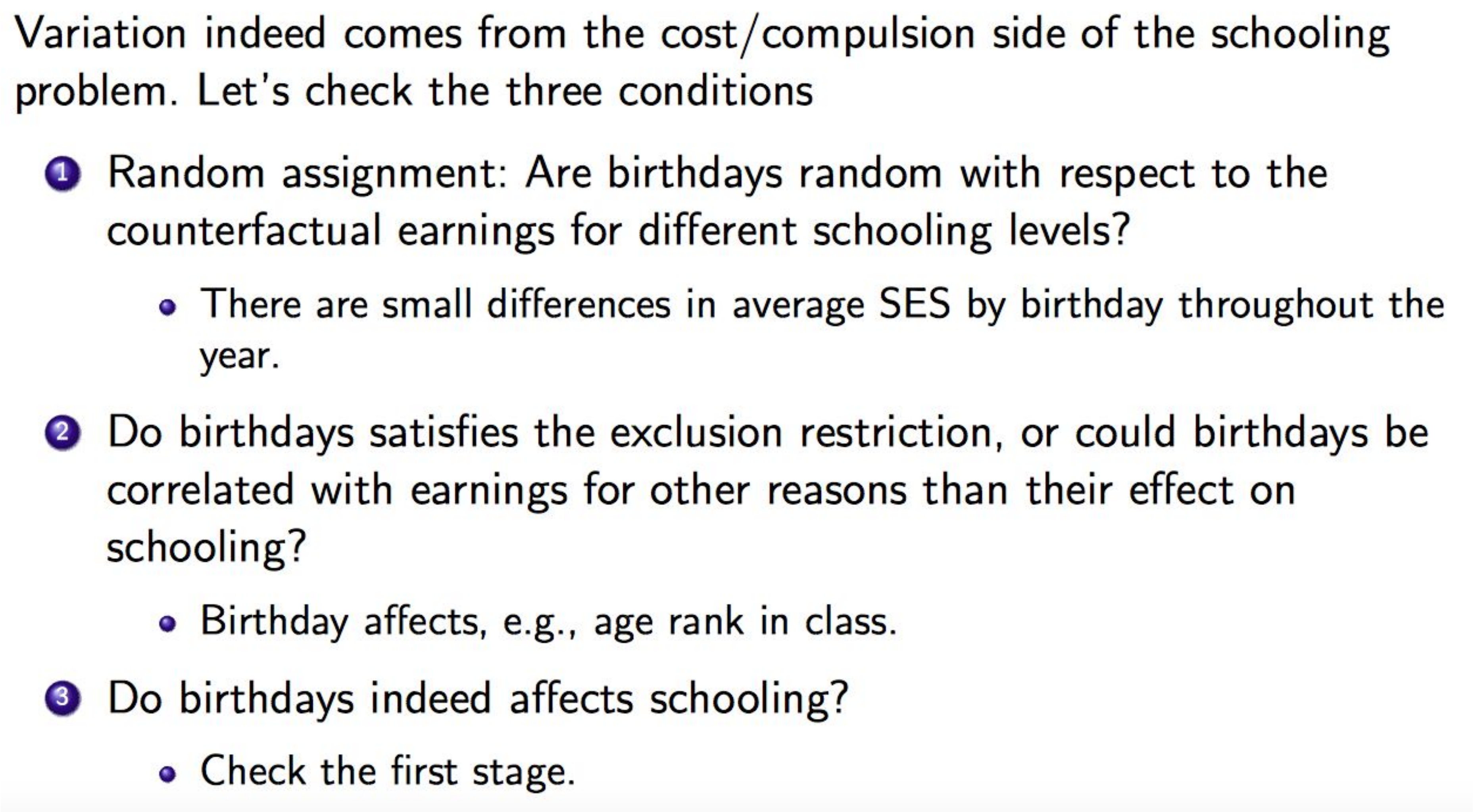


FIGURE IV
Season of Birth and Years of Schooling
Deviations from $MA(+2,-2)$

Why would quarter of the year be a good or bad IV (from Jorn-Steffen Pischke, LSE)?



Variation indeed comes from the cost/compulsion side of the schooling problem. Let's check the three conditions

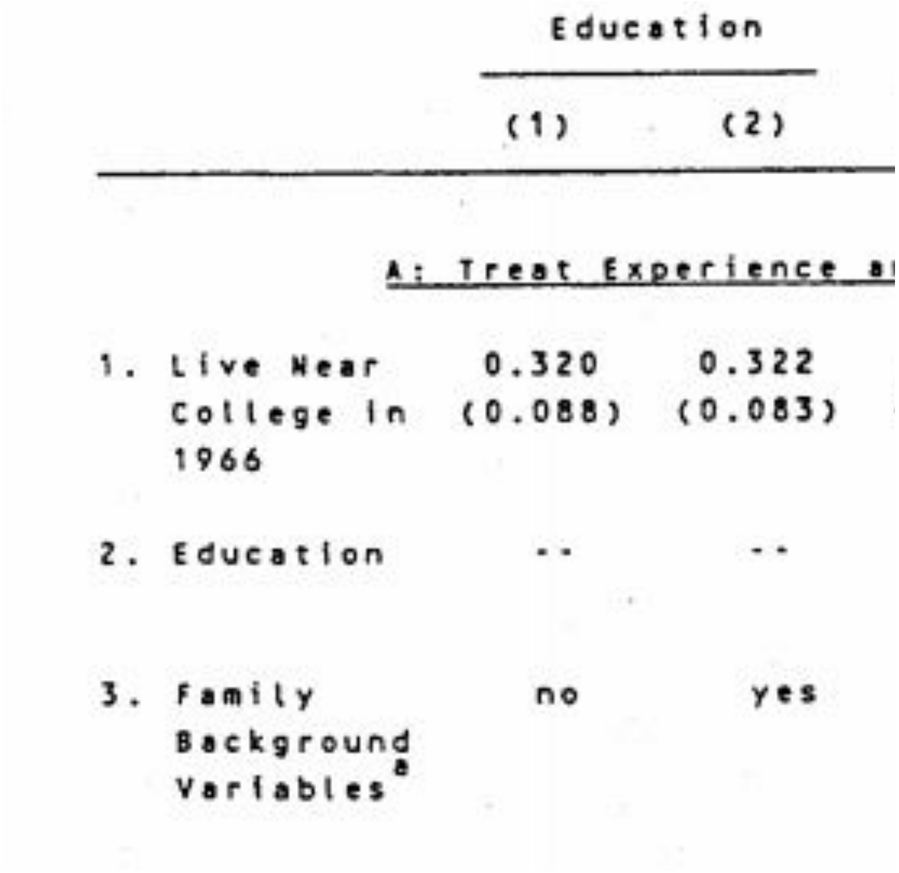
1. Random assignment: Are birthdays random with respect to the counterfactual earnings for different schooling levels?
   - There are small differences in average SES by birthday throughout the year.
2. Do birthdays satisfies the exclusion restriction, or could birthdays be correlated with earnings for other reasons than their effect on schooling?
   - Birthday affects, e.g., age rank in class.
3. Do birthdays indeed affects schooling?
   - Check the first stage.

Why would distance of college to home be a good or bad IV?

1. Something that has no direct effect on attendance. Okay here, I think.
2. Something that is not correlated with a person's underlying personality/temperament. Maybe. Where you're born should make no difference – but rural vs. urban, big vs. small, may play some role.
3. Something that is correlated with a person's education. Due to human scale, when things are closer, we use them more and more often (and they might be cheaper in some senses, too).

Distance from college as IV

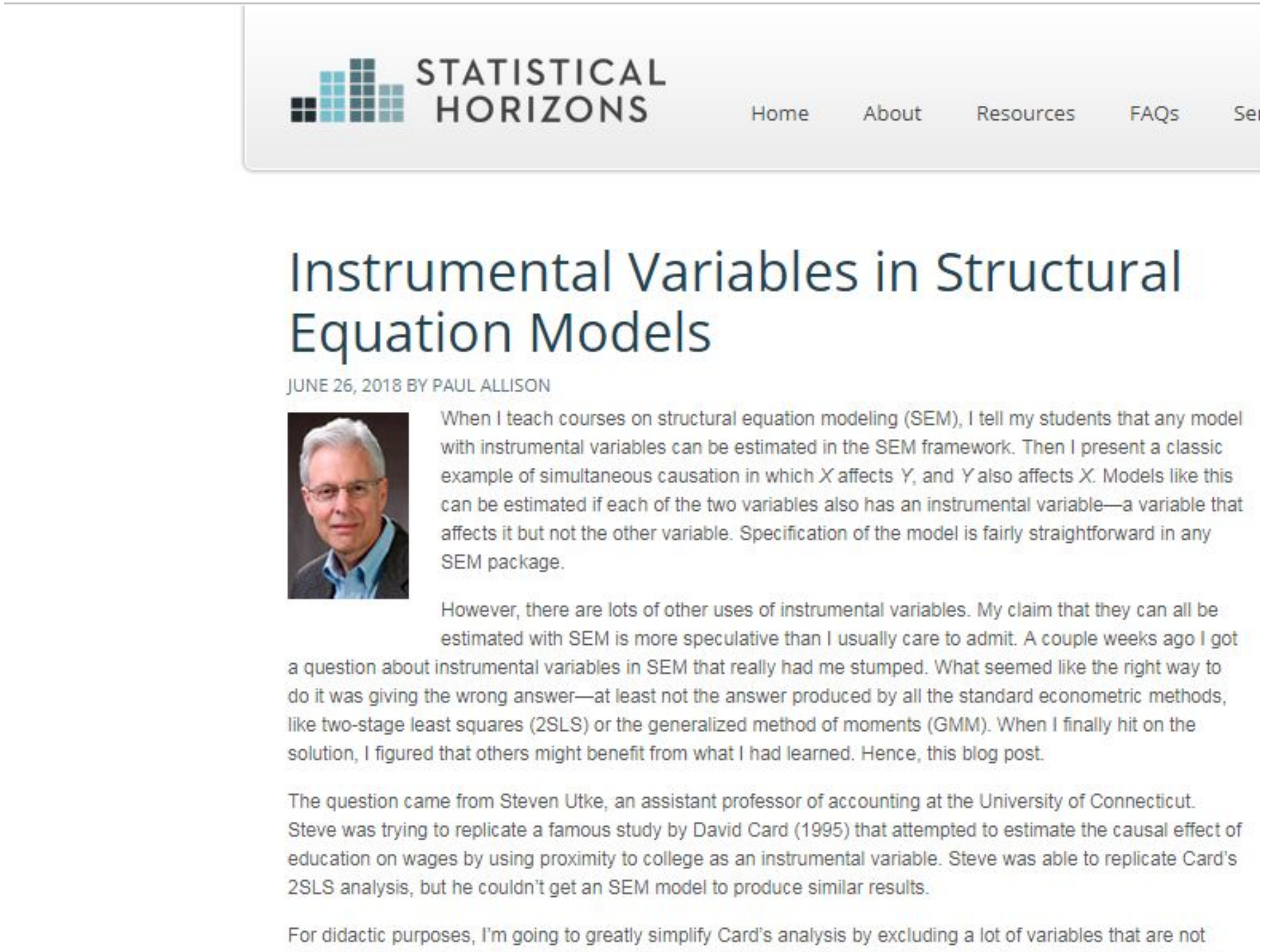
Card, David (1995) “Using geographic variation in college proximity to estimate the return to schooling.” Aspects of Labour Economics: Essays in Honour of John Vanderkamp. University of Toronto Press.



	Education	
	(1)	(2)
A: Treat Experience a		
1. Live Near College in 1966	0.320 (0.088)	0.322 (0.083)
2. Education	--	--
3. Family Background Variables[a]	no	yes

One other issue with Card’s IV

Allison takes issue with Card’s IV here:



Instrumental Variables in Structural Equation Models

JUNE 26, 2018 BY PAUL ALLISON

When I teach courses on structural equation modeling (SEM), I tell my students that any model with instrumental variables can be estimated in the SEM framework. Then I present a classic example of simultaneous causation in which *X* affects *Y*, and *Y* also affects *X*. Models like this can be estimated if each of the two variables also has an instrumental variable—a variable that affects it but not the other variable. Specification of the model is fairly straightforward in any SEM package.

However, there are lots of other uses of instrumental variables. My claim that they can all be estimated with SEM is more speculative than I usually care to admit. A couple weeks ago I got a question about instrumental variables in SEM that really had me stumped. What seemed like the right way to do it was giving the wrong answer—at least not the answer produced by all the standard econometric methods, like two-stage least squares (2SLS) or the generalized method of moments (GMM). When I finally hit on the solution, I figured that others might benefit from what I had learned. Hence, this blog post.

The question came from Steven Utke, an assistant professor of accounting at the University of Connecticut. Steve was trying to replicate a famous study by David Card (1995) that attempted to estimate the causal effect of education on wages by using proximity to college as an instrumental variable. Steve was able to replicate Card’s 2SLS analysis, but he couldn’t get an SEM model to produce similar results.

For didactic purposes, I’m going to greatly simplify Card’s analysis by excluding a lot of variables that are not

One other issue with Card’s IV

Allison suggests running IV models as explicitly Structural Equation Models:

“Interestingly, the estimated correlation between the error terms (not shown in the table) was -.66. That means that the collective impact of omitted variables is to affect **educ** and **lwage** in opposite directions. Whatever raises **lwage** lowers **educ**, and vice versa.

It’s hard to imagine what variables would behave like this. And that difficulty raises questions about the plausibility of the model. (Card suggested that measurement error in education could produce a negative correlation, but the degree of error would have to be unreasonably large to produce the result we just saw.) In any case, that’s not really our concern here. It does point out, however, that one of the advantages of the SEM approach is that you get an estimate of the error correlation—something you typically don’t see with 2SLS.

What’s the lesson here? If you want to use SEM to estimate an instrumental variable model, it’s essential to think carefully about why you need instruments in the first place. You should make sure that the specified model reflects all your beliefs about the causal mechanism. In particular, if you suspect that error terms are correlated with observed variables, those correlations must be built into the model.”

Why would number of siblings be a good or bad IV?

1. Something that has no direct effect on attendance. May not be okay here. More siblings may signal commitment to “be fruitful and multiply” – so that might increase likelihood of attedning somehow, maybe.
2. Something that is not correlated with a person’s underlying personality/motivation. More siblings may have meant a different home environment, which could affect people’s habits, etc.
3. Something that is correlated with a person’s education. Fine there.

Why would parental education be a good or bad IV?

1. Something that has no direct effect on attendance. Okay here, probably. Parental education does not determine someone’s current attendance patterns.
2. Something that is not correlated with a person’s underlying personality/temperament. Uh-uh. Genetics may play a role. Or more educated parents provide better skills for mingling and listening.
3. Something that is correlated with a person’s education. Fine there.

What are we doing?

We go from: “$B = \mathrm{Cov}(Y, X) / \mathrm{Var}(X)$”
to: “$B = \mathrm{Cov}(Y,Z) / \mathrm{Cov}(X,Z)$”

where the $\mathrm{Cov}(Z,u)=0$

(What would happen if $Z=X$?)

Let’s try one …

Some set-up

```
GSS=read.csv(file.choose()) ## use the 2006 GSS file ##

library(plyr)

vars <- c("attend", "educ", "maeduc", "age", "region", "relig")
sub <- GSS[, vars]

sub <- na.omit(sub)
```

My simple OLS model …

For every year more education, people on average go up 0.059*** categories of religious service attendance, net of

```
> lm.attend <- lm(attend ~ educ + age + as.factor(region) + as.factor(relig), data = sub)
> summary(lm.attend)

Call:
lm(formula = attend ~ educ + age + as.factor(region) + as.factor(relig),
    data = sub)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         2.485520   0.367971   6.755 1.76e-11 ***
educ                0.059261   0.015449   3.836 0.000128 ***
[omitted]
---------------------------------------------------------
Residual standard error: 2.439 on 2582 degrees of freedom
Multiple R-squared:  0.2494,  Adjusted R-squared:  0.243
F-statistic:    39 on 22 and 2582 DF,  p-value: < 2.2e-16
```

The IV model

More set-up ...

```
install.packages("AER")
library(AER)

iv.attend <- ivreg(attend ~
                     # first write equation without instrument (so educ instead of paeduc)
                     educ + age + as.factor(region) + as.factor(relig)
                   # then a vertical bar followed by the same equation again but this time with
                   # the instrument (so paeduc instead of educ)
                   | maeduc + age + as.factor(relig) + as.factor(region) ,
                   data = sub)

summary(iv.attend)
```

The IV model ...

Instrumenting mom’s education for R’s education leads to: For every year more education, people on average go down 0.02 categories of religious service attendance, net of …

```
> summary(iv.attend)

Call:
ivreg(formula = n.attend ~ educ + year + age + as.factor(relig) +
    as.factor(region) + as.factor(dwelown) | paeduc + year +
    age + as.factor(relig) + as.factor(region) + as.factor(dwelown),
    data = sub)

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         130.96792   45.80637   2.859 0.004253 **
educ                 -0.020861   0.01206  -1.351 0.320157
[omitted]
as.factor(dwelown)3  -0.76155    1.31042  -0.581 0.561144
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.62 on 2582 degrees of freedom
Multiple R-Squared: 0.1334,     Adjusted R-squared: 0.1321
Wald test: 113.2 on 25 and 17895 DF,  p-value: < 2.2e-16
```

But what is the IV actually doing?

How did I do that IV regression?

- First: I ran a regression predicting R’s education as a function of some stuff + mom’s education
- Then: I gathered the predicted Ys (*educ_pre*) from this model

```
> stage1 <- lm(educ ~ maeduc + age + as.factor(relig) + as.factor(region), data = sub)
> summary(stage1)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.100719   0.393778  20.572  < 2e-16 ***
maeduc              0.395992   0.014125  28.036  < 2e-16 ***
[omitted]
--------------------------------------------------------
Residual standard error: 2.72 on 2582 degrees of freedom
Multiple R-squared:  0.2659,     Adjusted R-squared:  0.2597
F-statistic: 42.52 on 22 and 2582 DF,  p-value: < 2.2e-16

---------------------------------------------------------

sub$educ.pred <- predict(stage1) # fitted values
```

How did I do that IV regression?

Mom’s education strongly predicts R’s education: For every year more educated mom is, the child gains 0.39 years more education on average, net of …

```
> stage1 <- lm(educ ~ maeduc + age + as.factor(relig) + as.factor(region), data = sub)
> summary(stage1)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.100719   0.393778  20.572  < 2e-16 ***
maeduc              0.395992   0.014125  28.036  < 2e-16 ***
[omitted]
---------------------------------------------------------
Residual standard error: 2.72 on 2582 degrees of freedom
Multiple R-squared:  0.2659,     Adjusted R-squared:  0.2597
F-statistic: 42.52 on 22 and 2582 DF,  p-value: < 2.2e-16
```

But wait! -- Avoiding “weak” instruments

For IV to work well, we must avoid “weak” instruments in our first stage regressions

That is: We must identify a Z which predicts X strongly enough (usually requiring an F-statistic of >10)

```
> summary(iv.attend, diagnostics = T)

Call:
ivreg(formula = attend ~ educ + age + as.factor(region) + as.factor(relig) |
    maeduc + age + as.factor(relig) + as.factor(region), data = sub)

Diagnostic tests:
                  df1  df2 statistic p-value
Weak instruments    1 2582   786.004  <2e-16 ***
Wu-Hausman          1 2581     8.211  0.0042 **
Sargan              0   NA        NA      NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.452 on 2582 degrees of freedom
Multiple R-Squared: 0.2416,      Adjusted R-squared: 0.2352
Wald test: 37.96 on 22 and 2582 DF,  p-value: < 2.2e-16
```

Where did “786.004” come from?

Where did “786.004” come from?

```
> stage1 <- lm(educ ~ maeduc + age + as.factor(relig) + as.factor(region), data = sub)
> summary(stage1)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         8.100719   0.393778  20.572  < 2e-16 ***
maeduc              0.395992   0.014125  28.036  < 2e-16 ***
[omitted]
---------------------------------------------------------
Residual standard error: 2.72 on 2582 degrees of freedom
Multiple R-squared:  0.2659,     Adjusted R-squared:  0.2597
F-statistic: 42.52 on 22 and 2582 DF,  p-value: < 2.2e-16
```

$$28.036 * 28.036 = 786.004$$

What it really means

- The part of someone’s education that is due to the effect of the “random” amount of education their mom had (on their own education), i.e., *educ_pred*, does not stat. sig. predict religious attendance (B=-02, p=0.32)

```
> stage2 = lm(attend ~ educ.pred + age + as.factor(relig) + as.factor(region), data = sub)
> summary(stage2)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.613593   0.540767   6.682 2.87e-11 ***
educ.pred          -0.020861   0.012068  -1.351 0.320156
--------------------------------------------------------
Residual standard error: 2.446 on 2582 degrees of freedom
Multiple R-squared:  0.2453,     Adjusted R-squared:  0.2389
F-statistic: 38.14 on 22 and 2582 DF,  p-value: < 2.2e-16
```

But wait! -- Shouldn’t we have asked if schooling is endogenous to begin with?

- The Hausman-Wu test consists of regressing the variable in question (schooling) on the other variables in the attendance equation plus the additional variables for identification (i.e., mom’s education). The residual from this estimate is included in the attendance regression along with the actual value for schooling. The absolute value on the t-statistic above 2 indicates schooling is endogenous.

The test

The endogeneity test

```
> summary(iv.attend, diagnostics = T)

Call:
ivreg(formula = attend ~ educ + age + as.factor(region) + as.factor(relig) |
    maeduc + age + as.factor(relig) + as.factor(region), data = sub)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.613593   0.542078   6.666 3.20e-11 ***
educ               -0.020861   0.032146  -0.649 0.516423
[omits]

Diagnostic tests:
                  df1  df2 statistic p-value
Weak instruments    1 2582   786.004  <2e-16 ***
Wu-Hausman          1 2581     8.211  0.0042 **
Sargan              0   NA        NA      NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Where did “8.211” come from?

Where did “8.211” come from?

- Get the residual from this first stage regression

```
> stage1 <- lm(educ ~ maeduc + age + as.factor(relig) + as.factor(region), data = sub)
> summary(stage1)

> sub$educ.res=resid(stage1) # residuals

> stage3 = lm(attend ~ educ + age + as.factor(relig) + as.factor(region) + educ.res, data = sub)
> summary(stage3)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.613593   0.538525   6.710 2.38e-11 ***
educ               -0.020861   0.031935  -0.653 0.513660
[omitted]
educ.res            0.104513   0.036474   2.865 0.004198
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 2.435 on 2581 degrees of freedom
Multiple R-squared:  0.2518,	Adjusted R-squared:  0.2451
F-statistic: 37.77 on 23 and 2581 DF,  p-value: < 2.2e-16
```

The residual from a previous model where we regressed schooling on the other variables in the attendance equation plus the additional variables for identification (i.e., mom’s education).

2.8655*2.8655=8.211=F

What does it mean?

A large t-statistic on the Wu-Hausman test means that which we cannot predict of someone’s education, even after including mom’s education as a covariate, is still quite predictive of our outcome, religious attendance. That means that there may be omitted variables present and we need to use mom’s education as an instrument and not only as a covariate. So our instrumental variable regression is needed.

```
> summary(iv.attend, diagnostics = T)

Diagnostic tests:
                  df1  df2 statistic p-value    
Weak instruments    1 2582   786.004  <2e-16 ***
Wu-Hausman          1 2581     8.211  0.0042 ** 
Sargan              0   NA        NA      NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

My conclusion is what then?

- Wow. My very strong initial results with OLS are completely undermined (and the sign has flipped) when I employ my instrumental variables approach

- So maybe education is endogenous to attendance after all

- Ok. What do I do about this? More to come in a minute …

Ah-hah! Omitted variable bias corrected, right?

Not necessarily. Why not?

1. Statistically, IV increases our standard errors and lowers our p-values (i.e., it is inefficient)
2. Even a small correlation of our IV with *u* can lead to biased results
3. We have to be able to make a case that paeduc is uncorrelated with u
4. If our IV is correlated with Y, we also have a problem

There’s more …

5. We need to check for endogeneity anyway
6. (If we use more than 1 instrument…) We need to check with overidentification too

1. Statistically, IV increases our standard errors and lowers our p-values

We may now have an unbiased, but inefficient, estimate of the return on education.

Compare 95% confidence intervals on *educ*:

OLS = (0.028, 0.090)

Vs.

IV = (-0.53607, 0.076)

2. Even a small correlation of our IV with *u* can lead to biased results

- We use instruments to lower our bias, but if corr(z,x)=0.2 even, the corr(z,u) must be less than 1/5 of corr(x,u) before IV has less asymptotic bias than OLS. We know this from simulation.
- The problem is we never know the magnitudes of the corr(x,u) and corr(z,u)

3. We have to be able to make a case that *maeduc* is uncorrelated with *u*

- We have already noted that this may be hard to do

4. If our IV is correlated with *Y*, we also have a problem

- Remember, we treat our IV as exogenous, except as it affects X (i.e., it only has indirect effects on Y through X)
- If our IV has direct effects on Y, then this violates one of its assumptions
- In certain specifications perhaps, *maeduc* would be correlated with attendance

5. We need to check for endogeneity anyway

To test: We include educ_res as another predictor of attendance and see if it is stat. sig.

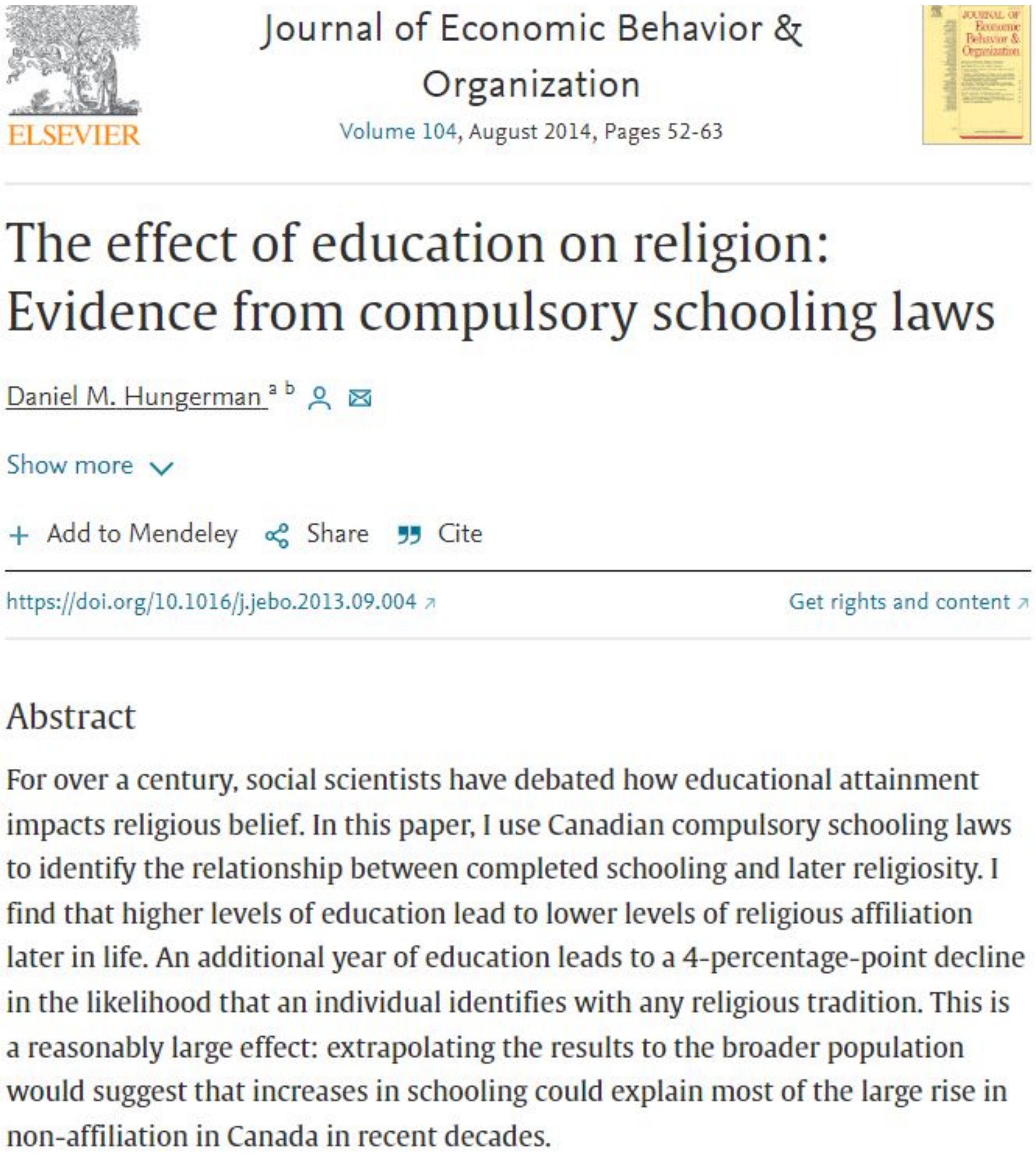
My conclusion is what then?

- Wow. My very strong initial results with OLS are completely undermined (and the sign has flipped) when I employ my instrumental variables approach

- So maybe education is endogenous to attendance after all

- It all rests on whether mom’s education is a valid instrument for R’s education in this context. What do we think?

IV for religion on education–



ELSEVIER

Journal of Economic Behavior & Organization

Volume 104, August 2014, Pages 52-63

The effect of education on religion: Evidence from compulsory schooling laws

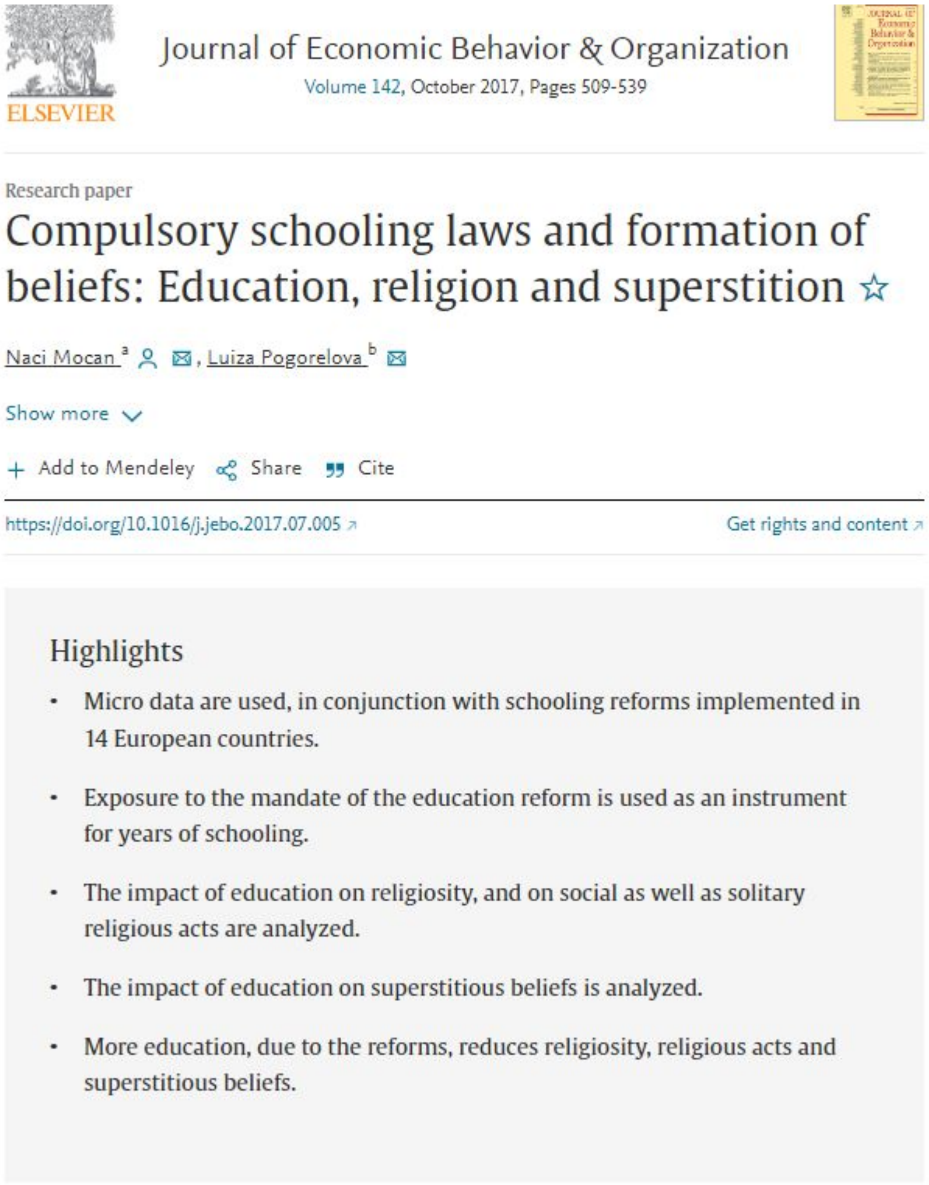
Daniel M. Hungerman [a] [b]

https://doi.org/10.1016/j.jebo.2013.09.004

Abstract

For over a century, social scientists have debated how educational attainment impacts religious belief. In this paper, I use Canadian compulsory schooling laws to identify the relationship between completed schooling and later religiosity. I find that higher levels of education lead to lower levels of religious affiliation later in life. An additional year of education leads to a 4-percentage-point decline in the likelihood that an individual identifies with any religious tradition. This is a reasonably large effect: extrapolating the results to the broader population would suggest that increases in schooling could explain most of the large rise in non-affiliation in Canada in recent decades.

IV for religion on education–



Journal of Economic Behavior & Organization
Volume 142, October 2017, Pages 509-539

Research paper

Compulsory schooling laws and formation of beliefs: Education, religion and superstition ☆

Naci Mocan [a], Luiza Pogorelova [b]

https://doi.org/10.1016/j.jebo.2017.07.005

Highlights

- Micro data are used, in conjunction with schooling reforms implemented in 14 European countries.
- Exposure to the mandate of the education reform is used as an instrument for years of schooling.
- The impact of education on religiosity, and on social as well as solitary religious acts are analyzed.
- The impact of education on superstitious beliefs is analyzed.
- More education, due to the reforms, reduces religiosity, religious acts and superstitious beliefs.

Quick context for the lab ...

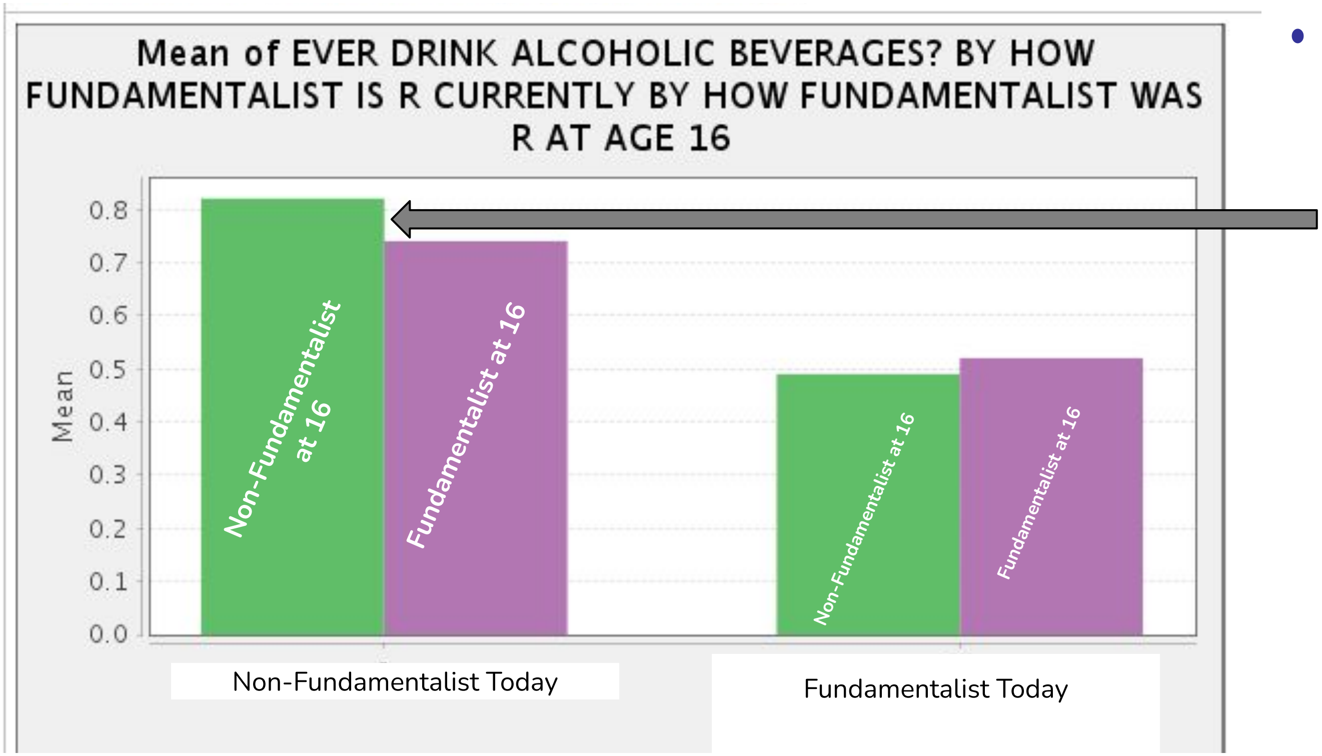
Where the lab example comes from...

- I want to see if drinking and violence are related, but people who drink may be inclined toward violence through genetics, personality, etc.

- I propose using the religion someone grew up in (which they didn't control) to instrument for "taste" for alcohol, since some religions prohibit alcohol (but it appears to be something of a historical accident which religions prohibit alcohol, not necessarily tied to genetics, personality, etc).

Where the lab example comes from...

- But religion does affect people’s likelihood of drinking, even if they leave that original religion (GSS data)



- People who grew up fundamentalist, but are not anymore are 7 points less likely to have ever drank, compared to people who were always non-fundamentalist

(c) Eirich 2013

*

Where the lab example comes from...

- Check out this article:

  French, Michael T., and Ioana Popovici. "That instrument is lousy! In search of agreement when using instrumental variables estimation in substance use research." Health Economics 20.2 (2011): 127-146.

Typical instruments

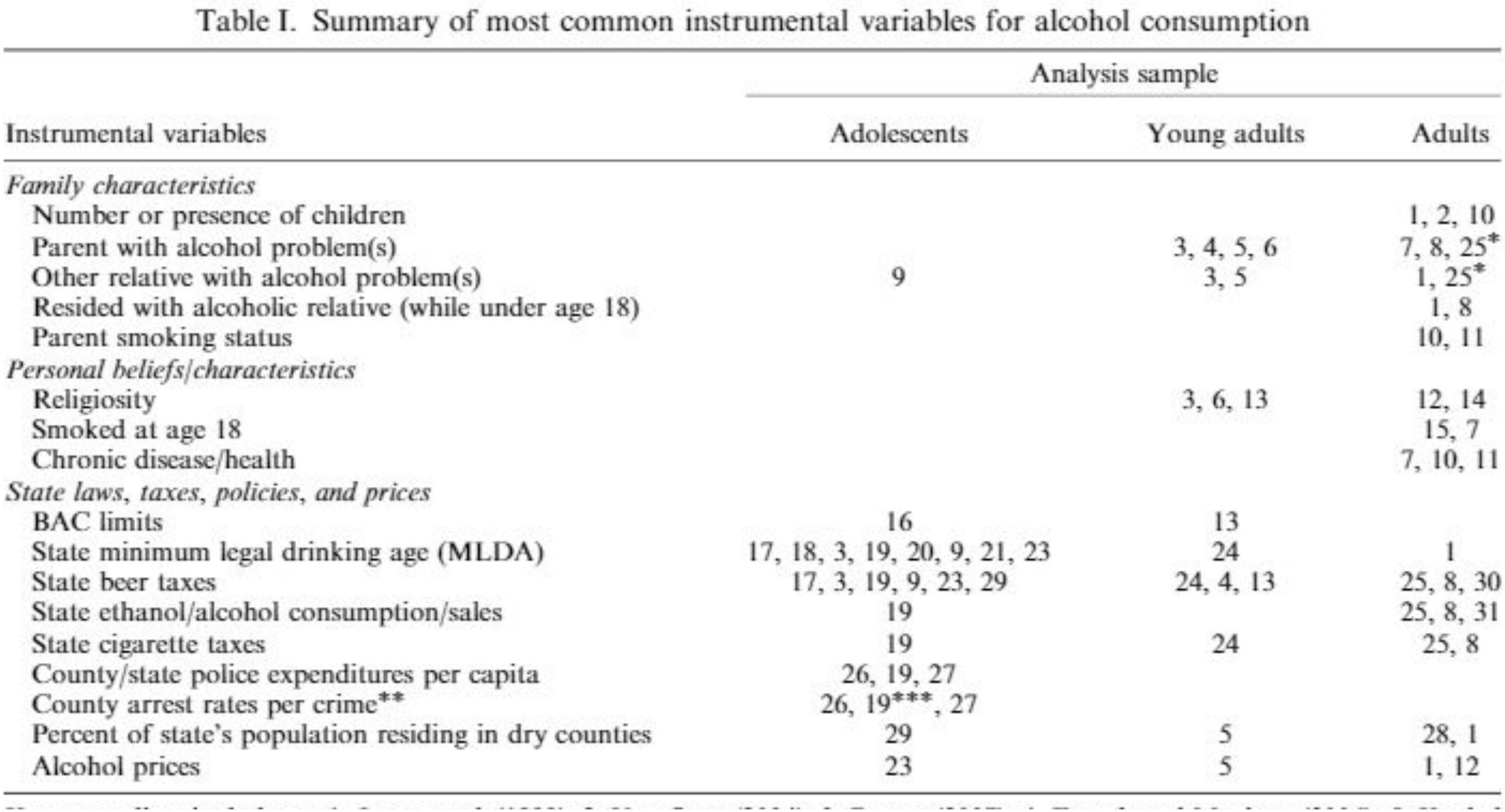


Table I. Summary of most common instrumental variables for alcohol consumption

	Analysis sample		
Instrumental variables	Adolescents	Young adults	Adults
Family characteristics			
Number or presence of children			1, 2, 10
Parent with alcohol problem(s)		3, 4, 5, 6	7, 8, 25*
Other relative with alcohol problem(s)	9	3, 5	1, 25*
Resided with alcoholic relative (while under age 18)			1, 8
Parent smoking status			10, 11
Personal beliefs/characteristics			
Religiosity		3, 6, 13	12, 14
Smoked at age 18			15, 7
Chronic disease/health			7, 10, 11
State laws, taxes, policies, and prices			
BAC limits	16	13	
State minimum legal drinking age (MLDA)	17, 18, 3, 19, 20, 9, 21, 23	24	1
State beer taxes	17, 3, 19, 9, 23, 29	24, 4, 13	25, 8, 30
State ethanol/alcohol consumption/sales	19		25, 8, 31
State cigarette taxes	19	24	25, 8
County/state police expenditures per capita	26, 19, 27		
County arrest rates per crime**	26, 19***, 27		
Percent of state's population residing in dry counties	29	5	28, 1
Alcohol prices	23	5	1, 12

2. Two Stage Least Squares (2SLS)

Two Stage Least Squares (2SLS)

- 2SLS is essentially the same as IV, but we can have multiple instruments for an endogenous X

Mom and dad educs as instruments

We get a very similar result

```
> vars <- c("attend", "educ", "maeduc", "paeduc", "age", "region", "relig")
> sub <- GSS[, vars]
> sub <- na.omit(sub)

> iv.attend <- ivreg(attend ~ educ + age + as.factor(region) +
      as.factor(relig)| maeduc + paeduc + age + as.factor(relig) +
      as.factor(region) ,data = sub)

> summary(iv.attend)

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.629447   0.548324   6.619 4.57e-11 ***
educ               -0.016552   0.030692  -0.539 0.589736
[omitted]
-----------------------------------------------
Residual standard error: 2.433 on 2103 degrees of freedom
Multiple R-Squared: 0.2425,   Adjusted R-squared: 0.2346
Wald test: 30.93 on 22 and 2103 DF,  p-value: < 2.2e-16
```

Two Stage Least Squares (2SLS)

- We should do all the stuff we did above, but now we have 1 additional test

- Test for Overidentification

- Overidentification is possible because we have more instruments than endogenous Xs

Mom and dad educs as instruments

Overidentification test:

The null hypothesis is that all of our instruments are valid (i.e., they are not correlated with u in some way)

- In this case, we cannot reject the null of valid instruments, so they are thought to be valid

```
> summary(iv.attend, diagnostics = T)
Diagnostic tests:
                  df1  df2 statistic p-value
Weak instruments    2 2102   466.497 < 2e-16 ***
Wu-Hausman          1 2102     7.416 0.00652 **
Sargan              1   NA     0.292 0.58919
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.433 on 2103 degrees of freedom
Multiple R-Squared: 0.2425,   Adjusted R-squared: 0.2346
Wald test: 30.93 on 22 and 2103 DF,  p-value: < 2.2e-16
```

Final thoughts ...

Maybe IVs work out where the “cure is worse than the disease”

- Check out this article:

  Young, Cristobal. "Model uncertainty in sociological research: an application to religion and economic growth." American Sociological Review 74.3 (2009): 380-397.

Maybe IVs work out where the “cure is worse than the disease”

- Problem #1 - Weak Instruments ...

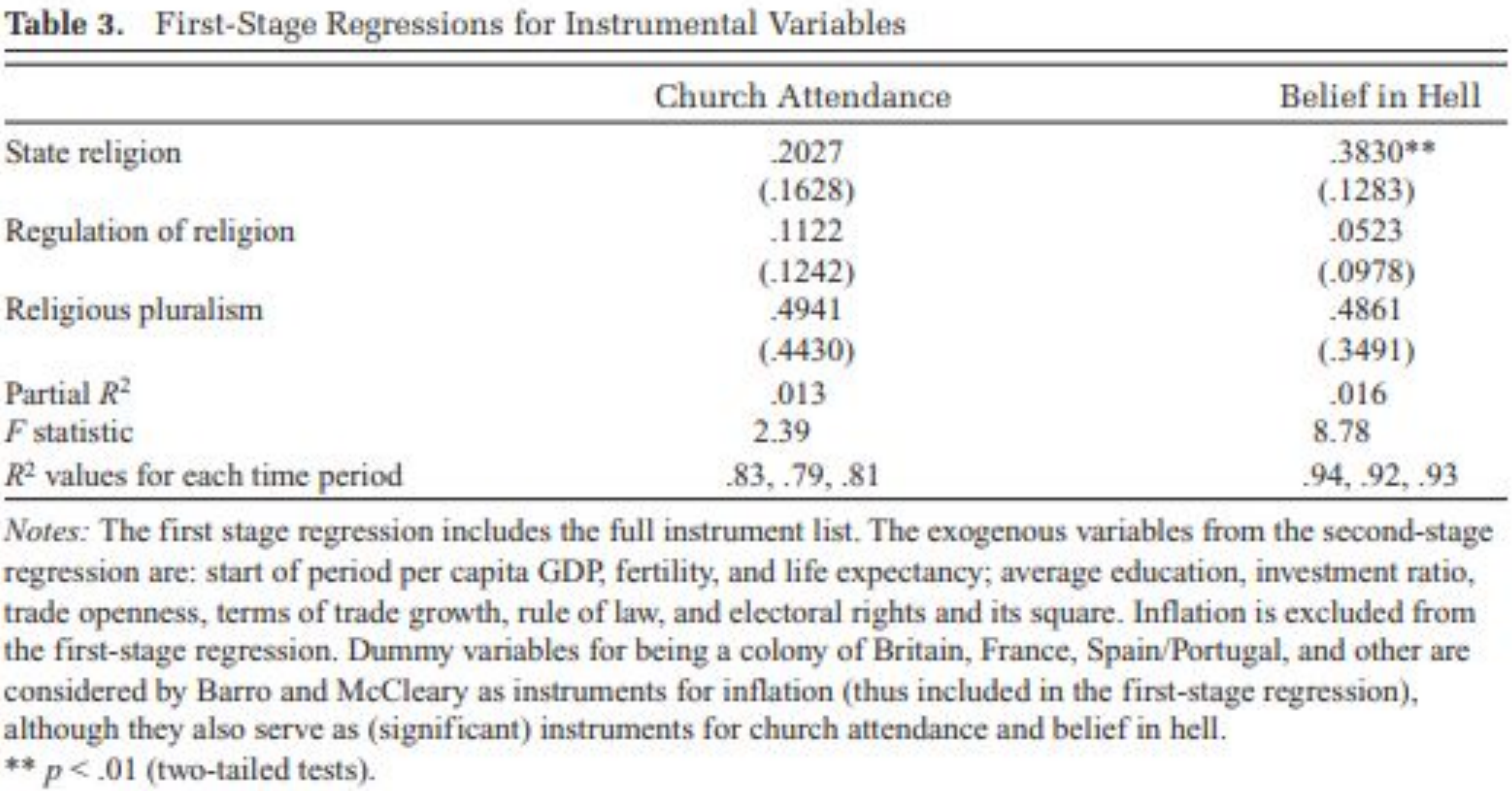


Table 3. First-Stage Regressions for Instrumental Variables

	Church Attendance	Belief in Hell
State religion	.2027	.3830**
	(.1628)	(.1283)
Regulation of religion	.1122	.0523
	(.1242)	(.0978)
Religious pluralism	.4941	.4861
	(.4430)	(.3491)
Partial R^2	.013	.016
F statistic	2.39	8.78
R^2 values for each time period	.83, .79, .81	.94, .92, .93

Notes: The first stage regression includes the full instrument list. The exogenous variables from the second-stage regression are: start of period per capita GDP, fertility, and life expectancy; average education, investment ratio, trade openness, terms of trade growth, rule of law, and electoral rights and its square. Inflation is excluded from the first-stage regression. Dummy variables for being a colony of Britain, France, Spain/Portugal, and other are considered by Barro and McCleary as instruments for inflation (thus included in the first-stage regression), although they also serve as (significant) instruments for church attendance and belief in hell.

** $p < .01$ (two-tailed tests).