IV

Bryan Kim

10/29/2020

```
library(tidyverse)  # ggplot(), %>%, mutate(), and friends
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(broom)  # Convert models to data frames
library(modelsummary)  # Create side-by-side regression tables
library(kableExtra)  # Add fancier formatting to tables
```

```
##
## Attaching package: 'kableExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
library(estimatr)  # Run 2SLS models in one step with iv_robust()

ed_fake = read.csv("/Users/bryankim/Documents/R/Instrumental Variables/Andrew Weiss/father_education.csv

# QUESTION: Does an extra year of education cause increased wages??
# PROBLEM:
    # Naive model
    # If we could actually measure ability, we could estimate this model, which closes the confounding
    # However, in real life we don't have ability, so we're stuck with a naive model: model_naive <- lm
    # The naive model overestimates the effect of education on wages (12.2 vs. 9.24) because of omitted

# Check instrument validity
# To fix the endogeneity problem, we can use an instrument to remove the endogeneity from education and
#
# For an instrument to be valid, it must meet three criteria:
#
# Relevance: Instrument is correlated with policy variable
# Exclusion: Instrument is correlated with outcome only through the policy variable
# Exogeneity: Instrument isn't correlated with anything else in the model (i.e. omitted variables)

#########################
```

```
## GENERAL IV PROCESS ##
########################

# 1. Is the instrument (in our case father's education) relevant? (e.g. is father's education correlate
#   -> Instrument correlated with policy/program (i.e. is there a significant relationship between the
# 2. Does instrument meet exclusion assumption?
#   -> Instrument causes outcome only through policy/program - lol good luck with that
# 3. Is the instrument exogenous?
#   -> No other variable is correlated with father's education from unmeasured things;error in uncorrel
# 4. 2-stage least squares (2SLS)
#   -> program ~ instrument; outcome ~ program_hat OR IV_robust()
```

```
# 1. (First Stage: Predicting your education based on father's education)
first_stage = lm(educ ~ fathereduc, data = ed_fake)
first_stage; tidy(first_stage); glance(first_stage) # "statistic is the F-stat"
```

```
##
## Call:
## lm(formula = educ ~ fathereduc, data = ed_fake)
##
## Coefficients:
## (Intercept)   fathereduc
##      2.2510       0.9162
```

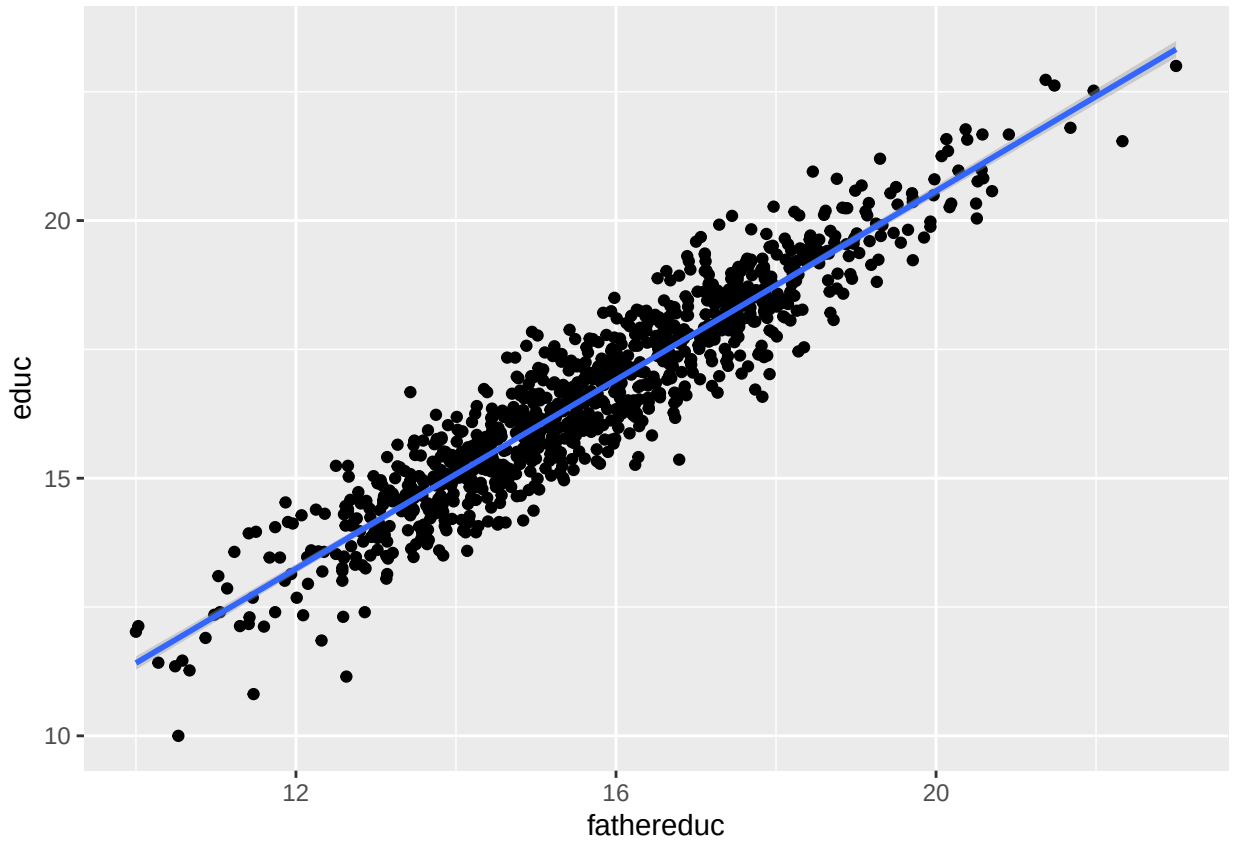
```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    2.25     0.172       13.1 3.67e-36
## 2 fathereduc     0.916    0.0108      84.5 0.
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.877         0.877 0.703     7136.       0     1 -1066. 2137. 2152.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
# first_stage shows us the instrument is relevant. Yay.

# lil' graph:
ed_fake %>% ggplot(aes(x = fathereduc, y = educ)) + geom_point() + geom_smooth(method = lm)
```

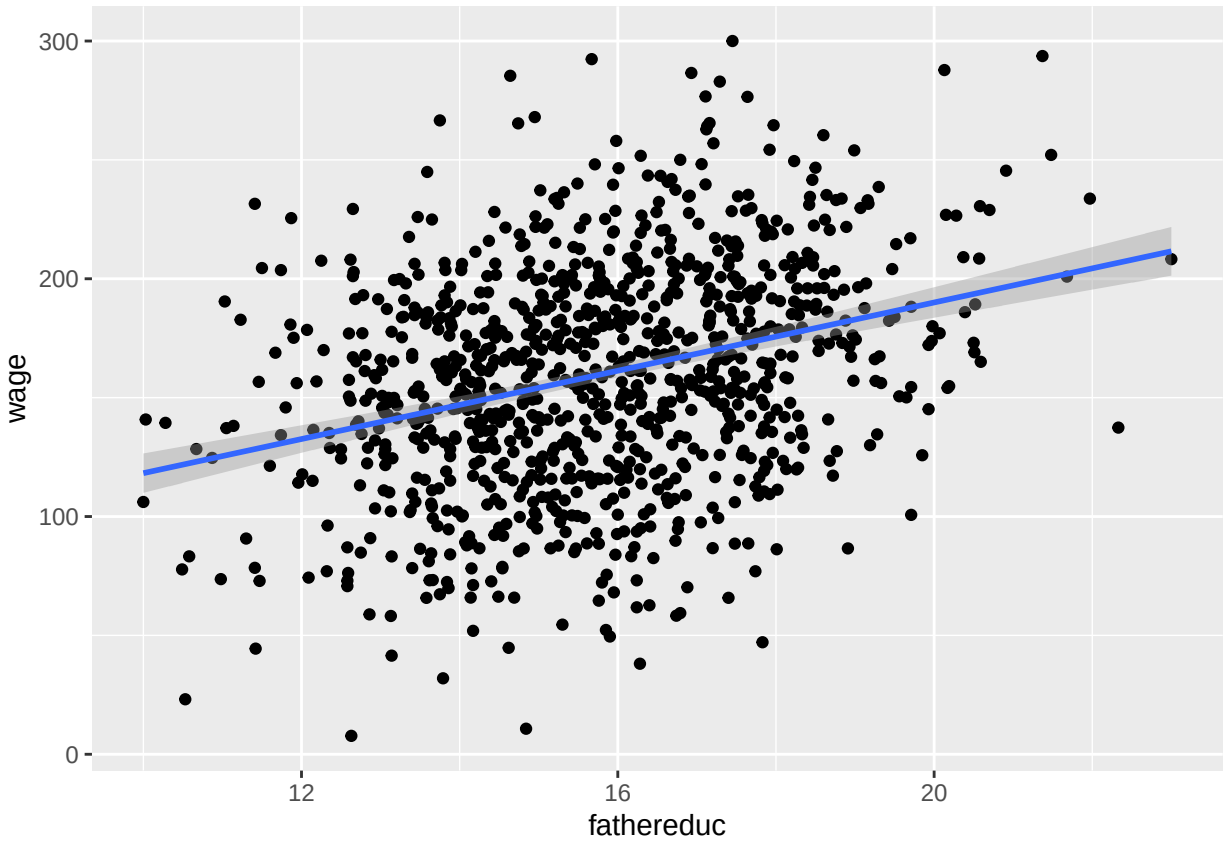
```
## `geom_smooth()` using formula 'y ~ x'
```




```
# 2. EXCLUSION (want fathereduc to be correlated with wages, but ONLY because of education)

ed_fake %>% ggplot(aes(x = fathereduc, y = wage)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# 3. CHECK FOR EXOGENEITY
# slejfnesfeslfe
```

```
# 2 STAGE LEAST SQUARES (2SLS)
```