Problem set 7: Education and wages

David Segovia

11-19/21


```
library(tidyverse)     # For ggplot, mutate(), filter(), and friends
library(broom)         # For converting models to data frames
library(estimatr)      # For lm_robust() and iv_robust()
library(modelsummary)  # For showing side-by-side regression tables
```

Task 1: Education, wages, and kids

Let's look once again at the effect of education on earnings. You'll use data from the 1976 Current Population Survey run by the US Census. The data is available as `wage` in the **wooldridge** R package—here is a subset of variables but are renamed. There are three columns:

Variable name	Description
`wage`	Average hourly earnings (in 1976 dollars)
`education`	Years of education
`n_kids`	Number of dependents living at home

You're interested in estimating β_1 in:

$$\text{Wage}_i = \beta_0 + \beta_1 \text{Education}_i + \epsilon_i$$

However, there is an issue with omitted variable bias and endogeneity. Instrumental variables can potentially help address the endogeneity.

Step 1

Load and look at the dataset

```
wages <- read_csv("wages.csv")
```

Step 2

We need an instrument for education, since part of it is endogenous. Do you think the variable `n_kids` (the number of children) would be a valid instrument? Does it meet the three requirements of a valid instrument? (Whether they (1) have *relevance*, (2) meet the *excludability* assumption, and (3) meet the *exogeneity* assumption.)

Answer: In terms of relevance, I think that the number of children does meet this because it does affect one's educational level. It can prevent people from going to school if they are forced to work to take care of their kid. So yes, this meets the relevance assumption.

For excludability, it makes sense that the number of kids is related only through the wages but this is really hard to prove. In order to prove that the number of kids impacts wages only through education will be hard to prove, so I don't think it meets this assumption.

For exogeneity, it is uncertain whether the number of kids variable is correlated with other endogenous variables in the model. I think that this variable can be correlated with other missing variables such as demographics- Black and Latino families are likely to have more kids, and demographics plays a huge role in one's wages.

Explain why it passes or fails each of the three requirements for a valid instrument. Test the requirements where possible using scatterplots and regression.

Relevance

```
relevance <- lm(education ~ n_kids, data = wages)
summary(relevance)
```

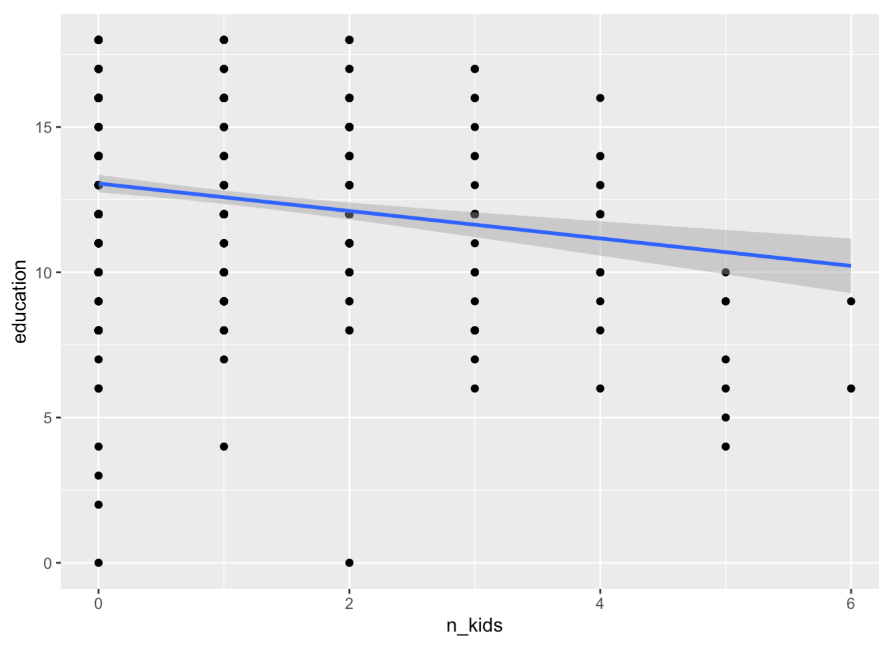
```
##
## Call:
## lm(formula = education ~ n_kids, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.056  -1.056  -0.111   1.889   5.889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.05582    0.15321  85.213  < 2e-16 ***
## n_kids      -0.47242    0.09361  -5.047 6.21e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.707 on 524 degrees of freedom
## Multiple R-squared:  0.04635,    Adjusted R-squared:  0.04453
## F-statistic: 25.47 on 1 and 524 DF,  p-value: 6.213e-07
```

```
glance(relevance)
```

```
## # 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    0.0464        0.0445  2.71      25.5 6.21e-7     1 -1269. 2544. 2557.
## # … with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
ggplot(wages, aes(x = n_kids, y = education)) +
  geom_point() + geom_smooth(method = "lm")
```

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

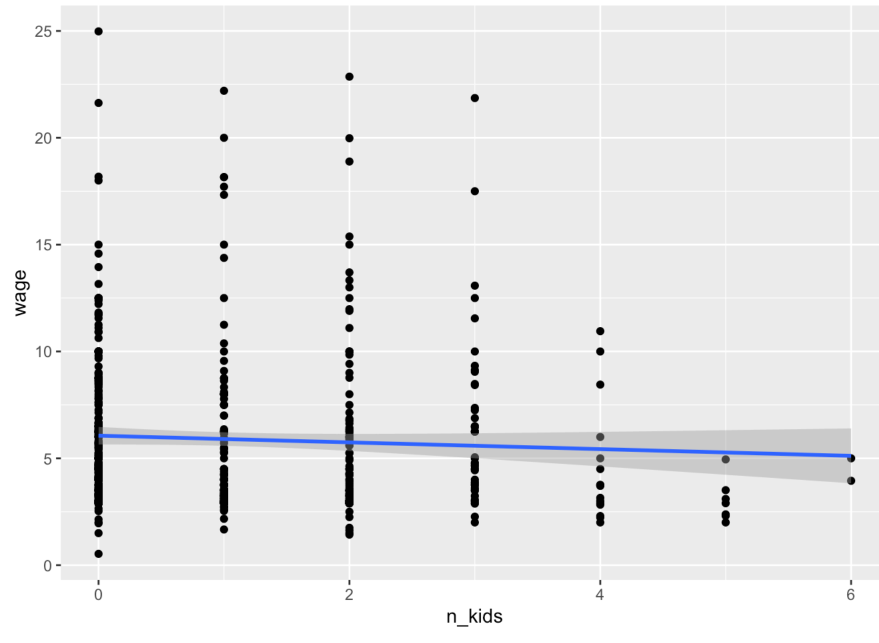


The F-statistic is 25.46, and this is above 10! It is both significant and the line looks OK. So relevance assumption is met, but I don't feel too confident about this.

Exclusion requirement is hard to see- we need to see if there is a relationship between number of kids and wages

```
ggplot(wages, aes(x = n_kids, y = wage)) +
  geom_point() + geom_smooth(method = "lm")
```

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




```
cor.test(wages$n_kids, wages$wage)
```

```
##
##  Pearson's product-moment correlation
##
## data:  wages$n_kids and wages$wage
## t = -1.2324, df = 524, p-value = 0.2184
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.13861643  0.03188105
## sample estimates:
##         cor
## -0.05375951
```

This requirement is not met, there is no correlation.

We cannot test for exogeneity since there is no other variables other than wage and number of kids in the model. Overall, I think that the number of kids is not a good instrument.

Step 3

Assume that the number of children is a valid instrument (regardless of whatever you concluded earlier). Using the number of children (`n_kids`) as an instrument for education (`education`), estimate the effect of education on wages via two-stage least squares (2SLS) instrumental variables (IV).

Do this by hand: create a first stage model, extract the predicted education, and use predicted education in the second stage.

Interpret the coefficient that gives the effect of education on wages (β_1) and its significance.

Manually

```
firststage <- lm(education ~ n_kids, data= wages)

prediction <- augment_columns(firststage, wages) %>% rename(educ_fitted = .fitted)
head(prediction)
```

```
## # A tibble: 6 x 10
##    wage education n_kids educ_fitted .se.fit .resid    .hat .sigma .cooksd
##   <dbl>     <dbl>  <dbl>       <dbl>   <dbl>  <dbl>   <dbl>  <dbl>   <dbl>
## 1  3.1         11      2        12.1   0.148 -1.11  0.00299   2.71 2.54e-4
## 2  3.24        12      3        11.6   0.218  0.361 0.00648   2.71 5.85e-5
## 3  3           11      2        12.1   0.148 -1.11  0.00299   2.71 2.54e-4
## 4  6            8      0        13.1   0.153 -5.06  0.00320   2.70 5.63e-3
## 5  5.3         12      1        12.6   0.118 -0.583 0.00190   2.71 4.44e-5
## 6  8.75        16      0        13.1   0.153  2.94  0.00320   2.71 1.91e-3
## # … with 1 more variable: .std.resid <dbl>
```

```
secondstage <- lm(wage ~ educ_fitted, data = prediction)
tidy(secondstage)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)    1.71      3.40      0.504   0.615
## 2 educ_fitted    0.333     0.270     1.23    0.218
```

one step

```
model <- iv_robust(wage ~ education | n_kids, data = wages)
tidy(model)
```

```
##          term  estimate std.error statistic   p.value   conf.low conf.high  df
## 1 (Intercept) 1.7122545  2.803071 0.6108494 0.5415642 -3.7943837 7.2188926 524
## 2   education 0.3330363  0.222132 1.4992725 0.1344052 -0.1033422 0.7694149 524
##   outcome
## 1    wage
## 2    wage
```

(Remember that you can also use the `iv_robust()` function from the **estimatr** package to run IV/2SLS models in one step with: `iv_robust(y ~ x | z, data = data)`, where `y` is the outcome, `x` is the policy/program, and `z` is the instrument. Try doing this to check your manual two-stage model.)

Education co-efficient interpretation: every year of education increases wages by about 0.33 in hourly earnings.

Step 4

Run a naive model predicting the effect of education on wages (i.e. without any instruments). How does this naive model compare with the IV model?

naive model

```
naive_model <- lm(wage ~ education, data = wages)
summary(naive_model)
```

```
##
## Call:
## lm(formula = wage ~ education, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3396 -2.1501 -0.9674  1.1921 16.6085
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.90485    0.68497  -1.321    0.187
## education    0.54136    0.05325  10.167   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.378 on 524 degrees of freedom
## Multiple R-squared:  0.1648, Adjusted R-squared:  0.1632
## F-statistic: 103.4 on 1 and 524 DF,  p-value: < 2.2e-16
```

The naive model's coefficient is 0.54, this is higher and overestimates the effect of education on wages because of omitted variable bias.There is endogenity at play here- variables in the model likely correlate with education

Show the results side-by-side here:

```
modelsummary(list("OLS" = naive_model, "2SLS(by hand)" = secondstage, "2SLS(automatic)" = model),
             gof_omit = "IC|Log|Adj|p\\.value|statistic|se_type",
             stars = TRUE )
```

```
## Warning: In version 0.8.0 of the `modelsummary` package, the default significance markers produced by the `sta
rs=TRUE` argument were changed to be consistent with R's defaults.
## This warning is displayed once per session.
```

	OLS	2SLS(by hand)	2SLS(automatic)
(Intercept)	−0.905	1.712	1.712
	(0.685)	(3.399)	(2.803)
education	0.541***		0.333
	(0.053)		(0.222)
educ_fitted		0.333	
		(0.270)	
Num.Obs.	526	526	526
R2	0.165	0.003	0.140
F	103.363	1.519	
Std.Errors			HC2

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Step 5

Explain which estimates (OLS vs. IV/2SLS) you would trust more (or why you distrust both)

The OLS model likely over-estimates the effect of education on wages due to omitted variable bias. The 2SLS model's coefficient of 0.33 at least removes the endogenous part of education and only has the exogenous part of education's effect on wages. However, the 2SLS model also does not have a good instrument, and the co-efficient of education is not significant.

Since the OLS model has education's coefficient both significant and the R^2 value is 2 points higher than the 2SLS model, I would trust the OLS model a little more. I would prefer multivariate regression model but there are not enough variables in the model to run this model.