Instrumental variables

Economists who are obsessed with the question of whether education causes an increase in income have collected yet another dataset of wages, education, and a host of other demographic characteristics (they seriously can't stop asking this question smh). There are seven variables in this new dataset:

Variable name	Description
`wage`	Hourly wage
`education`	Years of education
`distance`	Distance from a 4-year college (in 10 miles)
`gender`	Male or female (base case is male)
`ethnicity`	African American, Hispanic, or other (base case is other)
`unemp`	Unemployment rate in county
`home`	Does the family own their home

Your colleague knows that there are issues with omitted variable bias and selection bias in observational data—those who purposely seek out more education are likely going to purposely seek out higher paying jobs, for a host of reasons. To address this, your colleague thought that the distance to the closest 4-year college could be an instrument to address the endogeneity of education.

Your colleague attempted to measure the causal effect of more education on wages, using distance to a college as an instrument. They conducted some statistical analysis in R, but they forgot to interpret anything in the document, and now they've moved to a different office! (the audacity of this colleague and their constant moving smh)

Given the information provided below, interpret the results from this analysis, as well as any assumption checks or tests your colleague included. Does an additional year of education have a causal effect on wages? How much? Is it significant?

```
library(tidyverse)
library(estimatr)
library(broom)
library(modelsummary)

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

```
head(college_distance)
```

wage	education	distance	gender	ethnicity	unemp	home
8.09	12	0.2	male	other	6.2	yes
8.09	12	0.2	female	other	6.2	yes
8.09	12	0.2	male	other	6.2	yes
8.09	12	0.2	male	afam	6.2	yes
8.09	13	0.4	female	other	5.6	no
8.09	12	0.4	male	other	5.6	yes

↓ 1: What's going on here? ↓

```
# This is probably wrong...
model_naive <- lm(wage ~ education + gender + ethnicity + unemp,
                  data = college_distance)
tidy(model_naive)
```

term	estimate	std.error	statistic	p.value
(Intercept)	8.656	0.157	55.216	0.000
education	0.006	0.010	0.532	0.595
genderfemale	-0.086	0.037	-2.322	0.020
ethnicityafam	-0.539	0.051	-10.560	0.000
ethnicityhispanic	-0.534	0.048	-11.066	0.000
unemp	0.133	0.007	19.776	0.000

↓ 2: What's going on here? ↓

```
# Maybe distance to college is a good instrument?
first_stage <- lm(education ~ distance + gender + ethnicity + unemp,
                  data = college_distance)
tidy(first_stage)
```

term	estimate	std.error	statistic	p.value
(Intercept)	14.039	0.082	171.936	0.000
distance	-0.082	0.012	-6.974	0.000
genderfemale	-0.024	0.052	-0.454	0.650
ethnicityafam	-0.544	0.071	-7.644	0.000
ethnicityhispanic	-0.288	0.067	-4.278	0.000
unemp	0.010	0.010	1.005	0.315

```
glance(first_stage)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.023	0.022	1.77	21.9	0	5	-9427	18867	18912	14824	4733	4739

↓ 3: What's going on here? ↓

```
model_2sls <- iv_robust(wage ~ education + gender + ethnicity + unemp |
                          distance + gender + ethnicity + unemp,
                        data = college_distance)
tidy(model_2sls)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
(Intercept)	-0.766	1.985	-0.386	0.699	-4.657	3.125	4733	wage
education	0.677	0.142	4.779	0.000	0.399	0.955	4733	wage
genderfemale	-0.071	0.051	-1.390	0.165	-0.170	0.029	4733	wage
ethnicityafam	-0.201	0.096	-2.106	0.035	-0.389	-0.014	4733	wage
ethnicityhispanic	-0.336	0.077	-4.394	0.000	-0.486	-0.186	4733	wage
unemp	0.139	0.010	14.577	0.000	0.120	0.158	4733	wage

↓ 4: What's going on here? ↓

```
modelsummary(list("Naive OLS" = model_naive, "2SLS" = model_2sls))
```

	Naive OLS	2SLS
(Intercept)	8.656	-0.766
	(0.157)	(1.985)
education	0.006	0.677
	(0.010)	(0.142)
genderfemale	-0.086	-0.071
	(0.037)	(0.051)
ethnicityafam	-0.539	-0.201
	(0.051)	(0.096)
ethnicityhispanic	-0.534	-0.336
	(0.048)	(0.077)
unemp	0.133	0.139
	(0.007)	(0.010)
Num.Obs.	4739	
R2	0.109	-0.681
R2 Adj.	0.108	-0.682
AIC	15708.6	
BIC	15753.8	
Log.Lik.	-7847.289	
F	116.201	
N		4739
p.value.endogeneity		
p.value.overid		
p.value.weakinst		
se_type		HC2
statistic.endogeneity		
statistic.overid		
statistic.weakinst		