Problem set 7: Education and wages + public housing and health

Answer key - PMAP 8521, Spring 2021

April 5, 2021

Contents


```
library(tidyverse)      # For ggplot, %>%, mutate, filter, group_by, and friends
library(broom)          # For showing models as data frames
library(modelsummary)   # For side-by-side regression tables
library(estimatr)       # For iv_robust()
library(patchwork)      # For combining ggplots together
```

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 I've just taken a subset of variables and renamed them. 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("data/wages.csv")

# Show first few rows
head(wages)
```

wage	education	n_kids
3.10	11	2
3.24	12	3
3.00	11	2
6.00	8	0
5.30	12	1
8.75	16	0

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?

To be a valid instrument, a variable must meet three criteria:

1. ***Relevance*: Instrument is correlated with policy variable**
2. ***Exclusion*: Instrument is correlated with outcome *only through* the policy variable**
3. ***Exogeneity*: Instrument isn't correlated with anything else in the model (i.e. omitted variables)**

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

In order for number of kids to be a valid instrument, it must be relevant, exclusive, and exogenous. We can test some of these with statistics, but others need a compelling story or theory.

1. ***Relevance*: Number of kids is correlated with education**

   We can run a regression that predicts education based on the number of kids someone has. If there's a relationship, and if the F-statistic is greater than 104, we can safely claim relevancy.

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

term	estimate	std.error	statistic	p.value
(Intercept)	13.056	0.153	85.21	0
n_kids	-0.472	0.094	-5.05	0

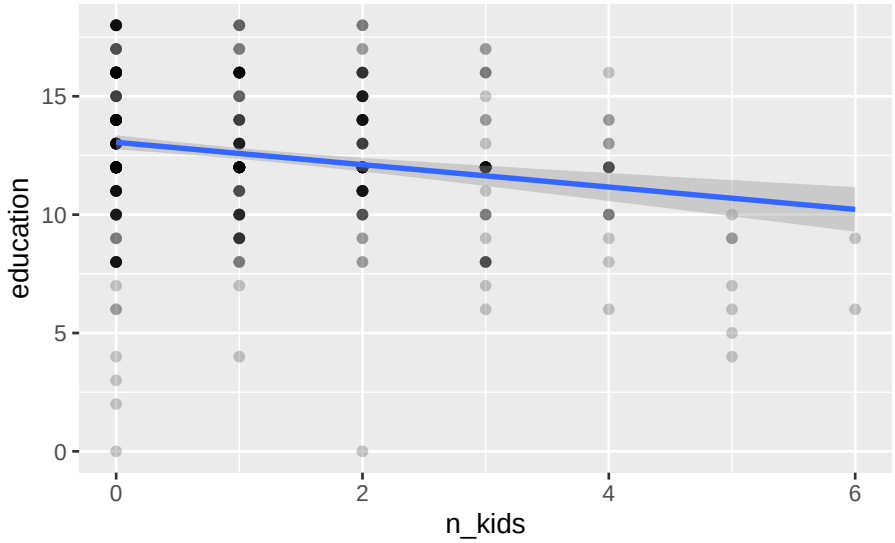
   ```
   glance(model_check_relevance)
   ```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.046	0.045	2.71	25.5	0	1	-1269	2544	2557	3839	524	526

Based on this model, having kids is negatively associated with years of education—an additional child is associated with 0.47 fewer years of education, on average, and this is statistically significant ($t = -5.05$; $p < 0.001$). The F-statistic is 25.5, which is above 10 and significant ($p < 0.001$). BUT it's not above 104, which nowadays is a more robust threshold, so it's not an incredibly strong or relevant instrument.

We can see this relationship in a plot too. The slope is definitely not zero, but there's a lot of variation in education that is not explained by kids:

```
ggplot(wages, aes(x = n_kids, y = education)) +
  # Make these points transparent since there's a lot of overplotting
  geom_point(alpha = 0.2) +
  geom_smooth(method = "lm")
```



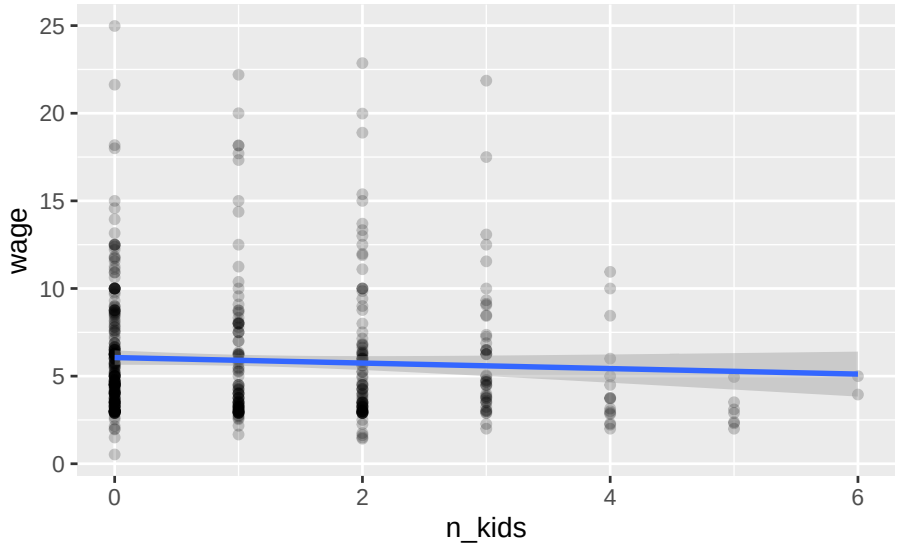
2. ***Exclusion*: Number of kids is correlated with wages *only through* education**

   We can check if there's a relationship between the number of kids someone has and their wages:

```
model_check_exclusion <- lm(wage ~ n_kids, data = wages)
tidy(model_check_exclusion)
```

term	estimate	std.error	statistic	p.value
(Intercept)	6.060	0.209	29.00	0.000
n_kids	-0.157	0.128	-1.23	0.218

```
ggplot(wages, aes(x = n_kids, y = wage)) +
  geom_point(alpha = 0.2) +
  geom_smooth(method = "lm")
```



It doesn't look like there's a strong relationship at all. Wages do go down as the number of kids goes up (β_1 = -0.157), but the relationship isn't statistically significant (p = 0.218). The trendline in the graph is pretty flat too. We probably don't meet the exclusion restriction.

Beyond this statistical test, we need to prove that the relationship between kids and wages *only* happens because of education. We have to think of possibilities where the number of kids you have influences your wages for non-educaiton reasons, and refute those with theory. For instance, maybe wages are lower because people self-select out of high paying venture capital-type jobs because they want to spend more time with their kids. That's a plausible story and we'd have to address it.

3. ***Exogeneity*: Number of kids isn't correlated with anything related to wages or education**

   Finally, we have to argue that the number of kids people have has nothing to do with wages or education. Like, people don't decide to have kids because of their education or becuase of their wages. lol.

   Moreover, the number of kids doesn't really meet Scott Cunningham's "weirdness" criterion, which means that it's highly plausible that the number of kids you have is closely linked to education and/or wages.

So, the number of kids people have is (1) kind of relevant, but not strongly, and it's not (2) exclusive or (3) exogenous, so it's likely not a great 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. (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 stages.)

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

So, if we pretend that number of kids *is* a good instrument, we can use it to remove the endogenous part of education and estimate the exogenous effect of education on wages. First we estimate the first stage model, which predicts education based on our instrument:

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

Then we use that model to calculated the predicted education for each person in the dataset, based on `n_kids`:

```
# I'm overwriting the wages dataset here. You can also make a new data frame
# named wages1 or whatever
wages <- augment_columns(first_stage_wages_educ, wages) %>%
  rename(education_hat = .fitted)  # Rename .fitted to make it easier to work with
head(wages)
```

wage	education	n_kids	education_hat	.se.fit	.resid	.hat	.sigma	.cooksd	.std.resid
3.10	11	2	12.1	0.148	-1.111	0.003	2.71	0.000	-0.411
3.24	12	3	11.6	0.218	0.361	0.006	2.71	0.000	0.134
3.00	11	2	12.1	0.148	-1.111	0.003	2.71	0.000	-0.411
6.00	8	0	13.1	0.153	-5.056	0.003	2.70	0.006	-1.871
5.30	12	1	12.6	0.118	-0.583	0.002	2.71	0.000	-0.216
8.75	16	0	13.1	0.153	2.944	0.003	2.71	0.002	1.089

Now we can use predicted education in the second stage of the model to predict wages:

```
second_stage_wages_educ <- lm(wage ~ education_hat, data = wages)
tidy(second_stage_wages_educ)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1.712	3.40	0.504	0.615
education_hat	0.333	0.27	1.232	0.218

According to this second stage model, an additional year of education causes an increase of $0.33 per hour in wages. This increase is not significant though (p = 0.218).

We can check this with `iv_robust()` and estimate both stages at the same time:

```
model_2sls_wages_educ <- iv_robust(wage ~ education | n_kids, data = wages)
tidy(model_2sls_wages_educ)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
(Intercept)	1.712	2.803	0.611	0.542	-3.794	7.219	524	wage
education	0.333	0.222	1.499	0.134	-0.103	0.769	524	wage

The coefficient is the same, and now the standard errors are more accurate. There's no significant effect of education on wages (p = 0.134).

Step 4a

For fun, we can calculate Anderson-Rubin confidence intervals, which show the range of possible variation in our β_1 estimate based solely on the strength or weakness of the instrument. Our F statistic was only 25, which again, in the good old days of looking for F statistics > 10 would be fine, but according to this new paper, is far less than 104 and is hardly sufficient be relevant.

To do this, we have to rerun the model with the `ivreg()` function from AER, and then feed that model to the `anderson.rubin.ci()` function from ivpack:

```
library(AER)
library(ivpack)

# Must include x = TRUE for this to work with anderson.rubin.ci()
model_again <- ivreg(wage ~ education | n_kids, data = wages, x = TRUE)
anderson.rubin.ci(model_again)
```

```
## $confidence.interval
## [1] "[ -0.242534063402667 , 0.830578358678587 ]"
```

Because `n_kids` is such a weak instrument (even though its F statistic is greater than 10!), the causal effect of education on wages could range anywhere from -$0.25 to $0.83 per hour. It could be negative, it could be positive. Who knows.

Step 5

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?

We can build a naive model without any instruments to show what the effect would be, not accounting for endogeneity:

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

Show the results side-by-side here:

```
# gof_omit here will omit goodness-of-fit rows that match any of the text. This
# means 'contains "IC" OR contains "Low" OR contains "Adj" OR contains "p.value"
# OR contains "statistic" OR contains "se_type"'. Basically we're getting rid of
# all the extra diagnostic information at the bottom
modelsummary(list("OLS" = model_naive_wage_educ, "2SLS" = second_stage_wages_educ,
                  "2SLS (robust)" = model_2sls_wages_educ),
             gof_omit = 'IC|Log|Adj|p\\.value|statistic|se_type',
             stars = TRUE)
```

	OLS	2SLS	2SLS (robust)
(Intercept)	-0.905	1.712	1.712
	(0.685)	(3.399)	(2.803)
education	0.541***		0.333
	(0.053)		(0.222)
education_hat		0.333	
		(0.270)	
Num.Obs.	526	526	
R2	0.165	0.003	0.140
F	103.363	1.519	
N			526

* p < 0.1, ** p < 0.05, *** p < 0.01

Comparing the 2SLS IV models with the naive OLS model is helpful here! Without accounting for the endogeneity in education, it looks like education has a positive significant effect on wages: a one year increase in education is associated with a $0.54 higher hourly wage, which is statistically significant (p < 0.001). However, after we remove the endogenous part of education using the instrument, that effect goes away. It's still positive (though muted at $0.33), but there's no guarantee it could also not be zero (i.e. it's not statistically significant).

Step 6

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

As mentioned above, the 2SLS models are arguably more accurate because they've removed the endogeneity from education and leave us with only the exogenous causal impact of education on wages. However, as we also saw above, the number of kids someone has is not the greatest instrument for this situation, since it doesn't meet the excludability or the exogeneity assumptions. I wouldn't trust any of these models.

Task 2: Public housing and health

Economic research shows that there is a potential (albeit weak) connection between health outcomes and residency in public housing. You are interested in finding the effect of public housing assistance on health outcomes. In the absence of experimental data, you must use observational data collected by the Georgia Department of Public Health. You have access to a dataset of 1,000 rows with the following columns:

Variable name	Description
`HealthStatus`	Health status on a scale from 1 = poor to 20 = excellent
`HealthBehavior`	Omitted variable (you can't actually measure this!)
`PublicHousing`	Number of years spent in public housing
`Supply`	Number of available public housing units in the city per 100 eligible households
`ParentsHealthStatus`	Health status of parents on a scale from 1 = poor to 20 = excellent
`WaitingTime`	Average waiting time before obtaining public housing in the city (in months)
`Stamp`	Dollar amount of food stamps (SNAP) spent each month
`Age`	Age
`Race`	Race; 1 = White, 2 = Black, 3 = Hispanic, 4 = Other
`Education`	Education; 1 = Some high school, 2 = High school, 3 = Bachelor's, 4 = Master's
`MaritalStatus`	Marital status; 1 = Single, 2 = Married, 3 = Widow, 4 = Divorced

(This is simulated data, but it's based on analysis by Angela R. Fertig and David A. Reingold)

Your goal is to measure the effect of living in public housing (`PublicHousing`) on health (`HealthStatus`). There is omitted variable bias, though, since people who care more about their health might be more likely to self-select into public housing and report a better health status score. The magic variable `HealthBehavior` measures this omitted variable, and you can use it as reference to make sure you get the models right (this is the same as "ability" in the examples in class), but don't include it in any of your actual models, since it's not real.

This data includes four potential instruments:

- `Supply`: Number of available public housing units in the city per 100 eligible households
- `ParentsHealthStatus`: Health status of parents on a scale from 1 = poor to 5 = excellent
- `WaitingTime`: Average waiting time before obtaining public housing in the city (in months)
- `Stamp`: Dollar amount of food stamps (SNAP) spent each month

You need to complete three tasks:

1. Evaluate the suitability of each of the four potential instruments. Check if they (1) have *relevance* with a scatterplot and model and F-test, (2) meet the *excludability* assumption, and (3) meet the *exogeneity* assumption. Choose one of these as your main instrument and justify why it's the best. Explain why the other three are not.

2. Estimate a naive model of the effect of public housing on health status (i.e. without any instruments). You can include any control variables you feel appropriate (i.e. that fit in your causal model). If you use variables that are categorical like race, education, or marital status, make sure you wrap them with `as.factor()` to treat them as categories instead of numbers (e.g. `as.factor(education)`).

3. Estimate the effect of public housing on health status using 2SLS IV (by hand with a first stage model, predicted public housing, and a second stage model using predicted public housing; don't use `iv_robust()` except to check your work). Compare the results with the naive model. Which model do you trust (if any), and why?

```
housing <- read_csv("data/public_housing.csv")
```

Task 1: Instrument suitability

For this policy, any instrument needs to meet these assumptions:

1. ***Relevance*: Instrument is correlated with public housing**
2. ***Exclusion*: Instrument is correlated with health *only through* public housing**
3. ***Exogeneity*: Instrument isn't correlated with anything else related to public housing**

We'll start by checking the relevancy of each of these potential instruments, looking for some sort of relationship between the instrument and public housing use. First we'll plot scatterplots of all four:

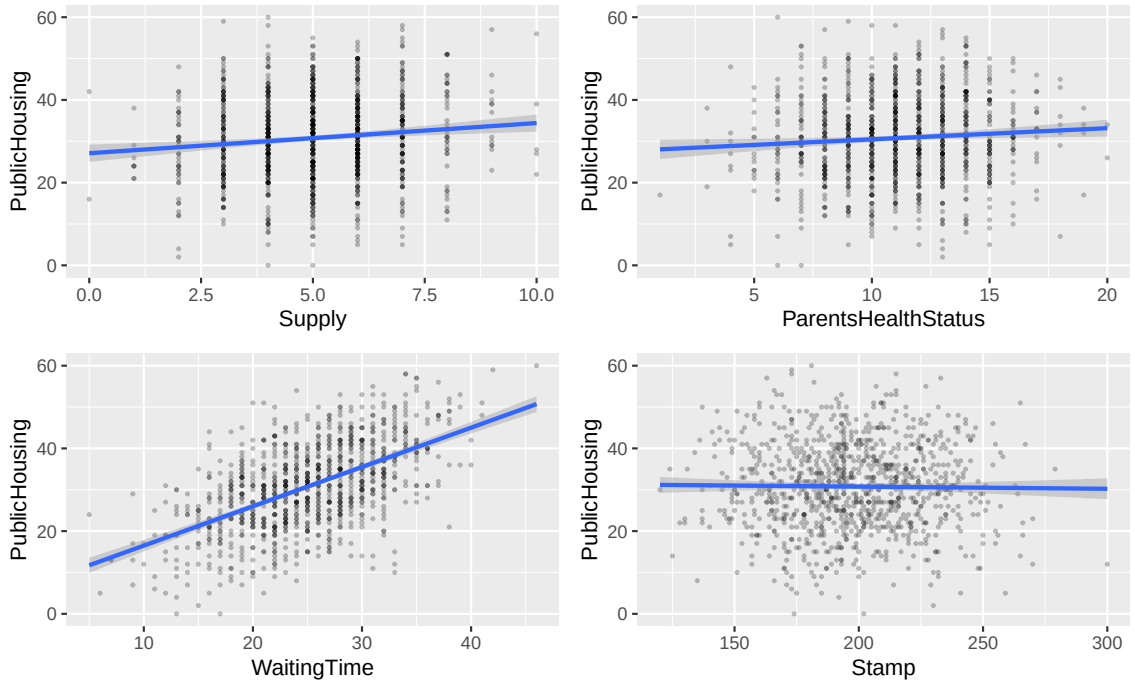
```
plot_supply <- ggplot(housing, aes(x = Supply, y = PublicHousing)) +
  geom_point(size = 0.5, alpha = 0.25) +
  geom_smooth(method = "lm")

plot_parents <- ggplot(housing, aes(x = ParentsHealthStatus, y = PublicHousing)) +
  geom_point(size = 0.5, alpha = 0.25) +
  geom_smooth(method = "lm")

plot_waiting <- ggplot(housing, aes(x = WaitingTime, y = PublicHousing)) +
  geom_point(size = 0.5, alpha = 0.25) +
  geom_smooth(method = "lm")

plot_stamp <- ggplot(housing, aes(x = Stamp, y = PublicHousing)) +
  geom_point(size = 0.5, alpha = 0.25) +
  geom_smooth(method = "lm")

plot_supply + plot_parents + plot_waiting + plot_stamp
```



Only one of these has a strong relationship with public housing: waiting time. We can run a bunch of simple regression models to check the size of the relationship, as well as the F-statistic (which we want to be above 104 (or 10, if you want to live in the good old days)):

```
check_supply <- lm(PublicHousing ~ Supply, data = housing)
check_parents <- lm(PublicHousing ~ ParentsHealthStatus, data = housing)
check_waiting <- lm(PublicHousing ~ WaitingTime, data = housing)
check_stamp <- lm(PublicHousing ~ Stamp, data = housing)

modelsummary(list(check_supply, check_parents, check_waiting, check_stamp),
             stars = TRUE, gof_omit = 'IC|Log|Adj')
```

	Model 1	Model 2	Model 3	Model 4
(Intercept)	27.107***	27.772***	6.991***	31.805***
	(1.067)	(1.288)	(1.141)	(2.384)
Supply	0.729***			
	(0.202)			
ParentsHealthStatus		0.270**		
		(0.112)		
WaitingTime			0.951***	
			(0.044)	
Stamp				-0.005
				(0.012)
Num.Obs.	1000	1000	1000	1000
R2	0.013	0.006	0.316	0.000
F	13.024	5.807	460.710	0.190

* p < 0.1, ** p < 0.05, *** p < 0.01

Housing supply, parental health status, and waiting time are all significantly associated with public housing use; food stamps are not. Parental health status has a low F-statistic that is below our threshold of 104. Housing supply and waiting time both have F-statistics that exceed 10, so they are both relevant. However, given the strong relationship between waiting time and public housing (and its large F-statistic), it's likely the more relevant instrument.

Next, we need to check to see if the instruments meet the exclusion assumption—that they only have a causal effect on health *through* public housing. We can check part of that assumption by looking at the effect of each instrument on health status:

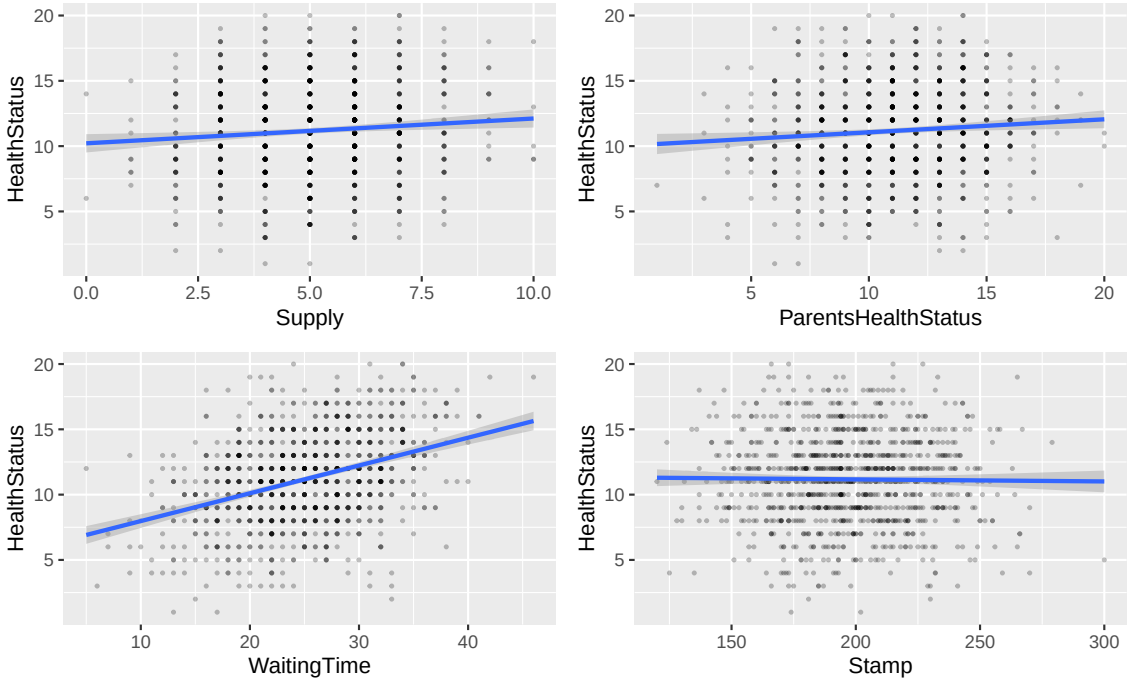
```
plot_supply_health <- ggplot(housing, aes(x = Supply, y = HealthStatus)) +
  geom_point(size = 0.5, alpha = 0.25) +
  geom_smooth(method = "lm")

plot_parents_health <- ggplot(housing, aes(x = ParentsHealthStatus, y = HealthStatus)) +
  geom_point(size = 0.5, alpha = 0.25) +
  geom_smooth(method = "lm")

plot_waiting_health <- ggplot(housing, aes(x = WaitingTime, y = HealthStatus)) +
  geom_point(size = 0.5, alpha = 0.25) +
  geom_smooth(method = "lm")

plot_stamp_health <- ggplot(housing, aes(x = Stamp, y = HealthStatus)) +
  geom_point(size = 0.5, alpha = 0.25) +
  geom_smooth(method = "lm")
```

```
plot_supply_health + plot_parents_health + plot_waiting_health + plot_stamp_health
```




```
check_supply_health <- lm(HealthStatus ~ Supply, data = housing)
check_parents_health <- lm(HealthStatus ~ ParentsHealthStatus, data = housing)
check_waiting_health <- lm(HealthStatus ~ WaitingTime, data = housing)
check_stamp_health <- lm(HealthStatus ~ Stamp, data = housing)

modelsummary(list(check_supply_health, check_parents_health,
                  check_waiting_health, check_stamp_health),
             gof_omit = 'IC|Log|Adj|p\\.value|statistic|se_type',
             stars = TRUE)
```

	Model 1	Model 2	Model 3	Model 4
(Intercept)	10.215***	10.063***	5.846***	11.480***
	(0.357)	(0.429)	(0.426)	(0.795)
Supply	0.190***			
	(0.068)			
ParentsHealthStatus		0.100***		
		(0.037)		
WaitingTime			0.213***	
			(0.017)	
Stamp				-0.002
				(0.004)
Num.Obs.	1000	1000	1000	1000
R2	0.008	0.007	0.142	0.000
F	7.910	7.110	165.618	0.154

* p < 0.1, ** p < 0.05, *** p < 0.01

Here's what we can conclude:

- ***Housing supply*: There's a positive significant relationship between housing supply and health status. The larger a city's supply of public housing, the healthier people report being. We need to make the argument that more public housing causes health only because it gives people access to housing. This argument is actually kind of reasonable.**
- ***Parent's health status*: There's a positive significant relationship between parental health and health status. The healthier one's parents are, the healthier one is. Making the case for exclusion here is tricky—we have to plausibly argue that parental health causes personal health *only* because of public housing, which is definitely not true.**
- ***Waiting time*: There's a positive significant relationship between waiting time for public housing and health status. The longer people wait for a spot in public housing, the healthier they report feeling. Here, we'd need to argue that longer wait times cause health only because of access to housing, which, might be the case?**
- ***Food stamps*: There's not much of a relationship here, and we'd need to argue that spending SNAP money causes you to be healthier (or less healthy) *only* because of access to public housing, which is a ridiculous idea.**

Finally, we need to examine how exogenous each of these instruments is:

- ***Housing supply*: This appears at first to be fairly exogenous. A city's housing supply might not be related to individual health. However, if a city has a lot of public housing, it probably spends money on lots of other public services, which then cause better health (and vice versa; low public housing spending probably accompanies low public health spending). This is likely not that exogenous.**
- ***Parent's health status*: This is most definitely not exogenous. If it were, your parents' health would have *no bearing whatsoever* on your health, which is not the case.**
- ***Waiting time*: This might be exogenous, though like housing supply, it might reflect underlying trends in the city's commitment to public service provision. Also, perhaps program administrators are purposely choosing the sickest people first and making healthier people wait longer.**
- ***Food stamps*: This is also definitely not exogneous. If it were, your spending on food would not be connected to your health.**

So, after all this checking, the variable that is likely the best instrument is waiting time. It is (1) highly relevant, (2) feels fairly exclusive, and (3) feels fairly exogenous. Housing supply might also be a good instrument too. Parental health and food stamp usage are both awful potential instruments and shouldn't be used.

Task 2: Naive model

We can get a naive estimate of the effect of public housing on health by running this model:

```
# Note the use of as.factor. Race, education, and marital status are not
# continuous variables---you can't report a marital status of 1.4, for instance
model_naive_housing <- lm(HealthStatus ~ PublicHousing + Age + as.factor(Race) +
                            as.factor(Education) + as.factor(MaritalStatus),
                          data = housing)
tidy(model_naive_housing)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-1.197	0.226	-5.287	0.000
PublicHousing	0.321	0.003	126.217	0.000
Age	0.056	0.004	12.865	0.000
as.factor(Race)2	-0.012	0.075	-0.164	0.870
as.factor(Race)3	0.052	0.075	0.695	0.487
as.factor(Race)4	-0.018	0.074	-0.243	0.808
as.factor(Education)2	0.023	0.074	0.308	0.758
as.factor(Education)3	-0.010	0.075	-0.128	0.898
as.factor(Education)4	-0.072	0.074	-0.967	0.334
as.factor(MaritalStatus)2	0.031	0.075	0.419	0.675
as.factor(MaritalStatus)3	-0.080	0.075	-1.072	0.284
as.factor(MaritalStatus)4	-0.012	0.075	-0.164	0.870

According to this model, controlling for age, race, education, and marital status, an additional year of living in public housing is associated with a 0.3 point increase in self-reported health status, and the difference is statistically significant ($p < 0.001$).

Task 3: Instrumental variables with 2SLS

This, however, is not a true causal effect, because there are endogenous elements of living in public housing. Perhaps people who care more about their health self-select into public housing. We can remove the endogenous part of public housing by using an instrument to predict the exogenous part of public housing use. We'll use waiting time, since it seemed to be the best instrument of the four we looked at.

In the instructions, I said you could just run `iv_robust()` to do both stages at once, but I'll do it by hand here too for the sake of showing the moving parts. First, we run the first stage model with all controls:

```
first_stage_housing <- lm(PublicHousing ~ WaitingTime + Age + as.factor(Race) +
                            as.factor(Education) + as.factor(MaritalStatus),
                          data = housing)
tidy(first_stage_housing)
```

term	estimate	std.error	statistic	p.value
(Intercept)	2.650	2.506	1.058	0.290
WaitingTime	0.946	0.044	21.372	0.000
Age	0.080	0.045	1.766	0.078
as.factor(Race)2	0.224	0.771	0.290	0.772
as.factor(Race)3	2.039	0.770	2.649	0.008
as.factor(Race)4	0.535	0.770	0.694	0.488
as.factor(Education)2	-0.294	0.768	-0.383	0.701
as.factor(Education)3	-0.715	0.771	-0.926	0.355
as.factor(Education)4	-0.100	0.770	-0.130	0.897
as.factor(MaritalStatus)2	0.141	0.774	0.182	0.856
as.factor(MaritalStatus)3	0.937	0.771	1.214	0.225
as.factor(MaritalStatus)4	0.752	0.772	0.974	0.330

```
glance(first_stage_housing)
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.residual	nobs
0.326	0.319	8.58	43.5	0	11	-3562	7150	7214	72700	988	1000

As expected, waiting time has a strong positive relationship with public housing (even after controlling for demographics), and the joint F-statistic for all these variables is 43.5, which is more than 10, but less than 104, which means our instrument is a lot weaker than we originally thought. OH NO. We should be wary of whatever effects we find.

Next we extract predicted public housing and add it to our original data frame:

```
housing <- augment_columns(first_stage_housing, housing) %>%
  rename(PublicHousing_hat = .fitted)
```

Now we run our second stage using `PublicHousing_hat` instead of `PublicHousing`, since the new "hatted" version has the endogeneity removed.

```
second_stage_housing <- lm(HealthStatus ~ PublicHousing_hat + Age + as.factor(Race) +
                             as.factor(Education) + as.factor(MaritalStatus),
                           data = housing)
tidy(second_stage_housing)
```

term	estimate	std.error	statistic	p.value
(Intercept)	1.419	0.948	1.496	0.135
PublicHousing_hat	0.224	0.017	12.964	0.000
Age	0.062	0.017	3.680	0.000
as.factor(Race)2	0.031	0.285	0.110	0.912
as.factor(Race)3	0.301	0.288	1.047	0.296
as.factor(Race)4	0.032	0.285	0.114	0.909
as.factor(Education)2	-0.011	0.284	-0.039	0.969
as.factor(Education)3	-0.116	0.286	-0.406	0.685
as.factor(Education)4	-0.049	0.285	-0.173	0.863
as.factor(MaritalStatus)2	0.079	0.286	0.275	0.783
as.factor(MaritalStatus)3	0.017	0.286	0.060	0.953
as.factor(MaritalStatus)4	0.097	0.286	0.337	0.736

Based on this model, an additional year in public housing causes a 0.224 increase in health status. We can talk about causation because we've dealt with endogeneity of public housing usage with our instrument.

Here's the same model with `iv_robust()`. Notice how the same controls go in both stages.

```
model_2sls_housing <-
  iv_robust(HealthStatus ~ PublicHousing + Age + as.factor(Race) +
              as.factor(Education) + as.factor(MaritalStatus) |
              WaitingTime + Age + as.factor(Race) +
              as.factor(Education) + as.factor(MaritalStatus),
            data = housing)
```

While this is neat, we should check for the effect of weak instruments since the F statistic is actually not in the 400s once we control for age, race, education, and marital status—it's now just 43. Let's look at the Anderson-Rubin confidence interval:

```
# I wish anderson.rubin.ci() worked with iv_robust() models, but alas
model_housing_again <-
  ivreg(HealthStatus ~ PublicHousing + Age + as.factor(Race) +
          as.factor(Education) + as.factor(MaritalStatus) |
          WaitingTime + Age + as.factor(Race) +
          as.factor(Education) + as.factor(MaritalStatus),
        data = housing, x = TRUE)

anderson.rubin.ci(model_housing_again)
```

```
## $confidence.interval
## [1] "[ 0.20905888610579 , 0.237135198367737 ]"
```

The effect here ranges between 0.209 and 0.237, which is a fairly narrow band, and it doesn't include 0, so even though the joint first stage F statistic is small-ish and might be weak and not relevant, it probably works here. Phew.

Also, just for fun, we can run the perfect model because I've included a column for the omitted health behavior measure, which explains all the endogeneity in public housing use. In real life this variable doesn't exist—we use instruments to get rid of this effect. How'd we do, using waiting time as an instrument?

```
model_perfect <- lm(HealthStatus ~ PublicHousing + HealthBehavior +
                      as.factor(Race) + as.factor(Education) +
```

```
                      as.factor(MaritalStatus),
                    data = housing)
```

```
library(kableExtra)  # For fancier table formatting

modelsummary(list("OLS" = model_naive_housing, "2SLS" = second_stage_housing,
                  "2SLS robust" = model_2sls_housing, "Perfect" = model_perfect),
             gof_omit = 'IC|Log|Adj|p\\.value|statistic|se_type',
             stars = TRUE) %>%
  row_spec(c(3, 25), background = "yellow")
```

	OLS	2SLS	2SLS robust	Perfect
(Intercept)	-1.197***	1.419	1.419***	0.623***
	(0.226)	(0.948)	(0.381)	(0.085)
PublicHousing	0.321***		0.224***	0.228***
	(0.003)		(0.007)	(0.003)
Age	0.056***	0.062***	0.062***	
	(0.004)	(0.017)	(0.007)	
as.factor(Race)2	-0.012	0.031	0.031	-0.040
	(0.075)	(0.285)	(0.120)	(0.055)
as.factor(Race)3	0.052	0.301	0.301**	0.049
	(0.075)	(0.288)	(0.120)	(0.055)
as.factor(Race)4	-0.018	0.032	0.032	-0.041
	(0.074)	(0.285)	(0.116)	(0.055)
as.factor(Education)2	0.023	-0.011	-0.011	0.015
	(0.074)	(0.284)	(0.118)	(0.054)
as.factor(Education)3	-0.010	-0.116	-0.116	-0.054
	(0.075)	(0.286)	(0.115)	(0.055)
as.factor(Education)4	-0.072	-0.049	-0.049	-0.066
	(0.074)	(0.285)	(0.113)	(0.054)
as.factor(MaritalStatus)2	0.031	0.079	0.079	0.028
	(0.075)	(0.286)	(0.115)	(0.055)
as.factor(MaritalStatus)3	-0.080	0.017	0.017	-0.125**
	(0.075)	(0.286)	(0.115)	(0.055)
as.factor(MaritalStatus)4	-0.012	0.097	0.097	-0.021
	(0.075)	(0.286)	(0.113)	(0.055)
PublicHousing_hat		0.224***		
		(0.017)		
HealthBehavior				0.269***
				(0.008)
Num.Obs.	1000	1000		1000
R2	0.943	0.172	0.861	0.970
F	1497.757	18.662		2867.248
N			1000	

* p < 0.1, ** p < 0.05, *** p < 0.01

Waiting time works pretty well! The true effect is 0.228ish, and after using an instrument to removed the endogeneity from `PublicHousing`, the effect in the 2SLS model is 0.224. Success!