14.32: Econometrics

Problem Set 6

due Wednesday, November 26, 2025 at 1pm

1. (a) A researcher plans to run an IV regression:

$$Y_i = \alpha + \beta X_i + e_i, \tag{1}$$

and is concerned that X_i may be correlated with e_i. He has heard that IV is a great strategy to estimate causal parameters. He generates a new variable

$$Z_i = X_i + \xi_i,$$

where ξ_i is randomly generated noise independent from the data. Prove that Z_i is a valid instrument if and only if X_i is exogeneous.

(b) In the previous question, assume that X_i is exogenous. Calculate the asymptotic variance of the IV and compare it to the OLS. Which estimator would you prefer?

(c) The IV estimator of the returns to education (β) in the following regression:

$$wage_i = \alpha + \beta educ_i + \epsilon_i,$$

with an instrument Z_i, which is a dummy variable for a person to be born in the first quarter of a year, is exactly the same as the following estimator:

$$\hat{\beta} = \frac{\overline{wage}_1 - \overline{wage}_0}{\overline{educ}_1 - \overline{educ}_0}$$

where $\overline{x}_1$ denotes the sample average of x in the group of individuals born in the first quarter, and $\overline{x}_0$ denotes the sample average of x in the other group.

2. Harvard economist Claudia Goldin attributes much of the rise of professional women in the U.S. labor force to their ability to engage in family planning after the introduction of the birth-control pill. In developing countries early childbearing is associated with lower levels of education and more dependency of women on their husband's earnings.

This question examines the effect of family size on female labor supply. The data set consists of observations on n = 254,654 married women, aged 21 – 35, who have at least two children. The data come from the 1980 U.S. Census of the Population (the data pertain to the full calendar year 1979).

Variables in the Female Labor Supply Data Set:

Variable	Definition
Wife's weeks worked	No. of weeks wife worked for pay in 1979
Husband's weeks worked	No. of weeks husband worked for pay in 1979
Same sex	= 1 if first two children are of the same sex, = 0 otherwise
2 boys	= 1 if first two children are boys, = 0 otherwise
2 girls	= 1 if first two children are girls, = 0 otherwise
Kids>2	= 1 if family has more than 2 children, = 0 otherwise
Boy first	= 1 if first child is a boy, = 0 otherwise
Current age of mother	age of mother in 1979
Age of mother at 1st birth	age of mother at birth of first child
Black	= 1 if black
Hispanic	= 1 if Hispanic
Other race	= 1 if nonwhite/nonblack/nonHispanic

(1) Consider the hypothesis that, on average, U.S. parents want to have children of both genders (that is, they prefer at least one girl and one boy to all girls or all boys). Does Table 1 provide evidence in favor of this hypothesis, against this hypothesis, or neither? Explain.

(2) To address a question on the effect of having more than two kids on the women's labor supply you run regression (3). Discuss why the answer provided by regression (3) may be invalid.

(3) Consider the following potential instrumental variables for *Kids*>2 in regression (3):

a) Whether the wife came from a large family (binary)

b) The teen pregnancy rate in the wife's city or town of residence

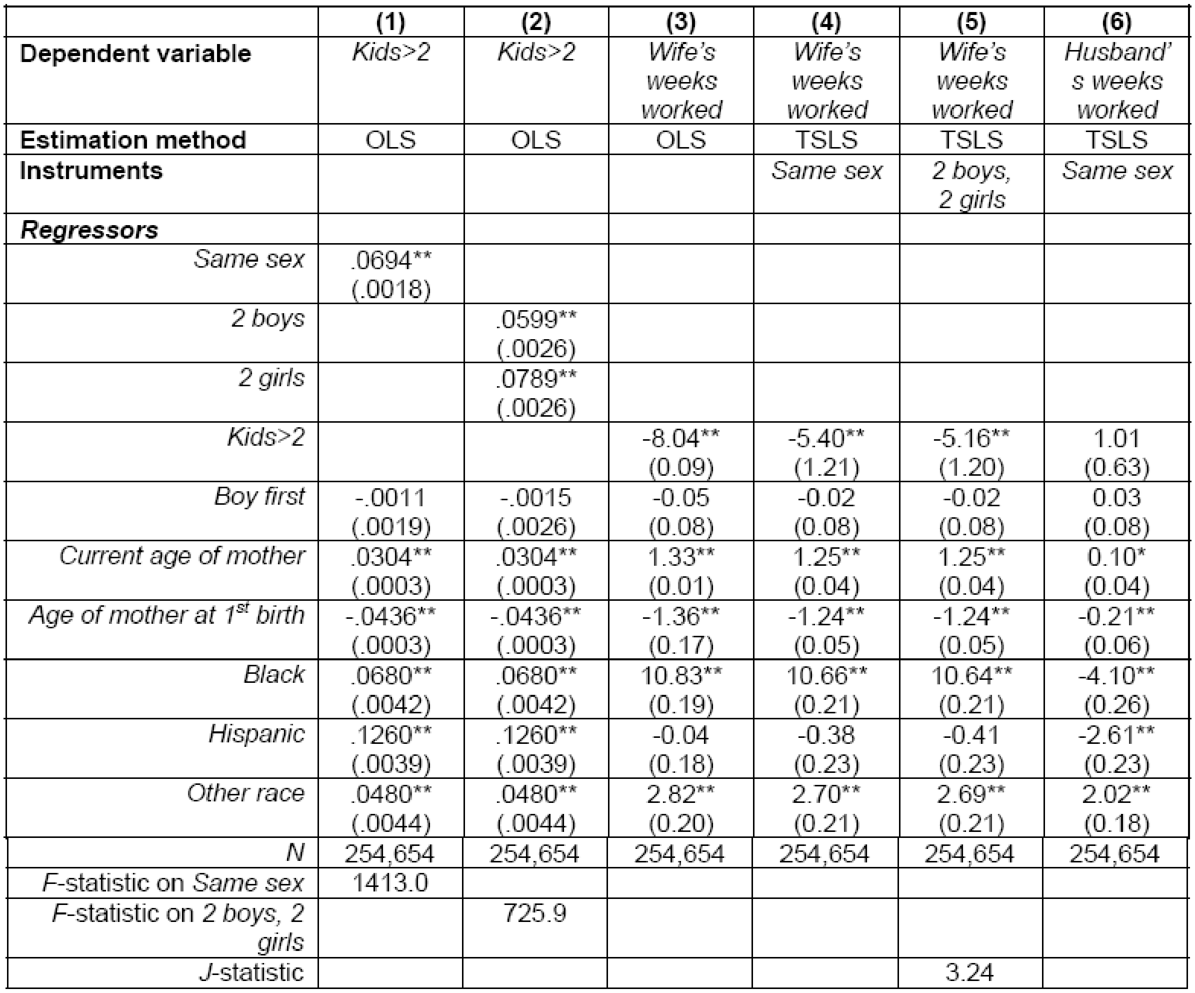
For each proposed instrument, is the variable arguably a valid instrumental variable? Briefly explain.

(4) Based on a combination of your judgment and the empirical results in Table 1, please, address the issue theoretically and test assumptions whenever it is possible:

a) Is *Same sex* a valid instrument in regression (4)?

b) Is the pair of variables, *2 boys* and *2 girls*, a valid set of instruments in regression (5)?

Table 1
Child Sex Composition, Family Size, and Labor Supply



	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Dependent variable	*Kids>2*	*Kids>2*	*Wife's weeks worked*	*Wife's weeks worked*	*Wife's weeks worked*	*Husband's weeks worked*
Estimation method	OLS	OLS	OLS	TSLS	TSLS	TSLS
Instruments				*Same sex*	*2 boys, 2 girls*	*Same sex*
Regressors						
Same sex	.0694** (.0018)					
2 boys		.0599** (.0026)				
2 girls		.0789** (.0026)				
Kids>2			-8.04** (0.09)	-5.40** (1.21)	-5.16** (1.20)	1.01 (0.63)
Boy first	-.0011 (.0019)	-.0015 (.0026)	-0.05 (0.08)	-0.02 (0.08)	-0.02 (0.08)	0.03 (0.08)
Current age of mother	.0304** (.0003)	.0304** (.0003)	1.33** (0.01)	1.25** (0.04)	1.25** (0.04)	0.10* (0.04)
Age of mother at 1st birth	-.0436** (.0003)	-.0436** (.0003)	-1.36** (0.17)	-1.24** (0.05)	-1.24** (0.05)	-0.21** (0.06)
Black	.0680** (.0042)	.0680** (.0042)	10.83** (0.19)	10.66** (0.21)	10.64** (0.21)	-4.10** (0.26)
Hispanic	.1260** (.0039)	.1260** (.0039)	-0.04 (0.18)	-0.38 (0.23)	-0.41 (0.23)	-2.61** (0.23)
Other race	.0480** (.0044)	.0480** (.0044)	2.82** (0.20)	2.70** (0.21)	2.69** (0.21)	2.02** (0.18)
N	254,654	254,654	254,654	254,654	254,654	254,654
F-statistic on *Same sex*	1413.0					
F-statistic on *2 boys, 2 girls*		725.9				
J-statistic					3.24	

Notes: Regressions (4), (5), and (6) are estimated by two-stage least squares (TSLS) regression, in which the included endogenous variable is *Kids>2*. Heteroskedasticity-robust standard errors appear in parentheses under the regression coefficients, and *p*-values appear in parentheses under *F*-statistics. All regressions include an estimated intercept, which is not reported. Regressions (1) – (5) are estimated using data on married women for 1979, regression (6) is estimated using data for the husbands of those married women.
Significant at the: **1%, *5% significance level.

(5) The estimated coefficient on *Kids*>2 differs in regressions (3) and (4) (the OLS estimate is more negative than the TSLS estimate). Provide a real-world explanation (an interpretation of the results) that explains why the OLS estimate is more negative than the TSLS estimate.

(6) Use Table 1 to comment on the following statement: Even though having large families reduces female labor force participation, there may be no impact on the household's economic wellbeing overall because their husbands will work more to compensate for the loss of the wife's earnings.

3. During the 1880s, a cartel known as the Joint Executive Committee (JEC) controlled the rail transport of grain from the Midwest to eastern cities in the United States. The cartel preceded the Sherman Antritrust Act of 1890, and it legally operated to increase the price of grain transportation above what would have been the competitive price. From time to time, cheating by members of the cartel brought about a temporary collapse of the collusive price-setting agreement. In this exercise, you will use variations in supply associated with the cartel's collapses to estimate the elasticity of demand for rail transport of grain. We will use a data file `JEC.dta` that contains weekly observations on the rail shipping price and other factors from 1880 to 1886.

   So, our main goal is to estimate the elasticity of **demand** for rail shipping of grain. In particular, we wish to regress $\ln(Q_i)$ where Q_i is the total tonnage of grain shipped in week i, on $\ln(P_i)$, where P_i is the price of shipping a ton of grain by rail.

   (a) Estimate the price elasticity of demand by using OLS to regress the log of the quantity of grain shipped on the log of the price and the full set of monthly binary indicators. (Since monthly binary indicators are named *seas1* ,..., *seas13* in this dataset, in Stata you can include them in the regression using *seas** as a shorthand.) What is the estimated value of the demand elasticity and its standard error?

   (b) Explain why the interaction of supply and demand could make the OLS estimator of the elasticity obtained in (a) biased.

   (c) Consider using the variable *cartel* as instrumental variable for $\ln(P)$. Use economic reasoning to argue whether *cartel* plausibly satisfies the two conditions for a valid instrument.

(d) One of the reasons for price fluctuations was that the Great Lakes periodically froze, making shipping grain by boat impossible and temporarily increasing the demand for rail. Ice_i is a binary variable that is equal to 1 if the Great Lakes are not navigable because of ice. Consider using the variable Ice_i as instrumental variable for $\ln(P_i)$. Use economic reasoning to argue whether Ice plausibly satisfies the two conditions for a valid instrument.

(e) Run 3 different TSLS regressions of the log of the quantity of grain shipped on the log of the price and the full set of monthly binary indicators using (i) *cartel*; (ii) *ice*; (iii)*cartel* and *ice* as instrument for $\ln(P_i)$.

(f) Check whether the instruments in regressions (e) are strong.

(g) Perform J-test of overidentifying restrictions in regression (e) (iii). What conclusions can you draw?

(h) What is your preferred regression? Why? What are the results of your study? What is your conclusion on the demand elasticity of railroad shipment?

DATA DESCRIPTION, FILE: JEC.dta

JEC contains weekly observations on prices and other factors from 1880-1886, for a total of n = 326 weeks.

Variable	Definition
week	week of observation: = 1 if 1/1/1880-1/7/1880, = 2 if 1/8/1880-1/14/1880, ..., = 328 for final week
price	= weekly index of price of shipping a ton of grain by rail
ice	= 1 if Great Lakes are impassable because of ice, = 0 otherwise
cartel	= 1 railroad cartel is operative, = 0 otherwise
quantity	= total tonnage of grain shipped in the week
seas1 ,..., *seas13*	= thirteen month binary variables. To match the weekly data, the calendar has been divided into 13 periods, each approximately 4 weeks long. Thus seas1 = 1 if date is January 1 through January 28, =0 otherwise seas2 = 1 if date is January 29 through February 25, =0 otherwise seas13 = 1 if date is December 4 through December 31, =0 otherwise