Causal Inference for Policy Evaluation
Assignment 3

Marco Gortan, Felix Schulz, Benjamin Weggelaar

May 2, 2025

	First Stage		Second Stage		OLS
	1(b)1	1(c)1	1(b)2	1(c)2	1(e)
(Intercept)	0.35***	0.35***	43.49***	43.44***	43.54***
	(0.00)	(0.00)	(0.42)	(0.41)	(0.04)
samesex	0.06***				
	(0.00)				
boys2		0.05***			
		(0.00)			
girls2		0.08***			
		(0.00)			
morekids			0.11	0.23	-0.03
			(1.10)	(1.07)	(0.07)
R^2	0.00	0.00	-0.00	-0.00	0.00
Adj. R^2	0.00	0.00	-0.00	-0.00	-0.00
Num. obs.	125725	125725	125725	125725	125725

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: Dad's hours worked and fertility: OLS and IV

	2(b)	2(c)
(Intercept)	6.17***	4.35***
	(0.88)	(1.24)
morekids	-7.81***	-2.28
	(0.16)	(2.65)
agem	1.41***	1.22***
	(0.03)	(0.09)
agefstm	-1.31***	-1.06***
	(0.04)	(0.13)
blackm	11.14***	10.79***
	(0.36)	(0.40)
hispm	1.54***	0.80
	(0.44)	(0.57)
othracem	3.13***	2.85***
	(0.49)	(0.52)
R^2	0.06	0.05
Adj. R^2	0.06	0.05
Num. obs.	69299	69299

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 2: Mother's weeks worked and fertility: OLS and IV

Instrumental Variables

Question 1

(a) The father has a much higher income than the mother. This hints a division in which the father works outside the household, while the mother is responsible for domestic chores and childcare.

(d) The instrument used in (b) is a weighted average of the variables used in (c). The assumption of exogeneity should therefore hold in both setups. Due to the large uncertainty in both the second stage estimates, we abstain from interpreting the small observed difference in point estimates. First stage results are highly significant for both, additionally showing that more couples seek to have another child after having two girls compared to two boys.

Table 3: Summary Statistics of Parents

Parent	Age at First Birth	Average Income
Mother	20.90	6232.10
Father	24.03	39030.66

Table 4: Share with >2 Children by Age Group

Age Group	Share with >2 Children	N
$age \leq 31$	0.34	69299
$age > 31$	0.42	56426

(e) Similar to the IV approach, the OLS regression does not yield a statistically significant result. While the sign of the estimate is the opposite of the above, this cannot be interpreted with confidence. Still, the IV approach provides an attractive causal identification and is therefore preferred.

Question 2

(b) The central endogeneity concern is that labor market participation might be caused by some confounding factors. The most likely candidates are controlled for here: Older women are more likely to have children, but also to work. Women that got children at a young age are more likely to have foregone education and therefore less likely to work. Race-specific culture is also likely to confound. However, there might still be endogeneity concerns. Although these variables are likely correlated with our controls, we do not completely observe general attitude toward work/leisure balance and access to child support, making the IV approach more attractive.

(c) While strongly negative and highly statistically significant when using OLS, the point estimate for fertility is, after instrumenting it with the samesex variable, statistically insignificant. The IV LATE shows that we cannot causally identify an impact of fertility on labor market participation and that effects are likely due to other characteristics of the mother causing both variables at the same time.

(d) Restricting the sample to women below the median age can weaken the first stage because younger women may not have completed their fertility decisions yet. This weakens the predictive power of the instrument (same-sex first two children) and risks weak instrument bias. It also introduces sample selection bias, making the 2SLS estimates less representative of the full population.

Question 3

(a) The female labor participation rate is 52.7% while the male labor participation rate is 97.7%. The share of parttime workers is 51.9% for moms and 6.2% for dads. Today there may be a higher expectation for dads to help with childcare and household tasks, due to weaker gender roles now compared to before. We can observe this in the data too, since more fathers do part-time work and have lower labor participation than before. As a consequence, running the IV model with current data, would more likely yield a negative causal effect on male labor supply as a result of their partner's fertility.

(b) We notice that the fertility rate in Switzerland today (1.4) is smaller compared to the estimates of the fertility rate in the US in the eighties (1.8). We therefore think it is more appropriate to use, as instrument in the first stage, their $Twins - 2$, since the rationale that the authors bring forward to support the *Same Sex* instrument, i.e. the willingness to diversify the sexes of the kids, might not apply anymore. However, the $Twin - 2$ channel might still be present. This is especially true if the limitations for parents to have more kids is to go through the entire pregnancy cycle, which causes temporary absence from job more. In the case of twins, there is only one pregnancy

cycle for two kids. Moreover, since women in Switzerland have children at a later age nowadays, we should observe a higher percentage (and so more power) of women with twins since, as the authors argue, the probability of having twins is positively correlated with age.

(c) Because women get children at much higher age (the average age of the mother at the birth of the first child is 31), restricting the age of the mother to be 35 maximum is excessively tight. We should raise this upper bound.

Question 4 (PhD Only)

(a) The (Hansen-Sargan) J-test assesses the exogeneity of instruments in a parametric IV setting, when multiple instruments are available. The test works by taking the fitted residuals from the two-stage least squares $\hat{u}_i^{TSLS}$, and run the following regression:

$$\hat{u}_i^{TSLS} = \delta_0 + \delta_1 Z_{1i} + ... + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + ... + \delta_{m+r} W_{ri} + e_i \tag{1}$$

Where the Z are the instruments and W are the exogenous regressors (control). The J-test has as null hypothesis: $H_0 : \delta_1 = ... = \delta_m = 0$, which states that all instruments are exogenous. The J-statistic takes the F-statistic such that $J = mF$, which follows a chi-square distribution $J \sim \chi^2_{m-k}$ with $m - k$ degrees of freedom. The intuition here is to test whether instruments explain any variation in the dependent variable (through the residuals) beyond the channel of the endogenous variable. We could observe this if any of the $\hat{\delta}_i$ are significantly different from zero.

(b) The test is informative about instrument validity if we assume that the model is correctly specified, and at least one of the instruments is valid. This is because the test looks into whether all instruments are jointly valid. If all are invalid, the test might fail to detect that. On the other hand, if all instruments are valid, it should correctly conclude that the instruments are exogenous.

Regression Discontinuity Design

Question 6

(a) 1990 Census data is used for a placebo test of the RD design. Since the 1990 outcomes predate the treatment, there should be no discontinuity at the cutoff if the design is valid. The observed smooth pattern supports the identification strategy by showing that any post-election discontinuities are likely due to Islamic rule, not pre-existing differences between municipalities.

(b) The Imbens-Kalyanaraman rule picks the RD window that balances bias and variance for a local-linear fit. Because women's outcome curve is more curved right at the cutoff, extending too far would add bias, so the formula squeezes their bandwidth to 0.24. The men's curve is flatter, so the rule can keep more data (bandwidth = 0.32) without sacrificing accuracy.

(c) The author explains the positive RD results is driven by areas which are both poorer and more religiously conservative. These areas exhibit higher barriers to entry for women compared to men, because secular restrictions such as the headscarf ban, and mixed classes, made it more likely for parents to not send their daughters to school. Islamist municipalities would then lower these barriers to entry by not enforcing secular restrictions. This is tested by assuming heterogeneous RD treatment effects, and thus re-estimating the RD models for subsamples, where the original sample is split in two based on being above or below the median for literacy share, share of religious buildings, and islamic vote share.