Assignment 6

Dmitrii Kuptsov

November 15, 2025

Task 3

Task 3.1

First, we need to estimate the following equation:

$$y_i = \beta_0 + \beta_1 \,\text{kidcount}_i + X_i'\beta + \varepsilon_i,$$

where y_i is the employment status of woman i, $kidcount_i$ is the number of children for woman i and X is the matrix of the controls. We need to decide, which variables from the dataset can be used as controls:

1. Sex of the first child (sexk). From the data we know, that our sample is limited to the women who have at least 2 children. Therefore, there wouldn't be the information from kidcount (because we know for sure that each woman has at least one child). As for the necessity of this control, it may be argued that not only number of children, but also their characteristics (such as sex, age etc.) can possibly affect the employment of their mother.

2. Mother's age and race should be included because they are key confounding variables that affect both fertility and labor supply. Age is strongly correlated with employment decisions and with the likelihood of having additional children. Race is correlated with labor market opportunities, discrimination, and cultural norms that influence both childbearing and work. Omitting these variables would lead to omitted variable bias, since kidcount would be correlated with unobserved determinants of employment.

To estimate OLS, we need to run the following regression:

reg workedm kidcount agem sexk blackm hispm othracem

Although OLS can be used with a binary dependent variable as a Linear Probability Model, it has well-known disadvantages. For example, it can predict probabilities outside the [0,1] range. Probit addresses these functional-form issues by modeling employment as a probability bounded between 0 and 1.

The results show that β_1 estimate is significant at any reasonable level. If we interpret this OLS as linear probability model, we can say that the probability of an employment decreases on 9% with an additional child. Also, it is important to mention that this model is measuring the degree of association, so it is not causal. OLS and probit do not recover a causal effect because kidcount is endogenous. Fertility decisions are correlated with unobserved preferences, labor market attachment, and household characteristics, all of which also affect employment. There is also reverse causality (labor supply influences fertility). Adding demographic controls does not eliminate this correlation. Therefore, both OLS and probit estimate associations, not causal effects.

```
reg workedm kidcount agem sexk blackm hispm othracem
```

Source	SS	df	MS			
				Number of obs	=	400,169
				F(6, 400162)	=	2374.95
Model	3380.97092	6	563.495154	Prob > F	=	0.0000
Residual	94944.7906	400,162	.237265884	R-squared	=	0.0344
				Adj R-squared	=	0.0344
Total	98325.7615	400,168	.245711205	Root MSE	=	.4871

workedm	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
kidcount	-.0910794	.0009621	-94.67	0.000	-.092965	-.0891937
agem	.0146908	.0002214	66.37	0.000	.014257	.0151247
sexk	.0009097	.0015405	0.59	0.555	-.0021095	.003929
blackm	.1506663	.0023869	63.12	0.000	.1459879	.1553446
hispm	-.0083029	.0045233	-1.84	0.066	-.0171684	.0005626
othracem	.0275119	.004631	5.94	0.000	.0184353	.0365885
_cons	.3372181	.0069022	48.86	0.000	.32369	.3507463

Figure 1: Results of the OLS regression

Next, we change the model from linear probability model to probit. Therefore, now the output of the model is limited from 0 to 1. However, this model is still measuring the association, and not causality. The results of the probit model are below.

Task 3.2

To make our model causal, we instrument the number of children with the indicator that the last birth was to twins. A twin birth mechanically increases the number of children, so twin_latest → kidcount, which satisfies the relevance condition. The first-stage regression confirms this: mothers whose last birth was a twin have significantly more children on average.

For the instrument to be valid, it must also satisfy the exclusion restriction, meaning that twin_latest should affect employment only through its impact on the number of children. The main justification for this assumption is that the occurrence of twins is largely random and not chosen by the mother. However, this assumption may be imperfect in practice. Twin births can influence the mother's employment directly through channels other than family size—for example, a twin birth may increase childcare burden, create additional expenses, or be associated with birth complications, all of which can directly affect labor supply. These channels violate the strict exclusion restriction because they represent effects of twin_latest that are not mediated solely through kidcount.

The results of the IV probit regression are shown below. Compared to the standard probit estimates, the magnitude of the coefficient on kidcount becomes smaller and is significant only at the 10% level. The IV probit coefficient on kidcount is –0.07, indicating that an additional child reduces the latent propensity for employment. Probit coefficients do not translate directly into probability changes, so this value should not be interpreted as a 7 percentage point decrease. The sign and magnitude indicate a negative association

```
. probit workedm kidcount agem sexk blackm hispm othracem

Iteration 0:  Log likelihood = -273933.15
Iteration 1:  Log likelihood = -266940.63
Iteration 2:  Log likelihood = -266932.27
Iteration 3:  Log likelihood = -266932.27

Probit regression                                      Number of obs =  400,169
                                                       LR chi2(6)    = 14001.76
                                                       Prob > chi2   =   0.0000
Log likelihood = -266932.27                            Pseudo R2     =   0.0256
```

workedm	Coefficient	Std. err.	z	P>|z|	[95% conf. interval]	
kidcount	-.2370876	.0025497	-92.99	0.000	-.2420848	-.2320903
agem	.038261	.0005794	66.03	0.000	.0371253	.0393967
sexk	.0025206	.0040251	0.63	0.531	-.0053684	.0104096
blackm	.403935	.006422	62.90	0.000	.3913482	.4165218
hispm	-.0203397	.0117698	-1.73	0.084	-.0434081	.0027287
othracem	.0717952	.0120918	5.94	0.000	.0480956	.0954947
_cons	-.4275721	.0179722	-23.79	0.000	-.4627971	-.3923472

Figure 2: Results of the probit regression

in the latent index, but the corresponding effect on employment probability must be evaluated using marginal effects.

Task 3.3

Here is the marginal effect of the number of children on the probability of mother's employment.

As we can see, the marginal effect becomes more negative when moving from two to four or five children, suggesting a stronger association between having additional children and lower employment probability. However, the pattern after the fifth child should not be interpreted substantively. Very few mothers in the sample have five or more children, so the confidence intervals widen substantially in that range and the apparent reversal is not statistically meaningful. In other words, the shape of the curve beyond four children is driven mostly by limited data and estimation noise, rather than by a real economic effect.

first-stage regression

Source	SS	df	MS
Model	10775.5354	6	1795.92257
Residual	255376.5	400,162	.638182785
Total	266152.035	400,168	.665100745

Number of obs	=	400,169
F(6, 400162)	=	2814.12
Prob > F	=	0.0000
R-squared	=	0.0405
Adj R-squared	=	0.0405
Root MSE	=	.79886

kidcount	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
twin_latest	.3850044	.0099461	38.71	0.000	.3655104	.4044983
agem	.030971	.0003598	86.08	0.000	.0302658	.0316761
sexk	.0138442	.0025263	5.48	0.000	.0088927	.0187958
blackm	.3235942	.0038811	83.38	0.000	.3159874	.3312011
hispm	.4370486	.0073863	59.17	0.000	.4225716	.4515256
othracem	.1210053	.0075927	15.94	0.000	.1061238	.1358868
_cons	1.557758	.0110493	140.98	0.000	1.536101	1.579414

Figure 3: Results of the first stage of IV probit regression

Probit model with endogenous regressors

Log likelihood = -744871.86

Number of obs	=	400,169
Wald chi2(6)	=	5126.47
Prob > chi2	=	0.0000

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
kidcount	-.0715877	.0413427	-1.73	0.083	-.1526179	.0094424
agem	.0328903	.0015186	21.66	0.000	.0299139	.0358667
sexk	.0002138	.0040526	0.05	0.958	-.0077292	.0081569
blackm	.3473656	.0161197	21.55	0.000	.3157715	.3789597
hispm	-.0915425	.0210614	-4.35	0.000	-.1328221	-.0502628
othracem	.0515402	.0131033	3.93	0.000	.0258581	.0772222
_cons	-.6801591	.0638538	-10.65	0.000	-.8053103	-.555008
corr(e.kidcount, e.workedm)	-.1310838	.0322812			-.1937386	-.0673641
sd(e.kidcount)	.7988564	.000893			.7971082	.8006085

Wald test of exogeneity (corr = 0): chi2(1) = 16.11 Prob > chi2 = 0.0001
Endogenous: kidcount
Exogenous: agem sexk blackm hispm othracem twin_latest

Figure 4: Results of the IV probit regression

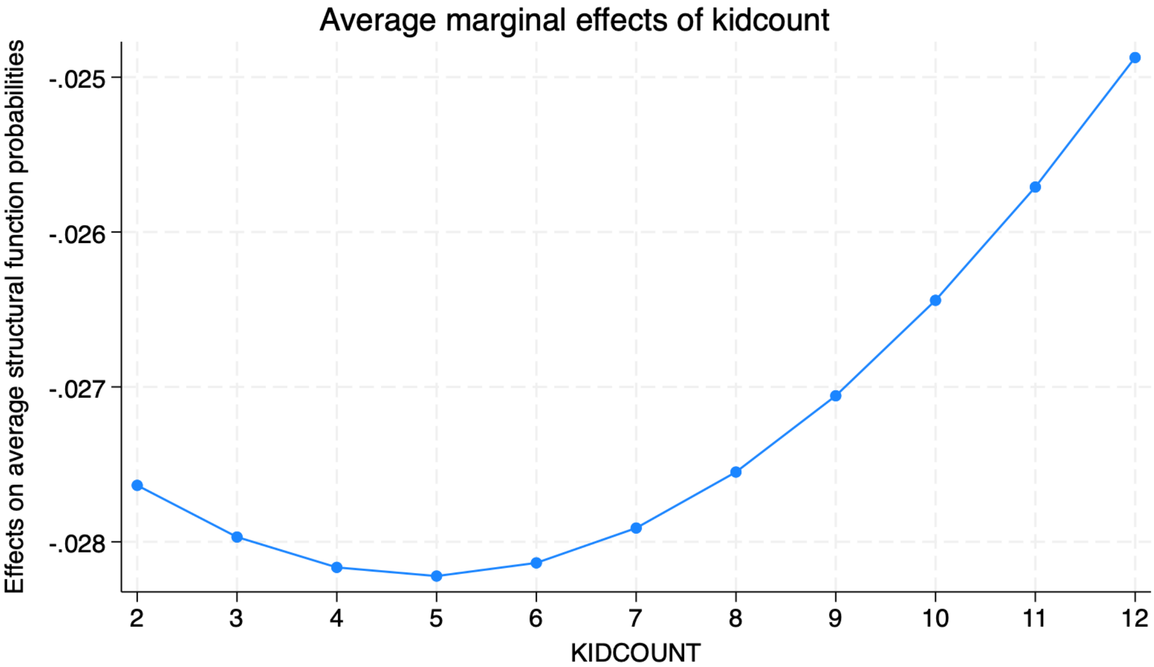

Figure 5: Marginal effect of the number of children on the probability of mother's employment