TD5 :
Endogeneity and instrumental variables (continued)
Maximum likelihood and limited dependent variable models

Exercise 1 (from Stock and Watson (2012))**:**
How does fertility affect labor supply? That is, how much does a woman's labor supply fall when she has an additional child? In this exercise you will estimate this effect using data in FERTILITY.XLS. The data set contains information on 30,000 randomly selected married women aged 21-35 with two or more children. More precisely, the variables in data set are:

Variable	Description
morekids	=1 if mom had more than 2 children
samesex	=1 if 1st two children same sex
agem1	age of mom
black	=1 if mom is black
hispan	=1 if mom is Hispanic
othrace	=1 if mom is not black, Hispanic or white
weeksm1	mom's weeks worked in 1979

a. Regress *weeksm1* on the indicator variable *morekids* using OLS. On average, do women with more than two children work less than women with two children? How much less?
b. Explain why the OLS regression estimated in (a) is inappropriate for estimating the causal effect of fertility (*morekids*) on labor supply (*weeksm1*).
c. The data set contains the variable *samesex*, which is equal to 1 if the first two children are of the same sex (boy-boy or girl-girl) and equal to 0 otherwise. Are couples whose first two children are of the same sex more likely to have a third child? Is the effect large? Is it statistically significant?
d. Explain why *samesex* is a valid instrument for the instrumental variable regression of *weeksm1* on *morekids*.
e. Is *samesex* a weak instrument?
f. Estimate the regression of *weeksm1* on *morekids* using *samesex* as an instrument. How large is the fertility effect on labor supply?
g. Do the results change when you include the variables *ageml* , *black*, *hispan*, and *othrace* in the labor supply regression (treating these variable as exogenous)? Explain why or why not.

Exercise 2 (from Wooldridge (2009))**:**

Use the data in 401KSUBS.XLS for this exercise. The equation of interest is a *linear probability model*:

$$pira = \beta_0 + \beta_1\, p401k + \beta_2\, inc + \beta_3\, inc^2 + \beta_4\, age + \beta_5\, age^2 + u.$$

The goal is to test whether there is a tradeoff between participating in a 401(k) plan and having an individual retirement account (IRA)[1]. Therefore, we want to estimate

a. Estimate the equation by OLS and discuss the estimated effect of *p401k.*

[1] Individual Retirement Account (IRA) and 401(k) are 2 types of retirement plans in US. The first is provided by many financial institutions, and the second, by employers.

b. For the purposes of estimating the ceteris paribus tradeoff between participation in two different types of retirement savings plans, what might be a problem with ordinary least squares?
c. The variable *e401k* is a binary variable equal to one if a worker is *eligible* to participate in a 401(k) plan. Explain what is required for *e401k* to be a valid IV for *p401k.* Do these assumptions seem reasonable?
d. Estimate the reduced form for *p401k* and verify that *e401k* has significant partial correlation with *p401k* (use a heteroskedasticity-robust standard error (see exercise 4 for explanation)).
e. Now, estimate the structural equation by IV and compare the estimate of β_1 with the OLS estimate. Again, you should obtain heteroskedasticity-robust standard errors.
f. Test the null hypothesis that *p401k* is in fact exogenous, using a heteroskedasticity robust test.

Exercise 3: (Maximum likelihood (ML) method)

We consider a discrete variable X with the following distribution:

X	**0**	**1**	**2**	**3**
Pr(X)	2θ/3	θ/3	2(1-θ)/3	(1-θ)/3

We draw a iid random sample for X : (3, 0, 2, 1, 3, 2, 1, 0, 2, 1). Find the ML estimate for θ.

Exercise 4: (Linear probability model)

We consider the linear probability model: $Y_i = \beta_0 + \beta_1 X_i + u_i$, where $\Pr(Y_i = 1) = \beta_0 + \beta_1 X_i$, i=1,…n.

a. Show that $E(u_i|X_i)=0$.
b. Compute $V(u_i|X_i)$ (***Reminder*** : variance of a discrete r.v. Y, with K values each with probability p_i, and with mean m : $V(Y) = E[(Y-m)^2] = \sum_{k=1}^{K}(y_k - m)^2 p_k$). Is u_i heteroskedastic?
c. Derive the likelihood function.