Microeconometrics

Spring Semester 2019/20

Problem set 2: Solution topics

Please ignore any obvious typos.

Exercise 1 (Discrete choice models)

For this exercise please create a do-file and corresponding log-file to hand in with your solution.

Use the dataset **laborparticipation.dta** that provides information on labor force participation of married women during 1975. To model labor market participation we consider the dependent variable $inlf$, which is equal to one if the woman reports working for a wage outside the home at some point during the year, and zero otherwise. The independent variables are the husband's income ($nwifeinc$, measured in thousands of dollars), years of education ($educ$), years of labor market experience ($exper$), age, number of children aged less than six years old ($kidslt6$), and number of kids between six and 18 years of age ($kidsge6$).

(a) Estimate a linear probability model (LPM) with heteroskedasticity-robust standard errors. Why might a linear probability model not be suitable for modelling the probability of labor force participation?

The two most important drawbacks of a linear probability model (LPM) are that the fitted probabilities can be less than zero or greater than one and the partial effect of any explanatory variable (included in level form) is constant. Also, due to the binary nature of the dependent variable, the LPM is intrinsically heteroskedastic (except in the case that the variance does not depend on any of the independent variables).

Stata command: reg inlf nwifeinc exper educ age kidslt6 kidsge6, robust

(b) Estimate the same model specification using the probit and the logit estimators. Interpret the results.

In a probit or logit model we define the probability of a success as: $\Pr[y_i = 1] = F(x_i'\beta)$*, where* $F(.)$ *is the cumulative density function of the standard normal distribution in the case of the probit model and the logistic distribution in the case*

of the logit model.

Assuming that $F(.)$ is differentiable with derivative $f(.)$, this implies that the marginal effect of the jth explanatory variable is given by:

$$\frac{\partial \Pr[y_i = 1]}{\partial x_{ij}} = f(x_i'\beta)\beta_j$$

where $f(.)$ is the normal density function in the case of the probit model and $f(.) = \Lambda(1 - \Lambda)$ in the case of the logit model (where $\Lambda(.)$ is the logistic cdf). This result means that the marginal effect of changes in the explanatory variables depends on the level of these variables and that the estimated coefficients $\widehat{\beta}$ give the sign of the impact because $F(.) > 0$.

Stata commands:
probit inlf nwifeinc exper educ age kidslt6 kidsge6
logit inlf nwifeinc exper educ age kidslt6 kidsge6

Then, according to the probit and logit estimated coefficients, the probability of participating in the labor force depends negatively on the husband's income, age, and number of kids aged less than six years old. Moreover, it depends positively on the number of years of experience, years of schooling, and the number of kids between six and 18 years old.

(c) Compare the LPM coefficients with the average marginal effects for the probit and the logit models.

The OLS estimates and the average marginal effects obtained from the probit and logit models are similar.

Stata commands:
probit inlf nwifeinc exper educ age kidslt6 kidsge6
margins, dydx(∗)
logit inlf nwifeinc exper educ age kidslt6 kidsge6
margins, dydx(∗)

(d) Estimate the probability of participation in the labor market for women with 0, 1, and 2 kids aged under 6 years using the probit model at the average value of the regressors. How do these probabilities compare with the probabilities estimated by the LPM?

Stata commands:
sum inlf nwifeinc exper educ age kidslt6 kidsge6
probit inlf nwifeinc exper educ age kidslt6 kidsge6

margins, at(kids=(0(1)3) nwifeinc= 20.12896 exper=10.63081 educ= 12.28685 age=42.53785 kidsge6=1.353254)
reg inlf nwifeinc exper educ age kidslt6 kidsge6

The estimated probabilities for women with 0, 1, and 2 kids aged under 6 years are approximately 0.6623, 0.3246, and 0.0919, respectively.

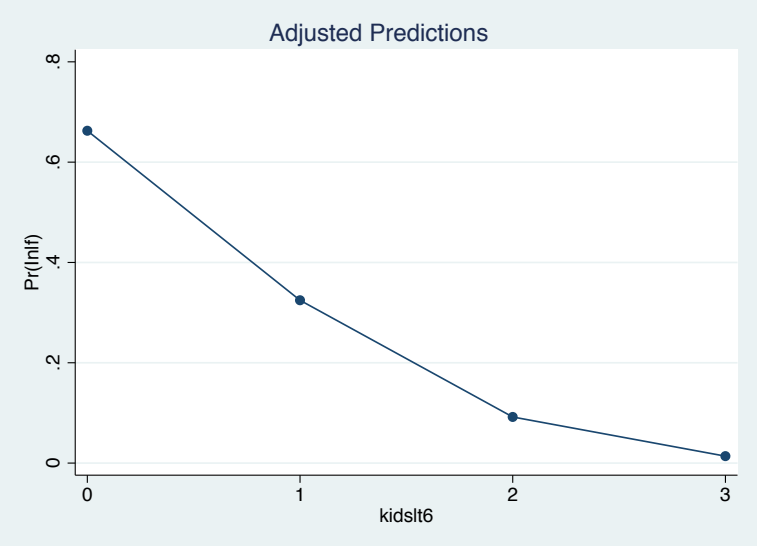
According to the OLS estimates, the probability of participating in the labor market for women with no kids:

$$\widehat{inlf} = 0.707 - 0.003 \times 20.12896 + 0.023 \times 10.63081 +$$
$$+ 0.0398 \times 12.28685 - 0.0177 \times 42.53785 - 0.272 \times 0 + 0.0125 \times 1.353254 = 0.6441$$

Similarly, the probability of labor market participation for women with one kid aged under six years old is 0.3721. For women with two years old, the probability is 0.1001.

(e) Given the results from part 1.b), set the regressors at the sample mean and draw a picture of the probability of participating in the labor market as a function of the number of kids.

Stata command:
probit inlf nwifeinc exper educ age kidslt6 kidsge6
margins, at(kids=(0(1)3) nwifeinc= 20.12896 exper=10.63081 educ= 12.28685 age=42.53785 kidsge6=1.353254)
marginsplot, recast(connected) noci



And similarly for the logit model.

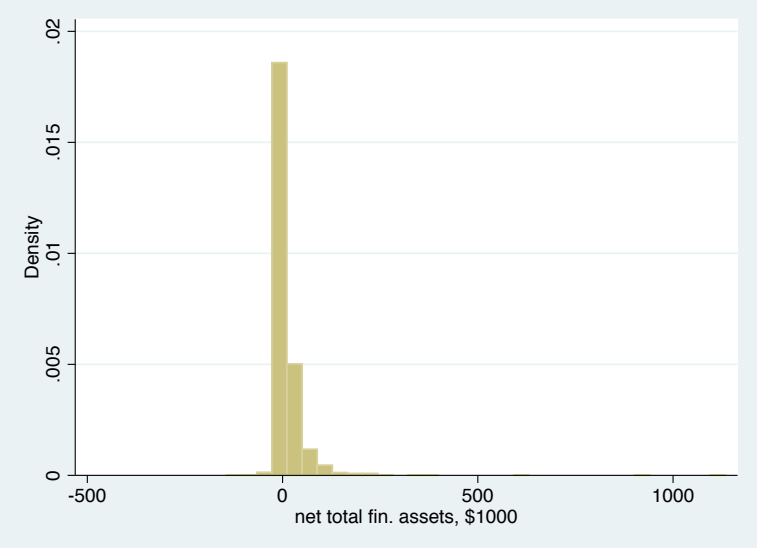
Exercise 2 (Quantile regression models)

For this exercise please create a do-file and corresponding log-file to hand in with your solution.

Consider the **netfinancialwealth.dta** dataset, which is a subset of the data used in Abadie (2003).

(a) Provide detailed descriptive statistics of the variables net total financial wealth $nettfa$, income inc, age, and a binary variable indicating whether an individual is eligible for a (401 (k)) pension fund through her employer. Financial wealth and income are measured in thousands of dollars. Draw a histogram of the variable $nettfa$. What do you find?

The mean of the net total financial wealth is 13.59 while the median is 1.4. This is consistent with a distribution that is skewed to the right.



Stata commands:
sum nettfa inc age e401k, detail
hist nettfa

(b) Run an OLS regression model with $netfa$ as dependent variable and inc, age, $agesq$, and $e401k$ as explanatory variables. Interpret the results. At what age

does net financial assets increase with age?

Stata commands:
reg nettfa inc age agesq e401k

A \$1000 increase in income is associated with a 7,826\$ in net total financial wealth, on average, ceteris paribus.
Getting older by one year is associated with a $-1.568 + 2 \times 0.0284 age$*, on average, ceteris paribus. The estimated impact of age is convex and depends on the age of each individual.*
The net total financial wealth of individuals who participate in the 401(k) pension fund is on average 6836\$ higher than that of those that do not participate in the pension fund, ceteris paribus.

Net financial assets increase with age from 28 years onwards
$age^* = -1.567/(-2 \times 0.0283)$

(c) Test the presence of heteroskedasticity in the model. State clearly the null and alternative hypotheses, the test statistic, and the decision rule. What do you conclude?

The Breusch-Pagan test is based on the following equation:

$$\widehat{u}^2 = \delta_0 + \delta_1 inc \delta_2 age + \delta_3 agesq + \delta_4 e401k + v$$

The null and alternative hypotheses are:

$$H_0 : \delta_1 = \delta_2 = ... = \delta_4 = 0$$
$$H_a : NotH_0$$

Stata commands:
reg nettfa inc age agesq e401k
predict uhat, res
gen uhatsq = uhat× uhat
reg uhatsq inc age agesq e401k

The F-statistic of joint significance is 3.49 and the p-value is 0.0075. Therefore we reject the null hypothesis in favor of the alternative hypothesis and the model is heteroskedastic. This means that the error term u and the explanatory variables cannot be independent.

(d) Obtain quantile regression estimates for the same model at the 0.1, 0.25, median, 0.75, and 0.9 quantiles. Interpret carefully the estimated coefficients on the *inc* variable and compare them with the OLS estimate.

The estimated impact of income at the first decile of nettfa equals -0.0179, at the median equals 0.324, and at the 9th decile is 1.291. These estimates suggest that income affects net financial wealth in an heterogenous way. Also, the impact of income is much stronger at higher quantiles. It is not statistically significant when evaluated at the first decile. The OLS estimate equals 0.7825.

Alternatively, we can interpret the estimated coefficients as the impact of a $1,000 increase in income at the qth quantile, for example, a $1,000 increase in income increases the median by 0.324.

Stata commands:
est clear
foreach tau in 0.1 0.25 0.5 0.75 0.9{
eststo: qreg nettfa inc age agesq e401k, quantile('tau')
}
esttab using ps2_qreg.tex, replace

	(1) q10	(2) q25	(3) q50	(4) q75	(5) q90
inc	-0.0179	0.0713***	0.324***	0.798***	1.291***
	(-0.74)	(7.39)	(13.72)	(15.86)	(13.54)
age	-0.0663	0.0336	-0.244	-1.386*	-3.579**
	(-0.22)	(0.28)	(-0.83)	(-2.20)	(-3.00)
agesq	0.00236	0.000372	0.00480	0.0242**	0.0605***
	(0.66)	(0.26)	(1.39)	(3.28)	(4.33)
e401k	0.949	1.281***	2.598**	4.460*	6.001
	(1.13)	(3.84)	(3.18)	(2.56)	(1.82)
_cons	-5.228	-4.373	-3.573	7.539	37.27
	(-0.86)	(-1.83)	(-0.61)	(0.60)	(1.57)
N	2017	2017	2017	2017	2017

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(e) Run the same quantile regressions using only $e401k$ as explanatory variable. Why did the coefficients change from those estimated in (d)?

	(1) q10	(2) q25	(3) q50	(4) q75	(5) q90
e401k	0.830	0.676**	6.549***	19.50***	32.10***
	(1.16)	(2.96)	(15.25)	(10.69)	(6.32)
_cons	-4.500***	-0.476***	0.350	6.149***	27.40***
	(-10.51)	(-3.48)	(1.36)	(5.62)	(8.98)
N	2017	2017	2017	2017	2017

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

In the model with additional controls, the impact of having a pension fund at the first decile of net financial assets is 0.949, at the median is 2.598, and at the 9th decile is 6.001.
Now the estimated impact of having a pension fund at the first decile of nettfa equals 0.83, at the median equals 6.549, and at the 9th decile is 32.10. It is not statistically significant at the first decile. This happens because we are omitting relevant variables that are surely correlated with having a 401(k) pension fund.

Stata commands:
est clear
foreach tau in 0.1 0.25 0.5 0.75 0.9{
eststo: qreg nettfa e401k, quantile('tau')
}
esttab using ps2_qreg_2.tex, replace

Exercise 3 (Multinomial discrete choice models)

For this exercise please create a do-file and corresponding log-file to hand in with your solution.

Consider the **travel_choice_MNL.dta** dataset on individual choice to travel between Sydney and Melbourne. The dataset is comprised of 210 observations on choice among four travel modes: air, train, bus, and car. The attributes used are choice-specific: *gc*, which is a measure of the generalized cost of the travel, *ttme*, which is the terminal time (zero for cars), and for the choice between air and the other modes *hinc*, the household income. The model specified is:

$$U_{ij} = \alpha_{air} d_{i,air} + \alpha_{train} d_{i,train} + \alpha_{bus} d_{i,bus} + \beta_g gc_{ij} + \beta_t ttme_{ij} + \gamma_h d_{i,air} hinc_i + \varepsilon_{ij}$$

(a) State the probabilities for a four-outcome conditional logit model.

The probabilities in a conditional logit model are given by the following expression:

$$p_{ij} = \frac{\exp \mathbf{x}_{ij}^{'}\beta}{\sum_{l=1}^{4} \exp \mathbf{x}_{il}^{'}\beta}, j = 1, ..., 4$$

Because $\sum_{j=1}^{m} p_{ij} = 1$, $m = 4$, *an equivalent model is obtained by defining* $\mathbf{x}_{ij}$ *to be deviations of regressors from values of alternative 1, for example, and setting* $\mathbf{x}_{i1} = 0$.

(b) Provide sample means of the attributes for the 210 observations and for the observations that made that choice.

Stata commands:
table mode_travel, c(mean ttme mean gc mean hinc freq)
table mode_travel if choice==1, c(mean ttme mean gc mean hinc freq)

(c) Estimate a Conditional Logit Model (CLM). Interpret the coefficient on household income.

Stata command:
asclogit choice ttme gc, case(id) alternatives(mode_travel) casevars(hinc) base(3).

The choice of airplane, bus and train depend negatively on households' income.

(d) Obtain the predicted probabilities of choice of each mode. Are they consistent with the actual frequencies?

Stata commands:
predict prob, pr
table mode_travel, c(mean choice)
table mode_travel, c(mean pr)

(e) Calculate the marginal effects of a change in household income at mean values. Interpret the results.

A one unit increase in households' income (at the mean) increases the probability of choosing airplane rather than bus, car or train by 0.0041 and increases the probability of choosing car by 0.0069 rather than airplane, bus, or train. Instead, a one unit increase in households' income (at the mean) decreases the probability

of choosing bus by 0.0010 rather than car, train or airplane and decreases the probability of choosing train by 0.010 rather than car, airplane or bus.

Stata command: estat mfx

(f) Reshape the data in wide format in order to end with a single line by individual. Estimate a multinomial logit model of *choicetravel* on the alternative-invariant variable. Comment on the estimated coefficients.

Stata command:
reshape wide ttme gc choice, i(id) j(mode_travel)
mlogit choicetravel hinc, b(3) nolog

The coefficients measure how the probability of choosing traveling by air, bus or train instead of traveling by car changes as households' income change. The estimates indicate that as households' income increase, individuals are less likely to travel by air, bus, or train relatively to car.

(g) Is the IIA likely to hold in this model? Perform a formal test.

Stata commands:
mlogit choicetravel hinc, b(1)
est store ML
mlogit choicetravel hinc if choicetravel!=4, b(1)
est store MLwo4
suest ML MLwo4, noomitted
test ([ML_BUS = MLwo4_BUS]) ([ML_CAR = MLwo4_CAR])

p-value=0.5766: this means that we fail to reject the H0 in favor of the alternative hypothesis for a 10% significance level and, therefore, the MNL model is consistent (H0 says that the coefficients are equal regardless of the presence of other alternatives).