

## Quantitative Methods in Finance

### **Tutorial, Part 14:**

#### ***Binomial discrete choice models. Maximum likelihood estimation.***

**Example 1:** A survey on eating habits concerning chicken meat was conducted on a sample of 419 individuals. Among others, the data on the following variables were collected (Stata data file `chicken.dta`):

- ♦ *butcher*: dichotomous variable for a person buying chicken meat at the butcher's instead of in the supermarket: 1 – the person is buying chicken meat at the butcher, 0 – the person is not buying chicken meat at the butcher;
  - ♦ *expense\_chicken*: money spent on chicken meat in an average week;
  - ♦ *safe\_butcher*: the butcher is selling safe meat (variable measured on a 7-level Likert scale: 1 – completely disagree, ..., 7 – completely agree);
  - ♦ *supermarket*: dichotomous variable for a person having trust in the information from the supermarket: 0 – no, 1 – yes;
  - ♦ *age*: person's age;
  - ♦ *hhsz*: number of household members.
- a) Explore the data. What kind of variables are there in the data set?
  - b) Assess the probability that a person is buying chicken meat at the butcher's. Assess the probability that a person is buying chicken meat at the butcher's if they completely agree that the butcher is selling safe meat. How about if they completely disagree? Assess the probability that a person is buying chicken meat at the butcher's if they do not trust the information from supermarket.
  - c) By applying the least squares estimator, estimate a simple linear probability model with *butcher* as the dependent variable and all the other variables as regressors. Interpret the coefficients. Does the trust into supermarket information affect the probability of buying chicken meat at the butcher's? Test the hypothesis that the probability of buying chicken meat at the butcher's does not depend on age.
  - d) Calculate the predicted probabilities for the regression in c). Do they have meaningful values? How does the distribution of residuals look like?
  - e) By applying the maximum likelihood estimator, estimate a probit model using the same variables as in c). Calculate and interpret the marginal effects at the means of explanatory variables (centroid).
  - f) By applying the maximum likelihood estimator, estimate a logit model using the same variables as in c). Again, calculate and interpret the marginal effects at the means of explanatory variables (centroid).
  - g) Compare the regression coefficients of probit and logit models. Compare the marginal effects of probit and logit models. What do you find out?
  - h) What is the hit ratio (classification accuracy or percentage correctly classified) of the logit model from f)?
  - i) Calculate the hit ratio for a logit model without insignificant explanatory variables. What do you find out?

### *Computer printout of the results in Stata:*

Part a)

**. summarize**

Variable	Obs	Mean	Std. Dev.	Min	Max
butcher	419	.3412888	.4747089	0	1
expense_ch~n	419	69.77804	218.6385	0	1199
safe_butcher	419	4.940334	1.728946	1	7
supermarket	419	.8186158	.3857969	0	1
age	419	45.70644	15.72711	18	87
hhsz	419	2.682578	1.287391	1	7

**. tab butcher**

Buying at			
butcher's	Freq.	Percent	Cum.
no	276	65.87	65.87
yes	143	34.13	100.00
Total	419	100.00	

**. tab safe\_butcher**

Meat at butcher's safe	Freq.	Percent	Cum.
Completely disagree	23	5.49	5.49
Mostly disagree	22	5.25	10.74
Somewhat disagree	28	6.68	17.42
Neither agree nor disagree	99	23.63	41.05
Somewhat agree	52	12.41	53.46
Mostly agree	102	24.34	77.80
Completely agree	93	22.20	100.00
Total	419	100.00	

Part b)

**. tab butcher**

Buying at			
butcher's	Freq.	Percent	Cum.
no	276	65.87	65.87
yes	143	34.13	100.00
Total	419	100.00	

**. tab butcher if safe\_butcher==7**

Buying at			
butcher's	Freq.	Percent	Cum.
no	50	53.76	53.76
yes	43	46.24	100.00
Total	93	100.00	

```
. tab butcher if safe_butcher==1
```

Buying at   butcher's	Freq.	Percent	Cum.
no	21	91.30	91.30
yes	2	8.70	100.00
Total	23	100.00	

```
. tab butcher if supermarket==0
```

Buying at   butcher's	Freq.	Percent	Cum.
no	46	60.53	60.53
yes	30	39.47	100.00
Total	76	100.00	

Part c)

```
. regress butcher expense_chicken safe_butcher supermarket age hhsize
```

Source	SS	df	MS	Number of obs =	419
Model	11.0014909	5	2.20029818	F( 5, 413) =	10.92
Residual	83.1942132	413	.201438773	Prob > F =	0.0000
Total	94.1957041	418	.225348574	R-squared =	0.1168
				Adj R-squared =	0.1061
				Root MSE =	.44882

butcher	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
expense_chicken	-.0001362	.0001017	-1.34	0.181	-.000336 .0000637
safe_butcher	.0782508	.0127748	6.13	0.000	.0531391 .1033625
supermarket	-.0711321	.0571179	-1.25	0.214	-.1834102 .0411459
age	.0041666	.00141	2.96	0.003	.001395 .0069382
hhsize	.0212159	.0173183	1.23	0.221	-.0128271 .055259
_cons	-.224919	.1118267	-2.01	0.045	-.4447395 -.0050985

```
. test supermarket=0
```

```
( 1) supermarket = 0
```

```
F( 1, 413) = 1.55
Prob > F = 0.2137
```

```
. test age=0
```

```
( 1) age = 0
```

```
F( 1, 413) = 8.73
Prob > F = 0.0033
```

Part d)

```
. predict prob_hat
```

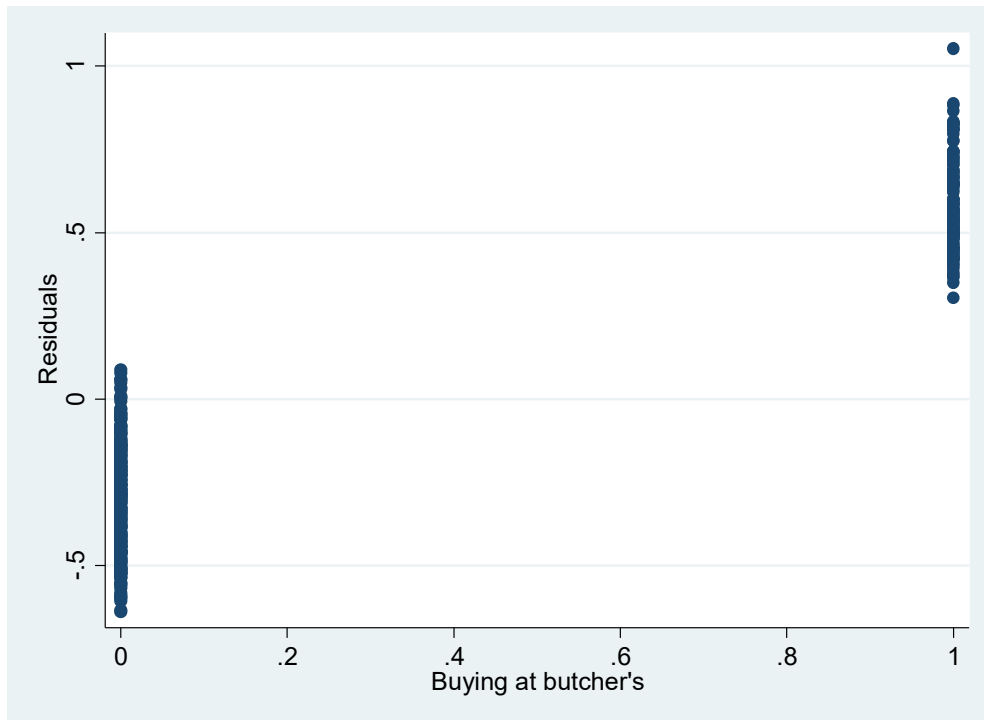
```
(option xb assumed; fitted values)
```

```
. summarize prob_hat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
prob_hat	419	.3412888	.1622324	-.0886773	.6956753

```
. predict resid, resid
```

```
. scatter resid butcher
```



Part e)

```
. probit butcher expense_chicken safe_butcher i.supermarket age hhsz
```

```
Iteration 0:  log likelihood = -268.95049
Iteration 1:  log likelihood = -242.52865
Iteration 2:  log likelihood = -242.40975
Iteration 3:  log likelihood = -242.40974
```

Probit regression	Number of obs	=	419
	LR chi2(5)	=	53.08
	Prob > chi2	=	0.0000
Log likelihood = -242.40974	Pseudo R2	=	0.0987

	butcher	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
expense_chicken		-.0004295	.0003209	-1.34	0.181	-.0010584 .0001994
safe_butcher		.2487339	.0426283	5.83	0.000	.1651839 .3322838
1.supermarket		-.2092815	.1699843	-1.23	0.218	-.5424446 .1238815
age		.0117166	.0042426	2.76	0.006	.0034012 .020032
hhsz		.0626871	.0520667	1.20	0.229	-.0393618 .164736
_cons		-2.195304	.3582204	-6.13	0.000	-2.897403 -1.493205

```
. margins, dydx(expense chicken safe butcher supermarket age hhsize) atmeans
```

```

Expression      : Pr(butcher), predict()
dy/dx w.r.t.    : expense_chicken safe_butcher 1.supermarket age hhsize
at              : expense_ch~n      = 69.77804 (mean)
                  safe_butcher      = 4.940334 (mean)
                  0.supermar~t      = .1813842 (mean)
                  1.supermar~t      = .8186158 (mean)
                  age                = 45.70644 (mean)
                  hhsize             = 2.682578 (mean)

```

Note:  $dy/dx$  for factor levels is the discrete change from the base level.

```
. logit butcher expense chicken safe butcher i.supermarket age hhsze
```

Logistic regression	Number of obs	=	419
	LR chi2(5)	=	53.10
	Prob > chi2	=	0.0000
Log likelihood = -242.39982	Pseudo R2	=	0.0987

```
. margins, dydx(expense chicken safe butcher supermarket age hhsize) atmeans
```

```

age          = 45.70644 (mean)
hhsize       = 2.682578 (mean)

```

		Delta-method				
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
expense_chicken	-.0001618	.0001234	-1.31	0.190	-.0004038	.0000801
safe_butcher	.0897076	.0153309	5.85	0.000	.0596595	.1197556
1.supermarket	-.0781375	.0648243	-1.21	0.228	-.2051908	.0489158
age	.0042743	.0015281	2.80	0.005	.0012793	.0072692
hhsize	.0231925	.0186461	1.24	0.214	-.0133531	.0597381

Note: dy/dx for factor levels is the discrete change from the base level.

Part h)

```
. qui logit butcher expense_chicken safe_butcher i.supermarket age hhsize
```

```
. predict probability_hat
(option pr assumed; Pr(butcher))
```

```
. summarize probability_hat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
probability_hat	419	.3412888	.1645072	.044734	.7285379

```
. generate butcher_hat=0
. replace butcher_hat=1 if probability_hat>=0.5
(80 real changes made)
```

```
. tabulate butcher butcher_hat
```

Buying at	butcher_hat		Total
butcher's	0	1	
no	239	37	276
yes	92	51	143
Total	331	88	419

```
. scalar hit_ratio=(239+51)/419
. display hit_ratio*100
69.212411
```

Part i)

```
. logit butcher safe_butcher age
```

```

Iteration 0: log likelihood = -268.95049
Iteration 1: log likelihood = -245.30396
Iteration 2: log likelihood = -244.86245
Iteration 3: log likelihood = -244.86208
Iteration 4: log likelihood = -244.86208

```

Logistic regression	Number of obs	=	419
	LR chi2(2)	=	48.18
	Prob > chi2	=	0.0000
Log likelihood = -244.86208	Pseudo R2	=	0.0896

butcher	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
safe_butcher	.4096279	.0729168	5.62	0.000	.2667135	.5525423
age	.0186308	.0069548	2.68	0.007	.0049996	.0322621
_cons	-3.634732	.5076558	-7.16	0.000	-4.629719	-2.639745

```
. predict probability_h
(option pr assumed; Pr(butcher))
```

```
. summarize probability_h
```

Variable	Obs	Mean	Std. Dev.	Min	Max
probability_h	419	.3412888	.1568919	.0595617	.7013106

```
. generate butcher_h=0
. replace butcher_h=1 if probability_h>=0.5
(80 real changes made)
```

```
. tabulate butcher butcher_h
```

Buying at butcher's	butcher_h		Total
	0	1	
no	238	38	276
yes	101	42	143
Total	339	80	419

```
. scalar hit_ratio1=(238+42)/419
. display hit_ratio1*100
66.825776
```

■

**Example 2:** In Stata data file `holiday.dta` there are data on several variables concerning the main annual holidays:

- ♦ *abroad*: dichotomous variable for a person spending holidays abroad: 1 – abroad, 0 – in the home country;
- ♦ *log\_income*: logarithm of annual family net income per household member;
- ♦ *age*: person's age;
- ♦ *pet*: dichotomous variable for the presence of pets in the family: 1 – yes, 0 – no.

- a) Estimate the probit model for the *abroad* variable as the dependent variable and all the other variables as regressors. Do the signs of regression coefficients make sense?
- b) Reproduce the likelihood ratio test statistic for testing the statistical significance of the regression model as a whole. How is the pseudo  $R^2$  coefficient calculated?
- c) Calculate automatically and manually the marginal effects at the means of explanatory variables (centroid). Interpret the calculated marginal effects.
- d) Estimate the logit model and compare the results to those from a).
- e) Again, calculate automatically and manually the marginal effects at the means of explanatory variables (centroid). Interpret the calculated marginal effects.
- f) Calculate (automatically and manually) and interpret the odds ratios from the logit model.

## Computer printout of the results in Stata:

Part a)

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
abroad	40	.55	.5038315	0	1
income	40	10779.5	4936.733	3360	27940
log_income	40	9.193383	.4326685	8.119697	10.23781
age	40	47.85	12.01185	26	72
pet	40	.3	.4640955	0	1

```
. probit abroad log_income age i.pet
```

```
Iteration 0: log likelihood = -27.525553
Iteration 1: log likelihood = -16.890274
Iteration 2: log likelihood = -16.783332
Iteration 3: log likelihood = -16.782923
Iteration 4: log likelihood = -16.782923
```

```
Probit regression                               Number of obs   =          40
                                                LR chi2(3)       =          21.49
                                                Prob > chi2      =          0.0001
Log likelihood = -16.782923                    Pseudo R2       =          0.3903
```

abroad	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
log_income	2.546419	.8138292	3.13	0.002	.9513429 4.141495
age	-.0610703	.0249326	-2.45	0.014	-.1099373 -.0122034
1.pet	-1.361899	.611138	-2.23	0.026	-2.559708 -.1640909
_cons	-19.8639	6.81757	-2.91	0.004	-33.22609 -6.501707

Part b)

```
. ereturn list
```

scalars:

```
      e(rank) = 4
      e(N) = 40
      e(ic) = 4
      e(k) = 4
      e(k_eq) = 1
      e(k_dv) = 1
      e(converged) = 1
      e(rc) = 0
      e(ll) = -16.78292322495001
      e(k_eq_model) = 1
      e(ll_0) = -27.52555254854354
      e(df_m) = 3
      e(chi2) = 21.48525864718707
      e(p) = .0000834653961897
      e(N_cdf) = 0
      e(N_cds) = 0
      e(r2_p) = .3902784260060915
```

macros:

```
      e(cmdline) : "probit abroad log_income age pet"
      e(cmd) : "probit"
      e(estat_cmd) : "probit_estat"
      e(predict) : "probit_p"
```



```

        e(title) : "Probit regression"
        e(chi2type) : "LR"
        e(opt) : "moptimize"
        e(vce) : "oim"
        e(user) : "mopt__probit_d2()"
        e(ml_method) : "d2"
        e(technique) : "nr"
        e(which) : "max"
        e(depvar) : "abroad"
        e(properties) : "b v"

matrices:
        e(b) : 1 x 4
        e(V) : 4 x 4
        e(mns) : 1 x 4
        e(rules) : 1 x 4
        e(ilog) : 1 x 20
        e(gradient) : 1 x 4

functions:
        e(sample)

. scalar LR=-2*(e(l1_0)-e(l1))
. display LR
21.485259

. display chi2tail(3,LR)
.00008347

. scalar pR2=1-e(l1)/e(l1_0)
. display pR2
.39027843

Part c)

. margins, dydx(log_income age pet) atmeans

Conditional marginal effects      Number of obs   =      40
Model VCE      : OIM

Expression      : Pr(abroad), predict()
dy/dx w.r.t.    : log_income age 1.pet
at
      log_income      =      9.193383 (mean)
      age              =      47.85 (mean)
      0.pet            =      .7 (mean)
      1.pet            =      .3 (mean)

-----
      |              Delta-method
      |              dy/dx      Std. Err.      z    P>|z|      [95% Conf. Interval]
-----+-----
      log_income |      .9925525      .3120236      3.18   0.001      .3809974      1.604108
      age        |     -.0238042      .0097122     -2.45   0.014     -.0428398     -.0047687
      1.pet       |     -.5034006      .190536     -2.64   0.008     -.8768442     -.129957
-----

Note: dy/dx for factor levels is the discrete change from the base level.

. qui probit abroad log_income age pet

. matrix B_probit=e(b)

```

```

. matrix list B_probit

B_probit[1,4]
      abroad:      abroad:      abroad:      abroad:
      log_income      age      pet      _cons
y1      2.546419      -.06107035      -1.3618994      -19.863898

. scalar b_probit_log_income=B_probit[1,1]
. scalar b_probit_age=B_probit[1,2]
. scalar b_probit_pet=B_probit[1,3]
. scalar b_probit_cons=B_probit[1,4]

. summarize log_income

      Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
      log_income |      40      9.193383      .4326685      8.119697      10.23781

. scalar mean_log_income=r(mean)

. summarize age

      Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
      age |      40      47.85      12.01185      26      72

. scalar mean_age=r(mean)

. summarize pet

      Variable |      Obs      Mean      Std. Dev.      Min      Max
-----+-----
      pet |      40      .3      .4640955      0      1

. scalar mean_pet=r(mean)

. scalar Xb_probit=b_probit_cons+b_probit_log_income*mean_log_income+
  b_probit_age*mean_age+b_probit_pet*mean_pet

. display Xb_probit
.21552118

. display normalden(Xb_probit)
.38978373

. display normalden(Xb_probit)*b_probit_log_income
.99255269

. display normalden(Xb_probit)*b_probit_age
-.02380423

. display normal(Xb_probit-b_probit_pet*mean_pet+b_probit_pet)-
  normal(Xb_probit-b_probit_pet*mean_pet)
-.50340064

```

Part d)

```

. logit abroad log_income age i.pet

Iteration 0:      log likelihood = -27.525553
Iteration 1:      log likelihood = -17.032774
Iteration 2:      log likelihood = -16.960203
Iteration 3:      log likelihood = -16.959864
Iteration 4:      log likelihood = -16.959864

```

```

Logistic regression
Number of obs   =          40
LR chi2(3)      =          21.13

Log likelihood = -16.959864
Prob > chi2     =          0.0001
Pseudo R2      =          0.3839

```

abroad	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
log_income	4.250643	1.449276	2.93	0.003	1.410114	7.091172
age	-.1004235	.0442162	-2.27	0.023	-.1870858	-.0137613
1.pet	-2.278437	1.075051	-2.12	0.034	-4.385498	-.171377
_cons	-33.26217	12.14418	-2.74	0.006	-57.06433	-9.460013

Part e)

```
. margins, dydx(log_income age pet) atmeans
```

```

Conditional marginal effects
Model VCE      : OIM
Number of obs   =          40

```

```

Expression      : Pr(abroad), predict()
dy/dx w.r.t.    : log_income age 1.pet
at              : log_income      =    9.193383 (mean)
                  age              =    47.85 (mean)
                  0.pet            =     .7 (mean)
                  1.pet            =     .3 (mean)

```

		Delta-method				
	dy/dx	Std. Err.	z	P> z	[95% Conf. Interval]	
log_income	1.034782	.350304	2.95	0.003	.3481992	1.721365
age	-.0244472	.0108509	-2.25	0.024	-.0457146	-.0031799
1.pet	-.5135038	.1927674	-2.66	0.008	-.891321	-.1356866

Note: dy/dx for factor levels is the discrete change from the base level.

```
. qui logit abroad log_income age pet
```

```
. matrix B_logit=e(b)
. matrix list B_logit
```

```

B_logit[1,4]
abroad:      abroad:      abroad:      abroad:
log_income   age      pet      _cons
y1  4.2506429  -.10042354 -2.2784374 -33.262172

```

```
. scalar b_logit_log_income=B_logit[1,1]
. scalar b_logit_age=B_logit[1,2]
. scalar b_logit_pet=B_logit[1,3]
. scalar b_logit_cons=B_logit[1,4]

```

```
. scalar Xb_logit=b_logit_cons+b_logit_log_income*mean_log_income+
b_logit_age*mean_age+b_logit_pet*mean_pet

```

```
. display Xb_logit
.32681861

```

```
. display logisticden(Xb_logit), logistic(Xb_logit)*(1-logistic(Xb_logit))
.24344141 .24344141

```

```
. display logisticden(Xb_logit)*b_logit_log_income
1.0347825

. display logisticden(Xb_logit)*b_logit_age
-.02444725

. display logistic(Xb_logit-b_logit_pet*mean_pet+b_logit_pet)-
  logistic(Xb_logit-b_logit_pet*mean_pet)
-.5135038
```

Part f)

```
. logistic abroad log_income age pet
```

```
Logistic regression                                Number of obs   =          40
                                                    LR chi2(3)      =         21.13
                                                    Prob > chi2     =         0.0001
Log likelihood = -16.959864                      Pseudo R2      =         0.3839
```

```
-----+-----
      abroad | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----
      log_income |    70.1505   101.6674     2.93   0.003    4.096423    1201.314
           age |    .9044543   .0399916    -2.27   0.023    .8293726    .9863329
           pet |    .1024442   .1101327    -2.12   0.034    .0124567    .8425039
           _cons |    3.58e-15   4.35e-14    -2.74   0.006    1.65e-25    .0000779
-----+-----
```

```
. scalar OR_log_income=exp(b_logit_log_income)
. display OR_log_income
70.150496
```

```
. scalar OR_age=exp(b_logit_age)
. display OR_age
.90445426
```

```
. scalar OR_pet=exp(b_logit_pet)
. display OR_pet
.10244416
```

