Эконометрика-2 ММАЭ
Семинар 17

Лекции: А.А. Пересецкий
Семинары: Е.С. Вакуленко
Составить: Е.Ю. Назруллаева

Binary Response Models

Probit: $P(y_i=1)=1-\Phi(-x'_i\beta)=\Phi(x'_i\beta)$, Φ is cumulative function of the standard normal distribution.

Logit: $P(y_i=1)=1-\frac{e^{-x_i\beta}}{1+e^{-x_i\beta}}=\frac{1}{1+e^{-x_i\beta}}=\frac{e^{x_i\beta}}{1+e^{x_i\beta}}$, which is based upon the cumulative function for the logistic distribution.

$$y^*=x\beta+u:\ u\sim N(0,1)\ \text{probit model},\ u\sim logistic\ \text{logit model}$$

$$y=\begin{cases}1,\ y^*>0\\0,\ y^*\le 0\end{cases}$$

- logit and probit models require more cases because they use maximum likelihood estimation techniques.
- keep in mind that when the outcome is rare, even if the overall dataset is large, it can be difficult to estimate a binary response model
- choice btw logit & probit: information criteria?
  - Akaike *AIC = - 2lnL/n + 2k/n*
  - Bayesian *BIC = - 2lnL/n + k ln(n)/n*

Problem 1

INLF – индикатор того, что замужняя женщина работает, EDUC – уровень образования, EXPER – опыт работы, AGE – возраст, KIDSLT6, KIDSGE6 – количество детей возраста до 6 и после 6 лет, MRT – ставка налога, HUSWAGE – зарплата мужа. Оценка logit модели участия на рынке труда приведена ниже.

Dependent Variable: INLF
Method: ML – Binary Logit (Quadratic hill climbing)
Included observations: 753

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	8.782271	1.786486	4.915984	0.0000
EDUC	0.141681	0.045215	3.133479	0.0017
EXPER	0.215674	0.032783	6.578852	0.0000
EXPER^2	-0.003423	0.001041	-3.287138	0.0010
AGE	-0.091060	0.014784	-6.159473	0.0000
KIDSLT6	-1.288610	0.203670	-6.326959	0.0000
KIDSGE6	0.101550	0.075930	1.337424	0.1811
MTR	-9.569563	1.822471	-5.250874	0.0000
HUSWAGE	-0.173875	0.034617	-5.022885	0.0000
McFadden R-squared	0.247083	Mean dependent var		0.568393
LR statistic (8 df)	254.4324			
Probability (LR stat)	0.000000			

1) Рассчитайте вероятность того, что женщина с данными (см. ниже) будет работать.
EDUC=12, EXPER=10, AGE=45, KIDSKT6=0, KIDSGE6=1, MTR=0.6, HUSWAGE=5.

2) Как изменится эта вероятность, если зарплата мужа HUSWAGE станет равна 10?

3) Найдите маржинальный эффект для точки из пункта (1) $\frac{\partial P(INLF=1)}{\partial HUSWAGE}$.

Problem 2

Estimating the Economic Model of Crime:

Becker (1968) [Becker, Gary. 1968. Crime and Punishment: An Economic Approach. *Journal of Political. Economy*. 78: 169-217] introduced an economic model explaining the number of crimes. The main implication of this model is that the number of crimes depends negatively on the probability to be arrested, the probability to be convicted conditional on being arrested, the probability to be imprisoned conditional on being convicted, and the average length of the imprisonment sentence. Since 1968, many empirical studies have tested the empirical implications of Becker's model, usually with cross-section data. Cornwell and Trumbull (1994) use panel data and their results suggest that the cross-section based estimates can be misleading.
The data are a random sample of 2725 Californian men, born in either 1960 or 1961, with at least one arrest prior to 1986 since age 18.

avgsen = average length of sentences served since age 18 (in months)
black =1 if black
born60 =1 if born in 1960
durat = recent unemployment duration (the number of quarters since the individual last had positive earnings or was released from prison)
hispan =1 if Hispanic
inc86 = reported legal earnings in 1986, tens of thousands $
inc86sq = inc86 squared
narr86 = the number of times a man was arrested
nfarr86 = # felony arrests, 1986
nparr86 = # property crime arrests, 1986
pcnv = the proportion of the prior arrests leading to conviction
pcnvsq = pcnv squared
pt86sq
ptime86 = months spent in prison in 1986
qemp86 = number of quarters in 1986 during which the man was legally employed
tottime = the number of months spent in prison prior to 1986 (since 18)

Exercise (Stata)

Estimate a binary model for *crime86*, where the explanatory variables are *pcnv, avgsen, tottime, ptime86, qemp86, inc86, durat, black, hispan, born60*

- logit / probit
- maximum score type

gen crime86=0
replace crime86=1 if narr86>0

Logit model:

logit crime86 pcnv avgsen tottime ptime86 qemp86 inc86 durat black hispan born60

predict xb1, xb
logit crime86 pcnv ptime86 inc86 black hispan
predict xb2, xb

Probit model:

probit crime86 pcnv avgsen tottime ptime86 qemp86 inc86 durat black hispan born60

AIC, BIC

estat ic

Goodness of fit – different measures for "pseudo R-squares" (http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm):

findit fitstat
fitstat, sav(r2_1)

Postestimation

logit crime86 pcnv avgsen tottime ptime86 qemp86 inc86 durat black hispan born60
predict plogit, pr
sum plogit

Classification table: correcty predicted y = 0 & y = 1

estat clas, cutoff(0.5)

The fraction of observations $y = 1$ that are correctly predicted is termed the sensitivity, while the fraction of observations $y = 0$ that are correctly predicted is known as specificity.

Sensitivity & specificity graph, LROC (receiver operating characteristic) curve

lsens
lroc
findit roccomp
roccomp crime86 xb1 xb2, graph summary

About roccomp: http://www.stata-journal.com/sjpdf.html?articlenum=st0023

Goodness-of-fit Hosmer-Lemeshov test

estat gof

Interpretation: marginal effects

$$\frac{\partial P(y=1|x)}{\partial x_k} = g\left(\bar{x}\hat{\beta}\right)\hat{\beta}_k$$ *(marginal effect in means, MEM)*

mfx
mfx, at(.357787 .632294 .838752 .387156 2.30903 54.967 2.25138 1 0 1)
mfx if black==1&hisp==0&born60==1

$$\frac{\partial P(y=1|x)}{\partial x_k} = \frac{1}{n}\sum g\left(x_i\hat{\beta}\right)\hat{\beta}_k$$ *(average marginal effect, AME)*

Alternative variant: calculating of marginal effects

findit margeff
margeff, dummies(black hispan born60)

Differences between the standard command "mfx" and the package "margeff" from Stata Journal: http://www.stata-journal.com/sjpdf.html?articlenum=st0086

- MEM (marginal effect in mean values) might both underestimate and overestimate AME (average marginal effect), depending solely on the sign of the second derivative of the density function
- The difference between AME and MEM is large when the parameter estimates are large
- MEM for dummy variables: dummy variables raise a more fundamental problem if the regression model includes several dummies
- Standard errors for AME calculated using delta-method