Applied Econometrics (MSc.)
Lecture 8
Binary Choice Models, Theory & Application

Yonas Alem

Department of Economics
University of Gothenburg

December 18, 2014

Maximum Likelihood Estimation Recap

Introduction

- The conditional distribution of an observed phenomenon is known, except for a finite number of unknown parameters.
- These parameters will be estimated by taking those values for them that give the observed values the highest probability, the highest likelihood.
- It provides an approach of estimating a set of parameters characterizing a distribution, if we know, or assume we know, the form of this distribution

Maximum Likelihood Estimation Recap

Introduction

Example -1

- Consider a large pool of balls filled with red and yellow balls
- One could be interested in the fraction p of red balls in this pool
- Take random sample of N balls only
- Let $y_i = 1$ if ball i is red and $y_i = 0$ otherwise
- Thus, $P\{y_i = 1\} = p$
- Suppose the pool of ball contains $N_1 = \sum_{i=1}^{N} y_i$ red and $N - N_1$ yellow balls

Maximum Likelihood Estimation Recap

Introduction Cont.

Example -1 Cont.

- The likelihood (probability) of obtaining such a sample is given by:

$$P\{N_1 \quad red \quad balls, \quad N - N_1 \quad yellow \quad balls\} = p^{N_1}(1-p)^{N-N_1} \tag{1}$$

- Equation 1 is what is called the **Likelihood Function** and it is a function of the unknown parameter p.
- In ML estimation, we choose a value for p such that the likelihood function is maximal, and obtain $\hat{p}$.
- The conventional practice is to maximize the log-likelihood, which is a simple monotonic transformation of equation [1] (for computational convenience)

$$logL(p) = N_1 log(p) + (N - N_1) log(1-p) \tag{2}$$

Maximum Likelihood Estimation Recap

Introduction Cont.

Example -1 Cont.

- FOC to maximize [1]:

$$\frac{dlogL(p)}{dp} = \frac{N_1}{p} - \frac{N - N_1}{1 - p} = 0 \tag{3}$$

- Solving [3] for p gives the ML estimator $\hat{p} = N_1 / N$
- It corresponds to the sample proportion of red balls, and most likely to your best guess for p based on the sample drawn
- SOC:

$$\frac{d^2 logL(p)}{dp^2} = \frac{N_1}{p^2} - \frac{N - N_1}{(1 - p)^2} < 0 \tag{4}$$

- Indicating that we indeed have a maximum
- Another example:

Maximum Likelihood Estimation Recap

Introduction Cont.

Example 2.

- Consider the simple regression model

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \tag{5}$$

- Keep assumptions [A1][A4]
- The assumptions imply that $E\{y_i|x_i\} = \beta_1 + \beta_s x_i$ & $V\{y_i|x_i\} = \sigma^2$
- To estimate the above model, we need to impose distributional assumption on ε, the most common being assumption [A5] (normal dist.)

Maximum Likelihood Estimation Recap

Introduction Cont.

Example 2.

- The contribution of the i^{th} observation to the likelihood function is the value of the density function at the observed point y_i. Which, for a normal distribution yields,

$$f(y_i|x_i;\beta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-\frac{1}{2}\frac{(y_i-\beta_1-\beta_2 x_i)^2}{\sigma^2}\} \quad (6)$$

- **Note**: y_i has a continues distribution, hence the likelihood of observing a particular outcome y for y_i is zero for any y
- Where $\beta = (\beta_1, \beta_2)$
- The joint density of $(y_1, ..., y_N)$ conditional on $X = (x_1, ..., x_N)'$ is stated as

$$f(y_1, ..., y_N|X;\beta,\sigma^2) = \prod_{i=1}^{N} f(y_i|x_i;\beta,\sigma^2) = (\frac{1}{\sqrt{2\pi\sigma^2}})^N \prod_{i=1}^{N} exp\{-\frac{1}{2}\frac{(y_i - \quad (7)$$

Maximum Likelihood Estimation Recap

Introduction Cont.

Example 2. Cont.

- The likelihood function and the joint density function of $y_1, ..., y_N$ are similar except the fact that the former is considered as a function of the unknown parameters β, σ^2
- The LL function is given by

$$logL(\beta, \sigma^2) = -\frac{N}{2} log(2\pi\sigma^2) - \frac{1}{2} \sum_{i=1}^{N} \frac{(y_i - \beta_1 - \beta_2 x_i)^2}{\sigma^2} \quad (8)$$

- Maximizing (8) w.r.t $\beta_1 \& \beta_2$ corresponds to minimizing the residual sum of squares $S(\beta)$, as shown in OLS. Do you see why?

Maximum Likelihood Estimation Recap

Introduction Cont.

Example 2. Cont.

- Meaning that the ML estimators of $\beta_1 \& \beta_2$ are identical to the OLS estimators!
- Denote these estimators by $\hat{\beta}_1$ and $\hat{\beta}_2$, and define the residuals $e_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i$ and maximize (8) w.r.t σ^2. FOC:

$$-\frac{N}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{1}{2}\sum_{i=1}^{N}\frac{e_i^2}{\sigma^4} = 0 \tag{9}$$

- Solve(9) for σ^2 to get the ML estimator for σ^2 given by

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} e_i^2 \tag{10}$$

Maximum Likelihood Estimation Recap

Introduction Cont.

Example 2. Cont.

- Note that however this estimator is consistent but not unbiased (a small sample problem) as the estimator in OLS which was given by

$$s^2 = \frac{1}{N-K}\sum_{i=1}^{N} e_i^2 \tag{11}$$

- In many cases, the ML estimator cannot be shown to be unbiased (unknown small sample properties)
- Its use generally is defended based on asymptotic grounds
- Analytical solution of the ML estimator is also difficult in many cases except in some general cases as shown above

Maximum Likelihood Estimation Recap

Specification Tests

- Three types of tests
  1. The Wald test: pretty much in line with t and F tests
  2. The likelihood ratio test: used to compare two alternative nested models
  3. The lagrange multiplier test: used to test restrictions imposed in estimation

Cross-sectional Binary Choice Models

- Used to model phenomenal that are of discrete nature
  - Do married women participate in the labor force?
  - Which sections of society are poor?
  - What are the determinants of an agricultural technology adoption?
- For such kinds of models, OLS is generally inappropriate - we rather use binary choice models
- Mostly (although not exclusively) the problems analyzed are micro-economic nature

Cross-sectional Binary Choice Models Cont.

- Suppose we want to study the impact of income (assumed as the only variable here) on the probability of owning a car:

$$y_i = \beta_1 + \beta_2 x_{i2} + \varepsilon_i = x_i'\beta + \varepsilon_i \tag{12}$$

- Where, $y_i = 1$ if family i owns a car, 0 if family i does not own a car
- $x_i = (x_{i1}, x_{i2})'$
- The standard assumptions:

$$E\{\varepsilon_i|x_i\} = 0 \quad such \ \ that \quad E\{y_i|x_i\} = x_i'\beta \Longrightarrow \tag{13}$$

$$E\{y_i|x_i\} = 1.P\{y_i = 1|x_i\} + 0.P\{y_i = 0|x_i\} \tag{14}$$

$$= 1.P\{y_i = 1|x_i\} = x_i'\beta \tag{15}$$

Cross-sectional Binary Choice Models Cont.

- Thus, the linear model implies that $x_i'\beta$ is a probability and should therefore lie between 0& 1.
- This is only possible if the x_i values are bounded and if certain restrictions on β are satisfied.
  - Hard to achieve this in practice
- Another fundamental problem:
  - ε_i in equation [12] has a highly non-normal distribution and suffers from heteroscedasticity
  - Because y_i has only two possible outcomes, so does the error term for a given value of x_i

Cross-sectional Binary Choice Models Cont.

- We therefore use binary choice models (or univariate dichotomous models)
- Describe the probability $y_i = 1$ directly (but derived from an underlying latent variable model (see next pages)
- The general formulation is:

$$P\{y_i = 1|x_i\} = G(x_i, \beta) \tag{16}$$

for some function $G(.)$

- Equation $[16]$ says that the probability of having $y_i = 1$ depends on x_i
- But, clearly, $G(.)$ should take on values in the interval $[0, 1]$ only
- Usually, we assume:

$$G(x_i, \beta) = F(x_i'\beta) \tag{17}$$

Cross-sectional Binary Choice Models Cont.

- Common choices of F are the standard normal distribution

$$F(x'\beta) = \Phi(x'\beta) = \int_{-\infty}^{x'\beta} \Phi(z)\,dz, \tag{18}$$

giving rise to the so-called **Probit Model**, and the standard logistic function given by:

$$L(x'\beta) = \frac{e^{x'\beta}}{(1 + e^{x'\beta})} \tag{19}$$

leading to the **Logit Model**

Cross-sectional Binary Choice Models Cont.

- A third option is a uniform distribution over the interval [0,1] with distribution function:

$$F(x'\beta) = 0, x'\beta < 0; \tag{20}$$

$$F(x'\beta) = x'\beta, 0 \leq x'\beta \leq 1; \tag{21}$$

$$F(x'\beta) = 1, x'\beta > 1. \tag{22}$$

- Leading to what is called the **Linear Probability Model**

Cross-sectional Binary Choice Models Cont.

- Probit and logit are more common on applied work.
- Both the standard normal and the standard logistic random variable have an expectation of zero, while the latter has a variance of $\pi^2/3$ instead of 1.
- Correcting for the scaling difference would give similar results
- Apart from their signs, the coefficients in these binary choice models are not easy to interpret directly
- One needs to compute the marginal effects of changes in the explanatory variables

Cross-sectional Binary Choice Models Cont.

- For a continuous explanatory variable, x_{ik}, say the marginal effect is defined as the partial derivative of the probability that y_i equals one.
- For the three models above, we obtain

$$\frac{\partial \Phi(x_i'\beta)}{\partial x_{ik}} = \phi(x_i'\beta)\beta_k; \tag{23}$$

$$\frac{\partial L(x_i'\beta)}{\partial x_{ik}} = \frac{e^{x_i'\beta}}{(1+e^{x_i'\beta})^2}\beta_k \tag{24}$$

$$\frac{\partial (x_i'\beta)}{\partial x_{ik}} = \beta_k; (or \;\; 0) \tag{25}$$

Cross-sectional Binary Choice Models

Estimation

- Very often binary choice modes are derived from underlying behavioral model - following the latent model approach.

$$y^* = x_i'\beta + \epsilon_i \tag{26}$$

- y^* is referred to as the latent variable because it is unobserved
- Assume a probability model of working where a person chooses to work if the utility difference exceeds a certain threshold level
- Thus, one observes $y_i = 1$ (working) if and only if $y_i^* > 0$, and $y_i = 0$ (not working) otherwise.
- Hence,

$$P\{y_i = 1\} = P\{y_i^* > 0\} = P\{x_i'\beta + \epsilon_i > 0\} = P\{-\epsilon_i \leq x_i'\beta\} = F(x_i'\beta) \tag{27}$$

  Where F denotes the distribution function of $-\epsilon_i$
- Thus, depending on the distributional assumptions of ϵ_i, one

Cross-sectional Binary Choice Models

Estimation Cont.

- The likelihood contribution of observation i with $y_i = 1$ is given by $P\{y_i = 1|x_i\}$ as a function of β. We do the same for $y_i = 0$
- We can write the likelihood function to be maximized for the entire sample as

$$L(\beta) = \prod_{i=1}^{N} P\{y_i = 1|x_i;\beta\}^{y_i} P\{y_i = 0|x_i;\beta\}^{1-y_i} \tag{28}$$

and the corresponding log-likelihood function (which is convenient to work with) will be given by

$$logL(\beta) = \sum_{i=1}^{N} y_i logF(x_i'\beta) + \sum_{i=1}^{N}(1-y_i)log(1-F(x_i'\beta)). \tag{29}$$

Cross-sectional Binary Choice Models

Estimation Cont.

- Substitute the appropriate for for F to get an expression to be maximized w.r.t β
- The values of β and their interpretation depends on the functional form used
- FOC: Differentiate [29] w.r.t. β to get

$$\frac{\partial logL(\beta)}{\partial \beta} = \sum_{i=1}^{N} [\frac{y_i - F(x_i'\beta)}{F(x_i'\beta)(1 - F(x_i'\beta))} f(x_i'\beta)] x_i = 0 \quad (30)$$

- Where $f = F'$ is the first derivative of the distribution function (so it is the density function)
- The term in the squared bracket is called the "generalized residual" of the model

Cross-sectional Binary Choice Models

Estimation Cont.

- It equals $f(x'\beta)/F(x_i'\beta)$ if $(y_i = 1)$ and $-f(x_i'\beta)/(1 - F(x_i'\beta))$ when $(y_i = 0)$
- The FOC says that each explanatory variable should be orthogonal to the generalized residual (over the whole sample)
- This is comparable with the OLS FOCs stating that the least square residuals are orthogonal to each variable in x_i
- For the Logit model, one can simplify [34] to

$$\frac{\partial logL(\beta)}{\partial \beta} = \sum_{i=1}^{N} [y_i - \frac{exp(x_i'\beta)}{1 + exp(x_i'\beta}]x_i = 0 \qquad (31)$$

- The solution to [31] is the ML estimator of $\hat{\beta}$. From this estimate one can estimate the probability that $y_i = 1$ for a given x_i as

$$\hat{p} = \frac{exp(x_i'\hat{\beta})}{1 + exp(x_i'\hat{\beta}} \qquad (32)$$

Cross-sectional Binary Choice Models

Estimation Cont.

- Consequently, the FOC for the logit model imply that:

$$\sum_{i=1}^{N} \hat{p}_i x_i = \sum_{i=1}^{N} y_i x_i \tag{33}$$

- Thus including the constant term in the regression, the sum of the estimated probabilities is equal to $\sum_i y_i$ or the number of observations in the sample for which $y_i = 1$ I.e., the predicted frequency is equal to the actual frequency
- Similarly, if x_i includes a dummy variable, say 1 for employed people and 0 for unemployed, then the predicted frequency will be equal to the actual frequency for each labor market status group
- Pretty much the same holds for the probit model.
- SOCs, will imply that the matrix of the second order derivatives is negative definite (with the assumption of no

Cross-sectional Binary Choice Models

Goodness-of-fit

- A goodness-of-fit measure is a summary statistic indicating the accuracy with which the model approximates the observed data, like R^2 in OLS
- In binary choice models, the dependent variable is qualitative and accuracy can be judged either in terms of the fit between the calculated probabilities and observed response frequencies or in terms of the model's ability to forecast observed responses
- Unlike the linear regression model, there is no single measure of goodness-of-fit for binary choice models
- Often, goodness-of-fit measures are based on comparison with a model that contains only a constant as explanatory variable
- Let $logL_1$ represent the ML value of the model of interest and let $logL_0$ denote the maximum value of the loglikelihood function when all parameters, except the intercept are zero

Cross-sectional Binary Choice Models

Goodness-of-fit Cont.

- Obviously, $logL_1 \geq logL_0 \Longrightarrow$ the larger the difference between teh two likelihood values, the more the extended model adds to the very restrictive model
- A formal likelihood ratio test can be based on the difference between the two values
- One popular measure goodness-of-fit measure is given by

$$pseudo - R^2 = 1 - \frac{1}{1 + 2(logL_1 - logL_0)/N} \tag{34}$$

Application
Bertrand & Mullainathan, 2004

- **Reference:** Bertrand, M., & Mullainathan, S. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 991ï£ · 1013.
- Significant racial inequality is reveled in every measure of economic success in the US
- Compared to Whites, African-Americans are twice as likely to be unemployed and earn nearly 25 percent less when they are employed
- A White name yields as many more callbacks as an additional eight years of experience on a resume
- The question is thus: "Do employers treat members of different races differentially?, in other words, do they discriminate?"

Application
Bertrand & Mullainathan, 2004

- Some say yes due to employer prejudice or employer perception that race signals lower productivity
- Others say, yes but due to a relic of the past
- There has been limited data to test these hypothesis (researchers don't have access to data that employers have)

Application
Bertrand & Mullainathan, 2004

- White and African-American workers that seem similar to researchers may look very different to employers.
- Consequently, any racial difference in labor market outcomes could just as easily be attributed to differences that are observable to employers but unobservable to researchers
- So how could one get over with such a problem?
- Bertrand & Mullainathan, 2004 conducted a field experiment between July 2001 -January 2002:
- Sent resumes to help-wanted ads in Chicago and Boston newspapers and measure call back for interview for each sent resume

Application
Bertrand & Mullainathan, 2004

- Experimentally manipulate perception of race via the name of the fictitious job applicant
- Randomly assigned White-sounding names (E.g., Emily Walsh or Greg Baker) to half of the resumes, and very African-American-sounding names (E.g., Lakisha Washington or Jamal Jones) to the other half
- They also experimentally varied the quality of the resumes used in response to a given ad. (higher-quality, and lower-quality)
- In total, they responded to 1300 employment adds, and sent 5000 resumes
- The jobs varied from cashier work at a retail establishments and clerical work in a mail room, to office and sales management positions

Application
Bertrand & Mullainathan, 2004

- Some descriptive results

Table: Descriptive Statistics

Variable	African American	White
Callback	0.064	0.097
Female	0.066	0.098
Male	0.058	0.089

- The relevant regression equation would be

$$D_{callback} = \beta_0 + \beta_1 Exp + \beta_2 Exp^2$$

$$\& \ +\beta_3 Volunt. + \beta_4 Millitary + \beta_5 Email$$
$$\& \ + \beta_6 Empl. + \beta_7 Schoolwork + \beta_8 Honors$$
$$\& \ +\beta_9 CompSkills + \beta_{10} SpecSkills + u \quad (35)$$

Application

Bertrand & Mullainathan, 2004

- What would be the problem if we use OLS to estimate [39]?

Table: Determinants of call back (African- Americans)

Variable	Coef.	SE	t-value	P-value
exp	0.003	0.003	0.83	0.406
exp2	-7.13e-06	0.000	-0.06	0.954
volunteer	0.013	0.014	0.95	0.340
military	-0.012	0.019	-0.63	0.529
email	-0.0050	0.014	-0.31	0.754
empholes	0.018	0.012	1.48	0.138
schoolwork	0.005	0.012	0.36	0.720
honors	0.046	0.023	1.97	0.049
computer skills	-0.014	0.032	-0.51	0.612
spec skills	0.047	0.010	4.28	0.000
Intercept	0.196	0.021	0.90	0.367
R-sq = 0.017				

Application

Bertrand & Mullainathan, 2004

Table: Determinants of call back (Whites)

Variable	Coef.	SE	t-value	P-value
exp	0.015	0.004	3.58	0.000
exp2	-0.0005	0.000	-2.85	0.004
volunteer	-0.011	0.017	-0.64	0.520
military	0.039	0.024	1.64	0.101
email	0.021	0.017	1.24	0.215
empholes	0.038	0.014	2.73	0.006
schoolwork	0.022	0.015	1.52	0.129
honors	0.067	0.027	2.48	0.013
computer skills	-0.043	0.016	-2.65	0.008
spec skills	0.075	0.013	5.73	0.000
Intercept	-0.016	0.026	-0.60	0.547
R-sq = 0.033				
Adj R-sq = 0.029				

Application
Bertrand & Mullainathan, 2004

- The R^2 is very low.
- The error term (as discussed previously) is heteroskedastic!
- The case is so, because the dependent variable is binary, and hence the model is no more linear!
- The predicted values of the dependent variable should lie in the interval [0,1] but OLS would lead to values outside this interval
- The appropriate models are thus binary choice models (either probit or logit)

Table: Determinants of Callback, Probit Marginal Effects

Variable	All.	Whites	African Americans
exp	0.07***	0.13***	0.02
exp2	-0.02**	-0.04***	-0.00
volunteer	-0.01	-0.01	0.01
military	-0.00	0.02	-0.01
email	0.02	0.03	-0.00
empholes	0.02***	0.03***	0.01
schoolwork	0.01	0.02	-0.00
honors	0.05***	0.06	0.03
computer skills	-0.02***	-0.04***	-0.00
spec skills	-0.05***	0.06***	0.04***

Application
Bertrand & Mullainathan, 2004

- Job applicants with African-American names get far fewer call-backs for each resume they send out
- Applicants with African-American names find it hard to overcome this hurdle in callbacks by improving their observable skills or credentials
- I.e, they face differential treatment when searching for jobs
    - This could be one factor explaining why they do poorly in the labor market

Application
Bertrand & Mullainathan, 2004

- **Policy Implications:** Training programs alone may not be enough to alleviate the racial gap in labor market outcomes
- For such programs to work, some general-equilibrium force outside the context of the experiment done
- If African Americans recognize that employers would discriminate against them, they may be less willing than Whites to even participate in these programs (it would be rational to do so)