DARDEN
BUSINESS PUBLISHING
UNIVERSITY of VIRGINIA

UV6335
Nov. 7, 2011

MODELING DISCRETE CHOICE: CATEGORICAL DEPENDENT VARIABLES, LOGISTIC REGRESSION, AND MAXIMUM LIKELIHOOD ESTIMATION

Consider an individual choosing between two or more discrete alternatives: a shopper in a grocery store deciding between apple or orange juice, or a prospective student determining which of several university offers he ought to accept. For the juice manufacturer and the university, the ability to predict the outcome of such choices is of vital importance.

In this note, we will discuss how this might be done. The process we will follow bears some similarity to a regular linear regression but also has substantial differences, primarily due to the fact that the choices are discrete; that is, they correspond to a categorical dependent variable in regression.

The Concept of Utility

A fundamental construct in estimating choice behavior is the concept of *utility*—a measure of one's relative satisfaction or pleasure resulting from a particular action (in the above examples, consumption of apple or orange juice and studies at various universities, respectively).

Suppose that, for a given individual, the utility from consumption of apple juice equals uA and utility from consumption of orange juice equals uO. To get a sense of utility, I can rate, on a 100% scale, how much I like apple juice and how much I like orange juice. Let us say I like apple at 60% and orange at 50%. (Note that these values are relative to one another.) My utility would be 0.6 and 0.5 for apple and orange juices, respectively. (Later in the note, we will estimate these utilities using a linear model: $Utility = a + b1 \times x1 + b2 \times x2 + \ldots$ as a function of certain attributes—fruit, size, packaging, price, etc.—but for now, assume that uA and uO are known.)

A very simple choice model could say that if $uA > uO$, the individual chooses apple juice; otherwise, he or she chooses orange. If $uA = uO$, one would be indifferent between the two. The difference $uA - uO$ is called *surplus*. Such a model is called *deterministic* utility.

Random Utility

The problem with the above model is that, as long as *uA* > *uO*, the individual *always* chooses apple juice, regardless of the magnitude of the surplus: The case of *uO* = 0.5 and *uA* = 0.6 is no different from the case of *uO* = 0.1 and *uA* = 0.9. This contradicts the way people typically behave: When surplus is small, people tend to be indifferent regarding the two juices, but when it is large, the preference for apple is stronger. It is reasonable to suppose that, in the former case, one might still occasionally purchase orange juice—for the sake of variety, lack of attention to the choice, or other reasons—even though one generally prefers apple. In the latter case, however, such instances should be much rarer.

The goal of *random* utility models is to capture the above behavior. Underlying such models is the assumption that, rather than being a set number, the utility is a draw from a particular distribution. In this case, *uA* and *uO* could be means of such distributions. If the means of the distributions are close, then one would see a situation that resembles indifference: One would choose apple somewhat more often but still occasionally would choose orange. If the means are far apart, then apple juice would be chosen much more often than orange.

A Logit Model

A particular form of random utility model that has gained wide acceptance in business analytics is a *logit* model (e.g., conjoint analysis a popular marketing research methodology is based on a logit model). Underlying the logit model is an assumption that utilities follow a Gumbel distribution. This distribution fits the actual choice data from numerous empirical studies well and results in an analytically appealing form for the choice probabilities.

In particular, given the expected utilities *uA* and *uO*, the Gumbel distribution suggests that the choice probabilities equal

Prob (Apple juice is chosen) = exp (*uA*) / [exp (*uO*) + exp (*uA*)]

and, correspondingly,

Prob (Orange juice is chosen) = exp (*uO*) / [exp (*uO*) + exp (*uA*)].

That is, if *uA* = 0.6 and *uO* = 0.5 as in the previous example, then

Prob (Apple juice is chosen) = exp (0.6) / [exp (0.5) + exp (0.6)] = 1.822 / 3.471 = 52.5%

and

Prob (Orange juice is chosen) = 47.5%.

Further, since utility is a measure of *relative* satisfaction/pleasure, without loss of generality, one can assume that either of the two utilities equals zero and rescale the other. For example, if $uO = 0$, then $uA = 0.6 - 0.5 = 0.1$. Then, recalling that exp (0) = 1, we obtain

Prob (Apple juice is chosen) = exp (uA) / [1 + exp (uA)]

and

Prob (Orange juice is chosen) = 1 / [1 + exp (uA)].

It is easy to verify that such substitution had no impact on the resulting choice probabilities (i.e., with $uO = 0$ and $uA = 0.1$, the choice probabilities are still 47.5% and 52.5%).

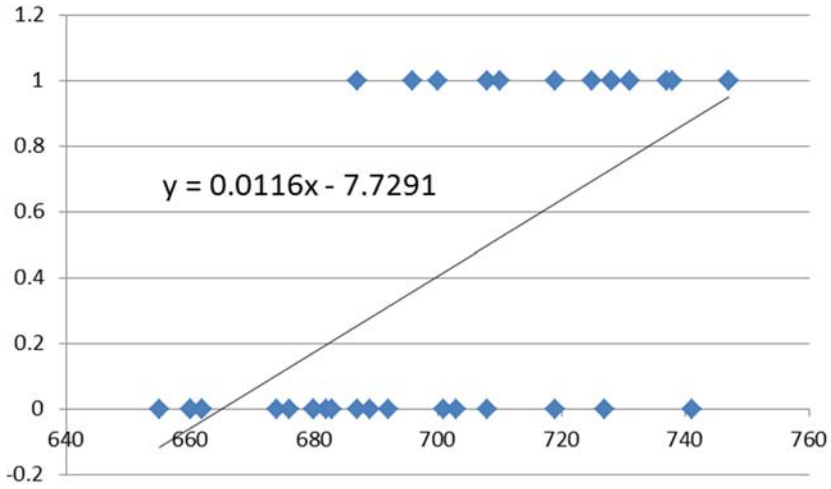
Estimating A Logit Model: Dummy Dependent Variables, Logistic Regression, and Maximum Likelihood Estimation

A process of statistical estimation of a logit model is conceptually similar to a "standard" linear regression (hence the name, logistic regression), yet it involves substantial technical differences. To illustrate the idea and the process, we abandon the juice example and instead consider the following situation:

> *An administrator at a business school* D *(name disguised for confidentiality purposes) collected data about each applicant's GMAT score and choice of business school* D *versus business school* H*—one of the* D*'s closest competitors.* (The data is presented in **Exhibit 1** and depicted in **Figure 1**.) *Given a student's GMAT score, what can be said about that particular student's choice?*

For someone who is familiar with linear regression, a natural tendency would be to regress the *dependent* variable **choice** (*D* versus *H*) onto the *independent* variable **GMAT score**. But the dependent variable is categorical, not continuous, so one needs to introduce a dummy variable, say, 1 for *H* and 0 for *D*. The result of such linear regression is presented in **Figure 1**.

Figure 1.Fitting the *D* versus *H* choice data to a linear model.



Using the equation for the line in **Figure 1**, it is not difficult to obtain a point prediction. For example, for independent variable **GMAT** = 700, **Figure 1** would suggest that the dependent variable **choice** equals 0.0116 × 700 – 7.7291 = 0.4044. But because the dependent variable is categorical (i.e., corresponds to a dummy variable that can be either 0 or 1), the prediction of 0.4044 is meaningless—the dependent variable can be only 0 or 1. In this situation, a natural desire is to interpret the 0.4044 number as a probability—in this case, given our definition of the dummy variable, that an applicant chooses H. That, however, immediately leads to a problem: For GMAT = 650, for example, the predicted "probability" is −0.1754 (i.e., it is negative, which makes no sense).

This inconsistency happens because the desire to fit a line to a categorical dependent variable violates linearity and homoscedasticity assumptions of linear regression. The logit model does not rely on these assumptions, so it ultimately results in better predictions.

The process of estimating a logit model is called *maximum likelihood estimation* (MLE). MLE is a "counterpart" to the least squares minimization used in estimating linear regression, but it accounts for the fact that the estimated quantities are probabilities. MLE proceeds as follows:

1. Express the utility of the choice given a GMAT score as a linear model $uH|GMAT = a + b \times$ GMAT and anchor on the alternative choice at 0; that is, set $uD = 0$.
2. Compute the corresponding choice probabilities:
   i. Prob (H is chosen given GMAT) = exp ($uH|GMAT$) / [1 + exp ($uH|GMAT$)]
   ii. Prob (D is chosen given GMAT) = 1 / [1 + exp ($uH|GMAT$)] = 1 – Prob (H is chosen given GMAT)

3. Given these probabilities, estimate the likelihood of observing the data you have.
   i. Applicant ID1 chose D, and the likelihood of that is Prob (D is chosen given GMAT = 655) = 1 / [1+ exp ($a + b \times 655$)].
   ii. Applicant ID2 also chose D and the likelihood of that is again Prob (D is chosen given GMAT = 660). The likelihood of ID1 and ID2 choosing what they chose is

       Prob (D is chosen given GMAT = 655) × Prob (D is chosen given GMAT = 660)

   iii. …and so on.
4. Compute the total likelihood (a product of the likelihoods for each data point).
5. Solve an optimization problem that will maximize the total likelihood by changing coefficients *a* and *b* (e.g., using Excel Solver).

Hint: Because the objective function (the total likelihood) involves a product of non-linear quantities, this is obviously a highly non-linear optimization model. It can be simplified if one considers logarithms of the likelihoods, instead of the likelihoods themselves. A helpful property of the logarithms is that log (L1 × L2 × L3×…) = log (L1) + log (L2) + log (L3) +… Further, log [exp (uH)] = $uH = a + b \times$ GMAT, which is a linear function of coefficients *a* and *b*. These features make a model with log-likelihoods easier to solve.

Exhibit 2 presents a snapshot of the Solver model that performs the above procedure. **Figure 2** presents the resulting probability, obtained using a logistic regression. From **Exhibit 2**, the resulting equation is

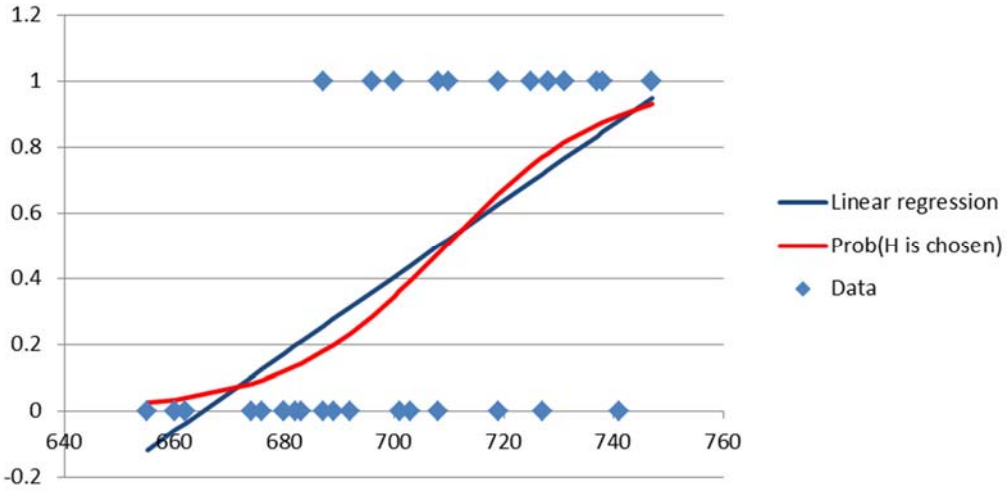
$$\text{Prob (H is chosen given GMAT)} = \exp(-48.47 + 0.0683 \times \text{GMAT}) / [1+ \exp(-48.47 + 0.0683 \times \text{GMAT})].$$

Returning to the previous example, for GMAT = 700 the resulting probability is

$$\text{Prob (H is chosen given GMAT = 700)} = \exp(-48.47 + 0.0683 \times \text{GMAT}) / [1+ \exp(-48.47 + 0.0683 \times \text{GMAT})]$$
$$= 0.5258 / 1.5258 = 34.46\%.$$

Likewise, for GMAT = 650, it equals 1.7% (**Figure 2**).

Figure 2. Fitting the D vs H choice data to linear and logistic models.



The procedures for estimating logistic regression coefficients are embedded in many statistical packages. For example, the StatTools add-on to Excel allows one to run a binary logistic regression (a case when there are two alternatives to choose from, as considered in this note). The resulting output is presented in **Exhibit 3**; not surprisingly, the coefficients are the same as in the model we obtained "manually" (i.e., with Solver).

Several statistical methods and software exist that can estimate much more complex logit models, such as multinomial logit (MNL, a case when there are three or more alternatives to choose from), latent class logits (when, in addition to estimating an MNL, one also wants to determine whether the respondents come from different subgroups that have different underlying preferences/utilities), nested logit (a case when several choices are embedded in one another, as when a customer first chooses a store from which to buy and then chooses the brand to buy), and many others.

Exhibit 1

MODELING DISCRETE CHOICE: CATEGORICAL DEPENDENT VARIABLES, LOGISTIC REGRESSION, AND MAXIMUM LIKELIHOOD ESTIMATION

Sample GMAT Score Data

ID	GMAT	Choice	Dummy
1	655	D	0
2	660	D	0
3	660	D	0
4	662	D	0
5	662	D	0
6	674	D	0
7	676	D	0
8	680	D	0
9	680	D	0
10	682	D	0
11	683	D	0
12	687	H	1
13	687	D	0
14	689	D	0
15	692	D	0
16	696	H	1
17	700	H	1
18	701	D	0
19	703	D	0
20	708	H	1
21	708	D	0
22	710	H	1
23	719	D	0
24	719	H	1
25	725	H	1
26	727	D	0
27	728	H	1
28	728	H	1
29	731	H	1
30	731	H	1
31	737	H	1
32	738	H	1
33	741	D	0
34	747	H	1
35	747	H	1

Data source: Sample data created by author.

Exhibit 2

MODELING DISCRETE CHOICE: CATEGORICAL DEPENDENT VARIABLES, LOGISTIC REGRESSION, AND MAXIMUM LIKELIHOOD ESTIMATION

$uH|GMAT = a + b \times GMAT$ $a=$ −48.471082 $b=$ 0.068326198

ID	GMAT	Choice	Dummy	*uH*	exp*uH*)	Prob (*H* is chosen)	Likelihood	Log (Likelihood)
1	655	D	0	−3.7174	0.0243	0.0237	0.9763	−0.0240
2	660	D	0	−3.3758	0.0342	0.0331	0.9669	−0.0336
3	660	D	0	−3.3758	0.0342	0.0331	0.9669	−0.0336
4	662	D	0	−3.2391	0.0392	0.0377	0.9623	−0.0384
5	662	D	0	−3.2391	0.0392	0.0377	0.9623	−0.0384
6	674	D	0	−2.4192	0.0890	0.0817	0.9183	−0.0853
7	676	D	0	−2.2826	0.1020	0.0926	0.9074	−0.0971
8	680	D	0	−2.0093	0.1341	0.1182	0.8818	−0.1258
9	680	D	0	−2.0093	0.1341	0.1182	0.8818	−0.1258
10	682	D	0	−1.8726	0.1537	0.1332	0.8668	−0.1430
11	683	D	0	−1.8043	0.1646	0.1413	0.8587	−0.1524
12	687	H	1	−1.5310	0.2163	0.1778	0.1778	−1.7268
13	687	D	0	−1.5310	0.2163	0.1778	0.8222	−0.1958
14	689	D	0	−1.3943	0.2480	0.1987	0.8013	−0.2215
15	692	D	0	−1.1894	0.3044	0.2334	0.7666	−0.2658
16	696	H	1	−0.9160	0.4001	0.2858	0.2858	−1.2526
17	700	H	1	−0.6427	0.5258	0.3446	0.3446	−1.0653
18	701	D	0	−0.5744	0.5630	0.3602	0.6398	−0.4466
19	703	D	0	−0.4378	0.6455	0.3923	0.6077	−0.4980
20	708	H	1	−0.0961	0.9083	0.4760	0.4760	−0.7424
21	708	D	0	−0.0961	0.9083	0.4760	0.5240	−0.6462
22	710	H	1	0.0405	1.0414	0.5101	0.5101	−0.6731
23	719	D	0	0.6555	1.9260	0.6582	0.3418	−1.0736
24	719	H	1	0.6555	1.9260	0.6582	0.6582	−0.4182
25	725	H	1	1.0654	2.9020	0.7437	0.7437	−0.2961
26	727	D	0	1.2021	3.3270	0.7689	0.2311	−1.4649
27	728	H	1	1.2704	3.5622	0.7808	0.7808	−0.2474
28	728	H	1	1.2704	3.5622	0.7808	0.7808	−0.2474
29	731	H	1	1.4754	4.3726	0.8139	0.8139	−0.2060
30	731	H	1	1.4754	4.3726	0.8139	0.8139	−0.2060
31	737	H	1	1.8853	6.5885	0.8682	0.8682	−0.1413
32	738	H	1	1.9537	7.0544	0.8758	0.8758	−0.1326
33	741	D	0	2.1586	8.6593	0.8965	0.1035	−2.2679
34	747	H	1	2.5686	13.0474	0.9288	0.9288	−0.0738
35	747	H	1	2.5686	13.0474	0.9288	0.9288	−0.0738

Total −15.4808

Please refer to UVA-QA-0779X (spreadsheet supplement) for the setup of the Solver model.

Data source: Created by author.

Exhibit 3

MODELING DISCRETE CHOICE: CATEGORICAL DEPENDENT VARIABLES, LOGISTIC REGRESSION, AND MAXIMUM LIKELIHOOD ESTIMATION

StatTools Report

Analysis: Logistic Regression
Performed By: XXXXXXX
Date: XXXXXXX
Updating: Static

Summary Measures	
Null Deviance	47.80356733
Model Deviance	30.96154543
Improvement	16.8420219
p-Value	< 0.0001

Regression Coefficients	Coefficient	Standard Error	Wald Value	p-Value	Lower Limit	Upper Limit	Exp(Coef)
Constant	-48.47037424	15.38526195	-3.150441923	0.0016	-78.62548765	-18.31526082	8.90398E-22
GMAT	0.068325198	0.021740921	3.142700311	0.0017	0.025712994	0.110937402	1.070713446

Classification Matrix	1	0	Percent Correct
1	11	4	73.33%
0	3	17	85.00%

Summary Classification	Percent
Correct	80.00%
Base	57.14%
Improvement	53.33%

Note: Each cell contains a comment that explains the entry. Please refer to UVA-QA-0779X.

Data source: Created by author.