

Testing the Multinomial Logit Model

Knut BARTELS¹

Yasemin BOZTUG²

Marlene MÜLLER³

¹ Department of Stochastics II

² Institute of Marketing II

³ Institute of Statistics and Econometrics

¹⁺²⁺³ Sonderforschungsbereich 373

¹ Institute of Mathematics

University Potsdam, Germany

²⁺³ Faculty of Economics

Humboldt–University at Berlin, Germany

February 12, 1999

1 Overview

Modeling the individual behavior of consumers is one of the main topics in marketing research. This individual behavior is influenced by socio-economic characteristics, marketing instruments or latent variables. The connection between these influencing variables and the choice of a product is typically studied by using a statistical choice model for disaggregated data.

A classic choice model is the conditional logit model of McFadden (1974). It is widely discussed and a standard in marketing (Guadagni & Little, 1983). This model however has some disadvantages, in particular the IIA (Independence of Irrelevant Alternatives) and a very restrictive assumption about the errors. This led to many approaches for relaxing these assumptions. For overviews see Ben-Akiva et al. (1997) and Horowitz et al. (1994).

All these approaches present alternative ways for modeling consumer purchase and obtain results which adapt better to the data than the classic approach. However, to our knowledge no general statistical test to check adequateness of the logit model was applied to marketing data until now. The present paper introduces a test procedure which will help in finding an appropriate consumer purchase model. The test is based on a nonparametric test statistic which makes it a very flexible and general tool. We apply the test to scanner panel data.

The paper is organized as follows: Section 2 reviews different types of logit models. The following Section 3 presents the test. Section 4 introduces the data used and presents the relevant results of the test. Finally, Section 5 concludes the paper with a summary and a short look on the next steps.

2 The Multinomial Logit Model

The logit model is a choice model between two or more alternatives. It belongs to the disaggregated choice models of consumer research. Let us start from the model with only two alternatives. Suppose, the consumer will make his choice based on the utility maximization rule (Ben-Akiva & Lerman, 1985). According to this rule,

the consumer i chooses the alternative, which maximizes his utility U_i . The two alternatives j and k create the choice set \mathcal{C} with $\mathcal{C} = \{j, k\}$. The probability that consumer i chooses alternative j is

$$P_i(j) = P(U_{ij} \geq U_{ik}). \quad (1)$$

Assume now, the utility function U_{ij} from equation (1) can be separated into two parts $U_{ij} = V_{ij} + \varepsilon_{ij}$, V_{ij} being a systematic utility component and ε_{ij} a stochastic component (Guadagni & Little, 1983). In the simplest case, the stochastic components are assumed to be i.i.d. and extreme value distributed. For other choice models, e.g. the probit model, another distribution for the error term is assumed (Ben-Akiva & Lerman, 1985). The systematic utility component is typically specified by a linear function

$$V_{ij} = \theta^T x_{ij}, \quad (2)$$

where θ is a vector of parameters to be estimated and x_{ik} the vector of explanatory variables. Equation (1) can now be rewritten as

$$\begin{aligned} P_i(j) &= P(U_{ij} \geq U_{ik}) = P(V_{ij} + \varepsilon_{ij} \geq V_{ik} + \varepsilon_{ik}) \\ &= P(\varepsilon_{ik} - \varepsilon_{ij} \leq V_{ij} - V_{ik}). \end{aligned} \quad (3)$$

Assuming that $\varepsilon_{ik} - \varepsilon_{ij}$ has a logistic distribution, the probability from equation (3) can be written as

$$\begin{aligned} P_i(j) &= \frac{1}{1 + \exp\{-(V_{ij} - V_{ik})\}} = \frac{\exp(V_{ij})}{\exp(V_{ij}) + \exp(V_{ik})} \\ &= \frac{\exp(\theta^T x_{ij})}{\exp(\theta^T x_{ij}) + \exp(\theta^T x_{ik})}. \end{aligned} \quad (4)$$

The coefficients in θ are typically estimated by Maximum Likelihood (see e.g. Ben-Akiva & Lerman, 1985). The absolute values (if all variables are on the same scale) and the signs of the estimated θ are of great interest. In particular, if the sign is positive, an increase in the explanatory variable results in an increase of the response variable. For a negative sign this effects turns to the opposite. The absolute values give information about the strength of the connection between the explanatory and the response variable.

The binary logit model from equation (4) can be generalized to a case with J alternatives in a straightforward way. Utility maximization is again the basic decision rule here. The choice set \mathcal{C} contains now J alternatives. Each consumer chooses the alternative that gives him maximal utility. With this decision rule, the multinomial case can be reduced to the binary model. This is possible, because the maximal utility is taken against the other alternatives, and these other alternatives can be grouped as one possible 'rest choice'. Formally this can be written as

$$\begin{aligned} P_i(j) &= P\left(U_{ij} = \max_{k \in \mathcal{C}} U_{ik}\right) = P(U_{ij} \geq U_{ik} \text{ for all } k \in \mathcal{C}) \\ &= P(\varepsilon_{ik} - \varepsilon_{ij} \leq V_{ij} - V_{ik} \text{ for all } k \in \mathcal{C}). \end{aligned} \quad (5)$$

In this framework, the systematic utility component is typically specified by the linear function

$$V_{ij} = \theta^T x_{ij} = \beta^T z_{ij} + \gamma_j^T w_i \quad (6)$$

where x_{ij} is split into $x_{ij} = [z_{ij}, w_i]$ with z_{ij} denoting the alternative specific part and w_i the individual specific part of the explanatory variables. w_i does not vary over the alternatives, because the household size or the number of children is independent of the purchase. With equation (5) and the assumption about the i.i.d. and logistic distributed error differences $\varepsilon_{ik} - \varepsilon_{ij}$, the probability of the i -th consumer to purchase alternative j is

$$P_i(j) = \frac{\exp(V_{ij})}{\sum_{k \in \mathcal{C}} \exp(V_{ik})} = \frac{\exp(\beta^T z_{ij} + \gamma_j^T w_i)}{\sum_{k \in \mathcal{C}} \{\exp(\beta^T z_{ik} + \gamma_k^T w_i)\}} \quad (7)$$

This model is the most general case of a multinomial logit model. The parameter values in β and γ_j , $j \in \mathcal{C}$ can again be estimated by Maximum Likelihood.

If only product specific variables z_{ij} are used as explanatory variables, the probability $P_i(j)$ is given by

$$P_i(j) = \frac{\exp(\beta^T z_{ij})}{\sum_{k \in \mathcal{C}} \exp(\beta^T z_{ik})}. \quad (8)$$

In this case, the model is called the conditional logit model. We will concentrate on this conditional logit case for the application to the data (cf. Section 4).

Multinomial logit models have some obvious lacks. One problem is the assumption of the logistic distribution of the error differences. Another structural problem lies in the linear assumption for the systematic part of the utility function, which is a very strong restriction. There is no need for the data to follow this linear modeling, also all other types for modeling the explanatory variables should be allowed.

These weak points are the reason for approaches to improve the model (e.g. Ben-Akiva et al. (1997) or Horowitz et al. (1994)). But all these new models are given without testing the multinomial logit model against an alternative. This substantial gap should be filled by this article.

3 The Test

In this section we introduce a formal specification test for the multinomial logit model. The test is based on a general test for the parametric specification of a regression function (Bartels, 1998).

The problem of testing the adequacy of a parametric model class in a regression context against the general nonparametric alternative can be formulated as follows:

$$\mathbf{H}_0 : P\{E(Y|X) = G(X, \theta_0)\} = 1 \text{ for a } \theta_0 \in \Theta, \quad (9)$$

versus

$$\mathbf{H}_1 : P\{E(Y|X) = G(X, \theta)\} < 1 \text{ for all } \theta \in \Theta. \quad (10)$$

Here Y and X denote random variables describing the dependent and explanatory variables in the regression, respectively. The function $G(\cdot, \theta)$ models the relation between Y and X and is specified up to a p -dimensional parameter $\theta \in \Theta$.

The idea of the test is to compare a parametric fit $G(\cdot, \hat{\theta})$ to the given observations $(y_i, x_i), i \in \{1, \dots, n\}$ with a nonparametric fit $\hat{G}(\cdot)$. In particular, the test statistic should be determined by the appropriately weighted integrated squared difference $\{G(\cdot, \hat{\theta}) - \hat{G}(\cdot)\}^2$. Based on this observation, we consider the test statistic

$$\hat{T}_n = \frac{1}{n} \sum_{1 \leq i, \ell \leq n} K_{i\ell} \hat{r}_i \hat{r}_\ell, \quad (11)$$

where

$$\widehat{r}_i = y_i - G(x_i, \widehat{\theta})$$

are the parametrically estimated residuals and $K_{i\ell} = k(x_i, x_\ell)$ are weights obtained from a nonnegative and symmetric kernel. For example, the weights can be defined by a multiplicative kernel $k(x_i, x_\ell) = \prod_{\nu=1}^d \kappa(x_{i\nu} - x_{\ell\nu})$ where κ denotes the Quartic kernel function $\kappa(t) = (1 - t^2)^2 I_{[-1,1]}(t)$ (Härdle, 1991).

Tests of this kind have been studied by several authors, e.g. Härdle & Mammen (1993), Fan & Li (1996), Rodrigues-Campos, Manteiga & Cao (1998). These approaches are based on a kernel $k(\cdot, \cdot)$ that depends on a bandwidth h , as usual in the nonparametric framework. For obtaining a normal limiting distribution of the test statistic in equation (11) this bandwidth must necessarily vanish with increasing sample size n . The choice of the bandwidth is a delicate issue in applying this test, since its influence on the results of the test is not covered by the theory.

Here, we consider the test for a fixed kernel, i.e. one that does not depend on any such vanishing bandwidth, and obtain the distribution of a weighted infinite sum of independent χ_1^2 random variables as limiting distribution. This approach is related to that of Bierens (1990) but much easier to apply.

The limiting distribution and its quantiles can be approximated by bootstrap methods. The bootstrap is also the preferred procedure even in the case of a vanishing bandwidth, since the convergence to the normal limit is rather slow. Details on the theory and regularity conditions are found in Bartels (1998).

To demonstrate the power for finite sample sizes n of a test based on equation (11), some simulation studies have been performed. For example the simple linear model $f(x) = \theta \cdot x + \varepsilon$ has been tested for artificial data $(y_1, x_1), \dots, (y_{25}, x_{25})$, generated by $y_i = \theta \cdot x_i + a \cdot x_i^2 + \varepsilon(x_i)$ with true parameter $\theta_0 = 5$. The variables x_1, \dots, x_{25} are i.i.d. and uniform on $[0, 1]$, and $\varepsilon(x_i)$ are i.i.d. normally distributed with mean zero and variance x_i . The coefficient a determines the amount of quadratic disturbance. The estimator $\widehat{\theta}$ is obtained by least squares. Table 1 reports the empirical power on 1000 iterations with 500 bootstrap replications each.

This test also applies to logit models and multidimensional dependent variables. Denote Y_{ij} the random variable being 1 if choice j has been made by individual i

a	0.0	1.0	2.0	5.0
power	0.044	0.137	0.396	0.964

Table 1: Empirical power at nominal level 0.05, Quartic kernel with bandwidth $h=0.4$

and 0 otherwise. Then

$$P_i(j) = E(Y_{ij}|x_{ij}) = \frac{\exp(\beta^T z_{ij} + \gamma_j^T w_i)}{\sum_{k \in \mathcal{C}} \exp(\beta^T z_{ik} + \gamma_k^T w_i)}. \quad (12)$$

Thus, the null hypothesis of testing whether the choice of the j -th alternative can be adequately described by (7) means to test the adequateness of

$$G(x_{ij}, \theta) = G_j([z_{ij}, w_i], [\beta, \gamma_j]) = \frac{\exp(\beta^T z_{ij} + \gamma_j^T w_i)}{\sum_{k \in \mathcal{C}} \exp(\beta^T z_{ik} + \gamma_k^T w_i)}.$$

The alternative consists of all possible deviations from the logit model.

The results of a simulation study for the simple binomial logit model are given in Table 2. The true model was $E(Y|X) = G_0(X, \theta)$ and alternatively binomial data were simulated according to $G_1(x, \theta)$ and $G_2(x, \theta)$. For x_1, \dots, x_{500} distributed i.i.d. uniformly on $[0, 1]$ and on 1000 iterations, each with 500 bootstrap replications, the observed proportions of rejections are shown in the right column of Table 2.

Model	Percentage of Rejections
$G_0(x, \theta) = \frac{\exp(\theta_1 + \theta_2 x)}{1 + \exp(\theta_1 + \theta_2 x)}, \quad \theta = (0.5, 3)^T$	0.048
$G_1(x, \theta) = 1 - \exp(-\exp(\theta_1 + \theta_2 x)), \quad \theta = (0.05, 3)^T$	0.089
$G_2(x, \theta) = \frac{\exp(\theta_1 + \theta_2 x + \theta_3 x^2)}{1 + \exp(\theta_1 + \theta_2 x + \theta_3 x^2)}, \quad \theta = (0.5, -6, 7)^T$	0.969

Table 2: Rejections at nominal level 0.05, Quartic kernel with bandwidth $h=0.4$

The conditional logit model with choice set $\mathcal{C} = \{1, \dots, J\}$ can be tested by combining the J univariate variables Y_1, \dots, Y_J to one multivariate variable $Y =$

$(Y_1, \dots, Y_J)^T$. The null hypothesis \mathbf{H}_0 is satisfied if $E(Y|X) = G(X, \theta_0)$ almost surely for some $\theta_0 \in \Theta$, where $G(X, \theta) = (G_1(X, \theta), \dots, G_J(X, \theta))^T$. Thus the test applies to this model, too. The test statistic in this case is

$$\hat{T}_n = \frac{1}{n} \sum_{1 \leq i, \ell \leq n} K_{i\ell} \hat{r}_i^T \hat{r}_\ell \quad , \quad (13)$$

where \hat{r}_i denotes the vector of residuals for individual i .

4 Applying the Test to Data

The presented test should now be applied to a data set. The data are from the GfK BehaviorScan. They describe purchases of one type of health care products over 104 weeks in a scanner panel data set. The data set includes information about the brand choice, the date of purchase, the actual marketing–mix–constellation (display and feature) at the purchase and the paid price for the product. We built two data sets from the base data: One with the nine main brands and one dummy brand for the others. Here were 1377 households making 5532 purchases (Table 3). In the second data set, we included only three main brands. There were 964 households with 2651 purchases (Table 4).

Because the variable display and feature are strongly correlated, they were put together in a new variable **Promotion** with the following specification:

$$\mathbf{Promotion} = \begin{cases} 0 & \text{neither display nor feature available} \\ 1 & \text{otherwise.} \end{cases}$$

Also a new variable was implemented, to measure **Loyalty** to the brand, defined as in Guadagni & Little (1983). **Loyalty** should represent the feedback effect in the model (Ailawadi, Gedenk & Neslin, 1997) and is a continuous variable. Tables 3 and 4 summarize some descriptive statistics for both data sets, the 3 and the 10 brands sample. Note that **Loyalty** always sums up to 1 over all brands in the model.

We applied the test procedure from Section 3 to both samples. Recall that the test is based on weighted residuals, such that explanatory variables close to each

10 Brands						
Brand	Purchase (in %)	Loyalty		Price		Promotion (in %)
		Mean	(S.D.)	Mean	(S.D.)	
1	4.79	0.0781	(0.1057)	0.7284	(0.0252)	15.89
2	8.97	0.0944	(0.1408)	0.6629	(0.0328)	14.95
3	6.78	0.0896	(0.1115)	0.5871	(0.0443)	23.83
4	11.59	0.1065	(0.1298)	0.6523	(0.0587)	25.96
5	15.67	0.1304	(0.1849)	0.9033	(0.1153)	34.07
6	3.34	0.0694	(0.0982)	0.6143	(0.0134)	1.14
7	19.11	0.1397	(0.1753)	0.6942	(0.0362)	54.52
8	13.14	0.1169	(0.1457)	0.5781	(0.0281)	39.44
9	14.37	0.1199	(0.1557)	0.6903	(0.0322)	39.15
10	2.24	0.0552	(0.0588)	0.8162	(0.0030)	16.72

Table 3: Descriptive statistics for the 10 brand data set

3 Brands						
Brand	Purchase (in %)	Loyalty		Price		Promotion (in %)
		Mean	(S.D.)	Mean	(S.D.)	
5	32.71	0.3413	(0.1916)	0.8943	(0.1250)	40.89
7	39.87	0.3451	(0.1737)	0.6864	(0.0401)	56.17
8	27.42	0.3137	(0.1539)	0.5754	(0.0317)	43.30

Table 4: Descriptive statistics for the 3 brand data set

other obtain higher kernel weights. The kernel which has been used for calculating the kernel weights is a multiplicative kernel, composed from univariate Quartic kernels for each of the continuous variables (**Loyalty** and **Price** and a kernel function for the discrete variable **Promotion** (see Silverman (1986), p. 126). The smoothing parameter for this variable is denoted by λ .

Tables 5 and 6 summarize the test results for different choices of h and λ . The columns give the test statistic \hat{T}_n and the critical value obtained from 250 bootstrap

simulations. In all cases, the conditional logit hypothesis is clearly rejected. Let us remark that the significance level for rejection is < 0.01 . But still a higher value of \hat{T}_n indicates a more significant rejection. As can be seen, the test statistics \hat{T}_n and the test decisions are not very sensitive with respect to the choice of both smoothing parameters. This is in accordance with the theory explained in Section 3. Also, the model for the 3 brands is rejected more significantly than the 10 brand model. This is as expected, since we have less parameters to describe the behavior of the consumers in the former case.

10 Brands		
h, λ	\hat{T}_n	critical value
0.05, 0.90	0.4159	0.2098
0.05, 0.95	0.4151	0.2103
0.05, 0.99	0.4146	0.2101
0.10, 0.90	0.4152	0.2100
0.10, 0.95	0.4137	0.2109
0.10, 0.99	0.4126	0.2094
0.20, 0.80	0.4338	0.2108
0.20, 0.90	0.4239	0.2093
0.20, 0.95	0.4210	0.2119
0.20, 0.99	0.4193	0.2107

Table 5: Test results for 10 brand data set, nominal level 0.05 and bootstrap sample size 250

To get more information, in which way the model could be improved, we applied the test on a number of modifications of the conditional logit model. In particular, higher order terms (up to quadratic and cubic) for **Loyalty** and **Price** and interaction terms were included. Also we studied the results of the test when **Loyalty** or **Price** were left out, respectively. Table 7 summarizes the tests for these models for bandwidth $h = 0.1$ and $\lambda = 0.95$. The value of the test statistic decreases with increasing numbers of parameters. From the last two lines of Table 7, we can conclude that the variable **Price** seems to be responsible for the lack of fit of the model.

3 Brands		
h, λ	\hat{T}_n	critical value
0.05, 0.90	1.3285	0.3736
0.05, 0.95	1.3004	0.3731
0.05, 0.99	1.2801	0.3685
0.10, 0.90	1.4367	0.3878
0.10, 0.95	1.3982	0.3868
0.10, 0.99	1.3699	0.3803
0.20, 0.80	1.7372	0.3946
0.20, 0.90	1.6052	0.3878
0.20, 0.95	1.5505	0.3972
0.20, 0.99	1.5111	0.3802

Table 6: Test results for 3 brand data set, nominal level 0.05 and bootstrap sample size 250

Model for V_{ij}	\hat{T}_n	critical value
Linear in all Regressors	1.3982	0.3868
Quadratic in Loyalty and Price	1.3810	0.3663
Cubic in Loyalty and Price	1.2587	0.3420
Bivariate Interactions	1.1250	0.3622
Model without Loyalty	8.3676	0.5976
Model without Price	0.9927	0.5041

Table 7: Results for different conditional logit models for the 3 brands case at nominal level 0.05, smoothing parameters $(h, \lambda) = (0.10, 0.95)$, bootstrap sample size 250

5 Summary

We have tested the goodness of fit of a multinomial logit model to explain consumer choice behavior on the base of a scanner panel data set. All variations of the logit model considered were rejected clearly. One possible explanation is that the data

set considered is inappropriate for the multinomial models. Another reason for the rejections observed could be a general misspecification of logit models for consumer choice. This should be tested for different data sets using the method presented here. The results also induce to search for alternative models, e.g. a different link function or a different form of the index, that better fit to this kind of consumer behavior. A non- or semiparametric formulation of the model should be considered as well.

References

- Ailawadi, K. L., Gedenk, K., and Neslin, S. A. (1997). Purchase Event Feedback and Heterogeneity in Choice Models: A Review of Concepts and Methods with Implications for Model Building.
- Bartels, K. (1998). A model specification test. Discussion paper 109, SFB 373, Humboldt-Universität zu Berlin.
- Ben-Akiva, M. and Lerman, S. R. (1985). *Discrete Choice Analysis*. The MIT Press.
- Ben-Akiva, M., McFadden, D., Abe, M., Böckenholt, U., Bolduc, D., Gopinath, D., Morikawa, T., Ramawamy, V., Rao, V., Revelt, D., and Steinberg, D. (1997). Modeling Methods for Discrete Choice Analysis. *Marketing Letters* 8(3), 273–286.
- Bierens, H. J. (1990). A consistent conditional moment test of functional form. *Econometrica* 58(6), 1443–1458.
- Fan, Y. and Li, Q. (1996). Consistent model specification tests: nonparametric versus Bierens' test. Working paper, Department of Economics, University of Windsor.
- Guadagni, P. M. and Little, J. D. C. (1983). A Logit Model of Brand Choice Calibrated on Scanner Data. *Marketing Science* 2(3), 203–238.
- Härdle, W. (1991). *Smoothing Techniques, With Implementations in S*. New York: Springer.

- Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics* 21(4), 1926–1947.
- Horowitz, J. L., Bolduc, D., Divakar, S., Geweke, J., Gönül, F., Hajivassiliou, V., Koppelman, F. S., Keane, M., Matzkin, R., Rossi, P., and Ruud, P. (1994). Advances in Random Utility Models. *Marketing Letters* 5, 311–322.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press.
- Rodrigues-Campos, M. C., Manteiga, W. G., and Cao, R. (1998). Testing the hypothesis of a generalized linear regression model using nonparametric regression estimation. *Journal of Statistical Planning and Inference* 67, 99–122.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Volume 26 of *Monographs on Statistics and Applied Probability*. London: Chapman and Hall.