IAPR TECHNICAL PAPER SERIES

ROBUST ANALYSIS OF DISCRETE CHOICE IN TRANSPORT WITH AN APPLICATION TO ALBERTA CYCLISTS

J.D. Hunt
Department of Civil Engineering and
Institute for Advanced Policy Research
University of Calgary

W.D. Walls
Department of Economics and
Institute for Advanced Policy Research
University of Calgary

December 2006

Technical Paper No. TP-06013

Institute for Advanced Policy Research
University of Calgary
Calgary, Alberta
Canada

http://www.iapr.ca
iapr@ucalgary.ca

Correspondence: J.D. Hunt, Department of Civil Engineeing, University of Calgary, jdhunt@ucalgary.ca

ROBUST ANALYSIS OF DISCRETE CHOICE IN TRANSPORT WITH AN APPLICATION TO ALBERTA CYCLISTS

J. D. Hunt[†] and W. D.Walls[‡]

[†]*Department of Civil Engineering and* [‡]*Department of Economics*
University of Calgary, Calgary, Alberta, Canada T2N 1N4

ABSTRACT

In this paper we apply recently-developed nonparametric conditional kernel density estimation to model discrete choice in a transportation context. Our empirical application, for the purpose of demonstrating the technique in a transport context, is to the route choices of cyclists in the Canadian province of Alberta. The empirical analysis employs the nonparametric multivariate kernel with mixed data types (continuous and discrete). This approach permits the inclusion of continuous and discrete variables where the rate of convergence only depends on the number of continuous variables; it allows for interactions between the covariates, which may be important determinants in discrete transport decisions; it allows the estimation of marginal effects without assuming a constant response regardless of the level of the explanatory variable as it would be in a linear model; and the non-parametric estimation does not impose assumptions in the underlying economic behavior such as IIA. Using a nonparametric index-free approach, the data are permitted to model the full range of relationships among variables. In our simple application of cyclist route decisions, the kernel-density-based estimator correctly classifies 76.24% of decisions in contrast to a correct classification rate of 61.28% for the probit model and 61.44% for the logit model.

1 INTRODUCTION

Discrete choice models are used widely in a number of disciplines, and occur with high frequency in transport economics and engineering. The purpose of modeling such choices is to alternative chosen and to assess the response of the choice probability to changes in exogenous variables hypothesized to influence choice. Binary and multinomial choice models are either parametric, semi-parametric, or density based. The parametric and semiparametric models typically assume a single threshold given by an index function beyond which it is more likely than not that a choice will be made; these straightforward models are especially useful in practice and they are widely used even though a single index imposes nontrivial restrictions on the underlying process. More recent semiparametric models permit multiple indexes, though these are less flexible than index-free models that estimate the conditional probabilities directly using kernel estimation. In this short paper we apply Racine's (2002b) index-free method of directly estimating the conditional density using kernel-based methods as refined by Hall, Racine, and Li (2004).

2 DISCRETE CHOICE MODELING

2.1 Parametric Models

Logit and probit are the most commonly used binary choice models. Both of these models are predicated on a known index which is assumed to influence choice, and a known parametric form for a distribution function that yields choice probabilities. By assumption a particular choice results when the index value exceeds a threshold level. These relatively simple models are readily estimated and interpreted and there are numerous outstanding papers that survey estimation and inference based upon parametric binary choice models; see, for example, Amemiya (1981), McFadden (1984), Blundell (1987), Greene (1997), Ben-Akiva and Lerman (1985) and Train (1986). These models are often derived using the concept of an unobserved or *latent* variable y^*. Decisonmakers make a marginal benefit-marginal cost calculation based on the utilities received from a binary choice. Since we can not observe their marginal utilities, they are modeled statistically by the unobserved variable $y^* = \beta' x_a + e_a$, where e is the stochastic disturbance, β and x are column vector parameters and associated exogenous variables, and $a = \{1, 2\}$ indexes the alternatives in the choice set. What we do observe is the decision y on a subject's choice where $y = 1$ (choose option 1) if $y_1^* > y_2^*$, and $y = 2$ (choose option 2) otherwise.

The probability that a consumer chooses option 1 is $\text{Prob}(y_1^* > y_2^*)$. Assuming the distribution of e is of known parametric form, we can estimate the parameters by the method of maximum likelihood method. For example, assuming a Gaussian distribution for e leads to the probit model and assuming a Logistic distribution leads to the logit model. Empirical choice probabilities result directly from $F(-)$, and the gradient of choice probability with respect to the conditioning variables is given by $dF(-)/dx$, where the maximum likelihood estimator of β is used in both expressions.

In this approach the choice of $F(-)$ is independent of the distribution of the explanatory variables. The assumed symmetry of $F(-)$ imposes symmetry on the choice probability gradient; it follows that the shape of the error distribution can not be estimated since e and y^* are not observed. As Racine (2002) notes, this is entirely different from standard regression model in which the vector of residuals can be used to nonparametrically estimate the density of e because the underlying dependent variable is observed. In the parametric approach a (parametric) distribution for e must be assumed, where the location and scale are determined by the observed choices.

2.2 Semiparametric Models

The literature has many single-index semiparametric discrete choice models. See, for example, the works of Chen and Randall (1997), Coslett (1983, 1991), Ichimura and Thompson (1998) Klein and Spady (1993), Lee (1995), Manski (1985), and Rudd (1986). Pagan and Ullah (1999) survey nonparametric estimation of discrete choice models and Racine and Ullah (2006) survey nonparametric econometrics. Several approaches use nonparametric techniques to estimate the conditional expectation of y resulting in ratios of nonparametric density estimates which may require trimming both to deal with behavior of the numerator at boundary points and to obtain asymptotic results. These approaches are similar to univariate Nadaraya-Watson regression (Nadaraya (1965), Watson (1964)) with the index function itself serving as conditioning

information. Existing semiparametric approaches usually assume an underlying latent variable specification, employ a scalar index which restricts the choice threshold to be a hyperplane, and model the (univariate) distribution of the scalar index function in order to generate empirical choice probabilities. Data-driven methods are typically not used for selection of function forms or index choice.

Discussing a particular semiparametric model briefly will be instructive. Racine (2002a) has proposed a semiparametric approach to the estimation of generalized binary choice models—generalized in the sense that there are separate indexes for each conditioning variable. The resulting choice probability distribution is the joint distribution across these indexes in contrast to the univariate distribution of single index models. Racine's (2002a) approach is like the Priestly-Chow regression estimator (Priestley and Chao, 1972) and it considers the case in which choices are governed in general by multiple thresholds defined over the choice variables, one for each explanatory variable. The probability density function of an individual choice is given by $f(y_i) = [F(g^1 - g_1(x_{i1}, \theta^1), \ldots, g^k - g_k(x_{ik}, \theta^k))]^{yi} \times [1 - F(g^1 - g_1(x_{i1}, \theta^1), \ldots, g^k - g_k(x_{ik}, \theta^k))]^{1-yi}$ where the functions $g_j(x_{ij}, \theta^j)$, $j = 1,\ldots, k$ are (unknown) functions which influence choices, θ^j is a vector of parameters, the g^i are thresholds above which choices tend to be made, and $F(-)$ is a joint distribution function. In this notational setting, the joint density function is given by $f(y_1, \ldots, y_n) = \Pi^n_{i=1} f(y_i)$. This approach focuses on nonparametric estimation of a joint distribution function which has as its arguments generalized known indexes of each conditioning variable. Estimation proceeds uses the method of kernels based on a nonparametric likelihood function which is maximized using cross-validatory techniques. Cross-validation is employed for both bandwidth and parameter selection and, in addition, the parametric form of the index itself can also be selected via cross-validation by letting the nature of the index be an argument of the cross-validatory likelihood function. Kernel estimation of distribution functions can be based upon an estimator such as the Nadaraya-Watson (Nadaraya (1965), Watson (1964)) estimator of a joint density function. Once a specific parametric function for the indexes $g_j(x_{ij}, \theta^j)$, $j = 1,\ldots, k$ a kernel estimator of the joint distribution function of these indexes can be obtained.

2.3 Index-Free Density-Based Models

Nonparametric methods have the advantage of flexibility. The data model the relations among variables and therefore nonparametric methods have the ability to detect nonlinearities that might go unnoticed in parametric models. Index-free models of discrete choice are based on direct density estimation. Racine's (2002b) model is an example of index-free multinomial choice. The densities can be estimated either parametrically, assuming a known joint distribution, or nonparametrically using the method of kernels. Racine (2002b) enumerates several advantages of this approach versus traditional approaches toward discrete choice modeling. First, there is no need to specify an index; an index can create identification and specification problems. Second, estimated probabilities are non-negative and sum to one over the choice set for a given realization of the covariates. Third, traditional approaches use a scalar index which reduces the dimensionality of the joint distribution to a univariate one. This restricts the types of situations that can be modeled to those in which there is only one threshold which is given by the scalar index. The density-based approach explicitly models the joint distribution. Fourth, the gradient of the choice probability, often of primary interest, is restricted in nature by scalar-index models.

Li and Racine (2003) and Racine and Li (2004) develop a method of non-parametric estimation

involving mixed discrete and continuous variables. Their method has been extended to the estimation of conditional density functions by cross-validation by Hall, Racine and Li (2004). Estimation of the conditional probability density function $f(y|x^c, x^d)$ of the choice variable y based on the discrete and continuous variables x^d and x^c is carried out by using a generalized product kernel function based on products of the appropriate kernel for each type of data using a data-driven selection of bandwidths. Using both kernel functions for continuous and discrete components, the conditional density function can be estimated by taking the ratio of the joint density $f(y, x^c, x^d)$ and the marginal density $f(x^c, x^d)$. The choice of bandwidths is crucial in the estimation of kernel-based density functions. The bandwidth determines the size of the neighborhood and it involves a trade-off between bias and variance (Pagan and Ullah, 1999). See Hall, Racine, and Li (2004) for details of the estimation algorithm.

While a nonparametric approach does not depend on functional form assumptions, in a multivariate setting it is subject to the so-called "curse of dimensionality": As the number of independent variables increases, the number of observations must increase exponentially to get accurate estimates of a density. This decreases the speed of convergence or rate at which the nonparametric estimator converges around the true value of the density (Pagan and Ullah, 1999). The rate of convergence of the estimator with mixed data depends only on the number of continuous regressors involved in the estimation (Hall, Racine, and Li 2004). Discrete covariates, which are especially common in transport modeling, do not intensify the curse of dimensionality problem.

We now employ this direct density-based approach in our application to Alberta cyclist choices.

3 APPLICATION TO ALBERTA CYCLIST ROUTE CHOICE

The kernel-density-based model of discrete choice can generally be applied when sufficient observations are available. We apply the robust binary choice model to Alberta cyclists' decision of route choice using a subset of the data collected and thoroughly analyzed by Hunt and Abraham (in press). The empirical example in this paper has the sole purpose of illustrating the potential usefulness of index-free density-based choice modeling. For a complete analysis of influences on cyclist route choice, a thorough description of the data set, and for a standard parametric analysis of cyclist route choice, the reader is referred to the Hunt and Abraham (in press) paper.

For the estimation in this paper, it will be sufficient to state that cyclists have a binary choice of route. The explanatory variables in modeling cyclists choice of route were time spent cycling on roads in mixed traffic; time spent cycling on designated bike lanes on roads; and time spent cycling on bike paths shared with pedestrians. A more complicated model, such as that used in the Hunt and Abraham paper could be estimated, but we have chosen to keep the model quite simple for the purpose of exposition.

The same explanatory variables were used to estimate choice behavior in an index-free kernel-density-based model as well as standard parametric probit and logit models. Tables 1, 2, and 3 show the prediction classification tables corresponding to each estimation. It is clear that the kernel-density-based model dominates the logit and probit models in generating predictions, with an overall correct classification rate of 76.24% as compared to 61.28% for the probit model and 61.44% for

the logit model.

Table 1 Classification Table for Kernel-Density-Based Estimator

A/P	0	1
0	344	171
1	97	516

CR(0-1) 76.24%
CR(0) 66.80%
CR(1) 84.18%

Table 2 Classification Table for Kernel-Density-Based Estimator

A/P	0	1
0	484	129
1	308	207

CR(0-1) 61.28%
CR(0) 61.11%
CR(1) 61.61%

Table 3 Classification Table for Kernel-Density-Based Estimator

A/P	0	1
0	486	127
1	308	207

CR(0-1) 61.44%
CR(0) 61.21%
CR(1) 61.98%

4 CONCLUSION

Logit and probit regressions are and will probably remain the most popular approaches for the conditional prediction of discrete variables. However, these models of discrete choice have a number of disadvantages that result from specifying a known distribution function and a known index function. Recent developments in semiparametric modeling generalize the choice model to an unknown and unspecified distribution function, but even the most generalized model still uses (multiple) index functions. The most generalized discrete choice model uses index-free density-based approach for obtaining choice probabilities. While the kernel-density-based models of discrete choice are data intensive, they appear to yield improved predictions relative to the commonly used logit and probit models. Applied transport researchers may find these kernel-density-based models

of discrete choice useful in practice, especially where data are plentiful and policy analysis demands more accurate predictions.

ACKNOWLEDGEMENT

Support from the Spatial Activities Working Group of the University of Calgary's Institute for Advanced Policy Research is acknowledged.

REFERENCES

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature*, 19:1483–1536.

Ben-Akiva, M. and Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, MA.

Blundell, R., editor (1987). *Journal of Econometrics*, volume 34. North Holland.

Chen, H. and Randall, A. (1997). Semi-nonparametric estimation of binary response models with an application to natural resource valuation. *Journal of Econometrics*, 76:323–340.

Coslett, S. (1983). Distribution-free maximum likelihood estimation of the binary choice model. *Econometrica*, 51:765–782.

Coslett, S. (1991). Semiparametric estimation of a regression model with sampling selectivity. In Barnett, W., Powell, J., and Tauchen, G., editors, *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, volume 5 of International Symposia in Economic Theory and Econometrics. Cambridge University Press, Cambridge.

Greene, W. H. (1997). *Econometric Analysis*. Prentice-Hall, New York, third edition.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(486):1015–1026.

Hunt, J. D. and Abraham, J. E. (in press). Influences on bicycle use. *Transportation*.

Ichimura, H. and Thompson, T. (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics*, 86:269–295.

Klein, R. W. and Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, 61:387–421.

Lee, L. F. (1995). Semiparametric maximum likelihood estimation of poly-chotomous and sequential choice models. *Journal of Econometrics*, 65:381–428.

Li, Q. and Racine, J. S. (2004). Cross-validated local linear nonparametric regression. *Statistica Sinica*, 14(2):485–512.

Manski, C. (1985). Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 27:313–333.

McFadden, D. (1984). Econometric analysis of qualitative response models. In Griliches, Z. and Intriligator, M., editors, *Handbook of Econometrics*, pages 1385–1457. North Holland.

Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge University Press, Cambridge.

Priestly, M. B. and Chao, M. T. (1972). Nonparametric function fitting. *Journal of the Royal Statistical Society*, 34:385–392.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, 119:99–130.

Racine, J. and Ullah, A. (2006). Nonparametric econometrics. In *Palgrave Handbook of Econometrics*., volume 1, Theoretical Econometrics. Palgrave Macmillan, London.

Racine, J. S. (2002a). Generalized semiparametric binary prediction. *Annals of Economics and Finance*, 3:131–147.

Racine, J. S. (2002b). Index-free density-based multinomial choice. In Ullah, A., Wan, A., and Chaturvedi, A., editors, *Handbook of Applied Econometrics and Statistical Inference*, pages 115–142. Marcel Dekker.

Ruud, P. (1986). Consistent estimation of limited dependent variable models despite misspecification of distribution. *Journal of Econometrics*, 32:157–187.

Train, K. (1986). *Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand*. MIT Press, Cambridge, MA.