

NSF GRANT # 0541610

NSF PROGRAM NAME: Engineering Design

Validating discrete choice models for use in engineering design optimization

Eleanor M. Feit

Marketing Department, Stephen M. Ross School of Business, University of Michigan
701 Tappan Street, Ann Arbor, MI 48109

Mark A. Beltramo

Vehicle Development Research Lab, General Motors Research & Development Center
30500 Mound Road, Warren, MI 48090

Fred M. Feinberg

Marketing Department, Stephen M. Ross School of Business, University of Michigan
701 Tappan Street, Ann Arbor, MI 48109

Abstract: The ultimate objective of product development is to create designs that are desirable to customers. In order to incorporate this objective into engineering design optimization, it is essential to develop models of consumer demand. One method for developing empirical models of demand is discrete choice modeling. Since most engineering applications of discrete choice models involve novel product features, the models are usually estimated using data from conjoint surveys where respondents make hypothetical product choices. Because a conjoint survey involves hypothetical choices, it is important to validate these models. We propose a new validation approach that compares the parameters of a discrete choice model estimated from survey data to the parameters of a discrete choice model estimated from real-world sales data. This approach allows us to assess the performance of the demand model at the attribute level. The validation approach is currently being applied to a test case involving minivan purchases in the US new-vehicle market.

1. Discrete choice demand models in engineering design optimization: Design optimization projects often struggle with how to define the design objective. Traditionally, projects choose either to maximize a particular performance measure defined by the engineer or to minimize deviation from some pre-determined requirements. However, the ultimate objective of product development is to create new designs that will have high value to customers and ultimately to maximize profit. In order to identify high-profit designs using design optimization, we need analytical models of consumer demand [1, 2].

Although modeling consumer choice is outside the traditional domain of engineering, marketing and economics have established and developed methods for estimating statistical models of consumer demand based on observed choice behavior. These models, called *discrete choice models*, predict market share as a function of product attributes and price. The key parameters of these models represent the value the consumer places on each product attribute. Discrete choice models are a foundational tool in market research, widely used both in academia and industry. They are also frequently applied in other areas where the analysis of individual choice is important, such as public policy, health care, education, environmental economics and transportation [3, 4].

Because discrete choice models can capture consumer preference analytically, they have great potential to be integrated into design optimization frameworks. By doing this, we can improve the quality of design decisions by ensuring that designs are selected which maximize consumer preference yet are still feasible from an engineering perspective [2]. Unfortunately, validation methods for discrete choice models are not as well-developed as those commonly used in engineering. The purpose of this research is to develop a new method for validating discrete choice models.

2. Model formulation: The standard discrete choice model assumes that there are J distinct product alternatives in the market, indexed by j in $\{1, \dots, J\}$. Each of these products is described by a vector of attributes x_j . The product attributes are typically described at the consumer level. For example, attributes

of a vehicle might include acceleration, fuel economy, handling, price and reliability.

The standard formulation of the discrete choice model postulates that a consumer has a certain *utility* for each of the J products, and will choose the product offering the greatest utility. The utility of product j is modeled as

$$U_j = V(x_j) + \varepsilon_j,$$

where V is a function of the attributes of product j and ε_j is a random variable. Because this sum includes the random variable ε_j , the utility of product j is itself a random variable; this type of discrete choice model is thus sometimes called a *random-utility model*. Under this formulation, the probability that a consumer chooses alternative j is equal to

$$P_j = \Pr[V(x_j) + \varepsilon_j > V(x_{j'}) + \varepsilon_{j'} \forall j' \neq j]. \quad (1)$$

To complete the formulation of the model, we need to specify the function $V(x_j)$ and the distributional form of the random variable, ε_j . For simplicity, we assume that $V(x_j)$ is a linear function of the product attribute vector x_j . This leads to a model that is compensatory; that is, a product's deficiencies on one attribute can be made up for by superior performance on another attribute. Several authors have proposed using a non-linear form for $V(x_j)$ [1, 5], which can result in a non-compensatory choice model. This approach is desirable when consumers have hard cut-offs on attributes; e.g., "I won't buy a car that costs more than \$30,000."

If one assumes that ε_j follows a Gumbel distribution (i.e., $\Pr[\varepsilon < y] = \exp[-\exp(-y)]$), the resulting formulation is called the multinomial logit model (MNL). Assuming that ε_j is normally distributed results in the multinomial probit model (MNP). The MNP model is less computationally tractable, since it precludes a closed-form expression for P_j .

To summarize, the consumer will buy the product j that offers the greatest utility. We assume that the utility of a product j is given by U_j , where,

$$U_j = V(x_j) + \varepsilon_j = \beta^* x_j + \varepsilon_j$$

$$\varepsilon_j \sim \text{Gumbel with scale factor } \lambda$$

The parameters of this model are β^* and λ . The vector β^* describes how much each product attribute influences the likelihood that the consumer will choose the product. The scale factor λ is related to the variance of ε_j and describes how much variation there is in the

utility for product j that is not explained by the product attributes.

Since we do not observe U_j directly, it is impossible to identify λ separately from β^* [6]. For this reason, the scale of ε_j is usually normalized to 1. The parameters that are statistically estimated are $\beta = \beta^*/\lambda$. Because the scale of the random component of the utility is fixed, the absolute magnitude of the estimated elements of β is an indicator of the error variance in the data. If all of the elements of β have large absolute value, then λ is small which means the product attributes explain much of the choice behavior. If the elements of β are all small in magnitude, then the model explains little of the variation in the observed choice behavior.

For the multinomial logit model, we can solve for P_j in terms of β as

$$P_j = \frac{\exp[\beta x_j]}{\exp\left[\sum_{j'=1}^J \beta x_{j'}\right]}.$$

(See [4] for details.) A discrete choice model can be estimated given data for a number of consumers on 1) which product was chosen by the consumer and 2) the attribute vector x_j for all product alternatives $j = 1, \dots, J$ that were available to the consumer. Discrete choice models are typically estimated by maximum-likelihood estimation (MLE); that is, the parameters β are selected so that they maximize the likelihood of the observed choices.

3. Alternative data sources for estimating discrete choice models: Many discrete choice models appearing in the marketing literature are estimated from data based on purchase records. This is very common practice in the packaged goods industry, where cash registers linked to UPC scanners make purchase data easily available [7]. Choice data like this, which reflect real-world purchases, are referred to as *revealed preference data* (RP). Another way to collect RP data that is common in the transportation community is to survey consumers on their past product choices. Wassenaar et al. use an RP discrete choice model within an engineering design optimization [1]. Their model is estimated based on data collected from the accounting systems of new car dealers.

Arguably, models fitted to RP data will most accurately reflect real-world behavior. Purchase data, however, can only tell us how customers react to features and combinations of features that already exist in the marketplace. This makes RP data of limited use in

making new product development decisions that typically involve innovative features or new levels of existing attributes. For example, because there are few hybrid engine vehicles available in today's automotive market, RP data can not be used to predict how new products with hybrid engines will ultimately be received in the marketplace.

Revealed preference data often has a serious problem with colinearity among the attributes of different alternatives. For example, used vehicles typically have poorer emissions and are also less expensive. This correlation between price and emissions across the alternatives makes it difficult to determine statistically whether people buy these cars because they are cheap or because they actually prefer cars that pollute [8]. The attributes that are typically included in a model used for design optimization are often correlated due to design constraints, e.g., roominess and fuel economy. (Models from marketing that use RP data typically include attributes like pricing and advertising, which are less likely to co-vary across product alternatives.)

would most likely buy?". This is called *stated choice data*.

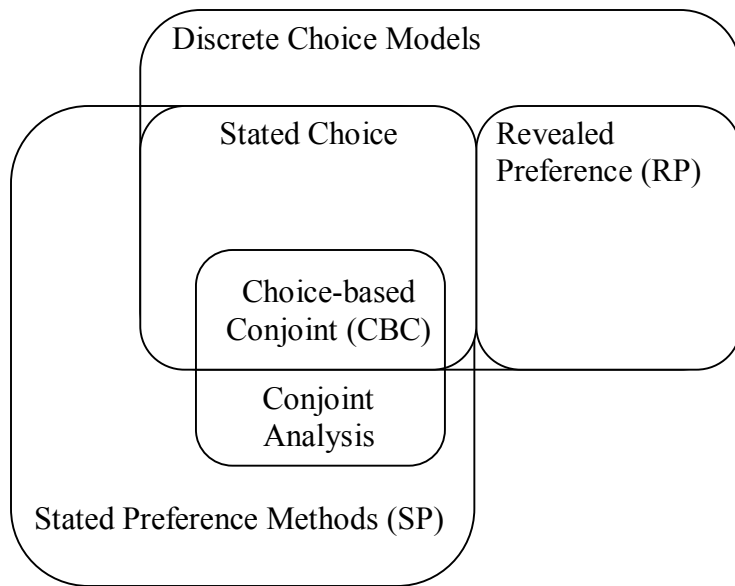
In most stated choice surveys, the product alternatives are selected according to an experimental design that maximizes the amount of information obtained about the parameters given a fixed sample size [See 3 for details.] Because the survey uses hypothetical alternatives, it is easy to incorporate novel features and attributes that are not currently available in the market. Michalek et al. give an example of an SP discrete choice model applied in an engineering design context [2].

Many of the survey methods that have grown up under the name *conjoint analysis* are techniques for eliciting stated choice data. For example, *choice-based conjoint* models are one type of discrete choice model estimated from SP data. However there are some conjoint analysis methods that are not analyzed using a discrete choice model (see Figure 1).

Given that we want to use a discrete choice model in the context of engineering design optimization, we argue that it will almost always be necessary to use SP data. However, there is some dispute about whether SP data accurately reflects real-world choice behavior [9]. A key concern is that respondents may not behave the same in the survey setting as they do when they make real purchase decisions. In particular, there are concerns that for expensive consumer durables survey respondents systematically overestimate what they would be willing to pay for improved designs. If this is true, design optimizations based on SP discrete choice models could erroneously suggest designs that are too expensive to be successful in the marketplace.

It is likely that this problem extends to attributes other than price. Brownstone et al. compared consumers' responses in a conjoint survey to their real-world vehicle purchases and found that some groups of consumers were much more likely to choose sports cars in the survey than in the real-world [8]. Uncorrected, this bias would cause a conjoint model based on the survey data to overestimate market demand for sports cars. Such biases would be easier to address were they diametric, always overstating hedonic choices, but the opposite can also be true. For example, survey respondents have been shown to overstate how much they value socially

Figure 1: Discrete Choice Model Types



Stated preference (SP) data describes any data about product preferences that are collected in a way that does not require the decision maker to make binding choices. Discrete choice models are often estimated from survey data that is collected by asking consumers to make hypothetical choices between product alternatives—i.e., “Which of these hypothetical cars do you think you

desirable attributes (e.g., green attributes, safety features) relative to their real-world purchases [10].

In summary, SP surveys produce rich information about how customers will react to combinations of features that are not currently available in the marketplace, but they are based on survey data that may only partially reflect real-world purchase behavior. RP datasets are not subject to this bias, but do not contain the information we need to predict customer response to innovative designs.

4. Validating SP discrete choice models using RP

data: Before we incorporate any model into a design optimization framework, it is important to ask whether that model has been shown to make accurate predictions. For instance, a model of beam strength based on FEA can be validated using physical testing. The parameters of an SP discrete choice model are estimated from the hypothetical choices made by survey respondents. These empirically estimated parameters need to be validated, yet current validation methods are lacking.

The most common validation strategy for SP models, hold-out validation, involves predicting the response to additional SP questions [3]. Hold-out validation gauges the internal validity of the survey, but it gives no reassurance that the SP model is consistent with real-world purchase behavior.

A second approach to validating SP models is to use the SP model to predict past (real-world) sales. This approach does give some reassurance that the SP model is consistent with real-world purchase behavior, but if the SP model does not predict the sales well, then we have no indication of which particular attribute parameter (i.e., which element of β) is inconsistent between the SP setting and the real-world.

We propose an alternative approach to validation that involves comparing parameters of a SP model to parameters estimated from RP data (i.e., observed purchases). Unlike the other approaches to external validation, this method can assess the validity of the conjoint model at the attribute level.

If respondents answer the SP questions in a way that is consistent with their real-world behavior, then we would expect the parameters β^* in an SP model to be the same as those in a related RP model. Suppose we have a set of $t = 1, \dots, T$ observations, some of which were collected in an RP setting and some of which were collected in an SP setting. Assuming consumers value the attributes similarly in both choice settings, then

utility is given by

$$U_{jt} = \begin{cases} \beta^* x_{jt} + \varepsilon_{jt}^{RP} & \text{if } t \in RP \\ \beta^* x_{jt} + \varepsilon_{jt}^{SP} & \text{if } t \in SP \end{cases} \quad (2)$$

$$\varepsilon_{jt}^{RP} \sim \text{IID Gumbel with scale factor } \lambda^{RP}$$

$$\varepsilon_{jt}^{SP} \sim \text{IID Gumbel with scale factor } \lambda^{SP}$$

Note that in equation 2, we do not assume that the random component ε_{jt}^{RP} has the same distribution as the random component ε_{jt}^{SP} . Even if the parameter vector, β^* is common across the two data sources, it is unreasonable to assume that the random utility variance is common across the two choice settings [6]. In the traditional econometric interpretation of the discrete choice model, the random utility component represents unobserved attributes; that is, aspects of the product that are known to the consumer, but not to the researcher. It is unlikely that these unobserved attributes are the same across the two choice contexts. The random utility component may also represent noise in the consumer's choice process which has been shown to change over choice settings [11].

Just as in the case of a discrete choice model estimated from a single data source, λ^{SP} and λ^{RP} are not separately identified from β^* and we typically normalize the scale factor of ε_{jt}^{RP} to 1. However, it is possible to estimate the ratio of the two scale parameters, $\mu = \lambda^{SP} / \lambda^{RP}$ as long as the two data sets have at least one parameter in common. The resulting model is

$$U_{jt} = \begin{cases} \beta x_{jt} + \varepsilon_{jt} & \text{if } t \in RP \\ \mu \beta x_{jt} + \varepsilon_{jt} & \text{if } t \in SP \end{cases}$$

$$\varepsilon_{jt} \sim \text{IID Gumbel with scale factor} = 1$$

where μ accounts for the difference in error variance across the two data sets [3,4].

It is not necessary to assume that all of the elements of β are common across the two data sources. A likelihood ratio test can be used to determine if a particular attribute coefficient is common across two data sets after a difference in scale is accounted for [3,4]. If the likelihood ratio test fails to reject the hypothesis that the parameters are equal, then we conclude that the SP choices are consistent with the RP data with respect to that attribute. This test allows us to validate the

parameters of the SP choice model directly against the parameters of an RP model.

Note that it is only possible to make the comparison between the estimated parameter for the two datasets after we have accounted for the scale ratio, μ . That is, we have to estimate the μ and adjust the parameters of the SP model by that factor before we can make comparisons between SP and RP model parameters.

RP and SP parameters have been compared for discrete choice models in several products including laundry detergent [12] and automobiles [8]. Generally, these studies show that the majority of parameters are consistent across the two choice contexts. There are some exceptions, for instance, Brownstone et al. found that the value consumers place on emissions in a discrete choice model for vehicle choice is different across the RP and SP data sets [8].

5. Bayesian methodology for estimating discrete choice models: A limitation of the discrete choice models we have described is that they assume that the parameters of the utility function are common among all consumers. Recent work in marketing has demonstrated that better model performance can be achieved by introducing a hierarchical (or random coefficients) model specification that allows for heterogeneity across consumers. For example, we can assume that each individual, i , has a unique parameter vector β_i and that this parameter vector follows a multivariate normal distribution across the population of consumers. I.e.,

$$\beta_i \sim MVN(\mu_\beta, \Sigma_\beta)$$

where μ_β is the population mean parameter vector and Σ_β is the population covariance matrix.

Hierarchical multinomial logit models are common in marketing and have been shown to perform substantially better than models which assume the relative importance of the attributes is the same for all consumers [4, 13]. One common approach for estimating these models is Bayesian inference. We are currently developing methods to estimate joint RP/SP hierarchical-Bayes discrete choice models.

6. Planned application for US-market minivans: We plan to apply these methods to SP data obtained from a conjoint survey conducted in 2003. This survey collected 12 SP choice questions for each of 199 potential minivan buyers. We have constructed a matched RP data set consisting of approximately 7000 minivan purchases that was obtained from a survey of

new minivan buyers conducted during the 2004 model year.

7. Acknowledgements: This material is based on work supported by General Motors and the National Science Foundation under Grant No. DMI-0541610. Any opinions, findings and conclusions or recommendations are those of the authors and do not necessarily reflect the views of the National Science Foundation.

8. References:

- [1] H.J. Wassenaar, W. Chen, J. Cheng, A. Sudhianto, "Enhancing Discrete Choice Demand Modeling for Decision-Based Design", *Journal of Mechanical Design*, vol. 127, pp. 514-523, 2005
- [2] J.J. Michalek, F.M. Feinberg, P.Y. Papalambros, "Linking Marketing and Engineering Product Design Decisions via Analytical Target Cascading", *J. Product Innovation Management*, vol. 22, pp. 42-62, 2005
- [3] J.J. Louviere, D.A. Hensher and J.D. Swait, "Stated Choice Methods: Analysis and Application", Cambridge University Press, Cambridge, UK, 2000
- [4] K. Train, "Discrete Choice Methods with Simulation", Cambridge University Press, Cambridge, UK, 2003
- [5] J.G. Kim, U. Menzefricke, and F.M. Feinberg, "Capturing Heterogeneous Utility Curves: A Bayesian Spline Approach," *Management Science*, forthcoming
- [6] J. Swait and J.J. Louviere, "On the Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models", *J. of Marketing Research*, vol. 30, n. 3, pp. 305-314, 1993
- [7] P.E. Guadagni and J.D.C. Little, "A Logit Model of Brand Choice Calibrated on Scanner Data", *Marketing Science*, vol. 2, n. 3, pp. 203-238, 1983
- [8] D. Brownstone, D.S. Bunch and K. Train, "Joint mixed logit models of stated and revealed preferences for alternative fuel vehicles", *Trans. Research Part B*, vol. 34, pp. 315-338, 2000
- [9] G. Allenby, G. Fennell, J. Huber, T. Eagle, T. Gilbride, D. Horsky, J. Kim, P. Lenk, R. Johnson, E. Ofek, B. Orme, T. Otter, J. Walker, "Adjusting Choice Models to Better Predict Market Behavior", *Marketing Letters*, vol. 6, n. 3, pp. 197-208, 2005

[10] R. Blamey and J. Bennett, "Yea-saying and Validation of a Choice Model of Green Product Choice", in *The Choice Modeling Approach to Environmental Evaluation*, J. Bennett and R. Blamey, eds., Edward Elgar Publishing, 2001

[11] M. Bradley and A. Daly, "Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data", *Transportation*, vol. 21, pp. 167-184, 1994

[12] J. Swait and R.L. Andrews, "Enriching Scanner Panel Models with Choice Experiments", *Marketing Science*, vol. 22, n. 4, 2003

[13] J.J. Michalek, F.M. Feinberg, F. Adiguzel, P. Ebbes and P.Y. Papalambros, "Realizable product line design optimization: Coordinating marketing and engineering models via analytical target cascading, Working Paper, University of Michigan, 2005