Accepted Version of Manuscript by *Journal of Applied Statistics*. Cite as:

Bergtold J.S., and E. Onukwugha. 2014. The probabilistic reduction approach to specifying multinomial logistic regression models in health outcomes research. *Journal of Applied Statistics* 41(10): 2206 – 2221.

Published Version at: https://doi.org/10.1080/02664763.2014.909785

The probabilistic reduction approach to specifying multinomial logistic regression models in health outcomes research

Jason S. Bergtold[1] and Eberechukwu Onukwugha[2]

[1] Associate Professor, Department of Agricultural Economics, Kansas State University, 304G Waters Hall, Manhattan, KS 66506-4011, USA, Phone: 785-532-0984, Fax: 785-532-6925, Email: bergtold@ksu.edu (Corresponding Author)

[2] Assistant Professor, Pharmaceutical Health Services Research, University of Maryland School of Pharmacy, 220 Arch Street, Baltimore, MD, 21201, USA, Phone: 410-706-8981, Email: eonukwug@rx.maryland.edu

The probabilistic reduction approach to specifying multinomial logistic regression models in health outcomes research

Abstract

The paper provides a novel application of the Probabilistic Reduction approach to the analysis of multi-categorical outcomes. The Probabilistic Reduction (PR), which systematically takes account of heterogeneity and functional form concerns, can improve the specification of binary regression models. However, its utility for systematically enriching the specification of and inference from models of multi-categorical outcomes has not been examined, while multinomial logistic regression models are commonly used for inference and, increasingly, prediction. Following a theoretical derivation of the PR-based multinomial logistic model (MLM), we compare functional specification and marginal effects from a traditional specification and a PR-based specification in a model of post-stroke hospital discharge disposition and find that the traditional MLM is misspecified. Results suggest that the impact on the reliability of substantive inferences from a misspecified model may be significant, even when model fit statistics do not suggest a strong lack of fit compared to a properly specified model using the PR approach. We identify situations under which a PR-based MLM specification can be advantageous to the applied researcher.

Keywords: heterogeneity, interaction effects, marginal effects, model specification, multinomial logistic regression, probabilistic reduction approach

The Probabilistic Reduction Approach to Specifying Multinomial Logistic Regression Models and Health Outcomes Research

1. Introduction

The availability of large datasets from insurers, hospitals, and third-party groups simplifies the task of generating population-based evidence of health differences and outcomes within a population. Statistical tools employed to analyze such datasets (and other similar datasets) should account for the complexity of the data generation process. There is often significant heterogeneity that exists in such data. 'Main effects'-only models assume the relationship between the factor being examined and the health outcome is homogeneous across population subgroups defined, for example, by race, disease severity or age. Understanding health trends across population subgroups defined by race is particularly important to the national public health agenda, as evidenced by the Healthy People 2010 and 2020 initiatives (http://www.healthypeople.gov/2020/default.aspx) targeting racial disparities in health status.

Main effects models are misspecified when heterogeneity across population subgroups or levels of analysis are present. To account for heterogeneity, statistical refinements presented in the literature include the incorporation of interaction effects (e.g. accounting for effect modification)[16, 8, 19]; stratification [7, 16, 18, 28] and/or extensions to address variance heterogeneity via multi-level models [29, 34, 35]. In the case of ethnic/racial disparities authors have reversed or otherwise revised [29, 34, 35] initial conclusions regarding the existence of a race disparity after enriching the model

using these techniques.[1] Evidence from the literature confirms that model specification can lead to differences in the conclusions regarding heterogeneity in outcomes. Prior work [4] discussed specification of binary logistic models for non-randomized designs. However, no systematic approach to model specification has been developed for the analysis of observational data for statistical models with multiple nominal outcomes.

This paper utilizes the Probabilistic Reduction (PR) approach [32, 33] to model specification to develop a systematic approach to empirical modeling using observational data that takes account of the probabilistic information and heterogeneity present in the data. As applied to models with multiple nominal outcomes, the PR approach leads to a model specification that accords a primary role to the inverse conditional distribution of the explanatory variables conditioned on the nominal outcomes. The specification based on the PR approach is consistent with that discussed in earlier work for binary logistic regression models [4, 20, 30]. Prior studies provide a rationale for the consideration of using inverse conditional distributions in specifying multinomial logistic regression models [2 , 23], but do not provide a systematic approach for model specification that extends much beyond linearity in the variables for the index or predictor functions of the model. The PR approach provides a justification for considering the role of a myriad of parametric inverse conditional distributions for specifying multinomial logistic regression models that may be nonlinear in the parameters, explanatory variables, or both (see Bergtold *et al*. [4] for examples for the binary case). Differences in the functional form can have important

[1] Heterogeneity and differences due to ethnic and racial disparity has been a significant topic in the health outcomes research literature [13,17,26,27].

implications for substantive inferences (e.g. marginal effects or odds ratios) estimated using the model.

The purpose of this paper is to examine the specification of the multinomial logistic regression model using the Probabilistic Reduction approach with an empirical application: examining hospital discharge disposition. The paper has three objectives: (1) to derive the multinomial logistic regression model using the PR approach, emphasizing the role the inverse conditional distribution plays in model specification; (2) to compare the PR approach to more traditional specifications; and (3) to examine the impact of model misspecification on inference via marginal effects when the traditional approach leads to a misspecified model. Model comparisons are conducted in the context of an empirical example of post-stroke hospital discharge disposition of Maryland patients from 2000 to 2005. Results from the empirical example comparing the multinomial logistic regression model specified using the PR and more traditional approaches suggest that the impact on the reliability of substantive inferences (including racial and ethnic disparities) from a misspecified model may be significant, even when model fit statistics do not suggest a strong lack of fit compared to the properly specified model. The novelty of the paper is the presentation of a systematic approach for modeling multinomial regression models that incorporates heterogeneity across subpopulations and provides derivations for marginal effect calculations that provide substantive inferences.

2. Theory and Methods

Existing frameworks for specifying the multinomial logistic regression models include the latent variable approach [37] and the transformational (generalized linear

model) approach [10]. Bergtold *et al*. [4] show that these two approaches can lead to model specifications that are either misspecified or do not take advantage of all the statistical information inherent in the data, by ignoring the distributional properties of covariates or explanatory variables. The generalized linear model approach lacks specific guidance on the functional form of the predictor (or log odds) function in the multinomial logistic regression model past linearity in the parameters, variables or both. The latent variable approach often imposes linearity to meet theoretical considerations. These requirements can result in model specifications which may not always be appropriate or complete [3]. The probabilistic structure of the explanatory variables can play an important role in the functional form of the predictor or index function. Arnold *et al*. [3] state that "many of the logistic regression models discussed in the applied literature are questionable in the light of these observations (p. 134)." This is not to mention the inclusion of covariate terms or structure to model heterogeneity across sub-populations. This leads to the novelty provided by examining the model specification of statistical models with nominal outcomes using the Probabilistic Reduction approach.

2. 1 ***The probabilistic reduction approach***

The Probabilistic Reduction approach to model specification specifies a statistical model based on the reduction of the joint distribution of the dependent binary variable and categorical and continuous explanatory variables. The importance accorded to the joint distribution leads to recognition of the role of inverse conditional distributions in providing relevant statistical information for model specification [4]. Inverse conditional

distributions (the distribution of the independent variables conditional on the discrete outcome) can provide important information about the functional form of the statistical model and in turn choice or outcome probabilities being estimated. This has been explored for the binary logistic regression model [4, 20, 30]. Using the joint distribution, via the inverse conditional distribution, to inform the model selection process can bring transparency to model-building and help ensure statistical adequacy and reliable substantive inferences.

Consider the finite nominal choice or outcome set $B = \{1, \dots, M\}$ for a set of N individuals and M choices. Furthermore, consider a set of K observed variables, denoted as $\boldsymbol{X}_i = \left(X_{1,i}, \dots, X_{K,i}\right)$, associated with the i^{th} individual. Let the distribution of $\boldsymbol{X}_i$ be given by $f_{\boldsymbol{X}}(\boldsymbol{X}_i; \theta)$, where θ is an appropriate set of parameters. Let Y_i denote a polychotomous index of the set B, such that $Y_i = j$ when individual i chooses alternative $j \in B$. Assume that $E(Y_i) = \mathbf{P}(Y_i = j) = p_j$, which makes the distribution of Y_i:

$$f_Y(Y_i; \boldsymbol{p}) = \prod_{j=1}^{M} p_j^{\mathbf{1}(Y_i = j)}, \tag{1}$$

where $\mathbf{1}(.)$ denotes the indicator function and $\sum_{j=1}^{M} p_j = 1$. Equation (1) is known as the multinomial distribution.

Assume that the joint vector stochastic process $\{(Y_i, \boldsymbol{X}_i), i = 1, \dots, N\}$ is independent (I). The stochastic process can be identically distributed (ID), but can be assumed to vary across sub-populations (e.g. heterogeneity among sub-groups) without loss of generality. Following Spanos [33], the PR approach starts with the decomposition of the joint distribution of the vector stochastic process $\{(Y_i, \boldsymbol{X}_i), i = 1, \dots, N\}$, which depends on the existence of the multivariate distribution of $(Y_i, \boldsymbol{X}_i)$. That is:

$$f(Y_1,...,Y_N,\mathbf{X}_1,...,\mathbf{X}_N;\phi)\overset{I}{=}\prod_{i=1}^{N} f_i\left(Y_i,\mathbf{X}_i;\varphi_i\right)\overset{ID}{=}\prod_{i=1}^{N} f\left(Y_i,\mathbf{X}_i;\varphi\right),$$

where ϕ is an appropriate set of parameters. The multivariate distribution $f(Y_i,\boldsymbol{X}_i;\varphi)$ can be represented as the product of conditional and marginal distributions, i.e.:

$$f(Y_i,\boldsymbol{X}_i;\varphi)=f_{Y|\boldsymbol{X}}(Y_i|\boldsymbol{X}_i;\boldsymbol{\beta})\cdot f_{\boldsymbol{X}}(\boldsymbol{X}_i;\theta)=f_{\boldsymbol{X}|Y}\left(\boldsymbol{X}_i|Y_i;\theta_j\right)\cdot f_{\boldsymbol{Y}}(Y_i;\boldsymbol{p}), \qquad (2)$$

where φ is an appropriate set of parameters, $f_{Y|\boldsymbol{X}}(Y_i|\boldsymbol{X}_i;\boldsymbol{\beta})$ is the conditional distribution of Y_i given $\boldsymbol{X}_i$; $\boldsymbol{\beta}$ is an appropriate set of parameters; $f_{\boldsymbol{X}|Y}\left(\boldsymbol{X}_i|Y_i;\theta_j\right)$ is the inverse conditional distribution of $\boldsymbol{X}_i$ given Y_i; and θ_j is an appropriate set of parameters and a function of $Y_i=j$. The existence of $f(Y_i,\boldsymbol{X}_i;\varphi)$ and in turn a proper statistical model is dependent on the compatibility of $f_{Y|\boldsymbol{X}}(Y_i|\boldsymbol{X}_i;\boldsymbol{\beta})$ and $f_{\boldsymbol{X}|Y}\left(\boldsymbol{X}_i|Y_i;\theta_j\right)$ [3].

Now assume that $f_{Y|\boldsymbol{X}}(Y_i|\boldsymbol{X}_i;\boldsymbol{\beta})$ is a conditional multinomial mass function that takes the form:

$$f_{Y|\boldsymbol{X}}(Y_i|\boldsymbol{X}_i;\boldsymbol{\beta})=\prod_{j=1}^{M} h_j(\boldsymbol{X}_i;\boldsymbol{\beta})^{\mathbf{1}(Y_i=j)}, \qquad (3)$$

where $\boldsymbol{\beta}=(\beta_1,\dots,\beta_M)$; $h_j(\boldsymbol{X}_i;\boldsymbol{\beta}):\Theta_{\boldsymbol{\beta}}\rightarrow[0,1]$ for $j=1,\dots,M$; $\Theta_{\boldsymbol{\beta}}$ is the parameter space associated with $\boldsymbol{\beta}$; and $\sum_{j=1}^{M} h_j\left(\boldsymbol{X}_i;\boldsymbol{\beta}\right)=1$. The probability of the i^{th} individual choosing choice j can be determined by letting $Y_i=j$; substituting equations (1) and (3) into equation (2); and rearranging terms, giving:

$$\mathbf{P}(Y_i=j|\mathbf{X}_i=\mathbf{x}_i)=E(Y_i=j|\mathbf{X}_i=\mathbf{x}_i)=h_j(\boldsymbol{X}_i;\boldsymbol{\beta})=\frac{f_{\boldsymbol{X}|Y=j}\left(\boldsymbol{X}_i|Y_i=j;\theta_j\right)\cdot p_j}{f_{\boldsymbol{X}}(\boldsymbol{X}_i;\theta)}. \qquad (4)$$

The objective now is to determine the functional form of $h_j(\boldsymbol{X}_i;\boldsymbol{\beta})$, which will incorporate the inverse conditional distribution $f_{\boldsymbol{X}|Y}\left(\boldsymbol{X}_i|Y_i;\theta_j\right)$, ensuring compatibility with $f_{Y|\boldsymbol{X}}(Y_i|\boldsymbol{X}_i;\boldsymbol{\beta})$ and the existence of a statistical model.

Following Arnold *et al*. ([3], p. 17), a sufficient condition for the compatibility of $f_{Y|\boldsymbol{X}}(Y_i|\boldsymbol{X_i};\boldsymbol{\beta})$ and $f_{\boldsymbol{X}|Y}\left(\boldsymbol{X}_i|Y_i;\theta_j\right)$ is that the ratio:

$$\frac{f_{Y|\boldsymbol{X}}(Y_i=j|\boldsymbol{X_i};\boldsymbol{\beta})\cdot f_{\boldsymbol{X}|Y=m}(\boldsymbol{X}_i|Y_i=m;\theta_m)}{f_{Y|\boldsymbol{X}}(Y_i=m|\boldsymbol{X_i};\boldsymbol{\beta})\cdot f_{\boldsymbol{X}|Y=j}(\boldsymbol{X}_i|Y_i=j;\theta_j)} \tag{5}$$

does not depend on $\boldsymbol{X}_i$ for any combination of choices j and m such that $j \neq m$. Using equation (2), condition (5) must be equal to $\frac{p_j}{p_m}$. Now let m represent the "all other" choice, such that $p_m = 1 - \sum_{s\neq m} p_s$ and $h_m(\boldsymbol{X}_i;\boldsymbol{\beta}) = 1 - \sum_{s\neq m} h_s\,(\boldsymbol{X}_i;\boldsymbol{\beta})$. Substituting equation (3) into condition (5) gives the following set of sufficient conditions:

$$\frac{f_{\boldsymbol{X}|Y=m}(\boldsymbol{X}_i|Y_i=m;\theta_m)}{f_{\boldsymbol{X}|Y=j}(\boldsymbol{X}_i|Y_i=j;\theta_j)} \cdot \frac{h_j(\boldsymbol{X}_i;\boldsymbol{\beta})}{1-\sum_{s\neq m} h_s(\boldsymbol{X}_i;\boldsymbol{\beta})} = \frac{p_j}{p_m} \text{ for } j = 1,\ldots,\text{M and } j \neq m.^{2} \tag{6}$$

That is, the conditions given by (6) must be satisfied for $f_{Y|\boldsymbol{X}}(Y_i|\boldsymbol{X_i};\boldsymbol{\beta})$ and $f_{\boldsymbol{X}|Y}\left(\boldsymbol{X}_i|Y_i;\theta_j\right)$ to be compatible. Note that the base choice m can be chosen arbitrarily, as well. Solving the set of M -1 equations given by condition (6) for $h_j(\boldsymbol{X}_i;\boldsymbol{\beta})$, j = 1,…,M and $j \neq m$, gives:

$$h_j(\boldsymbol{X}_i;\boldsymbol{\beta}) = \frac{g_j\left(\boldsymbol{X}_i;\beta_j(\theta_j,\theta_m,p_j,p_m)\right)}{1+\sum_{s\neq m} g_s\left(\boldsymbol{X}_i;\beta_s(\theta_s,\theta_m,p_j,p_m)\right)}, \tag{7}$$

where

$$g_j\left(\boldsymbol{X}_i;\beta_j(\theta_j,\theta_m,p_j,p_m)\right) = \frac{f_{\boldsymbol{X}|Y=j}(\boldsymbol{X}_i|Y_i=j;\theta_j)\cdot p_j}{f_{\boldsymbol{X}|Y=m}(\boldsymbol{X}_i|Y_i=m;\theta_m)\cdot p_m} \text{ for } j = 1,\ldots,\text{M and } j \neq m. \tag{8}$$

The parameter vector $\beta_j = \beta_j(\theta_j,\theta_m,p_j,p_m)$ is a function of the parameters of the inverse conditional distributions for $Y = j$ and $Y = m$.

[2] This approach is related to that used by Maddala [22].

Equations (3), (7) and (8) give us the general specification for the multinomial model with nominal responses in terms of the inverse conditional distribution. A more intuitive choice for $g_j\left(\boldsymbol{X}_i;\beta_j(\theta_j,\theta_m,p_j,p_m)\right)$ can be found by using the identity $f(.)=\exp\left(\ln f(.)\right)$, giving:

$$g_j\left(\boldsymbol{X}_i;\beta_j(\theta_j,\theta_m,p_j,p_m)\right)=\exp\left(\eta_j\left(\boldsymbol{X}_i;\beta_j(\theta_j,\theta_m,p_j,p_m)\right)\right)$$

$$=\exp\left(\ln\left[\frac{f_{\boldsymbol{X}|Y=j}(\boldsymbol{X}_i|Y_i=j;\theta_j)}{f_{\boldsymbol{X}|Y=m}(\boldsymbol{X}_i|Y_i=m;\theta_m)}\right]+\kappa\right), \qquad (9)$$

where $\kappa=\ln(p_j)-\ln(p_m)$; $\eta_j\left(\boldsymbol{X}_i;\beta_j(\theta_j,\theta_m,p_j,p_m)\right)$ is the index function (or predictor) for outcome j; and $p_m=1-\sum_{s\neq m}p_s$.

Taking equations (3), (7) and (9) together gives rise to the logistic multinomial regression model, but it should be emphasized that the functional form of the index/predictor functions $\eta_j\left(\boldsymbol{X}_i;\beta_j(\theta_j,\theta_m,p_j,p_m)\right)$ are dependent on the inverse conditional distributions $f_{\boldsymbol{X}|Y}(\boldsymbol{X}_i|Y_i;\theta_j)$. As will be seen below and emphasized by Bergtold *et al*. [4] for the binary case, specification of the index/predictor functions may not be linear in the variables (or parameters) as is usually found in applied literature. Furthermore, the choice of using the logistic cumulative distribution function in the specification of the model arises naturally from the derivation of the model and makes estimation of the model more compatible with existing statistical software packages. That is, model specification primarily deals with the specification of the index/predictor function, as the logistic formulation arises naturally from the distributional assumption about the dependent variable.

Given heterogeneity is defined in terms of the moments of a distribution [33] and the multinomial distribution is completely characterized by its mean vector, heterogeneity in the multinomial logistic regression model is defined in terms of the parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)$**.** The inclusion of indicator or binary variables to define sub-populations in the data has the effect of allowing $\boldsymbol{\beta}$ to vary across sub-populations defined by the indicator variables. Interactions of these indicator variables in the above framework can allow researchers to capture multilevel effects. The inclusion of higher order terms (beyond interaction terms) allows for more complex heterogeneity to be modeled. A significant benefit of this approach is that by invoking the use of the inverse conditional distribution for model specification, the approach inherently considers these potential interactions. Thus, the modeler may discover unknown or unsuspected heterogeneity in the observed data using the PR approach.

This unmodeled or unsuspected heterogeneity would include stratification, as the covariates would include multiplicative terms (e.g. age x race or age x race x income grouping) that directly capture the stratification in the model or the (marginal) effect of different sub-populations on a health outcome from a particular explanatory variable or covariate of interest. Additional forms of unmodeled heterogeneity may be captured using appropriate methods that allow for variation in $\boldsymbol{\beta}$ such as the mixed logit or other latent class models [4, 35]. Furthermore, following Fahrmeir and Tutz [10], the conditional covariance matrix across choices or outcomes is given by $\text{diag}(\boldsymbol{h}) - \boldsymbol{h}\boldsymbol{h}'$, where $\boldsymbol{h} = (h_1, \ldots, h_M)$ and h_j is given by equations (7) and (9). Thus, the covariance matrix is explicitly a function of $\boldsymbol{\beta}$. Following Bergtold *et al*. [4], heterogeneity in the

variance/covariance matrix has to be captured by variations in the parameter vector $\boldsymbol{\beta}$, as the variance/covariance matrix is explicitly defined in terms of the conditional mean vector $\boldsymbol{h}$.

2.2 *Functional form specification*

Specification of the index functions is accomplished by specifying the functional form for the inverse conditional distribution $f_{\boldsymbol{X}|Y}(\boldsymbol{X}_i|Y_i;\theta_j)$. Bergtold *et al*. [4]; Kay and Little [20]; and Scrucca and Weisberg [30] provide guidance on the specification of the index function for the binary choice logistic regression model using the inverse conditional distribution. The results from these papers have direct application here. For example, as shown by McFadden [23], the index functions will be linear in the explanatory variables if the inverse conditional distribution is multivariate normal with common covariance matrix for all $j \in B$. If the covariance matrices exhibit heterogeneity over $j \in B$, then following Kay and Little [20] the index functions would be quadratic in the explanatory variables. Bergtold et al. [4] provides some examples when $g_j\left(\boldsymbol{X}_i;\beta_j(\theta_j,\theta_m,p_j,p_m)\right)$ is nonlinear in $\boldsymbol{X_i}$. For example, consider the four variate case where X_1 and X_2 are jointly distributed bivariate beta, X_3 is binary, and X_4 is exponential conditional on X_3. Then the index function of the model or $g_j\left(\boldsymbol{X}_i;\beta_j(\theta_j,\theta_m,p_j,p_m)\right)$ would be linear in the parameters and include an intercept, $\ln(X_1)$, $\ln(X_2)$, $\ln(1-X_1-X_2)$, $X_3, X_4,$ and X_3X_4 [4]. Many of the direct distributional assumption examined in these papers are useful with a small number of covariates or covariates that are conditionally independent of each other. Kay and Little [20] emphasize the increasing difficulty in model specification when the inverse

conditional distribution must be treated as a multivariate distribution when the covariates are not conditionally independent. To help deal with this problem, the modeler can turn to approximations or other variable transformations to arrive at a manageable model specification approach. The latter approach is developed further in this paper.

A general functional form for the index function can be determined by following Bergtold *et al.* [4]. Consider the following decomposition of the inverse conditional distribution:

$$f_{\boldsymbol{X}|Y}\left(\boldsymbol{X}_i|Y_i;\theta_j\right) = f_{\boldsymbol{X}_1|\boldsymbol{X}_2,Y}\left(\boldsymbol{X}_{1,i}|Y_i,\boldsymbol{X}_{2,i};\theta_{1,j}\right)\cdot f_{\boldsymbol{X_2}|Y}\left(\boldsymbol{X}_{2,i}|Y_i;\theta_{2,j}\right), \tag{10}$$

where $\boldsymbol{X}_{1,i}$ is a set of continuous variables, $\boldsymbol{X}_{2,i}$ is a set of binary variables, $\theta_{1,j}$ is a set of parameters dependent upon $\boldsymbol{X}_{2,i}$ and Y_i, and $\theta_{2,j}$ is a set of parameters dependent upon Y_i. To obtain this decomposition, any ordinal variables may need to be recoded as binary or treated as continuous.

Using a conditional distribution proposed by Day and Kerridge [8], assume that $f_{\boldsymbol{X}_1|\boldsymbol{X}_2,Y}\left(\boldsymbol{X}_{1,i}|Y_i,\boldsymbol{X}_{2,i};\theta_{1,j}\right)$ takes the functional form:

$$f_{\boldsymbol{X}_1|\boldsymbol{X}_2,Y}\left(\boldsymbol{X}_{1,i}|Y_i,\boldsymbol{X}_{2,i};\theta_{1,j}\right) = \alpha_{1,j,\delta}\cdot\exp\left\{-\frac{1}{2}\left(\boldsymbol{X_{1,i}}-\Lambda_{1,\mathrm{j},\delta}\right)'\mathrm{A}_{1,\mathrm{j},\delta}^{-1}\left(\boldsymbol{X_{1,i}}-\Lambda_{1,\mathrm{j},\delta}\right)\right\}\cdot\delta\left(\boldsymbol{X_{1,i}}\right), \tag{11}$$

where $\Lambda_{1,\mathrm{j},\delta}$ are the mean vectors and $\mathrm{A}_{1,\mathrm{j}}$, δ is the covariance matrix conditional on j = 1,…,M, and for $\boldsymbol{X_{2,i}} = \boldsymbol{\delta}$, where $\boldsymbol{\delta}$ is a particular realization of the binary random variables $\boldsymbol{X_{2,i}}$.That is, the parameters of the inverse conditional distribution given by equation (11) are dependent on the values of Y_i and $\boldsymbol{X}_{2,i}$. The term $\delta\left(\boldsymbol{X_{1,i}}\right)$ is a non-negative scalar function of $\boldsymbol{X_{1,i}}$**.** When $\delta\left(\boldsymbol{X_{1,i}}\right) = 1$, the density given by equation (11) is

the multivariate normal distribution. When $\delta(\boldsymbol{X_{1,i}}) \neq 1$, the density function can represent a wide range of alternatives, including skewed distributions [6]. The advantage of this distributional assumption is that $\delta(\boldsymbol{X_{1,i}})$ does not have to be specified explicitly to arrive at an estimable model [8].

Letting $\boldsymbol{X}_{2,i}$ be a $(K_2 \times 1)$ vector of binary variables, $f_{\boldsymbol{X_2}|Y}(\boldsymbol{X}_{2,i}|Y_i;\theta_{2,j})$ will be a multivariate Bernoulli distribution. The multivariate Bernoulli distribution can be represented in log-linear form as:

$$f_{\boldsymbol{X_2}|Y}(\boldsymbol{X}_{2,i};\theta_{2,j}) = \exp\begin{pmatrix} u_{0,j} + \sum_{k=1}^{K_2} u_{k,j}X_{2,k,i} + \sum_{k=1}^{K_2}\sum_{r>k}^{K_2} u_{k,r,j}X_{2,k,i}X_{2,r,i} \\ + \ \ldots + u_{1,.2,\ldots,K_{2,j}}X_{2,1,i}\cdots X_{2,K_2,i} \end{pmatrix} \quad (12)$$

[21]. Combining equations (11) and (12) into equation (10) and then substituting this into the index function in equation (9), gives the following functional form for the index/predictor function (following [4]):

$$\eta_j(\boldsymbol{X}_i;\beta_j) = \left(\alpha_0 + \sum_{s=1}^{K_1}\alpha_s X_{1,s,i} + \sum_{s=1}^{K_2}\sum_{t\geq s}^{K_2} u_{k,r,j}X_{1,s,i}X_{2,t,i}\right) \times \left(u_0 + \sum_{k=1}^{K_2} u_k X_{2,k,i} + \sum_{k=1}^{K_2}\sum_{r>k}^{K_2} u_{k,r}X_{2,k,i}X_{2,r,i} + \ \ldots + u_{1,.2,\ldots,K_2}X_{2,1,i}\cdots X_{2,K_2,i}\right), \quad (13)$$

where $\boldsymbol{X_{1,i}}$ is a $(K_1 \times 1)$ vector of continuous variables. The index function given in equation (13) can be made linear in the parameters via a reparameterization of the model by letting $\beta_j = \beta(\boldsymbol{\alpha}, \boldsymbol{u})$. To further improve the tractability of the model, the order of interaction terms in equation (12) may need to be limited to $\overline{K} \leq K_2$.

As an example, consider the case of two continuous variables (X_1, X_2) and three binary variables (X_3, X_4, X_5), where only second order interaction terms are considered in equation (12). In this case, the index function would take the form (in terms of β_j):

$$\begin{aligned}\eta_j(\boldsymbol{X}_i;\beta_j) = {} & \beta_{0,j} + \beta_{1,j}x_{1,i} + \beta_{2,j}x_{2,i} + \beta_{3,j}x_{3,i} + \beta_{4,j}x_{4,i} + \beta_{5,j}x_{5,i} + \beta_{11,j}x_{1,i}^2 \\ & + \beta_{12,j}x_{1,i}x_{2,i} + \beta_{22,j}x_{2,i}^2 + \beta_{13,j}x_{1,i}x_{3,i} + \beta_{14,j}x_{1,i}x_{4,i} + \beta_{15,j}x_{1,i}x_{5,i} \\ & + \beta_{23,j}x_{2,i}x_{3,i} + \beta_{24,j}x_{2,i}x_{4,i} + \beta_{25,j}x_{2,i}x_{5,i} + \beta_{34,j}x_{3,i}x_{4,i} \\ & + \beta_{35,j}x_{3,i}x_{5,i} + \beta_{45,j}x_{4,i}x_{5,i} + \beta_{113,j}x_{1,i}^2x_{3,i} \\ & + \beta_{114,j}x_{1,i}^2x_{4,i} + \beta_{115,j}x_{1,i}^2x_{5,i} + \beta_{123,j}x_{1,i}x_{2,i}x_{3,i} + \beta_{124,j}x_{1,i}x_{2,i}x_{4,i} \\ & + \beta_{125,j}x_{1,i}x_{2,i}x_{5,i} + \beta_{223,j}x_{2,i}^2x_{3,i} + \beta_{224,j}x_{2,i}^2x_{4,i} + \beta_{225,j}x_{2,i}^2x_{5,i} \\ & + \beta_{134,j}x_{1,i}x_{3,i}x_{4,i} + \beta_{135,j}x_{1,i}x_{3,i}x_{5,i} + \beta_{145,j}x_{1,i}x_{4,i}x_{5,i} \\ & + \beta_{234,j}x_{2,i}x_{3,i}x_{4,i} + \beta_{235,j}x_{2,i}x_{3,i}x_{5,i} + \beta_{245,j}x_{2,i}x_{4,i}x_{5,i} \\ & + \beta_{1134,j}x_{1,i}^2x_{3,i}x_{4,i} + \beta_{1135,j}x_{1,i}^2x_{3,i}x_{5,i} + \beta_{1145,j}x_{1,i}^2x_{4,i}x_{5,i} \\ & + \beta_{1234,j}x_{1,i}x_{2,i}x_{3,i}x_{4,i} + \beta_{1235,j}x_{1,i}x_{2,i}x_{3,i}x_{5,i} + \beta_{1245,j}x_{1,i}x_{2,i}x_{4,i}x_{5,i} \\ & + \beta_{2234,j}x_{2,i}^2x_{3,i}x_{4,i} + \beta_{2235,j}x_{2,i}^2x_{3,i}x_{5,i} + \beta_{2245,j}x_{2,i}^2x_{4,i}x_{5,i}. \end{aligned} \quad (14)$$

The example illustrates the importance of appropriately choosing the order of interactions in the multivariate Bernoulli distribution. As $\overline{K}$ increases the number of terms grows exponentially, requiring greater degrees of freedom in the dataset. In addition, the inclusion of binary terms that represent indicator variables capturing sub-populations in the data allows the researcher to examine the differences in effects from covariates across alternative sub-populations defined by combinations or interactions of these different indicator variables. Furthermore, the significance of higher-order terms can be tested for

using asymptotic techniques, such as the Likelihood-Ratio or Lagrange Multiplier tests, to assess if the functional form is statistically appropriate and supported by the observed data.

2.3 *Marginal effects*

The coefficients of the multinomial logistic regression model are difficult to interpret given the nonlinear nature of the model. This arises due to the fact that coefficients enter the probabilities (or conditional means) for all choices in the choice or outcome set [14]. Marginal effects are a useful substantive measure of the impact of a relative change in an explanatory variable on the probability of a particular choice or outcome. For the multinomial model specified above, the marginal effects for continuous explanatory variables would be calculated as:

$$\frac{\partial \mathbf{P}(Y_i=j|\mathbf{X}_i=\mathbf{x}_i)}{\partial \boldsymbol{x_k}} = \frac{\partial h_j(\boldsymbol{X}_i;\boldsymbol{\beta})}{\partial \boldsymbol{x_k}} = h_j(\boldsymbol{X}_i;\boldsymbol{\beta})\left[\frac{\partial \eta_j(\boldsymbol{X}_i;\beta_j)}{\partial \boldsymbol{x_k}} - \sum_{s\neq m}\frac{\partial \eta_s(\boldsymbol{X}_i;\beta_s)}{\partial \boldsymbol{x_k}} \cdot h_s(\boldsymbol{X}_i;\boldsymbol{\beta})\right] . (15)$$

For binary variables, the marginal effect is the change in probability when changing the value of the binary variable from '0' to '1', *ceteris paribus* [14]. These latter marginal effects are determined by taking discrete differences.

Another marginal effect of interest is the marginal effect of interactions between explanatory variables. In the present framework, these marginal effects (or interaction effects) let the modeler examine how the effect of one explanatory variable on the probability for making choice *j* depends on the magnitude of another explanatory variable. In nonlinear models, marginal effects of interaction terms are more difficult to interpret, as they are not equivalent to the value of the coefficient on a particular interaction term. Ai and Norton [1] derive the marginal effect for the interaction between two continuous

variables (and the case when at least one variable is a binary variable) for the binary logistic case. In the equation below, we extend the marginal effect derivation to the case of an interaction between two continuous variables (x_k and x_r) in multinomial logistic regression for choice alternative j :

$$\frac{\partial^2 h_j(\boldsymbol{X}_i;\boldsymbol{\beta})}{\partial \boldsymbol{x_k}\partial \boldsymbol{x_r}} = h_j(\boldsymbol{X}_i;\boldsymbol{\beta})\left[\frac{\partial^2 \eta_j(\boldsymbol{X}_i;\beta_j)}{\partial \boldsymbol{x_k}\partial \boldsymbol{x_r}} - \sum_{s\neq m}\left(\frac{\partial h_s(\boldsymbol{X}_i;\boldsymbol{\beta})}{\partial \boldsymbol{x_r}}\cdot\frac{\partial \eta_s(\boldsymbol{X}_i;\beta_s)}{\partial \boldsymbol{x_k}} + h_s(\boldsymbol{X}_i;\boldsymbol{\beta})\cdot \frac{\partial^2 \eta_s(\boldsymbol{X}_i;\beta_s)}{\partial \boldsymbol{x_k}\partial \boldsymbol{x_r}}\right)\right] + \frac{1}{h_j(\boldsymbol{X}_i;\boldsymbol{\beta})}\cdot\frac{\partial h_j(\boldsymbol{X}_i;\boldsymbol{\beta})}{\partial \boldsymbol{x_k}}\cdot\frac{\partial h_j(\boldsymbol{X}_i;\boldsymbol{\beta})}{\partial \boldsymbol{x_r}} \quad (16).$$

If one or both of the explanatory variables are binary, then discrete differences are taken instead. For example, if x_r is a binary variable and x_k remains as a continuous variable, then the marginal effect is the discrete change in the marginal effect for x_k using equation (15) from changing the value of x_r from '0' to '1', *ceteris paribus*.

Marginal effects can be calculated at the mean of the random variable (i.e. the marginal effect for the average person) or calculated for each respondent and then averaged across respondents (i.e. the average partial or mean marginal effect). The corresponding marginal effects from either method are not necessarily asymptotically equivalent and the choice of method carries non- trivial implications for interpretation [39]. The standard error for both types of marginal effects can be found using the delta method [1, 14] or simulation methods, such as the bootstrap or jackknife [9]. The average marginal effect is employed in the empirical application examining discharge disposition following a hospital admission of stroke.

2.4 *Data*

The Maryland Health Services Cost Review Commission (HSCRC) (website: http://www.hscrc.state.md.us/) maintains a secure and proprietary database of all discharges from non-Federal short-stay hospitals in the state of Maryland. The database includes information on the demographics, admission diagnoses, procedures, payer characteristics, and hospital characteristics pertaining to each hospitalization. Study inclusion criteria were as follows: 1) hospital admissions with a discharge diagnosis of stroke as identified by the International Classification of Diseases, Ninth Revision, Clinical Modification (ICD9CM) codes 431-434 and 436-438; 2) patients discharged alive; and 3) patient age at admission greater than or equal to 18. The ICD-9 CM codes used in this study are the same ones used by Fang and Alderman [11] in their study of national trends in stroke hospitalizations. Discharges were excluded for the following reasons: nature of admission is listed as delivery; in-hospital death; no information on patient disposition among those discharged alive; invalid or missing provider Medicare number; and invalid or missing patient medical record number. The dataset consists of 69,921 hospital admissions for stroke over the January 2000 to December 2005 time period. Categories of discharge disposition include home (N = 7730), home health care (N = 525), rehabilitation (N = 9997), nursing home (including intermediate care) (N = 7323), discharges against medical advice (N = 4755), and all other (N = 39,591).

To examine factors affecting the type of discharge, the explanatory variables included patient age (centered), as well as binary variables indicating a transfer admission (equal to '1' if yes), gender (equal to '1' if male), marital status (equal to '1' if married),

insured status (equal to '1' if not insured or self-pay), race (equal to '1' if patient reports race/ethnicity other than non-Hispanic White), and hemorrhagic stroke (equal to '1' if yes).

2. 5 *Empirical model application*

To illustrate the methodology and importance of ensuring statistical adequacy through proper model specification, an empirical application examining the predictors of post-stroke hospital discharge disposition among live discharges of adult patients was conducted. A multinomial model was specified following the PR approach with the dependent variable being post-stroke hospital discharge disposition as described earlier. The reference category was set as the "all other" category.

The index function of the model was specified by examining the product of the inverse conditional distribution of the continuous covariates times the inverse conditional multivariate Bernoulli distribution of the binary covariates (e.g. following equation (10)). The only continuous covariate in the dataset is patient age. The distribution of patient age was found to be mesokurtic, but slightly skewed to the right. The sample kurtosis for patient age was 3.01 and the sample skewness was -0.602. This distributional shape with the given range of patient age does not readily fit any of the distributional forms presented in Bergtold et al. [4].[3] Thus, as suggested by Kay and Little [20], the functional

[3] An inverse conditional distribution that does fit the shape of the data and explored by Bergtold et al. [4] was the Weibull distribution. The Weibull distribution has two parameters (scale and location). If the scale parameter is not consistent across *j*, then this distribution will not provide a tractable option for defining an operational statistical model as no clear mapping between the parameters of the inverse conditional distribution and the multinomial logistic regression model (e.g. $\beta_j = \beta(\boldsymbol{\alpha}, \boldsymbol{u})$) can be established, making estimation extremely difficult. For the given empirical problem, when the Weibull distribution was fit to patient age for the different hospital discharge categories, the scale parameter varied between the *j* categories (i.e. from 4 to 8). Thus, a more flexible approach was deemed the optimal modeling strategy to pursue.

specification of the index function becomes more difficult. Given the nature of the skewness of the distribution of patient age and its tractability, the flexible distribution following Day and Kerridge [8] (see equation (11)) was utilized to model the inverse conditional distribution of patient age conditional on the other binary explanatory and dependent variables.

The multivariate inverse conditional distribution of the binary explanatory variables conditional on the dependent variable was modeled as a multivariate Bernoulli distribution. Likelihood ratio tests were utilized to determine the order of interactions to include in the index function. Results from asymptotic likelihood ratio tests indicated that: interactions of the binary variable up to order two with patient age and patient age squared, as well as third order interactions of the binary variables should be included in the specification of the index function. Likelihood ratio tests were conducted to test the significance of including third order interactions of binary variables with patient age and patient age-squared, as well as fourth order interaction terms of the binary variables. All tests conducted had p-values greater than 0.10. Thus, these higher order interaction terms were not included in the index function.

The model specified using the Probabilistic Reduction approach consisted of 405 estimable parameters. Given that the index function is linear in the parameters, the model was estimated using a multinomial logistic regression procedure in MATLAB.

For comparison, a traditional specification of the multinomial logistic regression model was estimated. The predictor or log odds function was assumed to be linear in the variables with the addition of three interaction terms: (i) patient age x race; (ii) insured

status x race; and (iii) hemorrhagic stroke x race. It is of interest to note that these interactions are included in the formulation following the PR approach. The first and third interaction terms reflect the potential for variation across age groups in the racial disparity in discharge outcomes [12] and the potential of a differential impact on discharge outcomes associated with the increased risk of hemorrhagic stroke among African Americans relative to Caucasians [25, 33]. The second interaction term is included to account for any differential effect on discharge outcome due to lower rates of insurance coverage for non-Caucasians compared to Caucasians [15]. Model specification was tested using Likelihood Ratio tests to compare the two models, as well as test for any functional misspecification in the model derived using the PR approach.

Of interest is the impact of potential model misspecification due to incorrect functional form on substantive inferences from the model. While odds ratios are commonly estimated, marginal effects facilitate interpretation of the regression results because they are measured in the same units as the dependent variable. Marginal effects for both the individual covariates and the interactions: (i) patient age x race; (ii) insured status x race; and (iii) hemorrhagic stroke x race, were estimated. Individual marginal effects were estimated as the mean marginal effect across respondents following Greene [14]. Estimates of marginal effects for interactions were calculated using equation (16) as the mean across respondents. Standard errors for all estimates were estimated using a delete-*d* jackknife estimator with *d* equal to 10 percent of the data selected randomly without replacement over 5000 pseudo-random samples [9].

3. Results and Discussion

3. 1 *Model Fit*

Model fit statistics for both estimated models for the post-stroke hospital discharge data are provided in Table 1. Comparison of the traditional specification against the PR approach based on the AIC and R-squared values does not provide compelling evidence of a significant lack of model fit by the traditional specification when compared to the specification of the model using the PR approach; however if a model is misspecified then the fit statistics, such as AIC and R-squared can be misleading, as they presume proper specification of the model [24,31]. Furthermore, the significance of higher order interaction terms was tested for the model using the PR approach to help identify a properly specified model. These tests results were presented earlier and establish the statistical adequacy of the model specified using the PR approach over the more traditional specification. The specification highlighted the need to include polynomial terms and interactions to account for potential nonlinearities and additional heterogeneity (via interactions of the binary covariates) in the model specification.

Table 1: Model fit statistics for the multinomial logistic regression of post-stroke hospital discharge disposition.

Statistic	Traditional Specification	PR Specification
Log Likelihood	-86,140.67	-85,544.87
R^2_{VZ} [1]	0.165	0.183
Correct Prediction	57.4 %	57.5 %
Correct Prediction (as 1st or 2nd likely outcome)	72.6 %	73.0 %
AIC[2]	172391.33	171899.75

[1] Veall and Zimmerman [38]
[2] Burnham and Anderson [5]

The estimated models were compared using a likelihood ratio test (distributed $\chi^2(350)$) with the null hypothesis that the traditional specification was correct (Table 1). The value of the test statistic was 1192 with an associated p-value of 0.000, indicating a strong lack of evidence for the null hypothesis, indicating that the traditional specification was misspecified. From Table 1, the percent correctly predicted was higher based on the PR specification; however there were only slight differences between both approaches, suggesting that the PR specification did not offer a significant improvement over the traditional specification. When the goal of the analysis is to develop substantive inference, model prediction may not represent a sufficient criterion for selecting the appropriate specification and comparing the models based on model fit statistics may be misleading. Spanos [31] suggests that misspecified models may provide misleading or erroneous conclusions and inferences when using model fit statistics.

Closer examination of the models in terms of significant covariates reveals that higher order nonlinear terms in the PR specification of the multinomial logistic regression model were statistically significant. For example, model coefficients for a third order interaction term between patient age, gender and hemorrhagic stroke ($P = 0.044$), as well as a fourth order term between patient age squared, gender and marital status ($P = 0.010$) were both statistically significant. While direct interpretation of these coefficients is not straightforward, the omission of nonzero nonlinear terms could have a significant impact on the estimation of marginal effects or other substantive inferences. Furthermore,

significance of higher order terms in the model may indicate the presence of unmodeled or previously unknown heterogeneity in the data.

The PR approach provides a systematic approach for obtaining a statistically adequate model, thereby capturing any potential statistical information in the observed data not anticipated by the applied modeler. Domain knowledge (or theory) provides a strong starting point for model specification, by indicating what explanatory variables or covariates should enter the model and guidance for initial model specification (e.g. using random utility theory). This would have likely lead to a potential model specification as given by the traditional multinomial logistic model specification indicated here, as most applied applications assume predictor (or index) functions linear in the covariates [3]. The above discussion indicates that the traditional specification was misspecified for the problem examined here, which can lead to erroneous inferences from the model [33]. The PR approach builds from the traditional specification by allowing for a more flexible modeling approach that starts with domain knowledge, but allows the statistical information in the observed data to be properly captured, so as to provide reliable inferences. Another way of viewing this is that the PR approach allows domain knowledge to be flexible by helping to determine the functional form and shape of the relationship being modeled. In addition, the PR approach helps to provide a modeling approach that can allow substantive inferences that provide evidence in support of claims or hypotheses arising from the domain knowledge [31, 33]. The impact on substantive inferences is highlighted in the next section of the paper.

3.2 *Marginal effects and substantive inferences*

Marginal effects for the individual and interaction effects ((i) patient age x race; (ii) insured status x race; and (iii) hemorrhagic stroke x race for the traditional and PR model specifications are provided in Tables 2 and 3, respectively. There are significant

Table 2: Marginal effect estimates for the traditional multinomial regression model.

Variable/ Discharge Category	**Home**	**Home Health Care**	**Rehabilitation**	**Nursing/ Intermediate Care**	**AMA Discharge**
			Individual Effects		
Patient Age	-0.0086*	0.0014*	-0.0002*	0.0063*	-0.0002*
	(0.0001)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Transfer	-0.0746*	0.0010	-0.0069*	0.0488*	0.0060*
Admission	(0.0017)	(0.0009)	(0.0010)	(0.0013)	(0.0004)
Male Gender	0.0158*	-0.0101*	0.0090*	-0.0105*	0.0032*
	(0.0013)	(0.0007)	(0.0008)	(0.0009)	(0.0002)
Married	0.0952*	-0.0006	-0.0163*	-0.0569*	-0.0044*
	(0.0013)	(0.0007)	(0.0008)	(0.0010)	(0.0002)
Uninsured/	-0.0102	0.0082*	-0.0282*	0.0001	0.0183*
Self-Pay (US)	(0.0069)	(0.0042)	(0.0037)	(0.0067)	(0.0017)
Non-Caucasian, (NC)	-0.3269*	0.0317*	0.0273*	0.2254*	0.0012
	(0.0053)	(0.0044)	(0.0039)	(0.0085)	(0.0009)
Hemorrhagic Stroke	-0.2312*	-0.0051*	0.0889*	0.0601*	-0.0014*
	(0.0026)	(0.0014)	(0.0022)	(0.0022)	(0.0005)
			Interaction Effects		
Patient Age x NC	0.0031*	-0.0004*	-0.0005*	-0.0012*	-0.0000*
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0000)
US x NC	0.1720*	-0.0138*	-0.0573*	-0.0678*	-0.0003
	(0.0086)	(0.0051)	(0.0040)	(0.0075)	(0.0020)
Hemorrhagic Stroke X NC	0.0793*	-0.0071*	-0.0400*	-0.0035	-0.0036*
	(0.0043)	(0.0023)	(0.0033)	(0.0036)	(0.0006)

Notes: AMA refers to 'against medical advice'. Individual marginal effects were estimated as the marginal effect averaged across respondents following Greene [14]. Estimates of marginal effects for interactions were calculated following Ai and Norton [1] and calculated as the marginal effect averaged across respondents. Standard errors for all estimates are in parentheses and were estimated using a delete-*d* jackknife estimator with *d* equal to 10 percent of the data selected randomly without replacement over 5000 pseudo-random samples [9]. '*' indicates statistical significance at the 10 percent level or above.

differences in the marginal effects estimates both in sign and magnitude between the two models. The signs for all the individual marginal effects for patient age and male gender

Table 3: Marginal effect estimates for multinomial regression model following the probabilistic reduction approach.

Variable/ Discharge Category	**Home**	**Home Health Care**	**Rehabilitation**	**Nursing/ Intermediate Care**	**AMA Discharge**
			Individual Effects		
Patient Age	0.0015*	-0.0002*	0.0069*	-0.0002*	0.0019*
	(3.20e-5)	(3.27e-5)	(4.75e-5)	(1.05e-5)	(3.92e-5)
Transfer	0.0021*	-0.0076*	0.0501*	0.0057*	0.0236*
Admission	(0.0010)	(0.0011)	(0.0013)	(0.0004)	(0.0012)
Male Gender	-0.0094*	0.0081*	-0.0122*	0.0034*	-0.0062*
	(0.0007)	(0.0008)	(0.0010)	(0.0002)	(0.0008)
Married	-0.0006	-0.0151*	-0.0485*	-0.0038*	-0.0104*
	(0.0007)	(0.0009)	(0.0010)	(0.0002)	(0.0009)
Uninsured/	0.0089*	-0.0531*	-0.0254*	0.0098*	0.0008
Self-Pay (US)	(0.0055)	(0.0036)	(0.0083)	(0.0014)	(0.0067)
Non-Caucasian, (NC)	0.0100*	0.0276*	0.0319*	0.0016*	0.0057*
	(0.0008)	(0.0009)	(0.0010)	(0.0003)	(0.0009)
Hemorrhagic	-0.0038*	0.0707*	0.0624*	-0.0025*	0.0806*
Stroke	(0.0012)	(0.0018)	(0.0018)	(0.0003)	(0.0018)
			Interaction Effects		
Patient Age x	-0.0004*	-0.0033*	0.0033*	-0.0003*	-0.0013*
NC	(7.53e-5)	(90.1e-5)	(0.0001)	(2.25e-5)	(0.0001)
US x NC	-0.0064	-0.0493*	-0.0612*	-0.0046*	-0.0291*
	(0.0094)	(0.0059)	(0.0126)	(0.0023)	(0.0105)
Hemorrhagic	-0.0142*	-0.0488*	-0.0107*	-0.0058	-0.0249*
Stroke X NC	(0.0074)	(0.0058)	(0.0051)	(0.0067)	(0.0043)

Notes: AMA refers to 'against medical advice'. Individual marginal effects were estimated as the marginal effect averaged across respondents following Greene [14]. Estimates of marginal effects for interactions were calculated following Ai and Norton [1] and calculated as the marginal effect averaged across respondents. Standard errors for all estimates are in parentheses and were estimated using a delete-*d* jackknife estimator with *d* equal to 10 percent of the data selected randomly without replacement over 5000 pseudo-random samples [9]. '*' indicates statistical significance at the 10 percent level or above.

are opposite in sign between the two models. For example, for each year older a patient is, the patient has a 0.86 percent lower likelihood of going home and 0.63 percent higher likelihood of receiving nursing/intermediate care under the traditional specification of the model. For the PR specification of the model, a patient has a 0.15 percent higher likelihood of going home and a 0.02 percent lower likelihood of receiving nursing/intermediate care. The signs are different for a number of the marginal effects from being a transfer admission and having a hemorrhagic stroke, as well. Under the traditional specification, a patient is 0.82 percent more likely to receive home health care if they are uninsured or self-pay; and 0.14 percent less likely to be discharged against medical advice if they have had a hemorrhagic stroke. Under the PR specification, a patient is 5.31 percent less likely to receive home health care if they are uninsured or self-pay; and 8.06 percent more likely to be discharged against medical advice if they have had a hemorrhagic stroke. The stark differences in the signs and some of the magnitudes of the marginal effects emphasize the significance in getting the model specification correct. Model misspecification may likely affect the sign or "direction" of an effect.

Of particular interest is the impact of ethnicity on post-stroke hospital discharge disposition, which is dramatically different between the two models for some categories. Following the traditional specification of the model, non-Caucasians are 32.7 percent less likely to be sent home following a stroke and 22.5 percent more likely to require additional

nursing or intermediate care (e.g. nursing home). Following the PR specification, non-Caucasians are only 1.0 percent more likely to be sent home and 0.16 percent more likely to be discharged to additional nursing or intermediate care. Furthermore, non-Caucasians are 0.57 percent more likely to be discharged against medical advice following a stroke under the PR specification, while the same marginal effect under the traditional specification is not statistically significant (from zero). The differences in magnitudes of these marginal effects across the two models are quite significant. While race/ethnicity plays a significant role in post-stroke discharge disposition in both model specifications, model misspecification resulted in significantly over-estimating the impact of this factor. Thus, model misspecification may significantly bias substantive inferences, resulting in over- or under- inflated estimates for the marginal effect of different individual factors. Including interaction terms in the traditional formulation to account for potential heterogeneity by race may not adequately capture the effects of interest or relevance.

The results for the interaction effect of insured status by race differ between the two models. The traditional specification indicates that uninsured non-Caucasians patients are 17 percent more likely to be sent home; 1.3 percent less likely to receive home health care; 6.8 percent less likely to receive nursing/intermediate care; 0.03 percent more likely to be discharged AMA (not statistically significant). In contrast, under the PR specification, uninsured non-Caucasian patients 0.64 less likely to be sent home (not statistically significant); 4.9 percent less likely to receive home health care; 0.46 percent less likely to receive nursing/intermediate care; and 2.9 percent less likely to be discharged AMA.

As with the previous interaction marginal effect, the two models differ with respect to the interaction of hemorrhagic stroke by race and patient age by race. Not including interactions that may capture unmodeled (and unexpected) heterogeneity or relationships in the observed data may result in model misspecifications that lead to biased estimates and erroneous inferences, that could wrongly influence decision-making and policy.

The PR approach to model specification results in the estimation of a model that is more complicated than one would traditionally estimate in practice. Thus, the modeler needs to consider the degrees of freedom available for estimation of the model and may need to limit the order of covariate terms used in the model. Traditional modeling approaches lead to models composed primarily of main effects (or terms linear in the variables). This assumption significantly reduces the number of parameters that need to be estimated, allowing for an easier interpretation of parameter estimates. However, what does the modeler do if the model is not properly specified? Several manipulations of the explanatory variables that can be considered include: nonlinear transformations of continuous covariates; inclusion of interaction terms between variables of interest; choosing a different link function; or assuming an alternative stochastic error distribution. In practice, how would the researcher systematically enrich a misspecified model? Neither the generalized linear model approach nor the latent variable approach provides a definite answer to this question that is implicit in the majority of applied studies. The PR approach provides an answer to this question and thus guidance through the maze of options by forcing the explicit consideration of nonlinearities, heterogeneity across sub-populations and dependence across included covariates in the *initial* model specification (via the use of

the inverse conditional distribution). Reductions from this initial specification (that are not related to sample size or multicollinearity concerns) can be systematically tested using likelihood ratio tests. The empirical example here provides evidence of the impact on substantive inferences that can occur when the relationship (both substantive and statistical) between the dependent and explanatory variables is misspecified. It should be emphasized that the use of model fit statistics requires proper model specification and remains unreliable in the presence of model misspecification.

4. Conclusion

Traditional approaches to specifying multinomial logistic regression models may ignore some of the probabilistic information in the observed data, potentially resulting in biased and inconsistent estimates, as well as unreliable substantive inferences. The probabilistic reduction approach to model specification provides an alternative that puts emphasis on using the inverse conditional distribution for model specification. Using the PR approach, the paper provides a systematic method for specifying multinomial logistic regression models using a flexible inverse conditional distribution that can account for various continuous and discrete combinations of explanatory variables applicable to a myriad of empirical settings.

An objective of this paper was to compare the PR approach to a more traditional model specification. The results from this paper suggest that proper model specification can strongly affect substantive inference. Furthermore, model fit statistics may not be reliable if the model is misspecified. Empirical results for the post-stroke hospital discharge analysis show that even if diagnostic tests do not show strong differences across

different models, nested tests such as the likelihood ratio test can be instructive for examining model specification to help arrive at a properly specified model, which in turn will lead to reliable substantive inference. Model misspecification may result in significant bias in substantive inferences (e.g. marginal effects) for covariates and interactions between covariates of interest. We show that the PR approach to the specification of multinomial logistic regression models leads to an *initial* model specification that differs from the initial model that follows from the traditional modeling approaches commonly specified using domain knowledge only at times. We show that these differences between the PR and traditional approach in terms of the initial model specification may be relevant for the results from model diagnostic tests and for inference on covariates of interest. Under the PR approach, the progress from the initial specification to the final model proceeds along a path that is transparent, reproducible, and based on internally consistent assumptions.

The PR approach provides a systematic approach for model specification. It builds on the use of domain knowledge for initial model specification by taking account of the relevant statistical information in the observed data used to estimate the model. The PR approach provides a methodology that can help in discovery of the shape of interactions between variables in a model (e.g. potential nonlinear relationships); the capturing of unexpected heterogeneity; and a modeling approach that provides reliable statistical and substantive inferences. It allows for domain knowledge to be expanded and tested. That is, domain knowledge (or theory) will indicate a particular relationship or hypothesis to be tested. The PR approach provides a statistical methodology that begins with, but is not

unduly constrained by, domain knowledge, so as to have the ability to provide reliable substantive inferences (which depend on statistical adequacy) to expand domain knowledge (or theory) by providing more rigorous evidence.

5. References

1. C. Ai, and E.C. Norton. *Interaction terms in logit and probit models*, Econom. Lett. 80 (2003), pp. 123 – 129.
2. J.A. Anderson, *Separate sample logistic discrimination*, Biometrika 59 (1972), pp. 19 – 35.
3. B.C. Arnold, E. Castillo, and M. Sarabia, *Conditional Specification of Statistical Models*, Springer Verlag: New York, NY, 1999.
4. J.S. Bergtold, A. Spanos, and E. Onukwugha, *Bernoulli regression models: revisiting the specification of statistical models with binary dependent variables*, J. Choice Modelling 3 (2010), pp. 1 - 28.
5. Burnham, K.P. and D.R. Anderson. *Model Selection and Multimodel Inference: A Practial Information-Theoretic Approach*, 2 ed. Springer Verlag, New York, 2002.
6. K. Byth and G.J. McLachlan, *Logistic regression compared to normal discrimination from non-normal populations*, Aust. and New Zealand J. Stat. 22 (1980), pp.188 – 196.
7. J.N. Cromwell, N.T. McCall, J. Burton, and C. Urato, *Race/ethnic disparities in utilization of lifesaving technologies by Medicare ischemic heart disease beneficiaries*, Medical Care 43 (2005), pp. 330 – 337.
8. N.E. Day and D.F. Kerridge, *A general maximum likelihood discriminant*, Biometrics 23 (1967), pp.313 – 323.
9. B. Efron and R.J. Tibshirani, *An Introduction to the Bootstrap*, Monograms on Statistics and Applied Probability 57, Chapman and Hall, New York, 1993.
10. L. Fahrmeir and G. Tutz, *Multivariate Statistical Modeling Based on Generalized Linear Models*, Springer-Verlag, New York, 2001.
11. F. Fang and M.H. Alderman, *Trend of stroke hospitalization, United States, 1988-1997*, Stroke 32 (2001), pp. 2221-2226.
12. W. Feng, P.J. Nietert, R.J. *Adams Influence of age on racial disparities in stroke admission rates, hospital charges, and outcomes in South Carolina*, Stroke 40 (2009), pp. 3096 – 3101.
13. P.S. Franks, S. Meldrum, and K. Fiscella, *Discharges against medical advice: are race/ethnicity predictors?* J. Gen. Internal Medicine 21 (2006), pp. 955 – 960.
14. W.H. Greene *Econometric Analysis*, 5th ed., Prentice Hall, Upper Saddle River, NJ, 2003.
15. L.J. Hargraves and J. Hadley, *The contribution of insurance coverage and community resources to reducing racial/ethnic disparities in access to care*, Health Services Research 38 (2003), pp. 809 – 829.

16. H.M. Harris, P. Golden-Larsen, K. Chantala, and R. Udry, *Longitudinal trends in race/ethnic disparities in leading health indicators from adolescence to young adulthood*, Arch. Pediatric Adolescence Medicine 160 (2006), pp. 74 – 81.
17. S.A. Ibrahim, C.K. Kwoh, and E. Krishnan, *Factors associated with patients who leave acute-care hospitals against medical advice*, Amer. J. Public Health 97 (2007), pp. 2204 – 2008.
18. A.K. Jha, D.O. Staiger, F.L. Lucas, and A. Chandra, *Do race specific models explain disparities in treatments after acute myocardial infarction?* Amer. Heart J. 153 (2007), pp. 785 – 791.
19. M. Jimenez, T. Dietrich, M.C. Shih, Y. Li, and K.J. Joshipura, *Racial/ethnic variations in associations between socioeconomic factors and tooth loss*, Community of Dental and Oral Epidemiology 37 (2009), pp. 267 – 275.
20. R. Kay and S. Little, *Transformations of the explanatory variables in the logistic regression model for binary data*, Biometrika 74 (1987), pp. 495 – 501.
21. K.Y. Liang, S.L. Zeger, and B. Qaqish, *Multivariate regression analyses for categorical data*, J. Roy. Stat. Soc. Ser. B Stat. Methodol. 54 (1992), pp. 3 – 40.
22. G.S. Maddala, *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge University Press, Cambridge, 1983.
23. D. McFadden, *A comment on discriminant analysis "versus" logit analysis*, Ann. Econom. Soc. Measure. 5/4 (1976): 511 – 523.
24. A.M. McGuirk, P. Driscoll, *The hot air in R^2 and consistent measures of explained variations*, American Journal of Agricultural Economics 77 (1995), pp. 319 – 328.
25. A.I. Qureshi, W.H. Giles, and J.B. Croft, *Racial differences in the incidence of intracerebral hemorrhage*, Neurology 52 (1999), pp. 1617.
26. E.C. Onukwugha, F.T. Shaya, E. Saunders, and M.R. Weir, *Ethnic disparities, hospital quality, and discharges against medical advice among patients with cardiovascular disease*, Ethnicity and Disease 19 (2009), pp. 172 – 178.
27. K.P. Pages, J.E. Russo, D.K. Wingerson, R.K. Ries, P.E. Roy-Byrne, and D.S. Cowley, *Predictors and outcomes of discharge against medical advice from the psychiatric units of a general hospital*, Psychiatric Services 49 (1998), pp. 1187 – 1192.
28. R.L. Sacco, B. Boden-Albala, G. Abel, I.F. Lin, M. Elkind, W.A. Hauser, M.C. Paik, and S. Shea, *Race-ethnic disparities in the impact of stroke risk factors: the northern Manhattan stroke study*, Stroke 32 (2001), pp. 1725 – 1731.
29. R.A. Scribner, K.P. Theall, N.R. Simonsen, K.E. Mason, and Y. Qingzhao, *Misspecification of the effect of race in fixed effects models of health inequalities*, Soc. Science Medicine 69 (2009), pp. 1584 – 1591.
30. L. Scrucca and S. Weisberg, *A simulation study to investigate the behavior of the log-density ratio under normality*, Comm. Statist. Simulation Comput. 33 (2004), pp. 159 – 178.
31. A. Spanos, *Akaike-type criteria and the reliability of inference*: *model selection versus statistical model specification*, J. Econometrics. 158 (2010), pp. 204 - 220
32. A. Spanos, *Statistical Foundations of Econometric Modeling*, Cambridge University Press, Cambridge, 1986.

33. A. Spanos, *Probability Theory and Statistical Inference: Econometric Modeling with Observational Data*, Cambridge University Press, Cambridge, 1999.
34. J.D. Sturgeon, A.R. Folsom, W.T. Longstreth , Jr., E. Shahar, W.D. Rosamond, and M. Cushman, *Risk factors for intracerebral hemorrhage in a pooled perspective study*, Stroke 38 (2007), pp. 2718 – 2725.
35. S. Subramanian, D. Acevedo-Garcia, and T. Osypuk, *Racial residential segregation and geographic heterogeneity in black/white disparity in poor self-rated health in the U.S.: a multilevel statistical analysis*, Soc. Science Medicine 60 (2005), pp. 1667 – 1679.
36. S.V. Subramanian, J.T. Chen, D.H. Rehkoph, P.D. Waterman, N. Krieger, *Racial disparities in context: a multilevel analysis of neighborhood variations in poverty and excess mortality among black populations in Massachusetts*, Amer. J. Public Health 95 (2005), pp. 260 – 265.
37. K.E. Train, *Discrete Choice Models with Simulation*, Cambridge University Press, Cambridge, 2003.
38. M. Veall and K. Zimmerman, *Pseudo-R^2 in the ordinal probit model*, J. Mathe. Sociol. 16 (1992), pp. 333 – 342.
39. J.A. Verlinda, *A comparison of two common approaches for estimating marginal effects in binary choice models*, Appl. Econom. Lett. 13 (2006), pp. 77–80