Marginal probabilities and point estimation for conditionally specified logistic regression

Curtis Miller*

February 4, 2015

Abstract

Keywords: Multiple binary responses; Conditionally specified logistic regression; Marginal probabilities

1 Introduction

In some situations we may need to model multiple binary outcomes:

$$\mathbf{Y} \sim \mathbf{X} \tag{1}$$

$\mathbf{Y}$ is a vector of p binary variables. $\mathbf{X}$ is a vector of n variables. Each of $X_1, \ldots, X_n$ is a potential predictor of $Y_j,\ 1 \leq j \leq p$. Problems described by Equation (1) tend to arise in medicine and public health. For examples, see Section 5, or those in [Joe and Liu (1996)] or [O'Brien and Dunson (2004)].

Conditionally specified logistic regression (to be referred to as CSLR) is a method for fitting the responses of Equation (1). CSLR was introduced in [Joe and Liu (1996)]. The equations that define a CSLR model are as follows. Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip})$ be the vector of responses and $\mathbf{X}_i = (X_{i1}, \ldots, X_{in})$ the vector of covariates for observation i. Then

$$\operatorname{logit}\left(P(Y_{ij} = 1 | \mathbf{X}_i, \{Y_{ik}, k \neq j\})\right) = \mu_j + \sum_{\ell=1}^{n} \alpha_{j\ell} X_{i\ell} + \sum_{k \neq j} \gamma_{jk} Y_{ik},\ 1 \leq j \leq p. \tag{2}$$

Consideration of Equation (2) leads to two questions:

- Equation(s) (2) implies that we must know the values of $\{Y_j :\ j \neq \ell\}$ to estimate the probability that $Y_\ell = 1$. This is not very useful, unless we have some reason for assuming values for all Y_j's except one. Often it is desirable to estimate $P(Y_\ell = 1)$, without conditioning on other responses.

*University of New Mexico, College of Pharmacy. email: cmille02@unm.edu

If doing prediction or interpolation, no values for any of the Y_j's are available. So in general the problem is to estimate $P(Y_\ell = 1)$ using only the covariate values $\mathbf{x}$.

- Equation (2) includes second order interaction terms γ_{jk}. However, there are p responses. Could we modify the formula by including terms for higher order interactions?

Item 1 is asking for first order marginal probabilities. These are the probabilities $\{P(Y_{ij} = s),\, s \in \{0,1\}|\mathbf{x}_i\}$. For any subset $S \subset \{1, \ldots, p\}$, the marginal probabilities for $\{Y_{ij}\}_{j \in S}$ are of form $P(\{Y_{ij} = s_j,\, j \in S\})$, where $\mathbf{s} \in \{0,1\}^{|S|}$. Often, in applied problems, we need to know the marginal probability that response $Y_j = 1$ is positive. We may also wish to estimate the probability that Y_j is positive when one or more covariates are set at some hypothetical level. Such counterfactual marginal probabilities are used to derive average treatment effects, used in econometrics and epidemiology (See [Imbens and Woolridge (2009]).

However, marginal probabilities have another interpretation. The Y_j's are binary, so $P(Y_\ell = 1|\mathbf{x})$ is a fitted value for Y_ℓ. If the coefficients in Equation (2) are derived from a data set, then in calculating marginal probabilities we are modeling the data for each response Y_ℓ. We must ask: How well has the data been modeled? Are there better methods for modeling data for multiple binary responses?

In Section 2 we will show a general method for deriving marginal probabilities $P(Y_\ell = 1|\mathbf{x})$.

In Section 3, models with third order interactions are shown to exist. This may seem a small innovation, but will turn out to be significant.

in Section 4, alternative methods for fitting multiple binary random variables are introduced and briefly described.

In Section 5, a data set is given, and all modeling methods are applied. The performances of the methods are assessed and compared.

The unifying concept is this: The ability to derive marginal probabilities (and model individual responses) is inseparable from assessment of goodness of modeling. CSLR can be made to yield fitted values $\{\hat{y}_{ij}\}$, but it must be seen how much these fitted values deviate from data values $\{y_{ij}\}$.

2 Marginal distributions and probabilities for CSLR

Assume that $Y_1, \ldots, Y_p$ is a set of p binary random variables, each taking values in $\{0,1\}$, and that $\mathbf{x} = (x_1, \ldots, x_n)$ is a vector of values for the covariates $X_1, \ldots, X_n$.

Notation. For $S \subset \{1, \ldots, p\}$, $Ev[S|]$ will denote the event $\{Y_\ell = 1 :\ \ell \in S\}$. Also, we will denote $\{1, \ldots, p\}$ by S_p.

Notice that $Ev[S]$ can be written as a disjoint union:

$$Ev[S] = \cup_{S'}\{Event\{Y_\ell = 1,\, \ell \in S',\, Y_\ell = 0,\, \ell \notin S'\} :\ S \subset S' \subset S_p\}. \tag{3}$$

Also, $Ev[\emptyset] = \{0,1\}^p$, so that $P(Ev[\emptyset]) = 1$.

We need one more preliminary result before presenting our main lemma. Suppose $S_1 \subset S_p$, $|S_1| < p$, and $a \in S^C = S_p/S_1$. Then

$$\begin{aligned} P(Y_\ell = 1,\, \ell \in S_1,\, Y_a = 0|\mathbf{x}) =& P(Y_\ell = 1,\, \ell \in S_1|\mathbf{x}) - P(Y_\ell = 1,\, \ell \in S_1 \cup \{a\}|\mathbf{x}) \\ =& P(Ev[S_1]|\mathbf{x}) - P(Ev[S_1 \cup \{a\}]|\mathbf{x}). \end{aligned}$$

If $\{a, b\} \subset S_p/S_1$, then

$$\begin{aligned} & P(Y_\ell = 1,\, \ell = inS_1,\, Y_a = Y_b = 0|\mathbf{x}) \\ =& P(Y_\ell = 1,\, \ell \in S_1|\mathbf{x}) - P(Y_\ell = 1,\, \ell \in S_1 \cup \{a\}|\mathbf{x}) \\ & - P(Y_\ell = 1,\, \ell \in S_1 \cup \{b\}|\mathbf{x}) + P(Y_\ell = 1,\, \ell \in S_1 \cup \{a,b\}|\mathbf{x}) \\ =& P(Ev[S_1]|\mathbf{x}) - P(Ev[S_1 \cup \{a\}]|\mathbf{x}) - P(Ev[S_1 \cup \{b\}]|\mathbf{x}) + P(Ev[S_1 \cup \{a,b\}]|\mathbf{x}). \end{aligned}$$

And in general, for $S_1, S_2 \subset S_p$. $S_1 \cap S_2 = \emptyset$,

$$\begin{aligned} & P(Y_\ell = 1,\, \ell \in S_1,\, Y_\ell = 0,\, \ell \in S_2|\mathbf{x}) \\ =& \sum_{S"} \{(-1)^{|S"|-|S_1|} P(Ev[S"]|\mathbf{x}) :\ S_1 \subset S" \subset S_1 \cup S_2\}. \end{aligned} \tag{4}$$

Lemma 1. *Let* $\mathbf{x}$ *be a vector of values of* $X_1, \ldots, X_n$*. Then the vector of probabilities* $\{P(Ev[S]|\mathbf{x}) :\ S \subset S_p,\, |S| < p\}$ *is preserved by an nontrivial affine transformation with coefficients determined by* $\mathbf{x}$ *and by the coefficients of the defining conditional equations (2).*

Proof. First, for any nonempty $S \subset \{1, \ldots, p\}$, choose one element $r(S) \in S$. The vector $\mathbf{r}$ is defined by these arbitrary choices.

Let $S \subset S_p$ and $\mathbf{x}$ be a vector of values of $X_1, \ldots, X_n$. By Equations (3) and (4),

$$\begin{aligned} & P(Ev[S]|\mathbf{x}) \\ =& P(Y_\ell = 1,\, \ell \in S|\mathbf{x}) \qquad (a) \\ =& \sum_{S'} \{P(Y_\ell = 1,\, \ell \in S',\, Y_\ell = 0, \ell \notin S'|\mathbf{x}) :\ S \subset S' \subset S_p\} \qquad (b) \\ =& \sum_{S'} \big[P(Y_{r(S')} = 1|Y_\ell = 1, \ell \in S'/\{r(S')\},\, Y_\ell = 0,\, \ell \notin S') \qquad (c) \\ & * P(Y_\ell = 1, \ell \in S'/\{r(S')\},\, Y_\ell = 0, \ell \notin S'|\mathbf{x})\big] \\ =& \sum_{S'} \big[P(Y_{r(S')} = 1|Y_\ell = 1, \ell \in S'/\{r(S')\},\, Y_\ell = 0,\, \ell \notin S') \qquad (d) \\ & * \big(\sum_{S''} \{(-1)^{|S''|-(|S'|-1)} P(Ev[S'']|\mathbf{x}) :\ S'/\{r(S')\} \subset S'' \subset S_p/\{r(S')\}\} \\ =& \sum_{S'' \subset S_p} q_{S,S''} P(Ev[S'']|\mathbf{x}) \end{aligned} \tag{5}$$

where

$$\begin{aligned}&q_{S,S''}\\ =&\sum_{S'}(-1)^{|S''|-|S'|+1}P(Y_{r(S')}=1|Y_\ell=1,\,\ell\in S'/\{r(S')\},\,Y_\ell=0,\ell\notin S',\mathbf{x})\end{aligned} \tag{6}$$

and the sum is over all subsets $S'\subset S_p$ such that (i) $S\subset S'$; (ii) $S'/\{r(S')\}\subset S''$; (iii) $r(S')\notin S''$. If no S' satisfy (i)-(iii), then $q_{S,S''}=0$.

In the Equation set (5), we used Equation (3) to go from (a) to (b) and Equation (4) to go from (c) to (d). The conditional probabilities on the right side of Equation (6) are determined by the coefficients in Equation (2).

Let S_p^* denote $S_p/\{r(S_p)\}$. By condition (iii), S'' cannot be S_p. Equation (5) holds if $S=S_p$, but conditions (i)-(iii) imply that $q_{S_p,S''}=0$ unless $S''=S_p^*$. In that case,

$$q_{S_p,S_p^*}=P(Y_{r(S_p)}=1|Y_j=1,j\neq r(S_p),\mathbf{x})$$

and Equation (5) simplifies to

$$\begin{aligned}&P(Y_j=1,1\le j\le p|\mathbf{x})\\ =&P(Ev[S_p]|\mathbf{x})\\ =&q_{S_p,S_p^*}P(Ev[S_p^*]|\mathbf{x})\\ =&P(Y_{r(S_p)}=1|Y_j=1,j\neq r(S_p),\mathbf{x})\cdot P(Y_j=1,j\neq r(S_p)|\mathbf{x}).\end{aligned}$$

This is given by the definition of conditional probability, and adds no new information. So we can reduce our system of equations of form Equation (5) by omitting those for which $|S|=p$ or $|S''|=p$.

If $S=\emptyset$, we have the trival equation $P(Ev[\emptyset]|\mathbf{x})=1$. S'' can be empty if and only if $|S|=1$, that is, $S=\{\ell\}$, some $\ell\in\{1,\dots,p\}$. Then S' must be $\{\ell\}$, and Equation (6) simplifies to

$$q_{\{\ell\},\emptyset}=P(Y_\ell=1|Y_j=0,j\neq\ell,\,\mathbf{x}). \tag{7}$$

This is the coefficient of $P(Ev[\emptyset])=1$. So in Equation (5), $P(Ev[S]|\mathbf{x})$ has a constant term if and only if $|S|=1$.

So by Equations (5), (6), and (7), the set $\{P(Ev[S]|\mathbf{x}):\ S\subset S_p,\,0<|S|<p\}$ satisfies the system

$$P(Ev[S]|\mathbf{x})=a_S+\sum\{q_{S,S''}P(Ev[S'']|\mathbf{x}):\ S''\subset S_p,\,0<|S''|<p\} \tag{8}$$

where

$$a_S=\begin{cases}P(Y_\ell=1|Y_j=0,j\neq\ell,\,\mathbf{x})\text{ if }S=\{\ell\}\\ 0\text{ otherwise}\end{cases}. \tag{9}$$

The Lemma holds with $\{a_S:\ S\subset S_p,0<|S|<p\}$ and $\{q_{S,S''}:\ S,S''\subset S_p,0<|S|,|S''|<p\}$ as the coefficients of the affine transformation. □

Notice that the proof used a vector $\mathbf{r}$, indexed by nonempty subsets of $\{1,\ldots,p\}$. $\mathbf{r}$ is chosen arbitrarily, so there is not a single canonical formula to calculate $\{q_{S,S''}\}$. However, $\mathbf{r}$ enters into Equation (5) at step (d). By the definition of conditional probability, the expression on line (c) equals that on line (d), no matter the choice of $\mathbf{r}$.

Let $U_p := \{S \subset S_p : \ 0 < |S| < p\}$. and let $\mathbf{V}$ denote the vector $\{P(Ev[S]) : S \in U_p\}$. Then Equation (8) can be expressed $\mathbf{V} = \mathbf{A} + \mathbf{QV}$, or

$$(\mathbf{I} - \mathbf{Q})\mathbf{V} = \mathbf{A}. \tag{10}$$

Choose a maximal subset $T \subset S_p$ such that rows of $\mathbf{I} - \mathbf{Q}$ indexed by T are linearly independent. For each $j \in T^C$, row j of $\mathbf{I}-\mathbf{Q}$ is a linear combination of rows indexed by T. That is, $P(Ev[S_j]|\mathbf{x})$ is determined by a linear combination of $\{P(Ev[S]|\mathbf{x}) : \ S \in T\}$ (plus a constant, if $|S_j| = 1$). So we have a new system

$$P(Ev[S]|\mathbf{x}) = a^*_S + \sum_{S"\in T} q^*_{S,S"} P(Ev[S]|\mathbf{x}), \ \forall S \in T$$

or $\mathbf{V}^* = \mathbf{A}^* + \mathbf{Q}^*\mathbf{V}^*$, so that

$$(\mathbf{I} - \mathbf{Q}^*)\mathbf{V}^* = \mathbf{A}^* \tag{11}$$

where $(\mathbf{I} - \mathbf{Q}^*)$ is nonsingular. If $\mathbf{A}^*$ is nonzero, there is a unique solution $\mathbf{V}^*$ of Equation (11), which implies a unique solution $\mathbf{V}$ of Equation (10). It is not clear as of now exactly when there is a unique solution of Equation (10), but so far there has been no difficulty in deriving the $\{P(Ev[S]|\mathbf{x})\}$.

Example. We show how the Lemma is applied for $p = 2$.

$$\begin{aligned}
&P(Y_1 = 1|\mathbf{x}) \\
=&P(Y_1 = 1, Y_2 = 1|\mathbf{x}) + P(Y_1 = 1, Y_2 = 0|\mathbf{x}) \\
=&P(Y_1 = 1|Y_2 = 1, \mathbf{x})P(Y_2 = 1|\mathbf{x}) + P(Y_1 = 1|Y_2 = 0, \mathbf{x})P(Y_2 = 0|\mathbf{x}) \\
=&P(Y_1 = 1|Y_2 = 1, \mathbf{x})P(Y_2 = 1|\mathbf{x}) + P(Y_1 = 1|Y_2 = 0, \mathbf{x})\left(1 - P(Y_2 = 1|\mathbf{x})\right) \\
=&P(Y_1 = 1|Y_2 = 0, \mathbf{x}) + \left(P(Y_1 = 1|Y_2 = 1, \mathbf{x}) - P(Y_1 = 1|Y_2 = 0, \mathbf{x})\right) P(Y_2 = 1|\mathbf{x}).
\end{aligned}$$

Similarly,

$$\begin{aligned}
&P(Y_2 = 1|\mathbf{x}) \\
=&P(Y_2 = 1|Y_1 = 0, \mathbf{x}) + \left(P(Y_2 = 1|Y_1 = 1, \mathbf{x}) - P(Y_2 = 1|Y_1 = 0, \mathbf{x})\right) P(Y_1 = 1|\mathbf{x})
\end{aligned}$$

so that

$$\begin{pmatrix} P(Y_1 = 1|\mathbf{x}) \\ P(Y_1 = 2|\mathbf{x}) \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}^{-1} \begin{pmatrix} P(Y_1 = 1|Y_2 = 0, \mathbf{x}) \\ P(Y_2 = 1|Y_1 = 0, \mathbf{x}) \end{pmatrix}$$

where

$$\begin{aligned}
m_{11} = m_{22} =&1 \\
m_{12} =&P(Y_1 = 1|Y_2 = 0, \mathbf{x}) - P(Y_1 = 1|Y_2 = 1, \mathbf{x}) \\
m_{21} =&P(Y_2 = 1|Y_1 = 0, \mathbf{x}) - P(Y_2 = 1|Y_1 = 1, \mathbf{x}).
\end{aligned}$$

2.1 Modeling the marginal probabilities

Choose $\ell \in S_p$. Let $\mathbf{x}_i$ the vector of covariates values at observation i. The marginal probability $P(Y_\ell = 1|\mathbf{x}_i)$ defines the marginal distribution of the binary variable Y_ℓ, given covariates $\mathbf{x}_i$. The $\{P(Y_\ell|\mathbf{x}_i) : \ 1 \leq i \leq N\}$ are in $(0, 1)$. We can model the marginal distributions by applying some suitable transformation $g : (0, 1) \to \mathbb{R}$ and then fitting a linear regression model to the $\{g\left(P(Y_\ell = 1|\mathbf{x}_i)\right)\}$. Each marginal probability is a weighted mean of conditional probabilities

$$\begin{aligned} &P(Y_\ell = 1|\mathbf{x}) \\ =&\sum\{P(Y_\ell = 1|Y_j = y_j,\, j \neq \ell, \mathbf{x}) \cdot P(Y_j = y_j,\, j \neq \ell|\mathbf{x}) : \ \mathbf{y}_{-\ell} \in \{0, 1\}^{p-1}\} \end{aligned} \tag{12}$$

and each conditional probability is defined by Equation (17). where $h(x) = e^x/(1+e^x)$. So a natural choice for transformation g is h^{-1}: $g(p) = \log\left(p/(1-p)\right)$. We then model the marginal distribution of Y_ℓ by fitting a linear regression model to the logits of the marginal probabilities.

What would the linear model for the marginal distribution look like? We can estimate this in certain cases. In Equation (17), there is no error term; the conditional probabilities are not give and must be derived from the linear models. In Equation (12), the conditional probabilities may have similar values (this happens if the effect of other Y_j's on Y_ℓ is small), or one weight may be very large for all observations, so that the weighted mean is always dominated by one conditional probability. In either case, $P(Y_{\ell,i}|\mathbf{x}_i)$ approximates one conditional probability, for all i. The linear model for the marginal probabilities will be very similar to that for the conditional probability. Further, the linear model for the conditional probability is without error, so the linear model for the marginal probabilities will have very little error.

3 Alternative methods for modeling multiple binary responses

3.1 Multivariate probit

Multivariate probit was probably the first method developed for multiple ordinal responses. A good introduction to fitting multivariate probit models with MCMC is given in [Chib and Greenberg, (1998)]. Multivariate probit and multivariate t-link both assume that each binary response Y_i arises from a continuous variable z_i, such that

$$Y_i = \begin{cases} 1 \text{ if } z_i > 0 \\ 0 \text{ if } z_i \leq 0. \end{cases} \tag{13}$$

Multivariate probit is based on the multivariate normal density $F(\mathbf{t}|\mu, \mathbf{\Sigma})$. The values of μ are given by linear regression. For observation i, the mean μ_{ij} of response j is

$$\mu_{ij} = \alpha_{j0} + \sum_{\ell=1}^{n} \alpha_{j\ell} X_{i\ell}. \tag{14}$$

Then for $\mathbf{y} \in \{0,1\}^p$, the probability that $Y_{ij} = y_j$, $1 \leq j \leq p$ is given by

$$P(\{Y_{ij} = y_j\}_{j=1}^{p} | \mathbf{X}_i) = \int_{A_i 1} \cdots \int_{A_i p} F(\mathbf{t}|\mu_i, \mathbf{\Sigma}) d\mathbf{t} \tag{15}$$

where

$$A_{ij} = \begin{cases} (0, \infty) \text{ if } y_j = 1 \\ (-\infty, 0) \text{ if } y_j = 0 \end{cases} .$$

3.2 Multivariate t-link

Multivariate t-link was introduced in [O'Brien and Dunson (2004)]. This model is based on the multivariate density $\mathcal{L}_{p,\nu}(\mathbf{z}|\mu, \mathbf{R})$. Regression by way of μ and probability $P(\mathbf{Y}_i = \mathbf{y})$ are defined as in Equations (14) and (15). The multivariate t-link is defined so that the marginal densities of z_j are given by the univariate logistic densities $\mathcal{L}(z_j|\mu_j)$, where

$$\mathcal{L}(z|\mu) = \frac{\exp\left(-(z-\mu)\right)}{[1 + \exp\left(-(z-\mu)\right)]^2} \tag{16}$$

The density $\mathcal{L}_{p,\nu}(\mathbf{z}|\mu, \mathbf{R})$ is difficult to work with. In [O'Brien and Dunson (2004)], actual calculations were done with the density F_ν of the multivariate F density with ν degrees of freedom. If $\nu = 7.3$, then F_ν closely approximates $\mathcal{L}_{p,\nu}(\mathbf{z}|\mu, \mathbf{R})$.

3.3 Marginal distributions and probabilities

For the multivariate probit and multivariate t-link, each Y_j is determined by a latent variable z_j. The marginal distribution of a single Y_j is determined by the marginal distribution of z_j. For $S \subset \{1, \ldots, n\}$, the marginal distribution of $\{Y_j\}_{j \in S}$ is determined by the joint distribution of $\{z_j\}_{j\ inS}$. A marginal probability (of the Y_j's) can be estimated by integrating the marginal p.d.f. of the z_j's over a quadrant. For a first order marginal probability, the marginal distribution is univariate.

In the multivariate probit, the joint distribution of $(z_1, \ldots, z_n)$ is multivariate normal. The marginal distribution of any z_j is univariate normal. For the multivariate t-link, the marginal distribution of a single z_j is the logistic distribution $\mathcal{L}(\cdot|\mu_j)$. This is easily integrated.

For a higher order marginal probability of $\{Y_j\}_{j \in S}$, the marginal distribution of $\{z_j\}_{j \in S}$ is multivariate normal for MVT. For MVTL, use F_ν to find an approximate value for the marginal probability. In either case, the multivariate density must be integrated over a quadrant. Such integrations are feasible, but need much more computation than the univariate integrals of first order marginal probabilities.

4 CSLR with third order interaction

Equation (17 is a variation of Equation (2). It defines a variation of CSLR that includes third or higher order interactions of the responses.

$$\begin{aligned}\operatorname{logit}\left(P(Y_{ij}=1|\mathbf{X}_i,\{Y_{ik},k\neq j\})\right)=&\mu_j+\sum_{\ell=1}^{n}\alpha_{j\ell}X_{i\ell}+\sum_{k\neq j}\gamma_{jk}Y_{ik}\\&+\sum_{k,\ell\not j,k\not\ell}\gamma_{jk\ell}Y_{ik}Y_{i\ell},1\leq j\leq p.\end{aligned}\tag{17}$$

By the way this is defined, $\gamma_{jk\ell}=\gamma_{j\ell k}$, any (j,k,ℓ) distinct in $1,\ldots,p)$.

Lemma 2. *Assume that* $\gamma_{jk\ell}=\gamma_{kj\ell}$. *Then Equation(s) (17) define compatible conditional distributions.*

Proof. Given in the Appendix. □

For $p=3$, this model has one new parameter, γ_{123}. A CSLR model with this interaction will be denoted CSLR.3int.

5 An example: Data from the DiNEH Project

Data comes from a survey carried out on the Navajo Nation by the DiNEH Project in 2004-2010. A description can be found in the Appendix. An analysis of data from this project has appeared; see [Hund et al., (2014)] There were 1304 participants. 32 observations were dropped, so $N=1272$. There are three binary responses, for occurrences of three diseases: Kidney disease, hypertension, and diabetes. Responses:

	Kidney disease	Hypertension	Diabetes
yes	70	456	316
no	1202	816	956

The covariates are: age, gender (1=female, 0=male), BMI,M, E, Navajo use, travel time to food store, education, income, and family history for each disease. These are the covariates for each response, in the CSLR model:

- Kidney disease: M, Navajo use, family history of KD.
- Hypertension: age, gender, BMI, E, family history of hypertension.
- Diabetes: age, gender, BMI, store travel time, education, income, family history of diabetes.

5.1 Modeling

These response-covariate relations were used for all models. Multivariate probit and multivariate t-link models were fit. A conditionally specified logistic regression model was fit using Equation (2). Another CSLR model was fit, with a third order interaction (See Equation (2).) A CSLR model with this interaction will be denoted CSLR.3int.

All models were fit by MCMC. Each fitted model gave a set of point estimates for coefficients. The coefficient estimates are shown in Table 1.

5.2 Results

The coefficient estimates for the CSLR shown in Table 1 were used in to calculate marginal probabilities. The coefficient estimates for the MVP and MVTL shown in Table 1 were used specify multivariate densities. Then marginal probabilities for MVP and MVTL were calculated by integration of marginal distributions, as described in Section 3.3. Then for each method there are marginal probabilities for all diseases (and pairs of diseases), for each of 1272 participants.

Correlations are shown in Table(s) 2. These are between first order marginal probabilities of different methods, with separate correlations for each disease response. Table 2 shows that for all three diseases: (a) The marginal probabilities from MVP and CSLR have low correlation; (b) the marginal probabilities for these two methods have higher correlation with the marginal probabilities from MVTL; and (c)the correlation between the m.p.'s from CSLR and those from MVTL is always slightly higher than the correlation between the m.p.'s from MVP and those from MVTL. For each disease, the marginal probabilities from MVTL have the highest correlation with the actual binary response, and those from MVP have the lowest correlation.

Plots of first order marginal probabilities are in the Appendix. The plots of marginal probabilities from CSLR vs. those from MVP (Figure 1), do not clearly indicate linear relations; for Figure 1(A), no linear relation can be seen. By contrast, the plots of m.p.s from CSLR against those from MVTL (Figure 2) show relations that are strongly monotonic, with little scatter. However, these plots imply curves, not straight lines. Also, in Figures 1 and 2, most points fall below the diagonal $x = y$, so that CSLR produces smaller marginal probabilities than do MVP or MVTL. So, for this data set, MVP and MVTL tend to overestimate marginal probabilities, or CSLR tends to underestimate. Figure 3 shows the m.p.s from CSLR.3int plotted against those from MVTL, for two diseases. These plots have little scatter and close approximation to the diagonal, and indicate near-linear relationships.

5.2.1 Goodness of fit

As explained in Section 2, the marginal probabilities model the binary responses and pairs of responses. We used two measures of goodness of fit for a binary response: Deviance and the Hosmer-Lemeshow statistic. For definitions, see

[Hosmer, Lemeshow, and Sturdivant (2013)]. The Hosmer-Lemeshow statistic was calculated by dividing a set of 1272 marginal probability values into deciles. Table 3 shows deviances and Hosmer-Lemeshow statistics. In Tables 3(A) and (B), $\mathbf{y}$ is successively the binary response for kidney disease, hypertension, and diabetes, and $\mathbf{p}$ is the vector of marginal probabilities for that response from each of: MVP, MVTL, CSLR, and CSLR.3int. In Tables 3(C) and (D), $\mathbf{y}$ is successively the product of binary responses for pairs of the diseases; $\mathbf{p}$ is successively the vector of second order marginal probabilities for those responses from MVTL, CSLR, and CSLR.3int.

Judging by Table 3, the models that best fit the responses are MVTL and CSLR.3int. These two perform almost equally in fitting second order events (of form $Y_{ij} = 1 \ and \ Y_{ik} = 1$). CSLR.3int is clearly the best for fitting individual binary responses with first order marginal probabilities. MVP does not fit first order marginal probabilities nearly as well as MVTL or CSLR.3int.

Goodness of fit statistics vary drastically by disease. Consider deviances for first order marginal probabilities. No model gives deviance greater than 800 for kidney disease, or less than 1200 for hypertension or diabetes. Methods did not show consistent patterns of superiority across responses. For two diseases, marginal probabilities from CSLR (not CSLR.3int) modeled responses as well as those from MVTL; for diabetes, CSLR was inferior.

5.3 Models of marginal distributions

The logits of the marginal probabilities were fitted to linear regression models. For each response variable, the covariates used were those used for that response in the CSLR. Table (4) shows the coefficient estimates for the resulting models. The standard errors and p-values for these estimates were all extremely small, so there seemed no point in showing them. Also shown are the corresponding coefficient estimates in the CSLR, and the overall adjusted R^2 of the linear model for the marginal probabilities. As can be seen, R^2_{adj} are high for all responses, though noticeably lower for kidney disease. The coefficients in the models for marginal probabilities quite close to those in the CSLR.

6 Discussion

We can find marginal probabilities for the conditionally specified logistic regression model. This means that CSLR can model multiple binary data, in the sense of making point estimates. An example showed that this method can model data well.

Consider the marginal probabilities of form $P(Y_{i,j} = 1, \ j \in S|\mathbf{x})$, for any subset S of $\{1, \ldots, p\}$. These satisfy a system of affine equations. If the system can be solved, the marginal probabilities are found. In addition, the marginal distribution of each response can be modeled using the first order marginal probabilities.

CSLR was applied to an example where $p = 3$, both with and without a third order interaction. The first order marginal probabilities from both CSLR models were very closely related to those from the multivariate t-link. Using standard goodness of fit criteria, CSLR with third order interaction modeled this data better than any other method. CSLR without third order interaction performed significantly worse than MVTL.

The results for this data set indicate what might be expected when applying CSLR to other data sets with multiple binary responses. If $p > 3$, interactions of order four or more may be needed. And it may be that CSLR generally needs higher order interactions to model data well if $p > 2$. Further investigation is needed to answer these questions.

Acknowledgments

The DiNEH survey and its analysis was funded by NIEHS grants R01 ES014565, R25 ES013208, and P30 ES-012072, a USEPA-ERRG pass through grant, with support from DHHS/NIH/NCCR #1UL1RR031977-01; and a contract from the Navajo Nation Environmental Protection Agency. I would like to thank Gabriel Huerta and Glenn Stark, who first applied CSLR to the DiNEH Project data; and Johnnye Lewis, who suggested adding a third order interaction to the model.

References

[Chib and Greenberg, (1998)] Chib, S. and Greenberg, E., 1998. Analysis of Multivariate Probit Models. Biometrika, 85, 347-361.

[Hosmer, Lemeshow, and Sturdivant (2013)] Hosmer, D., Lemeshow, S., Sturdivant, R., 2013. Applied Logistic Regression, third revised edition. Wiley, Hoboken.

[Hund et al., (2014)] Hund, L, Bedrick, E., Miller, C., Huerta, G., Nez, T., Cajero, M., Lewis, J., 2104. A Bayesian framework for estimating disease risk due to exposure to uranium mine and mill waste in the Navajo Nation. To appear in J. R. Stat. Soc. (B).

[Imbens and Woolridge (2009] Imbens, G. and Wooldridge, J., 2009. Recent developments in the econometrics of program evaluation. J. of Econ. Lit., 5-86.

[Joe and Liu (1996)] Joe, H., Liu, Y., 1996. A model for a multivariate binary response with covariates based on compatible conditionally specified logistic regressions. Statistics and Probability Letters 31, 113-120.

[O'Brien and Dunson (2004)] O'Brien, S., Dunson, D., 2004. Bayesian Multivariate Logistic Regression. Biometrics 60, 739-746.

	MVP	MVTL	CSLR	CSLR.3int
Intercept:KD	-2.279	-2.548	-4.68	-4.947
M:KD	0.2456	0.3009	0.401	0.4401
Navajo use:KD	0.1829	0.1641	0.2496	0.2067
FH.Kidney:KD	0.3118	0.4921	0.7557	0.7530
Intercept:HTN	-4.219	-4.413	-6.973	-6.803
age:HTN	0.04119	0.04187	0.05879	0.0598
gender:HTN	-0.1269	-0.03376	-0.06141	-0.1860
BMI:HTN	0.0398	0.04127	0.05981	0.0564
E:HTN	0.5578	0.6682	1.364	1.363
FH.HTN:HTN	0.5021	0.5388	0.8921	0.8194
Intercept:Di	-3.796	-4.014	-5.631	-5.206
age:Di	0.03072	0.03048	0.03393	0.0261
gender:Di	0.1212	0.1583	0.313	0.4314
BMI:Di	0.03226	0.03646	0.04234	0.0307
Store time:Di	0.1435	0.1357	0.2064	0.2495
Education:Di	-0.03606	-0.04536	-0.09478	-0.1017
Income:Di	-4.976e-06	-4.976e-06	-0.0001028	-1.027e-05
FH.Diabetes:Di	0.5757	0.6104	1.016	0.9630
γ_{12}	0.367	0.3386	0.808	0.7048
γ_{13}	0.485	0.466	0.9062	1.5800
γ_{23}	0.6314	0.6383	0.9746	1.9540
γ_{123}	NA	NA	NA	-0.0357

Table 1: Point estimates of coefficients for models of multiple binary responses. Methods used to fit models: Multivariate probit (MVP), multivariate t-link (MVTL), conditionally specified logistic regression (CSLR), and CSLR with a three-way interaction term.

(A)

	MVP	MVTL	CSLR	CSLR.3int	response
MVP	1.0000	0.5807	0.3027	0.2549	0.0891
MVTL	0.5807	1.0000	0.8852	0.8750	0.2040
CSLR	0.3027	0.8852	1.0000	0.9813	0.1933
CSLR.3int	0.2549	0.8750	0.9813	1.0000	0.2011
response	0.0891	0.2040	0.1933	0.2011	1.0000

(B)

	MVP	MVTL	CSLR	CSLR.3int	response
MVP	1.0000	0.8961	0.8374	0.8474	0.4053
MVTL	0.8961	1.0000	0.9762	0.9786	0.4589
CSLR	0.8374	0.9762	1.0000	0.9942	0.4481
CSLR.3int	0.8474	0.9786	0.9942	1.0000	0.4529
response	0.4053	0.4589	0.4481	0.4529	1.0000

(C)

	MVP	MVTL	CSLR	CSLR.3int	response
MVP	1.0000	0.7447	0.6130	0.6432	0.2705
MVTL	0.7447	1.0000	0.7956	0.9543	0.3827
CSLR	0.6130	0.7956	1.0000	0.8535	0.3020
CSLR.3int	0.6432	0.9543	0.8535	1.0000	0.3800
response	0.2705	0.3827	0.3020	0.3800	1.0000

Table 2: Correlations of first order marginal probabilities. (A): Correlations of marginal probabilities for kidney disease. (B): Correlations of marginal probabilities for hypertension. (C): Correlations of marginal probabilities for diabetes.

(A)

	kidney disease	hypertension	diabetes	total
MVP	564.0822	1695.413	1613.257	3872.753
MVTL	593.5593	1403.907	1275.771	3273.238
CSLR	511.3395	1403.138	1656.995	3571.472
CSLR.3int	506.569	1379.889	1239.914	3126.372

(B)

	kidney disease	hypertension	diabetes	total
MVP	29.06586	286.3585	293.4718	608.8961
MVTL	72.01331	44.11962	48.92854	165.0615
CSL	11.6615	62.60299	1406.991	1481.255
CSLR.3int	14.8617	33.1592	26.8517	74.8726

(C)

	HTN+Diabetes	Kidney+Diabetes	Kidney+HTN	total
MVP	1360.746	406.840	400.763	2168.350
MVTL	1042.239	369.821	374.840	1786.900
CSLR	1431.248	455.387	394.819	2281.454
CSLR.3int	1037.044	370.535	381.225	1788.804

(D)

	HTN+Diabetes	Kidney+Diabetes	Kidney+HTN	total
MVP	300.412	37.665	15.613	353.690
MVTL	13.167	22.608	12.513	48.287
CSLR	1106.523	424.1947	31.91071	1562.628
CSLR.3int	23.768	12.109	15.323	51.200

Table 3: Measures of goodness of fit. Lower is better. (A): Deviances of first order marginal probabilities from binary disease responses, for three methods, six hybrids, and three diseases. (B): Hosmer-Lemeshow statistics of first order marginal probabilities. (C): Deviances of second order marginal probabilities for two methods and two hybrids. (D): Hosmer-Lemeshow statistics of second order marginal probabilities.

(A)

coefficient	marginal	CSL
(Intercept)	-4.562	-4.946
M	0.5588	0.436
Navajo use	0.3475	0.2069
FH.Kidney	0.8304	0.7526
Adjusted R^2:	0.7856	

(B)

coefficient	marginal	CSL
(Intercept)	-6.93	-6.797
age	0.05897	0.05527
gender	-0.1592	-0.1858
BMI	0.05905	0.05616
E	1.537	1.362
FH.HTN	0.8626	0.8203
Adjusted R^2:	0.9928	

(C)

coefficient	marginal	CSL
(Intercept)	-5.948	-5.202
age	0.03832	0.02152
gender	0.3283	0.4311
BMI	0.04704	0.03105
Store time	0.2269	0.2494
Education	-0.0965	-0.1013
Income	-2.39e-05	-2.548e-05
FH.Diabetes	1.041	0.9623
Adjusted R^2:	0.9662	

Table 4: Coefficient estimates for linear regression models of logits of marginal probabilities, for three responses. (A) Kidney disease (B) Hypertension (C) Diabetes.