Applied Mathematics, **2013, 4, 1485-1489**
Published Online November 2013 (http://www.scirp.org/journal/am)
http://dx.doi.org/10.4236/am.2013.411200

Scientific Research

On Expressing the Probabilities of Categorical Responses as Linear Functions of Covariates

Tejas A. Desai
The Adani Institute of Infrastructure Management, Ahmedabad, India
Email: tejasdesai4@gmail.com

Received August 22, 2013; revised September 22, 2013; accepted September 29, 2013

ABSTRACT

Logistic regression is usually used to model probabilities of categorical responses as functions of covariates. However, the link connecting the probabilities to the covariates is non-linear. We show in this paper that when the cross-classification of all the covariates and the dependent variable have no empty cells, then the probabilities of responses can be expressed as *linear* functions of the covariates. We demonstrate this for both the dichotmous and polytomous dependent variables.

Keywords: Logistic Regression; Linear Regression; Maximum Likelihood Estimation; Least-Squares Estimation

1. Introduction

The probability of a dichotomous response is usually modelled as functions of covariates using the following:

$$\Pr\left(Y=1\middle|X_1=x_1,\cdots,X_p=x_p\right)=\frac{\exp\left(\alpha+x_1\beta_1+\cdots+x_p\beta_p\right)}{1+\exp\left(\alpha+x_1\beta_1+\cdots+x_p\beta_p\right)}$$

A feature of the above formulation is that the quantity on the right-hand side of the above equation is a fraction, and so the rule that probabilities have to lie in the interval [0, 1] is not violated assuming the estimates of $\alpha,\beta_1,\cdots,\beta_p$ exist. In this paper, we are interested in the following questions: under what conditions we can express the probabilities as the following:

$$\Pr\left(Y=1\middle|X_1=x_1,\cdots,X_p=x_p\right)=\alpha+x_1\beta_1+\cdots+x_p\beta_p$$

so that the quantities on the left-hand side of the above equations indeed lie in the interval [0, 1] once the estimates of the unknown parameters are known to be finite. We show in the remaining paper that the above, *linear* formulation will yield estimates of probabilities lying in [0, 1] if the cross-classification of all the covariates and the dependent variable has no empty cells. In Section 2, we formulate the problem and prove our main result. In Section 3, we work out a detailed example wherein the dependent variable is dichotomous. In Section 4, we work out a detailed example wherein the dependent variable is ordinally polytomous. In Section 5, we present a conjecture regarding the least-squares estimation of the parameters in our model. In Section 6, we end the paper with concluding remarks.

2. Problem Formulation and the Main Result

Let Y be a categorical variable with possible values $0,\cdots,q$. Y may be a dichotomous random variable, a nominal polytomous random variable, or an ordinal polytomous random variable. The covariates, $X_1,\cdots,X_p$, may be categorical or continuous. Let $\left(y_j;x_{j1},\cdots,x_{jp}\right)$, $1\le j\le n$, denote a data set with n outcomes of Y and of each of the p covariates. For $j=1,\cdots,n$, let

$$\begin{aligned}&\Pr\left(Y=y_j=i\middle|i>0;X_1=x_{j1},\cdots,X_p=x_{jp}\right)\\&=\alpha_i+x_{j1}\beta_{i1}+\cdots+x_{jp}\beta_{ip}\end{aligned}\tag{1.1}$$

and

$$\begin{aligned}&\Pr\left(Y=y_j=0\middle|x_{j1},\cdots,x_{jp}\right)\\&=1-\sum_{k=1}^{q}\Pr\left(Y=y_j=k\middle|X_1=x_{j1},\cdots,X_p=x_{jp}\right)\\&=1-\sum_{k=1}^{q}\left(\alpha_k+x_{j1}\beta_{k1}+\cdots+x_{jp}\beta_{kp}\right)\end{aligned}\tag{1.2}$$

Then we have the following result:

Theorem 1: Suppose that the cross-classification of the data $\left(y_j; x_{j1}, \cdots, x_{jp}\right)$, $1 \le j \le n$, has no empty cells. If the mle's obtained by specifying the likelihood using (1.1) and (1.2) exist, then the estimates of probabilities of the response given the covariates are constrained to lie in the interval (0, 1).

Proof: Let For $j = 1, \cdots, n$, let

$$I_0\left(y_j, x_{j1}, \cdots, x_{jp}\right) = \begin{cases} 0 & \text{if } y_j \neq 0 \\ 1 & \text{if } y_j = 0 \end{cases}, \cdots, I_q\left(y_j, x_{j1}, \cdots, x_{jp}\right) = \begin{cases} 0 & \text{if } y_j \neq q \\ 1 & \text{if } y_j = q \end{cases}$$

Consider the function

$$L = \prod_{j=1}^{n} \left[\left(1 - \sum_{k=1}^{q} \left(\alpha_k + x_{j1}\beta_{k1} + \cdots + x_{jp}\beta_{kp} \right) \right)^{I_0\left(y_j, x_{j1}, \cdots, x_{jp}\right)} \cdot \prod_{k=1}^{q} \left(\alpha_k + x_{j1}\beta_{k1} + \cdots + x_{jp}\beta_{kp} \right)^{I_k\left(y_j, x_{j1}, \cdots, x_{jp}\right)} \right]$$

Now suppose that $\alpha_1 = \cdots = \alpha_q = \dfrac{1}{2q}$ and $\beta_{i1} = \cdots = \beta_{ip} = 0$ for $1 \le i \le q$. Then the value of $L \ge \left(\dfrac{1}{2q}\right)^n$. This means that the maximum of L over the parameter space is either finitely positive or it is positive infinity. Suppose that the maximum of L is finitely positive. Then the maximization of $\log L$ must yield the same parameter values as the maximization of L. Let $\hat{\alpha}_i, \hat{\beta}_{i1}, \cdots, \hat{\beta}_{ip}$, $1 \le i \le q$, be the parameter estimates obtained by maximizing $\log L$. Then note that for any $i, 1 \le i \le q$, the term $\hat{\alpha}_i + x_{j1}\hat{\beta}_{i1} + \cdots + x_{jp}\hat{\beta}_{ip}$ cannot be less than or equal to 0 as that would mean that $\log\left(\hat{\alpha}_i + x_{j1}\hat{\beta}_{i1} + \cdots + x_{jp}\hat{\beta}_{ip}\right)$, and hence $\log L$, is undefined. Similarly, for any $i, 1 \le i \le q$, the term $\hat{\alpha}_i + x_{j1}\hat{\beta}_{i1} + \cdots + x_{jp}\hat{\beta}_{ip}$ cannot be greater than or equal to 1 because then again $\log\left(1 - \sum_{k=1}^{q}\left(\hat{\alpha}_k + x_{j1}\hat{\beta}_{k1} + \cdots + x_{jp}\hat{\beta}_{kp}\right)\right)$, and hence $\log L$, would be undefined. Furthermore, note that $0 < \sum_{k=1}^{q}\left(\hat{\alpha}_k + x_{j1}\hat{\beta}_{k1} + \cdots + x_{jp}\hat{\beta}_{kp}\right) < 1$, as otherwise, $\log L$ would again be undefined. Thus all the estimates of the probabilities in (1.1) and (1.2) are constrained to lie in the interval (0, 1). □

3. Detailed Example: Dichotomous Response

Consider the data in **Table 1**. The data comes from a study on coronary artery disease and is reported in [1]. The question of interest is whether gender and electrocardiogram (ECG) measurement have an effect on disease status.

Table 1. Coronary artery disease data.

Gender	ECG	Disease	No Disease
Female	<0.1 ST segment depression	4	11
Female	≥0.1 ST segment depression	8	10
Male	<0.1 ST segment depression	9	9
Male	≥0.1 ST segment depression	21	6

Let $Y = 1$ if disease is present, and $Y = 0$ if disease is absent. Let $SEX = 0$ if gender is female and $SEX = 1$ if gender is male. Let $ECG = 0$ if ST segment depression is less than 0.1 and $ECG = 1$ if ST segment depression is greater than or equal to 0.1. Consider the following relations:

$$\Pr\left(Y = 1 \middle| SEX = x_1, ECG = x_2\right) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

$$\Pr\left(Y = 0 \middle| SEX = x_1, ECG = x_2\right) = 1 - \alpha - \beta_1 x_1 - \beta_2 x_2$$

We want to estimate α, β_1, and β_2, and check whether the estimated probabilities lie in the interval $(0,1)$. We wish to use the Newton-Raphson method for the purpose of estimation. To use the Newton-Raphson method, we need good starting estimates. As starting estimates, we use the estimates provided by least-squares estimation of the following linear model:

$$Y = \delta + \lambda SEX + \tau ECG + \epsilon$$

The least-squares estimates are: $\hat{\delta} = 0.23563$, $\hat{\lambda} = 0.29023$, $\hat{\tau} = 0.23467$. We use these as starting estimates of α, β_1, and β_2, respectively. We stop the Newton-Raphson algorithm when the absolute difference of successive iterates is less than 0.00001 for all the three parameters. Using this criterion we notice that the Newton-Raphson algorithm converges and estimates we get are: $\hat{\alpha} = 0.2405112$, $\hat{\beta}_1 = 0.2892142$, $\hat{\beta}_2 = 0.2336847$. Note that we can now witness the effect of the covariates on the disease status. For example, as *SEX* goes from 0 to 1, the probability of being diseased goes up. Similarly, as *ECG* status goes from 0 to 1, the probability of being diseased goes up. The estimated probabilities, using our method and the least-squares method, are given in **Table 2**.

Note that the estimation of probabilities using the least-squares method is as follows:

$$\widehat{\Pr}\left(Y = 1 \middle| SEX = x_1, ECG = x_2\right) = \hat{\delta} + \hat{\lambda} x_1 + \hat{\tau} x_2$$

$$\widehat{\Pr}\left(Y = 0 \middle| SEX = x_1, ECG = x_2\right) = 1 - \hat{\delta} - \hat{\lambda} x_1 - \hat{\tau} x_2$$

Notice that all the estimates of probabilities in **Table 2** lie in the interval (0, 1). Also notice the striking similarity between the estimates using our method and the corresponding estimates using the least-squares method. However, it seems difficult to prove a least-squares analogue of Theorem 1.

Now we turn our attention to goodness of fit. The two traditional goodness-of-fit statistics are Pearson's chi-square and the likelihood ratio chi square, namely, Q_P and Q_L, respectively. The latter statistic is also known as *deviance*. Let $h=0$ if $SEX=0$ and $h=1$ if $SEX=1$. Let $i=0$ if $ECG=0$ and $i=1$ if $ECG=1$. Finally, let $j=0$ if $Y=0$ (disease absent) and $j=1$ if $Y=1$ (disease present). It then follows that

$$Q_P=\sum_{h=0}^{1}\sum_{i=0}^{1}\sum_{j=0}^{1}\left(n_{hij}-m_{hij}\right)^2\Big/m_{hij} \text{ and}$$

$$Q_L=\sum_{h=0}^{1}\sum_{i=0}^{1}\sum_{j=0}^{1}2n_{hij}\log\left(\frac{n_{hij}}{m_{hij}}\right)$$

where

$$m_{hij}=\begin{cases}n_{hi+}\widehat{\Pr}\left(Y=0\middle|SEX=h,ECG=i\right) & \text{if } j=0\\ n_{hi+}\widehat{\Pr}\left(Y=1\middle|SEX=h,ECG=i\right) & \text{if } j=1\end{cases}$$

For the present model, there are four subpopulations and three parameters, giving us $4-3=1$ degree of freedom for each of the Pearson's and likelihood-ratio statistics. The values of Q_P and Q_L and the respective p-values are given in **Table 3**.

The goodness-of-fit statistics thus indicate that the above model fits the data reasonably well. It must be noted that there are sample-size guidelines to be followed in order to ensure that the Pearson's and likelihood-ratio statistics approximately follow the chi-square distribution. These guidelines are mentioned in [1].

4. Detailed Example: Polytomous Response

Logistic regression is defined in terms of a dichotomous response variable. Therefore, for a polytomous response, one has to form cumulative logits in case of ordinal response, and generalized logits in the case of a nominal response. Thus, logistic regression is indirectly applied. However, the application of our model is direct in the sense that the possibility of a polytomous response is already accounted for. We illustrate with the following example.

Consider the following data in **Table 4**. The data is reported in [1] and it concerns an arthritis study wherein males and females were administered either a drug or placebo and their response (improvement) was measured as being one of "marked", "some" or "none".

The data in **Table 4** does not meet the requirements of Theorem 1 since there is one zero count in the cross-classification. Since our purpose here is to illustrate our model and estimation of model parameters, we will consider the fictional data set obtained by replacing the zero count with a count of 1. The fictional data is presented in **Table 5**.

There are no zero counts in the cross-classification in **Table 5**. Let $M=1$ if improvement is marked, and $M=0$ otherwise. Let $S=1$ if there is some improvement, and $S=0$ otherwise. Let $N=1$ if there is no improvement, and $N=0$ otherwise. We will denote the gender variable as SEX, and the treatment variable as TRT. Let $SEX=0$ if gender is female and $SEX=1$ if gender is male. Let $TRT=0$ if treatment is placebo and $TRT=1$ if treatment is active.Finally, let $Y=1$ if there is no improvement, $Y=2$ if there is some improvement, and $Y=3$ if there is marked improvement. Our model is as follows:

$$\Pr\left(Y=2\middle|SEX=x_1,TRT=x_2\right)=\alpha_2+\beta_{21}x_1+\beta_{22}x_2$$

$$\Pr\left(Y=3\middle|SEX=x_1,TRT=x_2\right)=\alpha_3+\beta_{31}x_1+\beta_{32}x_2$$

Table 2. Estimates of probabilities.

Estimates of Probabilities	Our Method	Least-Squares Method
$\widehat{\Pr}\left(Y=0\middle\|SEX=0,ECG=0\right)$	0.75949	0.76437
$\widehat{\Pr}\left(Y=1\middle\|SEX=0,ECG=0\right)$	0.24051	0.23563
$\widehat{\Pr}\left(Y=0\middle\|SEX=0,ECG=1\right)$	0.52580	0.52969
$\widehat{\Pr}\left(Y=1\middle\|SEX=0,ECG=1\right)$	0.47420	0.47031
$\widehat{\Pr}\left(Y=0\middle\|SEX=1,ECG=0\right)$	0.47027	0.47414
$\widehat{\Pr}\left(Y=1\middle\|SEX=1,ECG=0\right)$	0.52973	0.52586
$\widehat{\Pr}\left(Y=0\middle\|SEX=1,ECG=1\right)$	0.23659	0.23946
$\widehat{\Pr}\left(Y=1\middle\|SEX=1,ECG=1\right)$	0.76341	0.76054

Table 3. Goodness-of-fit Statistics and their respective p-values.

Pearson		Deviance	
Statistic Value	p-Value	Statistic Value	p-Value
0.215	0.643	0.214	0.644

Table 4. Arthritis data.

		Improvement		
Gender	Treatment	Marked	Some	None
Female	Active	16	5	6
Female	Placebo	6	7	19
Male	Active	5	2	7
Male	Placebo	1	0	10

Table 5. Fictional arthritis data.

		Improvement		
Gender	Treatment	Marked	Some	None
Female	Active	16	5	6
Female	Placebo	6	7	19
Male	Active	5	2	7
Male	Placebo	1	1	10

$$\Pr\left(Y=1\middle|SEX=x_1,TRT=x_2\right)$$
$$=1-\alpha_2-\beta_{21}x_1-\beta_{22}x_2-\alpha_3-\beta_{31}x_1-\beta_{32}x_2$$

To estimate the model parameters, we specify the log-likelihood and apply the Newton-Raphson algorithm. Once again, we use least-squares estimates as starting values. Consider the following two linear models:

$$S=\delta_S+\lambda_S SEX+\tau_S TRT+\epsilon_S$$
$$M=\delta_M+\lambda_M SEX+\tau_M TRT+\epsilon_M$$

The least-squares estimates are: $\hat{\delta}_S=0.20571$, $\hat{\lambda}_S=-0.08760$, $\hat{\tau}_S=-0.00507$, $\hat{\delta}_M=0.20589$, $\hat{\lambda}_M=-0.17161$, and $\hat{\tau}_M=0.36490$. These are also our starting estimates for α_2, β_{21}, β_{22}, α_3, β_{31}, and β_{32}, respectively. As before, we stop the Newton-Raphson algorithm when the absolute difference of successive iterates is less than 0.00001 for all the six parameters. Using this criterion we notice that the Newton-Raphson algorithm converges and estimates we get are: $\hat{\alpha}_2=0.2025164$, $\hat{\beta}_{21}=-0.098328$, $\hat{\beta}_{22}=0.0107827$, $\hat{\alpha}_3=0.2056062$, $\hat{\beta}_{31}=-0.138855$, and $\hat{\beta}_{32}=0.3494801$. Note, again, that from the preceding estimates, we can directly assess the effect of covariates on the probability of improvement. The estimated probabilities are given in **Table 6**.

Note that, once again, the probabilities in **Table 6** lie in the interval (0, 1). Also, once again, note the similarity between the estimated probabilities obtained using our method, and the ones obtained using the least-squares method. To take into account the ordinality in the response, read the probabilities across the rows in **Table 6**. The response levels are correlated with the row probabilities. Note that for any treatment, active or placebo, males perform poorly compared to females. As expected, both males and females respond better to active treatment than placebo in the sense that for both sexes, the probability of some or marked treatment goes up with active treatment. The least-squares estimates of probabilities were obtained as follows:

$$\widehat{\Pr}\left(Y=2\middle|SEX=x_1,TRT=x_2\right)=\hat{\delta}_S+\hat{\lambda}_S x_1+\hat{\tau}_S x_2$$

$$\widehat{\Pr}\left(Y=3\middle|SEX=x_1,TRT=x_2\right)=\hat{\delta}_M+\hat{\lambda}_M x_1+\hat{\tau}_M x_2$$

$$\widehat{\Pr}\left(Y=1\middle|SEX=x_1,TRT=x_2\right)$$
$$=1-\hat{\delta}_S-\hat{\lambda}_S x_1-\hat{\tau}_S x_2-\hat{\delta}_M-\hat{\lambda}_M x_1-\hat{\tau}_M x_2$$

The goodness-of-fit tests are conducted as in Section 3 except that the number of degrees of freedom for Q_P and Q_L is $(4-3)\times(3-1)=2$. The goodness-of-fit statistics and their respective p-values are given in **Table 7**.

So both Pearson's chi-square and the deviance statistics seem to support model-fit. The response in this example is ordinal, so the question arises whether an analogue of the proportional-odds model can be defined. It can be defined as follows:

$$\Pr\left(Y=2\middle|SEX=x_1,TRT=x_2\right)=\alpha_2+\beta_1 x_1+\beta_2 x_2$$

$$\Pr\left(Y=3\middle|SEX=x_1,TRT=x_2\right)=\alpha_3+\beta_1 x_1+\beta_2 x_2$$

$$\Pr\left(Y=1\middle|SEX=x_1,TRT=x_2\right)$$
$$=1-\alpha_2-\beta_1 x_1-\beta_2 x_2-\alpha_3-\beta_1 x_1-\beta_2 x_2$$

The problem with the above model is that the resulting likelihood is multi-modal, and no good starting estimates for the Newton-Raphson algorithm are available. Indeed, the author found that with some starting estimates, the resulting probabilities lay outside the interval [0, 1]. More research is needed on this front.

5. A Conjecture Regarding the Least-Squares Estimates

We saw in the preceding examples that the least-squares estimates of probabilities of responses lay in the interval [0, 1] if the cross-classification of the covariates and the responses contained no empty cells. The author believes that this is not a coincidence, but is unable to prove it. So we offer the following conjecture:

Conjecture 1: Let Y be a categorical variable with possible values $0,\cdots,q$. Y may be a dichotomous random variable, a nominal polytomous random variable, or an ordinal polytomous random variable. The covariates, $X_1,\cdots,X_p$, may be categorical or continuous. Let $\left(y_j;x_{j1},\cdots,x_{jp}\right)$, $1\le j\le n$, denote a data set with n outcomes of Y and of each of the p covariates. Let the matrix of covariate values have full rank. Let

Table 6. Estimates of probabilities.

	$\widehat{\Pr}\left(Y=1\middle\|SEX,TRT\right)$		$\widehat{\Pr}\left(Y=2\middle\|SEX,TRT\right)$		$\widehat{\Pr}\left(Y=3\middle\|SEX,TRT\right)$	
Stratum	Our Method	Least Squares	Our Method	Least Squares	Our Method	Least Squares
$SEX=0,TRT=0$	0.5918774	0.5884	0.2025164	0.20571	0.2056062	0.20589
$SEX=0,TRT=1$	0.2316145	0.22857	0.2132992	0.20064	0.5550863	0.57079
$SEX=1,TRT=0$	0.8290608	0.84761	0.1041885	0.11811	0.0667507	0.03428
$SEX=1,TRT=1$	0.468798	0.48778	0.1149712	0.11304	0.4162308	0.39918

Table 7. Goodness-of-fit statistics and their respective p-values.

Pearson		Deviance	
Statistic Value	p-Value	Statistic Value	p-Value
0.613	0.736	0.615	0.735

$$Z_1 = \begin{cases} 0 & \text{if } Y \neq 1 \\ 1 & \text{if } Y = 1 \end{cases}, \cdots, Z_q = \begin{cases} 0 & \text{if } Y \neq q \\ 1 & \text{if } Y = q \end{cases}$$

Consider the following model:

$$Z_1 = \delta_1 + \lambda_{11} X_1 + \cdots + \lambda_{1p} X_p + \epsilon_1$$
$$\vdots$$
$$Z_q = \delta_q + \lambda_{q1} X_1 + \cdots + \lambda_{qp} X_p + \epsilon_q$$

Let $\hat{\delta}_k$, $\hat{\lambda}_{k1}, \cdots, \hat{\lambda}_{kp}$, $k = 1, \cdots, q$, be the resulting estimates of parameters obtained using ordinary least-squares. Then the following estimates of probabilities lie in the interval [0, 1]:

$$\widehat{\Pr}\left(Y = k \middle| X_1 = x_1, \cdots, X_p = x_p\right)$$
$$= \hat{\delta}_k + \hat{\lambda}_{k1} x_1 + \cdots + \hat{\lambda}_{kp} x_p, k = 1, \cdots, q, \text{ and}$$
$$\widehat{\Pr}\left(Y = 0 \middle| X_1 = x_1, \cdots, X_p = x_p\right)$$
$$= 1 - \sum_{k=1}^{q} \left(\hat{\delta}_k + \hat{\lambda}_{k1} x_1 + \cdots + \hat{\lambda}_{kp} x_p\right).$$

6. Concluding Remarks

In this article, we demonstrated that probability estimates lying in the interval [0, 1] can be obtained if the probabilities themselves are modelled as *linear* functions of covariates, provided that the cross-classification of the covariates and the response has no empty cells. The main advantage of this formulation is that effects of covariates on the probabilities can be directly measured, unlike in logistic regression where the link function is non-linear. The emphasis of this article is on estimation. However, hypothesis-testing using the m.l. and least-squares estimates can be done routinely as is discussed extensively in the literature. See, for example, [2,3]. Also, the data sets we have considered in this paper are complete. When data are missing at random, one may multiply impute the data sets, say, *m* times, and then combine the *m* estimates to yield a single estimate. See [4] for more details. To be honest, our method does have its limitations. For example, when one of the covariates is continuous, there are likely to be several cells in the cross-classification that are empty. Consequently, our method will be usually applicable when the covariates as well as the response are categorical. Another limitation seems to be that the analogue of the proportional-odds model is not straightforward to implement. Also, both maximum-likelihood estimation and least-squares estimation find their utility when the underlying sample sizes are relatively large. For smaller sample sizes, one has to develop *exact* methods which will be a subject of one of the author's future articles.

REFERENCES

[1] M. E. Stokes, C. S. Davis and G. G. Koch, "Categorical Data Analysis Using the SAS System," SAS Institute and Wiley, Cary, 2001.

[2] C. R. Rao, "Linear Statistical Inference and Its Application," Wiley, New York, 1973. http://dx.doi.org/10.1002/9780470316436

[3] C. R. Rao and H. Toutenburg, "Linear Models: Least Squares and Alternatives," Springer, New York, 1999.

[4] J. L. Schafer, "Analysis of Incomplete Multivariate Data," Chapman & Hall/CRC, Boca Raton, 1999.