The Korean Communications
in Statistics Vol. 11 No. 3, 2004
pp. 519-527

Suppression and Collapsibility for Log-linear Models

Chong Sun Hong[1)]

Abstract

Relationship between the partial likelihood ratio statistics for logisitic models and the partial goodness-of-fit statistics for corresponding log-linear models is discussed. This paper shows how definitions of suppression in logistic model can be adapted for log-linear model and how they are related to confounding in terms of collapsibility for categorical data. Several $2 \times 2 \times 2$ contingency tables are illustrated.

Keywords : Confounding, Goodness-of-fit statistic, Logistic, Likelihood ratio statistic, Suppressor variable.

1. Introduction

Consider $H_0 : \beta_2 = 0$ for the following two models :

$$\begin{aligned} H_0 &: Y_i = \alpha + \epsilon_i \\ H_1 &: Y_i = \alpha + \beta_2 X_{2i} + \epsilon_i, \end{aligned} \tag{1.1}$$

$$\begin{aligned} H_0' &: Y_i = \alpha + \beta_1 X_{1i} + \epsilon_i \\ H_1' &: Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i. \end{aligned} \tag{1.2}$$

The sums of squares due to regression of the test statistics for the above two hypotheses would be the sum of squares from regression on X_2 alone, $SSR(X_2)$, and the increase in the regression sum of squares due to addition of X_2 to the model that already contains X_1, $SSR(X_2|X_1)$, respectively. Horst (1941) defined a suppressor variable as a predictor X_2 that is correlated with the first predictor X_1 but uncorrelated with the response Y ; that is, $r_{x_1x_2} \neq 0$ and $r_{yx_2} \approx 0$. Such a variable causes the squared multiple correlation coefficient R^2 to exceed the sum of two squared simple correlation coefficients with Y ;

$$R^2 > r_{yx_1}^2 + r_{yx_2}^2 , \tag{1.3}$$

1) Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, KOREA.
E-mail : cshong@skku.ac.kr

$$SSR(X_2 \mid X_1) > SSR(X_2). \tag{1.4}$$

X_2 is thus said to suppress some of the variance in X_1 not relevant to Y, thereby increasing X_1's importance in the regression. Note that (1.3) condition is identical with (1.4) condition. The relationships between $SSR(X_2 \mid X_1), SSR(X_2)$, and among the correlation coefficients r_{yx_1}, r_{yx_2} and $r_{x_1x_2}$ have been considered by Hamilton (1987, 1988) with further contributions by Mitra (1988) and Freund (1988) (called a classical suppression). Hamilton (1987) suggested a necessary and sufficient condition for (1.3) in terms of the partial correlation coefficient for X_2, $r_{yx_2.x_1}$ (a cooperative suppression).

Schey (1993) explained a geometrical description about the relationship between $SSR(X_2)$ and $SSR(X_2 \mid X_1)$ with respect to some correlation coefficients. Sharpe and Roberts (1997) investigated the condition under which suppression can occur algebraically and graphically in linear regression. Grayson (1987) discussed the analogous definitions of the suppression and the confounding for both linear and logistic regression models. However, Lynn (2003) argued that there are important distinctions in the conditions that lead to suppression and confounding in logistic regression versus those in linear regression. For logistic regression, he defined the conditions leading to confounding and suppression (classical and cooperative) via log odds ratios. Since likelihood ratio statistics play a similar role as regression sum of squares (actually SSE ; error sum of squares), Lynn (2003) defined a similar condition to (1.4) as

$$L(X_2 \mid X_1 Y) < L(X_2 \mid Y), \tag{1.5}$$

where $L(X_2 \mid Y)$ and $L(X_2 \mid X_1 Y)$ are defined as the log likelihood ratio statistics to test the hypotheses (2.3) and (2.4), respectively (Author regards these statistical models as logit models rather than logistic models. In this paper, logit model will be named instead of logistic model). Lynn (2003) demonstrated three $2 \times 2 \times 2$ categorical data in order to explain the relationship between $L(X_2 \mid X_1 Y)$ and $L(X_2 \mid Y)$ in the absence of confounding.

In this paper, we extend Lynn's work on suppression and confounding to log-linear models and discuss how the suppression and confounding relate to the collapsibility with confounding in log-linear models.

2. Suppression and Collapsibility with Confounding

In epidemiology, confounding is typically presented in terms of association between disease $(D) = Y$, exposure $(E) = X_1$ and confounder $(C) = X_2$. <Table 2.1> shows three $2 \times 2 \times 2$ contingency tables to explain the relationship between $L(X_2 \mid X_1 Y)$ and $L(X_2 \mid Y)$, which are designed by Lynn (2003).

These values of the log likelihood ratio statistics for the logit models fitted to data in

<Table 2.1> are obtained in <Table 2.2>. Note that one can get $L(X_2 \mid X_1 Y) = L(X_1X_2 \mid Y) - L(X_1 \mid Y)$, where $L(X_1 \mid Y)$ and $L(X_1X_2 \mid Y)$ are the statistics of logit models for the null and the alternative hypotheses in (2.4).

<Table 2.1> Suppression and collapsibility with confounding

Example 1 ($X_1 \perp X_2 \mid Y$)				Example 2 ($X_2 \perp Y \mid X_1$)				Example 3 ($X_1 \perp X_2 \mid Y$ and $X_2 \perp Y \mid X_1$)			
X_2 = yes				X_2 = yes				X_2 = yes			
		X_1				X_1				X_1	
		E	$\bar{E}$			E	$\bar{E}$			E	$\bar{E}$
Y	D	20	15	Y	D	15	30	Y	D	12	8
	$\bar{D}$	10	20		$\bar{D}$	20	15		$\bar{D}$	2	8
X_2 = no				X_2 = no				X_2 = no			
		X_1				X_1				X_1	
		E	$\bar{E}$			E	$\bar{E}$			E	$\bar{E}$
Y	D	40	30	Y	D	30	10	Y	D	6	4
	$\bar{D}$	15	30		$\bar{D}$	40	5		$\bar{D}$	1	4

<Table 2.2> log likelihood ratio statistics for the logit models

	Example 1	Example 2	Example 3
$L(X_1 \mid Y)$	10.0893	8.7946	6.7903
$L(X_2 \mid Y)$	**0.8400**	1.3960	0.0000
$L(X_1X_2 \mid Y)$	10.8827	8.7946	6.7903
$L(X_2 \mid X_1Y)$	**0.7934**	0.0000	0.0000

Since one could derived the logit models which correspond to log-linear models, the log likelihood ratio test statistics $L(X_2 \mid Y)$ and $L(X_2 \mid X_1Y)$ for the logit models are equivalent to the goodness-of-fit statistics for the corresponding log-linear models. Hence one can establish the following Theorem and Definition.

Theorem 2.1.

For dichotomous categorical variables Y, X_1, and X_2, the log-linear models in the following hypotheses

$$\begin{aligned} H_0 &: \log m_{ijk} = u + u_{y(i)} + u_{x_1(j)} + u_{x_2(k)} + u_{x_1x_2(jk)} \\ H_1 &: \log m_{ijk} = u + u_{y(i)} + u_{x_1(j)} + u_{x_2(k)} + u_{x_1x_2(jk)} + u_{yx_2(ik)} \end{aligned} \tag{2.1}$$

and

$$H_0' : \log m_{ijk} = u + u_{y(i)} + u_{x_1(j)} + u_{x_2(k)} + u_{yx_1(ij)} + u_{x_1x_2(jk)}$$
$$H_1' : \log m_{ijk} = u + u_{y(i)} + u_{x_1(j)} + u_{x_2(k)} + u_{yx_1(ij)} + u_{x_1x_2(jk)} + u_{yx_2(ik)} \tag{2.2}$$

are equivalent to the corresponding logit models in the following hypotheses :

$$H_0 : logit(jk) = w$$
$$H_1 : logit(jk) = w + w_2 \tag{2.3}$$

and

$$H_0' : logit(jk) = w + w_1$$
$$H_1' : logit(jk) = w + w_1 + w_2, \tag{2.4}$$

respectively, where $logit(jk) = log(m_{1jk}/m_{2jk})$, $w = 2u_{y(1)}$, $w_1 = 2u_{yx_1(1j)}$, and $w_2 = 2u_{yx_2(1k)}$.

proof :

When $m_{ijk}, i = 1, 2$ is substituted by appropriate log-linear models in (2.1) and (2.2), the corresponding logit models in (2.3) and (2.4) are easily obtained. For example, the alternative hypothesis model in (2.4) could be derived as

$$\begin{aligned} logit(jk) &= log m_{1jk} - log m_{2jk} \\ &= [u + u_{y(1)} + u_{x_1(j)} + u_{x_2(k)} + u_{yx_1(1j)} + u_{x_1x_2(jk)} + u_{yx_2(1k)}] \\ &\quad - [u + u_{y(2)} + u_{x_1(j)} + u_{x_2(k)} + u_{yx_1(2j)} + u_{x_1x_2(jk)} + u_{yx_2(2k)}] \\ &= [u_{y(1)} - u_{y(2)}] + [u_{yx_1(1j)} - u_{yx_1(2j)}] + [u_{yx_2(1k)} - u_{yx_2(2k)}] \\ &= w + w_1 + w_2. \end{aligned}$$

◇

The generalized likelihood ratio statistics to test the hypotheses (2.1) and (2.2) could be defined as the partial goodness-of-fit statistics

$$G^2(H_0|H_1) \equiv G^2(H_0) - G^2(H_1) \text{ and } G^2(H_0'|H_1') \equiv G^2(H_0') - G^2(H_1'),$$

where $G^2(H_0)$ and $G^2(H_0|H_1)$ are the goodness-of-fit statistics under certain models

$$G^2(H_0) = 2\sum_{ijk} x_{ijk} \log x_{ijk}/\hat{m}_{ijk}^{H_0} \text{ and } G^2(H_0|H_1) = 2\sum_{ijk} \hat{m}_{ijk}^{H_1} \log \hat{m}_{ijk}^{H_1}/\hat{m}_{ijk}^{H_0},$$

respectively, with the observed value x_{ijk} and the expected value $\hat{m}_{ijk}$ of a (i,j,k) cell.

As we mentioned in Section 1, the log likelihood ratio statistics $L(X_2 \mid Y)$ and $L(X_2 \mid X_1 Y)$ are the statistics for testing the hypotheses (2.3) and (2.4). And test statistics for the equivalent hypotheses (2.1) and (2.2) are well known as the partial goodness-of-fit statistics $G^2(H_0 \mid H_1)$ and $G^2(H_0'|H_1')$ (see Christensen (1990) and Agresti (1990) for more detail). Therefore, one obtains that the log likelihood ratio statistics for testing the hypotheses (2.3) and (2.4) are identical to the partial goodness-of-fit statistics for testing the hypotheses

(2.1) and (2.2), respectively, such as

$$L(X_2|Y) = G^2(H_0|H_1)$$
$$L(X_2|X_1Y) = G^2(H_0'|H_1'). \tag{2.5}$$

This coincident behavior between log likelihood ratio statistics and goodness-of-fit statistics might be confirmed at <Table 2.3> with comparing values in <Table 2.2>.

<Table 2.3> goodness-of-fit statistics for log-linear models

	Example 1	Example 2	Example 3
$G^2(H_0)$	10.8827	8.7946	6.7903
$G^2(H_1)$	10.0427	7.3986	6.7903
$G^2(H_0 \mid H_1)$	**0.8400**	**1.3960**	**0.0000**
$G^2(H_0')$	0.7934	0.0000	0.0000
$G^2(H_1')$	0.0000	0.0000	0.0000
$G^2(H_0' \mid H_1')$	**0.7934**	**0.0000**	**0.0000**

We knew that the log-linear models in the hypotheses (2.3) and (2.4) could be rewritten as

$$H_0 : [Y][X_1X_2]$$
$$H_1 : [YX_2][X_1X_2]$$
$$H_0' : [YX_1][X_1X_2]$$
$$H_1' : [YX_1][X_1X_2][YX_2].$$

Hence one obtains that the suppression condition for log-linear models might be established as the following Definition based on the condition (1.5) for logit models.

Definition 2.1.

For three dimensional contingency tables, the variable X_2 is defined as a suppressor if it satisfies the condition

$$G^2([YX_1][X_1X_2] \mid [YX_1][X_1X_2][YX_2]) < G^2([Y][X_1X_2]) \mid [YX_2][X_1X_2]). \tag{2.6}$$

The definitions of suppression in (1.5) and (2.6) could be interpreted that the suppressor variable X_2 is correlated with the variable X_1 but uncorrelated with the response Y in the log-linear model when there exists relationship between Y and X_1. In other words, the interaction terms between both X_1 and X_2, and Y and X_1 are statistically significant, but that between Y and X_2 is not significant.

Based on the definition of the collapsibility over the variable X_2 (confounder), the variable X_2 is uncorrelated to either the variable Y or the variable X_1 or both (Bishop, Fienberg, and Holland (1975, pp. 47), Agresti (1984, pp. 146), and Cristensen (1990, pp. 114) for more detail). It notes that the collapsible log-linear models with confounding over the variable X_2 are $[YX_1][YX_2], [YX_1][X_1X_2]$, and $[YX_1][X_2]$ models among three dimensional models. These models are also called as strictly and strongly collapsible defined by whittemore (1978), Durcharme and Lepage (1986), and Geng (1992). In this paper we use the general definition of the collapsibility defined by BFH 1975) etc. These $[YX_1][YX_2], [YX_1][X_1X_2]$, and $[YX_1][X_2]$ models are best fitted log-linear models to the three data sets in <Table 2.1>, respectively, and the notation $(X_1 \perp X_2 \mid Y)$, $(X_2 \perp Y \mid X_1)$, and $(X_1 \perp X_2 \mid Y \text{ and } X_2 \perp Y \mid X_1)$ for the logit models fitted to the data in Example 1 to Example 3 at <Table 2.1> are also expressed by $[YX_1][YX_2], [YX_1][X_1X_2]$, and $[YX_1][X_2]$ for the log-linear models. Hence, we can derive the fact that the categorical data satisfying the condition (2.6) are collapsible with confounding over the variable X_2, which turns out to be the suppressor variable. (see <Table 2.2 and 2.3>). Therefore, we might mention the following Theorem 2.2 :

Theorem 2.2

The three dimensional categorical data including a suppressor variable X_2 (confounder) is collapsible with confounding over the variable X_2 for the log-linear model.

proof :

Based on the definition of the collapsibility, there exists a distinct relationship between the variable Y and X_1, which is coincided with the definition of the suppression. Hence the circumstance for three dimensional log-linear model based on the relationship among the variable $Y, X_1, \text{and} X_2$ which are explained by the definition of the collapsibility over the variable X_2 is identical with the situation when the variable X_2 is suppressor for both the logit and regression models. ◇

Unfortunately, except in the trivial case when both $L(X_2 \mid X_1Y)$ and $L(X_2 \mid Y)$ equal zero in logit model, $L(X_2 \mid X_1Y)$ is not always less than $L(X_2 \mid Y)$ (Lynn, 2003). We find that this phenomenon happens to the log-linear models in <Table 2.2 and 2.3>.

3. Non-Collapsibility without Confounding

Confounding is typically presented in terms of odds ratios in epidemiology : the conditional odds ratio measuring the association between disease and confounder among the non-exposed,

and the conditional odds ratio measuring the association between exposure and confounder among the non-diseased (Kleinbaum, Kupper, and Morgenstern 1982). With these odds ratios, Boivin and Wacholder (1985) defined positive and negative confounding, and Lynn (2003) explained classical and cooperative suppressions. Based on these theories, one could check that the three data sets in <Table 2.1> satisfy all these suppression conditions.

In this section, other three data sets which do not satisfy the suppression conditions are generated and listed at <Table 3.1>, which are all non-collapsible over the variable X_2 (non-collapsibility without confounding). As we discussed in Section 2, similar results of log likelihood ratio statistics for logit and goodness-of-fit statistics for log-linear models are shown in <Table 3.2 and 3.3>.

<Table 3.1> Non-collapsibility without confounding

Example 1 $[Y][X_1X_2]$				Example 2 $[YX_2][X_1]$				Example 3 $[YX_2][X_1X_2]$			
X_2 = yes				X_2 = yes				X_2 = yes			
		X_1				X_1				X_1	
		E	$\bar{E}$			E	$\bar{E}$			E	$\bar{E}$
Y	D	17	9	Y	D	9	34	Y	D	68	25
	$\bar{D}$	21	11		$\bar{D}$	5	19		$\bar{D}$	29	11
X_2 = no				X_2 = no				X_2 = no			
		X_1				X_1				X_1	
		E	$\bar{E}$			E	$\bar{E}$			E	$\bar{E}$
Y	D	27	36	Y	D	11	44	Y	D	12	35
	$\bar{D}$	34	45		$\bar{D}$	16	62		$\bar{D}$	5	15

<Table 3.2> log likelihood ratio statistics for the logit models

	Example 1	Example 2	Example 3
$L(X_1 \mid Y)$	0.0002	0.0010	0.0039
$L(X_2 \mid Y)$	**0.0035**	9.3802	0.0011
$L(X_1X_2 \mid Y)$	0.0044	9.3831	0.0086
$L(X_2 \mid X_1Y)$	**0.0042**	9.3821	0.0047

Since these illustrated data are non-collapsible without confounding, neither (1.5) nor (2.6) condition is satisfied (see <Table 3.2 and 3.3>). In other words, the interaction term between Y and X_1 is not statistically significant in the best fitted log-linear models ($[Y][X_1X_2]$,

$[YX_2][X_1]$, and $[YX_2][X_1X_2]$ models) corresponding the data in <Table 3.1>. This coincides with the fact that we obtained in Section 2.

<Table 3.3> goodness-of-fit statistics for log-linear models

	Example 1	Example 2	Example 3
$G^2(H_0)$	0.0044	9.3855	0.0086
$G^2(H_1)$	0.0008	0.0053	0.0075
$G^2(H_0 \mid H_1)$	**0.0036**	**9.3802**	**0.0011**
$G^2(H_0')$	0.0041	9.3846	0.0047
$G^2(H_1')$	0.0000	0.0024	0.0000
$G^2(H_0' \mid H_1')$	**0.0041**	**9.3822**	**0.0047**

4. Conclusion

Previous sections explain how definition of suppression in logit model can be adapted for use in log-linear model and how suppression is related to concepts of confounding in terms of collapsibility for three dimensional categorical data.

If for three dimensional contingency table with X_2 being confounder,

$$G^2([Y][X_1X_2] \mid [YX_2][X_1X_2] > G^2([YX_1][X_1X_2] \mid [YX_1][X_1X_2][YX_2]),$$

then we find that the variable X_2 is a suppressor. And one could conclude that the data including the suppressor variable X_2 (confounder) might be collapsed over the variable X_2.

It notes that collapsible log-linear models over the confounder X_2 are $[YX_1][YX_2]$, and $[YX_1][X_2]$, $[YX_1][X_1X_2]$ models among three dimensional log-linear models. For the data explained by these collapsible log-linear models ($[YX_1][YX_2]$, and $[YX_1][X_2]$, $[YX_1][X_1X_2]$), the variable X_2 turns out to be a suppressor or confounder variable.

Since Schey (1993) and Sharpe and Roberts (1997) examined the relationship between regression sums of squares and the correlation coefficients by using geometric methods in regression model, one might study the analogus relationship about measure of associations among Y, X_1, and X_2 for further work.

References

[1] Agresti, Alan.(1984). *Analysis of Ordinary Categorical Data*, New York: John Wiley and Sons.

[2] Bishop, Yvonne M. M., Fienberg, Steve E. and Holland, Paul W.(1975). *Discrete Multivariate Analysis*, Cambridge, Massachusetts: MIT Press.

[3] Boivin, J-F. and Wacholder, S.(1975). Conditions for Confounding of the Risk Ratio and of the Odds Ratio, *American Journal of the Epidemiology*, Vol. 121, pp. 152-158.

[4] Christensen, Ronaldo.(1990). *Log-Linear Models*, New York: Springer-Verlag.

[5] Ducharme, G. R. and Lepage, Y.(1986). Testing Collapsibility in Contingency Tables, *Journal of the Royal Statistical Society*, B, 48(2), 197-205.

[6] Freund, R. J.(1988). When is $R^2 > r_{yx_1}^2 + r_{yx_2}^2$?(Revisited), *The American Statistician,* Vol. 42, pp. 89-90.

[7] Geng, Z.(1992). Collapsibility of Relative Risk in Contingency Tables with a Response Variable, *Journal of the Royal Statistical Society*, B, 54(2), 585-593.

[8] Grayson, D. A.(1987). Confounding Confounding, *American Journal of Epidemiology*, Vol. 126, pp. 546-553.

[9] Hamilton, D.(1987). Sometimes $R^2 > r_{yx_1}^2 + r_{yx_2}^2$, Correlated Variables are not Always Redundant, *The American Statistician*, Vol. 41, pp. 129-132.

[10] Hamilton, D.(1988). (Reply to Freund and Mitra), *The American Statistician*, Vol. 42, pp. 90-91.

[11] Horst, P.(1941). The Role of Prediction Variables Which are Independent of the Criterion, in *The Prediction Adjustment*, ed. P. Horst, New York: Social Science Research Council, pp. 431-736.

[12] Kleinbaum, D. G., Kupper, L. L. and Morgenstern, H.(1941). *Epidemiologic Research: Principles and Quantitative Methods*, California: Lifetime Learning Publications.

[13] Lynn, H. S.(2003). Suppression and Confounding in Action, *The American Statistician,* Vol. 57, pp. 58-61.

[14] Mitra, S.(1988). The Relationship Between the Multiple and the Zero-Order Correlation Coefficients, *The American Statistician*, Vol. 42, pp. 89.

[15] Schey, H. M.(1993). The Relationship Between the Magnitudes of $SSR(x_2)$ and $SSR(x_2 \mid x_1)$: A Geometric Description, *The American Statistician*, Vol. 47, pp. 26-30.

[16] Sharpe, N. R. and Roberts, R. A.(1997). The Relationship Among Sums of Squares, Correlation Coefficients, and Suppression, *The American Statistician*, Vol. 51, pp. 46-48.

[17] Whittemore, A. S.(1978). Collapsibility of Multidimensional Contingency Tables, *Journal of the Royal Statistical Society*, B, 40(3), 328-340.

[Received August 2004, Accepted September 2004]