

A STUDY OF MULTICOLLINEARITY EFFECT ON A BINARY RESPONSE

A. O. Adejumo and E. T. Jolayemi
Department of Statistics, University of Ilorin, Ilorin.

ABSTRACT

Regression analysis involves the study of existence of relationships between the dependent variable (the response) and sets of independent variables. If the correlation that exists between the independent variables is however high, then the estimates of the regression coefficients of these predictors are unstable and therefore unreliable. This is referred to as multicollinearity effect in this paper.

In this paper we consider the multicollinearity effect on a binary response using more than one independent variable. Without loss of generality we used the logistic function, which represents the family of monotonic functions, to model the response.

We observed that the expected response could differ sometimes remarkably but similarly monotonic when multicollinearity exists between the predictors. This fact is useful in obtaining a value of the response of the logistic function which could be greater than the values of any of the individual function when the levels of the independent variables are carefully chosen.

INTRODUCTION

In multiple regression, when independent variables are correlated, multicollinearity results of the coefficients of the regression having very high variances and therefore making them unstable. The consequence of this is that information about the individual true regression coefficients is not precise. The coefficients may be statistically insignificant even though a definite statistical relationship may exist between the dependent and a set of independent variables. In this context, independent variables that are as orthogonal as possible are desirable for the regression equation (Silvey, 1969; Neter and Wasserman, 1974).

In this paper, we study the multicollinearity effect on a binary response, which has the logistic function. The result is applicable to all monotonic functions of the binary response.

In section two we present the logistic function as an example of the family of monotonic functions of the binary response. A simulation procedure, which utilizes various levels of multicollinearity is presented in section 3. We conclude this work in section 4 presenting the results and possible applications.

LOGISTIC MODEL AND MULTICOLLINEARITY EFFECT

Assume that $y_i | x_i$ is binary and $x_i \geq 0$.

Let $\pi_i, 0 \leq \pi_i \leq 1$ be defined as $\pi_i = P(y_i = 1 | x_i)$. If π_i monotonically increases (decreases) with increase (decrease) in x_i , one of the models often considered [see for example Berkson (1953), Cox (1970), Finney (1971) and Jolayemi (1987)] is the logistic function defined as

$$\lambda_i = \ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta'x_i \quad \text{or}$$

$$\pi_i = \frac{e^{\beta'x_i}}{1 + e^{\beta'x_i}} \quad \dots\dots\dots 2.1$$

where β is a $1 \times (p+1)$ vector of unknown parameters and X is $(p+1) \times 1$ vector of predictors.

If π_i in equation 2.1 is plotted as a function of one X_i , a symmetrical stimulus-response sigmoid curve of the type shown in figures 2.1 is obtained, provided that $\beta \neq 0$. The case $\beta = 0$ serves as the equiprobability model.

The numerical value of β relates to the change in the response probability. In chemotherapy analysis the lethal dose LD₅₀ or IC₅₀, which is the dose for which the response is 50% is used to determine drug efficacy (Jolayemi and Okoro, 1995; Cobb and Church, 1982).

The parameters involved in the logistic function can be estimated in various ways, such methods include Maximum Likelihood, Least Square and Minimum Chi-square methods (Berkson, 1955; Cox, 1970; Finney, 1952; McCullagh and Nelder, 1983).

However, a comparative study of these methods conducted by Jolayemi (1995) showed that the iterative procedure using the Maximum Likelihood method is superior to others. This method is adopted in this work.

MULTICOLLINEARITY EFFECT

Assume initially that there exists two independent variables x_1 and x_2 . We further assume that multicollinearity effect can be caused by one predictor influencing the other, by changing the quantity of the other when both combine together. This is the synergistic effect found in some combined drugs, Jolayemi (1999). In this regard, assume that the correlation coefficient between x_1 and x_2 is ρ . The synergistic effect can then be said to be from only one of them: for example from x_1 where the expected additional induced quantity of x_1 , $E(x_1 | x_2)$ is given by the straight line equation (with no intercept):

$$E(X_1) = \gamma X_2 \quad \dots\dots\dots 2.2$$

where $\gamma = \rho \frac{\sigma_{x_2}}{\sigma_{x_1}}$, σ_{x_1} and σ_{x_2} are the standard deviations for the predictors X_1 and X_2

By equation 2.1

$$\begin{aligned} \lambda_i &= \beta'X_i \\ &= \beta_0 + (\beta_1 + \beta_2\gamma)x_2 \quad \dots\dots\dots 2.3 \end{aligned}$$

instead of

$$\lambda_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \dots\dots\dots 2.4$$

when ρ is assumed zero.

The estimate of the coefficients of the independent variables become unstable, so that estimate of any response varies very widely beyond expectation if ρ is not equal to zero (Neter and wassrman, 1974; Chartterjee and Price, 1977).

The focus of this paper is therefore to compare equations 2.3 and 2.4.

SIMULATION

Consider a logistic function which depends on two predictors x_1 and x_2 . Then the probability of success π_i is given by equation 2.1. Now for a given pair (x_{1i}, x_{2i}) and a fixed sample size n_i , let r_i be the number of successes which is binomially distributed with parameters (n_i, π_i) .

That is

$$r_i \sim b(n_i, \pi_i) \quad \dots\dots\dots 3.1$$

$$\text{where } \hat{\pi}_i = \frac{r_i}{n_i}$$

Suppose $\rho \neq 0$, then the expected additional quantity δ_1 for X_1 given X_2 is as provided in equation 2.1

Multicollinearity effect on a binary response

$$\delta_i = X_{1i} / X_{2i} = \rho \frac{\sigma_{x2}}{\sigma_{x1}} X_{2i} \dots\dots\dots 3.2$$

where all parameters ρ , σ_{x1} and σ_{x2} are known by careful choices of X_1 and X_2 .
The simulation procedure is as follows:

Assume that $X_{2i} \sim N(\mu, \sigma_{x2})$.

For preselected values of β_0 , β_1 and β_2 , obtain π_i using equations 3.1 and 3.2 in turns and then generate the number of successes r_{oi} and r_{ei} , which are assumed to be the observed and expected successes (Sanni and Jolayemi, 1998; Abdulkadir, 1997), respectively. For moderate values of r_{ei} ,

$$X^2 = \sum \frac{(r_{oi} - r_{ei})^2}{r_{ei}} \dots\dots\dots 4.3$$

has the chi-square distribution with $(k-1)$ degrees of freedom, where k is the number of groups spread sufficiently to have a good coverage for π_i .

For the simulation procedure in this study, we varied the values of β_1 and β_2 as follows.

Case 1: $\beta_1 \cong 0$ and $\beta_2 \cong 0$ (marginal effect of each on the response)

Case 2: $\beta_1 > 0$ and $\beta_2 \cong 0$ (only x_1 has significant effect on response)

Case 3: $\beta_1 > 0$ and $\beta_2 > 0$ (both x_1 and x_2 have significant effects)

For $\beta_1 \cong 0$ and $\beta_2 > 0$, we have the same result as in case 2. Also β is consider to be significant if its value is greater than 2.

RESULTS AND DISCUSSION

Tables 4 gives results obtained from the simulation procedure which has $\beta' = (-9, 2, 0.3)$ as an example when both predictors had marginal effects. (Figure 4.1 shows this situation when $\rho = 0$). There was no significant difference observed though the logistic functions were both monotonic. The observed logistic function from combine X_1 and X_2 had higher in values than the corresponding values for X_1 and X_2 separately. For case (ii), significant different values were observed only for very high multicollinearity: that is $|\rho| \rightarrow 1$. Monotonicity of the probabilities were also observed and in the same direction (Table 4.2 and Figure 4.2). Table 4.3 shows early discrepancy between observed and expected. That is, when the multicollinearity effect is moderate ($|\rho| > 0.3$) significant differences were identified.

CHI-SQUARE (X^2) CALCULATED WITH APPROPRIATE P-VALUE FOR MONOTONIC FUNCTION

Table 4.1: Case (1): $[\beta' = (-9, 2, 0.3)]$ Response Not Dependent on Any Predictor

$ \rho $	X^2 VALUE	P-VALUE
0.159	8.225	0.512
0.282	8.371	0.497
0.362	8.51	0.423
0.444	9.159	0.483
0.530	10.691	0.298
0.622	10.807	0.289
0.745	11.254	0.259
0.802	11.345	0.253
0.949	14.457	0.107

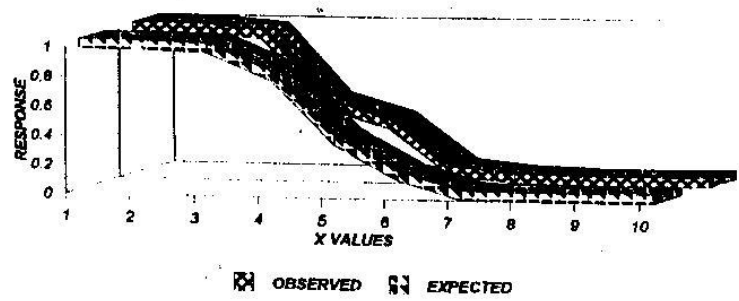


Fig. 4.1. Response not dependent on any predictor
($B_0 = 9$, $B_1 = 2$, $B_2 = 0.3$)

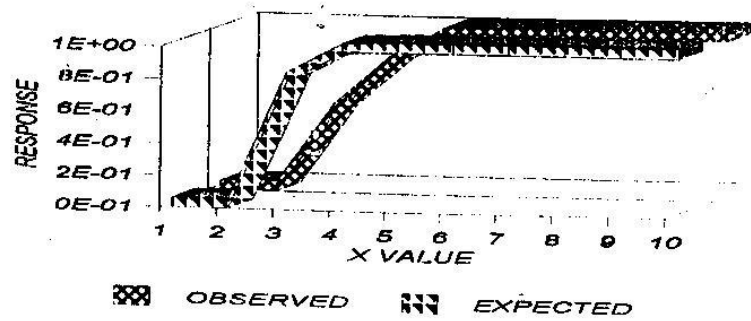


Fig. 4.2. Response not dependent on any predictor
($B_0 = 15$, $B_1 = 5.2$, $B_2 = 1$)

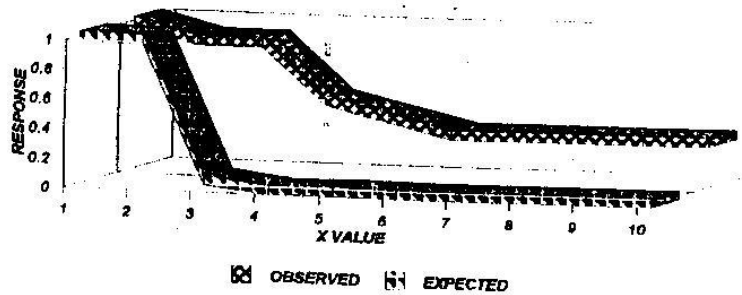


Fig. 4.3. Response not dependent on any predictor
($B_0 = 42$, $B_1 = 5$, $B_2 = 5$)

Multicollinearity effect on a binary response

Table 4.2: Case (I): $|\beta'| = (-15, 5.2, 1)$ Response Dependent on Only One Predictor

$ \rho $	X ² VALUE	P-VALUE
0.159	2.36	0.984
0.282	4.39	0.883
0.362	4.53	0.873
0.444	7.120	0.624
0.530	9.54	0.388
0.622	10.47	0.313
0.745	11.5	0.243
0.802	11.56	0.239
0.949	52.4	5.36E-07

Table 4.3: Case (II): $|\beta'| = (-14, 3.5, 3.5)$ Response Dependent on Both Predictors

$ \rho $	X ² VALUE	P-VALUE
0.17	12.57	0.183
0.20	13.42	0.145
0.37	16.02	0.0664
0.45	17.51	0.0414
0.53	18.91	0.0259
0.61	22.42	0.00764
0.72	23.88	0.00450
0.80	24.92	0.00306
0.94	26.95	0.00143

CONCLUSION

In conclusion, departure from what was expected in a multiple logistic model was found to be dependent on the significance of the predictor variables with the binary response as well as the level of the multicollinearity. The response was unaffected by any level of multicollinearity if none of the predictors has any appreciable effect on the response. However, when only one of the predictors influences the response in question, only a very high level of multicollinearity influences the response. In contrast, it was observed that if the two predictors individually influence the response, a very slight level of multicollinearity resulted in significant departure from the expected response. The consequence of these results is that two or more predictors that individually significantly influence the binary response (using the logistic function) may be carefully combined to produce a higher response against any of the individual predictors. This could also be higher than the expected when the reaction is not synergistic. This has immediate application in pharmacology where two drugs can be carefully combined to produce effects higher than expected.

REFERENCES

- Abdulkadir, S. S. (1997). On the Approximation of Pearson's X^2 by the chi-square distribution. A M. Sc. thesis submitted to Department of Statistics, Unilorin.
- Berkson, J. (1953). A Statistically precise and simple method of estimating the bio-assay with quantal response based on the logistic function. *Journal of Amer. Stat. Association*; 48: 456-598.
- Berkson, J. (1955). Maximum Likelihood and Minimum X^2 estimate of the logistic function. *Journal of Amer. Stat. Association*; 50:130-162.
- Chatterjee, S. and Price, B. (1977). *Regression Analysis by Example*. New York, Wiley.
- Cobb, E. B. and Church, J. D. (1982). Small Sample Quantal Response Methods for estimating the location

- parameter for Response curve. *Journal of Amer. Stat. Association*; **78:99-107**.
- Cox, D. R. (1970). Analysis of Binary Data. Chapman and Hall, London.
- Finney, D. J. (1971). Statistical Method in Biological Assay. 2nd edition ; N. Y. Hafner.
- Jolayemi, E. T. (1987). A Monte Carlo Study of the method of estimating parameters of the logistic model. *Nig. Jour. Pure and Applied Sciences*.
- Jolayemi, E. T. and Okoro, E. O. (1995). On the estimation of mean IC_{50} . *Bioscience Res. Comm.*; **7: 175-178**.
- Jolayemi, E. T. (1999). Drug Interaction Effects when Response has the Logistic Function. 6th SUSAN'99 International Biometricians Society (IBS) Conference, Ibadan.
- McCullagh, P. and Nelder, J. A. (1983). Generalized Linear Models. Chapman and Hall, London.
- Neter, J. and Wasserman, W. (1974). Applied linear statistical models.. Homewood, Illinois, Irwin.
- Sanni, O. O. M. and Jolayemi, E. T. (1998). On the use of some categorical test statistics on Sparse contingency table. *Jour. of Pure and Applied Science*. University of Ilorin.
- Silvey, S. D. (1969). Multicollinearity and imprecise estimation. *Jour. Royal Stat. Society*; **31:539-552**.