Effects of Cross-Product Ratio upon Prediction among Three Binary Variables

Sung-Ho Kim[1]

We show that when the three-way association level among the three binary variables, X, Y, Z is fixed, $D_P = P(X = 1|Y = 1) - P(X = 1|Y = 0)$ increases as the cross-product ratio (≥ 1) of Y and Z increases under the assumption that X is positively associated with Y and Z. We then discuss some implications of this property.

KEY WORDS: Conditional probability; discriminating ability; positive association.

1 Introduction and problem

The cross-product ratio (cpr) is a basic measure of association in 2×2 tables, and many of the measures of association for 2×2 tables as discussed in Goodman and Kruskal (1954, 1959, 1963, 1972) are simply monotone functions of the cross-product ratio. In addition to that, the cpr has been widely used in educational, genetic, and medical contexts during the last 50 years (Holland and Thayer, 1988; Kimura, 1965; Cornfield, 1956). The association measure for 2^k ($k \geq 2$) tables can also be represented in the form of nested ratios of conditional cpr's (Bishop et al., 1975). An example for 2^3 tables is given in expression (1). While an association measure for 2^k ($k \geq 2$) tables gives us an overall picture of association among the k variables involved, it is not clear how it is related to prediction for a variable given the values of the other variables. We will investigate this problem in this article.

Consider three binary variables, U_1, U_2, and X, and suppose that we are interested in predicting for U_1 given an outcome of X. Assuming that all the variables are binary taking on 0 or 1, we will explore how the cpr between U_1 and U_2 is related with

$$D_P = P(X = 1|U_1 = 1) - P(X = 1|U_1 = 0).$$

D_P may be regarded as a measure of discrimination between the two events, $X = 1$ and $X = 0$, based on the information from U_1. The larger the value of D_P becomes, the better discriminator U_1 becomes. This discriminating ability may change according to the association level of the two U variables.

[1]Division of Applied Mathematics, Korea Advanced Institute of Science and Technology, Daejeon, 305-701, South Korea. e-mail: Sung-Ho.Kim@kaist.ac.kr

The three-way interaction measure among U_1, U_2, and X can be expressed as

$$\frac{P_{X|U_1U_2}(1|1,1)P_{X|U_1U_2}(0|0,1)}{P_{X|U_1U_2}(1|0,1)P_{X|U_1U_2}(0|1,1)} \Bigg/ \frac{P_{X|U_1U_2}(1|1,0)P_{X|U_1U_2}(0|0,0)}{P_{X|U_1U_2}(1|0,0)P_{X|U_1U_2}(0|1,0)}. \tag{1}$$

This expression implies that the three-way interaction remains the same as long as the conditional probability $P_{X|U_1U_2}$ remains the same. Suppose that we have two probability distributions Q and R for U_1, U_2, X. It may sound reasonable that if the three-way interaction is the same between the two probability distributions, then $Q_{X|U_1} = R_{X|U_1}$. But it is not true in general, as we will see below.

To see how the association level between the U variables affect the discriminating ability, we will consider some restriction on the probability models as follows:

(i) $Q(X = 1|U_1 = u_1, U_2 = u_2) = R(X = 1|U_1 = u_1, U_2 = u_2)$, for every u_1 and u_2.

(ii) $Q(U_i = 1) = R(U_i = 1)$, $i = 1, 2$.

Conditions (i) and (ii) mean that the conditional probability of X conditional on the U variables are the same between the two models and so are the marginals of the U variables. The only possible differences between the models are on the association level between the two U variables and the marginal on X.

2 A theorem

We will now prove a theorem which shows a direct relation between the discriminating ability D_P and the association level among the U variables.

Theorem 1 *Consider a model of three binary variables, U_1, U_2, and X for which conditions (i) and (ii) are satisfied and the three variables are three-way interactive. Also assume that, whenever $u_1 \geq u_1'$ and $u_2 \geq u_2'$,*

$$Q_{X|U_1,U_2}(1|u_1, u_2) \geq Q_{X|U_1,U_2}(1|u_1', u_2') \tag{2}$$

and similarly for $R(\cdot)$. Then

(a) $$D_Q \geq D_R \text{ if and only if } cpr^Q_{U_1U_2} \geq cpr^R_{U_1U_2}, \tag{3}$$

where $cpr^Q_{U_1U_2}$ ($cpr^R_{U_1U_2}$) is the cross-product ratio of U_1 and U_2 with the probability distribution Q (R).

(b) For each value of $P_{X|U_1}(1|0)$,

$$cpr^Q_{XU_1} \geq cpr^R_{XU_1} \text{ if and only if } cpr^Q_{U_1U_2} \geq cpr^R_{U_1U_2}. \tag{4}$$

Equality holds simultaneously in (3) and (4).

Proof: First, we will prove the sufficiency of (a). After a simple algebra, we have

$$\begin{aligned} D_Q - D_R &= \sum_{u_2} Q_{X|U_1,U_2}(1|1,u_2)) \left\{ Q_{U_2|U_1}(u_2|1) - R_{U_2|U_1}(u_2|1) \right\} \\ &\quad - \sum_{u_2} Q_{X|U_1,U_2}(1|0,u_2)) \left\{ Q_{U_2|U_1}(u_2|0) - R_{U_2|U_1}(u_2|0) \right\} \\ &= \left\{ Q_{U_2|U_1}(1|1) - R_{U_2|U_1}(1|1) \right\} \left\{ Q_{X|U_1,U_2}(1|1,1) - Q_{X|U_1,U_2}(1|1,0) \right\} \\ &\quad + \left\{ R_{U_2|U_1}(1|0) - Q_{U_2|U_1}(1|0) \right\} \left\{ Q_{X|U_1,U_2}(1|0,1) - Q_{X|U_1,U_2}(1|0,0) \right\}, \end{aligned} \tag{5}$$

where the first equality follows from condition (i).

The inequality $cpr^{Q}_{U_1U_2} > cpr^{R}_{U_1U_2}$ is equivalent to that

$$\frac{Q_{U_2|U_1}(1|1)/Q_{U_2|U_1}(1|0)}{R_{U_2|U_1}(1|1)/R_{U_2|U_1}(1|0)} > \frac{Q_{U_2|U_1}(0|1)/Q_{U_2|U_1}(0|0)}{R_{U_2|U_1}(0|1)/R_{U_2|U_1}(0|0)} \tag{6}$$

Under condition (ii), inequality (6) leads to

$$Q_{U_2|U_1}(1|1) > R_{U_2|U_1}(1|1). \tag{7}$$

By condition (ii) and inequality (7), we have

$$Q_{U_2|U_1}(1|1) - R_{U_2|U_1}(1|1) > 0 \quad \text{and} \quad R_{U_2|U_1}(1|0) - Q_{U_2|U_1}(1|0) > 0.$$

Therefore, from the assumption (2) of the theorem, we have that $D_Q - D_R > 0$. Note that equality in (6) leads to, by condition (ii),

$$Q_{U_2|U_1}(1|1) - R_{U_2|U_1}(1|1) = 0 \quad \text{and} \quad R_{U_2|U_1}(1|0) - Q_{U_2|U_1}(1|0) = 0, \tag{8}$$

from which follows that $D_Q - D_R = 0$.

Now we will prove the necessity of (a). We can rewrite the inequality $D_Q - D_R > 0$, from (5), as

$$\begin{aligned} &\left\{ Q_{U_2|U_1}(1|1) - R_{U_2|U_1}(1|1) \right\} \left\{ Q_{X|U_1,U_2}(1|1,1) - Q_{X|U_1,U_2}(1|1,0) \right\} \\ &\quad > \left\{ R_{U_2|U_1}(1|0) - Q_{U_2|U_1}(1|0) \right\} \left\{ Q_{X|U_1,U_2}(1|0,1) - Q_{X|U_1,U_2}(1|0,0) \right\}. \end{aligned} \tag{9}$$

Note that this inequality is possible, under condition (ii), the three-way interactivity, and assumption (2), only when

$$Q_{U_2|U_1}(1|1) > R_{U_2|U_1}(1|1).$$

This inequality implies inequality (6) under condition (ii). To see if equality holds simultaneously, suppose that inequality (9) becomes an equality. By the three-way interactivity, at most one of $Q_{X|U_1,U_2}(1|1,1) - Q_{X|U_1,U_2}(1|1,0)$ and $Q_{X|U_1,U_2}(1|0,1) - Q_{X|U_1,U_2}(1|0,0)$ is equal to 0. If neither of them equals zero, the equality in (9) is not guaranteed because of condition (ii). If one of them equals zero, expression (8) must hold by condition (ii). This means that equality holds simultaneously in expression (3).

The proof of (b) is immediate from result (a). The cpr_{XU_1} is the same whether it is expressed in terms of $P_{X|U_1}$ or $P_{U_1|X}$, that is,

$$\frac{P_{X|U_1}(1|1)/P_{X|U_1}(1|0)}{P_{X|U_1}(0|1)/P_{X|U_1}(0|0)} = \frac{P_{U_1|X}(1|1)/P_{U_1|X}(1|0)}{P_{U_1|X}(0|1)/P_{U_1|X}(0|0)}. \tag{10}$$

And that, if we let $d = P_{X|U_1}(1|1) - P_{X|U_1}(1|0)$ and $a = P_{X|U_1}(1|0)$, it follows that

$$cpr_{XU_1} = \frac{1 + d/a}{1 - d/(1-a)}. \tag{11}$$

For each value of a, cpr_{XU_1} is strictly increasing in d. Therefore, result (b) follows from result (a), and so the equality holds simultaneously in (4). $\square$

Expression (3) means that, as long as $P_{X|U_1}(1|1) > P_{X|U_1}(1|0)$, the difference between these two values increases as $cpr^P_{U_1U_2}$ increases. Expression (3) is restated in terms of cpr_{XU_1} in expression (4). This theorem shows us the relationship between cpr_{XU_1} and $cpr_{U_1U_2}$ under the assumption that X is positively associated with U_1 and U_2.

Under condition (2), the U variables in the theorem are not necessarily positively associated. Those who are interested in a detailed description on the notion of positive association among categorical variables and its implications are referred to Holland and Rosenbaum (1986) and Junker and Ellis (1997).

The relation between cpr_{XU_1} and D_P as expressed in (11) is illustrated in Table 1, which is obtained with $P(U_1 = 1)$ fixed to 0.5. The table displays the $cpr_{U_1U_2}$ values and the D_P values. We can see that the D_P values increase as the cpr values increase. We consider only the cases where U_1 and U_2 are independent and positively associated, because such cases are meaningful in many real world problems in educational testing and medicine (Mislevy 1994; Kim 2003). We can expect a similar result when the association between the U variables is negative since U_1 and $1 - U_2$ are then positively associated. Figure 1 is a graphic display of the rows of Table 1 where $P(U_2 = 1) = 0.6$. Graphs for the other rows are ignored since they show similar patterns of monotone increase.

We have considered the relation between cpr_{XU_1} and $cpr_{U_1U_2}$ so far, but we can have a similar result as for the relationship between cpr_{XU_2} and $cpr_{U_1U_2}$ by exchanging U_1 and U_2 in the above argument.

Table 1: $cpr_{U_1U_2}$ and D_P values with $P(U_1 = 1)$ fixed to 0.5.

$P(U_2 = 1)$											
0.3	cpr	1.00	1.33	1.73	2.20	2.76	3.40	4.15	5.01	6.00	7.13
	D_P	0.28	0.39	0.41	0.43	0.44	0.46	0.47	0.48	0.49	0.50
0.4	cpr	1.00	1.40	1.90	2.51	3.27	4.18	5.27	6.57	8.11	9.91
	D_P	0.34	0.44	0.46	0.49	0.50	0.52	0.54	0.55	0.56	0.57
0.5	cpr	1.00	1.49	2.16	3.04	4.20	5.70	7.64	10.12	13.29	17.33
	D_P	0.40	0.49	0.52	0.54	0.57	0.59	0.60	0.62	0.64	0.65
0.6	cpr	1.00	1.40	1.90	2.51	3.27	4.18	5.27	6.57	8.11	9.91
	D_P	0.46	0.52	0.55	0.57	0.59	0.60	0.62	0.63	0.64	0.66
0.7	cpr	1.00	1.33	1.73	2.20	2.76	3.40	4.15	5.01	6.00	7.13
	D_P	0.51	0.56	0.58	0.60	0.61	0.62	0.63	0.65	0.66	0.66
0.8	cpr	1.00	1.28	1.62	2.00	2.43	2.93	3.50	4.14	4.86	5.67
	D_P	0.57	0.60	0.62	0.63	0.64	0.65	0.65	0.66	0.67	0.67
0.9	cpr	1.00	1.25	1.53	1.85	2.21	2.61	3.07	3.57	4.14	4.77
	D_P	0.63	0.65	0.66	0.66	0.67	0.67	0.68	0.68	0.68	0.69

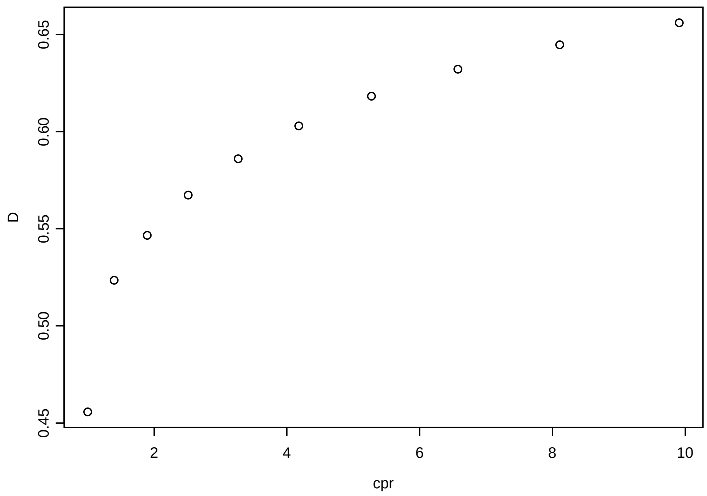

Figure 1: Plot of the $cpr_{U_1U_2}$ and D_P values for the case that $P(U_2 = 1) = 0.6$ in Table 1

3 Discussion

Conditions (i) and (ii) lead us to consider a model where the state of X is influenced by U_1 and U_2. In particular, when the ratio in expression (1) is equal to one, we have

$$logit(u_1, u_2) = \log \frac{P(X = 1|u_1, u_2)}{P(X = 0|u_1, u_2)} = \alpha + \beta_1 u_1 + \beta_2 u_2.$$

Note in (11) that $cpr_{U_1X} = 1$ if and only if $D_P = 0$. It thus follows from Theorem 1 that cpr_{U_1X} increases beyond 1 as $cpr_{U_1U_2}$ increases, as long as $D_P > 0$. We can have an analogous result when the ratio in (1) is not equal to 1, in which case the logit is given by

$$logit(u_1, u_2) = \alpha' + \beta_1' u_1 + \beta_2' u_2 + \beta_3' u_1 u_2.$$

From a regression point of view, we can say that whether the conditional probability of X depends upon the cross-product term of U_1 and U_2 or not has nothing to do with that the cpr_{U_1X} increases beyond 1 as the $cpr_{U_1U_2}$ increases as long as $D_P > 0$.

Inequality (2) does not imply that X, U_1, U_2 are positively associated each other. X and U_1 can be negatively associated and also the pair of X and U_2. This is an instance of the Simpson's paradox (Fienberg 1980, p. 45). However, if

$$X, U_1, U_2 \quad \text{are positively associated with each other,} \tag{12}$$

then under condition (2),
(a') the measure of discrimination D_P is non-negative and increases as U_1 and U_2 become more highly associated; and
(b') X and U_1 are positively associated and their association level increases as U_1 and U_2 become more highly associated.

(b') means that X becomes more informative for U_1 as $cpr_{U_1U_2}$ increases. This contributes to the stability of the prediction for U_1 which is made in terms of $P(U_1 = 1|X = x)$. This has to do with the robustness of classification for binary variables where the conditional probability that a binary variable takes on 1 given observed values of a set of binary variables is categorized and predictions are made in terms of the category level. When the predicted binary variables are more associated among themselves, the category levels for individual predicted variable become more stable. This phenomenon is found under a more general setting in Kim (2003).

The positive association among a set of variables is a common phenomenon in educational testing since abilities, knowledge units, and item scores are in general causally related or positively associated among themselves (Mislevy 1994; Tatsuoka 1990). In educational or medical testing, it is often the case that we build a causal model where some or most of the causal variables in the model are unobservable and most of the observables are on the effect side. Given the values of the observed variables, predictions are often made for causal or unobserved variables. According to the result of section 2, it is desirable that, when we want to predict for causal variables, we avoid any pair of independent variables that are causal to the same observable variable and at the same time try to have a set of variables that are highly associated each other as causal to one and the same variable. This point seems to be very important in designing an educational test or a medical examination among others that a set of variables which are causal to an observable variable be associated highly each other when we are interested in predicting for individual causal or unobservable variables.

Finally, we let $d' = P_{U_1|X}(1|1) - P_{U_1|X}(1|0)$, $a' = P_{U_1|X}(1|0)$, and $D'_P = P(U_1 = 1|X = 1) - P(U_1 = 1|X = 0)$. Since cpr_{XU_1} can be expressed in terms of either $P_{X|U_1}$ or $P_{U_1|X}$ as in (10), we have from (4) and

$$cpr_{XU_1} = \frac{1 + d'/a'}{1 - d'/(1 - a')},$$

that, for each pair of the values a, a',

$$D'_Q \geq D'_R \text{ if and only if } \quad cpr^Q_{U_1U_2} \geq cpr^R_{U_1U_2}.$$

Under the assumption (12), the D' values are not negative. So as the $cpr_{U_1U_2}$ increases, the difference between $P(U_1 = 1|X = 1)$ and $P(U_1 = 1|X = 0)$ gets larger, improving the overall prediction accuracy for U_1.

References

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA.: MIT Press.

Cornfield, J. (1956). A statistical problem arising from retrospective studies. In J. Neyman (Ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4, pp. 135-148. Berkeley, University of California Press.

Fienberg, S. E. (1980). *The Analysis of Cross-Classified Categorical Data.* MIT Press, Cambridge, MA.

Goodman, L.A. and Kruskal, W.H. (1954). Measures of association for cross-classifications, *Journal of The American Statistical Association*, **49**, 732-764.

Goodman, L.A. and Kruskal, W.H. (1959). Measures of association for cross-classifications, II: Further discussion and references, *Journal of The American Statistical Association*, **54**, 123-163.

Goodman, L.A. and Kruskal, W.H. (1963). Measures of association for cross-classifications, III: Approximate sampling theory, *Journal of The American Statistical Association*, **58**, 310-364.

Goodman, L.A. and Kruskal, W.H. (1972). Measures of association for cross-classifications, IV: Simplification of asymptotic variances, *Journal of The American Statistical Association*, **67**, 415-421.

Holland, P.W. and Rosenbaum, P.R. (1986). Conditional association and unidimensionality in monotone latent variable models, *The Annals of Statistics*, **14**, 4, 1523-1543.

Holland, P.W. and Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Junker, B.W. and Ellis, J.L. (1997). A characterization of monotone unidimensional latent variable models, *The Annals of Statistics*, **25**, 3, 1327-1343.

Kim, S.-H. (2004). Stochastic ordering and robustness in classification from a Bayesian network, *Decision Support Systems*. To appear.

Kimura, M. (1965). Some recent advances in the theory of population genetics. *Japanese Journal of Human Genetics*, **10**, 43-48.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika*, **59**, 4, 439-483.

Tatsuoka, K. K. (1990). Toward an integration of item response theory and cognitive error disgnosis. In N. Frederiksen, R. Glaser, A. Lesgold, and M. G. Shafto, (Eds.), *Diagnostic Monitoring of Skill and Knowledge Acquisition* (pp. 453-488). Hillsdale, Jew Jersey: Erlbaum.