MEASURES OF ASSOCIATION BETWEEN A DICHOTOMOUS AND A CONTINUOUS VARIABLE

Janet Dixon Elashoff, Stanford University

Introduction How should the degree of association between a dichotomous variable and a continuous variable be measured? The usual answer is to use the point biserial correlation coefficient. This coefficient, however, is specially designed for the case in which the conditional distribution of y, the continuous variable, given the value of x, the dichotomous variable, is normal, and the mean of the conditional distribution of y depends on x but the variance does not.

Goodman and Kruskal (1954) have argued persuasively that a measure of association for cross classifications should be chosen with a particular underlying model and a purpose in mind. Many different models could be proposed to describe a relationship between a dichotomous and a continuous variable; two general models will be discussed here. A measure of association might be examined with many different purposes in mind. In this paper, some measures of association are suggested which are appropriate for the purpose of screening y variables for use in predicting x. That is, we propose measures of association appropriate for the classification problem.

The basic model to be discussed is one in which the x variable takes on the values 1 and 2 with probabilities (1-p) and p respectively. The distribution of y given x is $F_x(y)$. The problem is to decide how useful the y variable would be in assigning new individuals to x categories, given n_1 observations with x = 1 and n_2 observations with x = 2, with $N = n_1 + n_2$ and observations y_{ij}, $i = 1,2$ and $j = 1,\ldots n_i$. We discuss two situations: (1) $F_1(y)$ and $F_2(y)$ differ only in the mean; (2) $F_1(y)$ and $F_2(y)$ may have different variances as well as different means.

Model 1 $F_1(y)$ and $F_2(y)$ differ only in the mean. If normality is assumed, the point biserial correlation coefficient, ρ, is appropriate. The probability of misclassification using y is a function of Δ, the distance between $F_1(y)$ and $F_2(y)$, where

$$(1) \qquad \Delta = \frac{\mu_2 - \mu_1}{\sigma}$$

and ρ is a function of Δ,

$$(2) \qquad \rho = \Delta\sqrt{\frac{p(1-p)}{1 + p(1-p)\Delta^2}} \ .$$

The maximum likelihood estimator of ρ is

$$(3) \qquad r_{pb} = \hat{\Delta}\sqrt{\frac{\hat{p}(1-\hat{p})}{1 + \hat{p}(1-\hat{p})\hat{\Delta}^2}}$$

where

$$(4) \qquad \hat{p} = n_2/N$$

$$(5) \qquad \hat{\Delta} = \frac{N(\bar{y}_2 - \bar{y}_1)}{\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j}-\bar{y}_2)^2} \ .$$

Conditional on n_1 and n_2, a test of $\rho = 0$ can be based on the usual t test;

$$(6) \qquad t = \frac{r_{pb}\sqrt{N-2}}{\sqrt{1-r_{pb}^2}}$$

has a t distribution with N-2 degrees of freedom when $\rho = 0$.

When Model 1 is true, but we are unwilling to assume normality, it is natural to consider nonparametric classification procedures based on ranks. Das Gupta (1964) suggests a classification procedure based on the sample cumulative distribution function. Define

$$(7) \qquad \hat{F}_i(a) = c_i/n_i$$

where c_i is the number of observations $y_{ij} \leq a$. Let y' be an observation to be classified. Then assign an individual to category x = 1 if

$$(8) \qquad \left|F_1(y') - \frac{1}{2}\right| < \left|F_2(y') - \frac{1}{2}\right| \ .$$

Using this classification procedure, the probability of misclassification is a function of π, where

$$(9) \qquad \pi = P(\ y_2 > y_1\ )$$

is the probability that a y observation from x = 2 is larger than a y observation from x = 1. For a dichotomous and a continuous variable, Goodman and Kruskal's measure of association γ reduces to

$$(10) \qquad \gamma = 2\pi - 1 \ .$$

The Mann-Whitney U statistic provides an estimator of π and

$$(11) \qquad \hat{\gamma} = 2U/n_1n_2 - 1 \ .$$

Conditional on n_1 and n_2, a test of $\gamma = 0$ can be made using tables for the U statistic.

Another method of classification using ranks was developed by Stoller (1954) for the situation where $F_x(y)$ is absolutely continuous and the optimal discrimination rule consists of classifying an individual into category 1 if $y \leq a^*$ and into category 2 otherwise. The probability of a correct classification using any cutoff point a is

$$(12) \qquad Q(\ a\ ) = (1-p)F_1(a) + p(1 - F_2(a))$$

and a natural measure of association is

$$(13) \quad \lambda_1 = 1 - \frac{P(\text{ misclassification}\,|\,y\text{ known })}{P(\text{ misclassification}\,|\,y\text{ unknown })} = \frac{Q(\ a^*\ ) - m}{1 - m}$$

where $\qquad m = \max(p,(1-p))$.

A distribution-free estimate of Q(a) for any a is obtained by substituting $\hat{p} = n_2/N$ and

$\hat{F}_1(a)$ in (12) to obtain

$$(14) \qquad \hat{Q}(\ a\ ) = \frac{1}{N}(n_2 + c_1 - c_2)\ .$$

The point a* is estimated using the point a for which $\hat{Q}(\ a\ )$ is maximized. Thus letting

$$(15) \qquad d^+ = \max_a(c_1 - c_2)$$

$$(16) \qquad \hat{Q}(\ a^*\ ) = \frac{1}{N}(n_2 + d^+)\ .$$

If it is not known a priori whether $\mu_2 > \mu_1$ or $\mu_2 < \mu_1$, the rule can be extended by letting

$$(17) \qquad \begin{aligned} d^- &= \max_a(c_2 - c_1) \\ d &= \max(d^+, d^-) \end{aligned}$$

and defining

$$(18) \qquad \hat{Q}(\ a^*\ ) = \begin{cases} \dfrac{n_2 + d}{N} & d^+ > d^- \\ \dfrac{n_1 + d}{N} & d^+ < d^- \end{cases}$$

and

$$(19) \qquad \hat{\lambda}_1 = \frac{\hat{Q}(\ a^*\ ) - \hat{m}}{1 - \hat{m}}\ .$$

This derivation assumes that we want to estimate p from the sample at hand. However, if $n_1 = n_2$, or if we can assume p = .5, the formulation is simplified and the distribution theory is known. Define

$$(20) \qquad \begin{aligned} D^+ &= \max_a(\hat{F}_1(a) - \hat{F}_2(a)) \\ D^- &= \max_a(\hat{F}_2(a) - \hat{F}_1(a)) \\ D &= \max(D^+, D^-)\ ; \end{aligned}$$

these are the well-known Kolmogorov-Smirnov statistics. Then

$$(21) \qquad \hat{\lambda}_1 = D$$

and a test of $\lambda_1 = 0$ conditional on n_1 and n_2 can be based on tables of the D statistic.

Some general properties of these three measures of association are obvious. The measures ρ and γ range from -1.0 to +1.0 while λ_1 must lie between 0 and +1.0. For fixed $F_1(y)$ and $F_2(y)$, γ is unaffected by the value of p, but ρ and λ_1 decrease as $|p - .5|$ increases. The estimator $\hat{\lambda}_1$ can be expected to have a positive bias. The measure $\hat{\rho}$ can be expected to be much more strongly affected by the presence of outliers.

The behavior of these association measures is illustrated in two examples. Example 1 is calculated on the data shown in Table 1 which is generated by a normal shift model with $\Delta = 1$. The estimates are $r_{pb} = .43$, $\hat{\gamma} = .55$, and $\hat{\lambda}_1 = .67$; all are significantly different from zero at the 5% level; note that r_{pb} has the smallest and $\hat{\lambda}_1$ the largest value. For example 2, the data from Table 1 was used again, except that the largest observation in category 1 was changed from 7.6 to 27.6. For example 2, the estimates are $r_{pb} = .34$, $\hat{\gamma} = .46$, and $\hat{\lambda}_1 = .60$; both $\hat{\gamma}$ and $\hat{\lambda}_1$ are still significant at the 5% level. Note that the effect of the outlier on $\hat{\lambda}_1$ was considerably smaller than on r_{pb} and $\hat{\gamma}$.

TABLE 1. DATA GENERATED FROM A NORMAL DISTRIBUTION WITH $\sigma^2 = 100$

$\mu_1 = 0$	$\mu_2 = 10$
-16.9	-19.0
-12.2	-9.9
-6.7	-5.8
-3.4	0.7
-2.1	1.6
-1.0	8.1
-0.9	10.4
-0.8	10.7
0.7	10.8
1.5	11.0
1.8	11.1
2.9	12.8
3.7	21.5
3.9	22.1
7.6	26.0

These three measures of association provide reasonable and interpretable measures of association for the classification problem where only a difference in means is of interest. But what about the situation in which the variances may differ also?

Model 2 The conditional distributions $F_1(y)$ and $F_2(y)$ may differ in variance as well as in mean. If normality is assumed, the probability of misclassification using a quadratic classification rule is a rather messy function of the means and variances which suggests no simple overall measure. As an ad hoc two-stage procedure, one could examine r_{pb} first and if it were not found to be significant, take a look at the F statistic.

A one-stage procedure can be obtained by extending the Stoller classification procedure to a rule in which an observation is assigned to category 1 if $y \leq a_1^*$, or $y > a_2^*$. Again, the cutoff points a_1^* and a_2^* are estimated by maximizing the estimated probability of a correct classification and the measure λ is used. When p is estimated from the data,

$$(22) \qquad \hat{\lambda}_2 = 1 - \frac{(a - d^+ - d^-)}{N(1 - m)}$$

where

$$(23) \qquad a = \begin{cases} n_2 & \hat{a}_1^* < \hat{a}_2^* \\ n_1 & \text{otherwise} \end{cases}\ .$$

For $p = .5$

$$\text{(24)} \qquad \hat{\lambda}_2 = D^+ + D^-$$

where D^+ and D^- are given in (20). Using definition (24), the distribution of $\hat{\lambda}_2$ conditional on n_1 and n_2 is given by Gnedenko (1954).

In examples 1 and 2, where $\hat{\lambda}_2 = .73$, $\hat{\lambda}_2 = .67$ respectively, $\hat{\lambda}_2$ is only slightly larger than $\hat{\lambda}_1$. It is larger than $\hat{\lambda}_1$ because of the -19.0 observation in category 2 which is smaller than all the observations in category 1. In small samples like these, $\hat{\lambda}_2$ will be overly sensitive to one observation.

Example 3 has been calculated on the data shown in Table 2 which was generated by a normal model with $\mu = 0$ and $\sigma_1 = 10$, $\sigma_2 = 40$. The estimates of the association measures are $r_{pb} = .19$, $\hat{\gamma} = .08$, $\hat{\lambda}_1 = .40$, $\hat{\lambda}_2 = .73$. The estimates of ρ and γ are small. Although not significant at the 5% level, $\hat{\lambda}_1$ is fairly sizeable by the standards one is used to with measures of association. Of course, if the variances are quite different, one could often expect to do better even with a one-sided classification rule than would be obvious from examination of a difference in means. The estimated $\hat{\lambda}_2$ is significant at the 5% level.

TABLE 2. DATA GENERATED FROM A NORMAL DISTRIBUTION WITH $\mu = 0$

$\sigma_1 = 10$	$\sigma_2 = 40$
-25.1	-57.0
-13.0	-50.6
-10.9	-23.6
-6.6	-17.7
-6.1	-15.4
-4.3	-14.0
-1.5	-10.6
-1.3	6.6
0.1	7.6
3.2	28.0
3.8	42.8
4.0	56.4
9.0	59.0
10.3	61.2
13.5	69.2

The measure $\hat{\lambda}_2$ would seem to be a useful measure of association for a dichotomous and a continuous variable for the classification problem in which means, variances, or both may differ.

Additional research to determine large and small sample properties of these sample measures of association for specific choices of $F_1(y)$ and $F_2(y)$ is underway. Investigation of other models for the relationship between dichotomous x and continuous y and other problems requiring a measure of association should lead to alternative measures.

REFERENCES

Das Gupta, Somesh (1964). "Non-parametric classification rules." Sankhyā, Series A, 26, 25-30.

Gnedenko, B. V. (1954). "Kriterien für die Unveränderlichkeit der Wahrscheinlichkeitsverteilung von zwei unabhängigen Stichprobenreihen." Mathematisches Nachrichten, 12, 29-66.

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

Stoller, David S. (1954). "Univariate two-population distribution-free discrimination." Journal of the American Statistical Association, 49, 770-777.