Users of regression analysis are well schooled in procedures for "shrinking" or adjusting the least squares sample value of $R^2$ which is known to be a biased overestimate of the population value of $R^2$ and that sample values of $R^2$ should be corrected for "shrinkage" as a routine part of the data analysis (Olkin & Pratt, 1958; Pedhazur, 1982; Wherry, 1931). It seems to be far less well known that sample measures of strength of association derived from $\chi^2$ tests of contingency tables are equally subject to overfitting. A brief perusal of several introductory, intermediate and advanced statistics textbooks popular in the behavioral sciences reveals that most, if not all, authors make no mention of the potential bias in measures of strength of association associated with $\chi^2$ tests (e.g., Glass & Stanley, 1970; Hays, 1973). Although many texts are clear about the fact that measures such as the phi coefficient and its square ($\chi^2$) are simply Pearson product-moment correlation coefficients applied to two dichotomous dummy variables, few introductory or intermediate textbook authors carry the logic one step further and show that the value of $\chi^2$ can easily be derived via least squares, as may the $\chi^2$ test of significance that accompanies such analyses. It is the least squares connection that provides the clearest understanding of the bias in sample estimates of categorical measures of association. Although many advanced texts and articles do make the connection between $\chi^2$ (or its variants $V^2$, $T^2$, and $C^2$), $\chi^2$, and $r^2$, none address the issue of overfitting or bias in the measures of association (Bishop, Feinberg, & Holland, 1975; Gittins, 1985; Kendall & Stuart, 1973; Knapp, 1978; Kshirsagar, 1972). The intuitive expectation that categorical measures of association would be subject to shrinkage problems in the same way that $r^2$ is affected, is suggested by the fact that $\chi^2$ analysis of contingency tables can be readily accomplished by solution of the canonical correlation model, which is itself a least squares procedure. The canonical solution to the problem of categorical association and/or hypothesis testing has a rather lengthy history in the statistical literature (Fisher, 1940; Hirschfeld, 1935; Kendall & Stuart, 1973; Kshirsagar, 1972; Gilula & Haberman, 1988), in the biometric literature (Gittins, 1985; Williams, 1952) and in the psychological literature (Cohen, 1988; Knapp, 1978).

In the remainder of this paper we shall (a) review the basic canonical correlation problem and its solution; (b) review the $r \times c$ categorical association problem and its standard solutions via $\chi^2$ analysis, and discuss its corollary canonical analysis solution; and (c) adopt a method suggested by Cohen & Nee (1984) to derive the expected values of several least squares measures of association which provide the basis for deriving shrinkage formulas for four common measures of association in the multicategory, two-way contingency table situation. We conclude with the results of a monte carlo study of the efficacy of the corrected estimators.

The Canonical Correlation Problem

Canonical correlation analysis is especially suited to addressing the problem of the nature and degree of relationships between a set of $Y(n \times p)$, $p \geq 1$, criterion or dependent variables and a set of $X(n \times q)$, $q \geq 1$, predictor or independent variables. The model is general enough to encompass both standard multivariate multiple regression problems and multivariate analysis of variance models (Finn, 1974). The general linear model statement of the problem,

$$Y(n \times p) = X(n \times q)B(q \times p) + E(n \times p)$$

can be solved via least squares minimization of $E'E$ (Finn, 1974; Johnson & Wichern, 1982) to yield the estimates of the parameters of the model,

$$B(q \times p) = S_{XY}^{-1}(q \times q)S_{XY}(q \times p)$$

and associated SSCP matrix due to regression of,

$$SSCP_{X \times Y} = S_{XY}(q \times q)B(q \times p)$$

in which $S_{XX}$, $S_{XY}$, and $S_{YY}$ are the deviation corrected sums of squares and cross products matrices between $X$ and $Y$ sets of variables and among $X$ variables, respectively.

A measure of overall multivariate
The association between Y and X is given by (Cramer & Nicewander, 1979),

\[ \eta^2_v = (S_{yy})^{-1}SSCP_{TOTAL} \]

in which \((S_{yy})^{-1}\) is the mean corrected SSCP TOTAL.

The multivariate measure of association in Equation 4 is defined by the multivariate test statistic (V) Pillai’s Trace (Cramer & Nicewander, 1979; Serlin, 1984) and can be obtained in two ways: as the trace of the quadruple matrix product,

\[ \text{Tr}(M) = \text{Tr}(S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}) \]

or as a function of the eigenvalues, \(\rho^2_1, \rho^2_2, \ldots, \rho^2_s\), of the determinantal equation,

\[ |S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy} - \rho^2 I| = 0 \]

where the \(\rho^2_i\) are the squared canonical correlations between the Y and X sets of variables. From equations 5 and 6 the definitions of \(\eta^2_v\) are \(\text{Tr}(M)/s = V/s\), and \(\rho^2_i/s = V/s\), respectively. The most commonly employed approximate test statistics applied to Pillai’s V are an \(F\)-test approximation (Pillai, 1960), or a \(X^2\) approximation (Kshirsagar, 1972). The \(X^2\) approximation is of central interest here and is defined on \((pq)\) degrees of freedom as,

\[ \chi^2_{(pq)} = n\nu = n\Sigma \rho^2_i = n\text{Tr}(M) \]

The data of any \(r \times c\) contingency table can be analyzed by the usual \(X^2\) analysis of cross classified observations. These same analyses can also be accomplished by application of canonical correlation analysis to a set of dummy coded variables for both classification variables of the problem. Two methods are available for solution in which the less than full rank model (Kendall & Stuart, 1973; Williams, 1952) can be performed on three 1/0 coded dummy variables for both effects, or on the full rank model (Gittins, 1985; Kshirsagar, 1972) in which the leading \((r-1)\) rows of the table are dummy coded, as are the leading \((c-1)\) columns of the table, with the remaining vector of each classification being unnecessary because of its linear dependence on the preceding vectors.

As shown by Kshirsagar (1972, pp. 379-385) the canonical correlation analysis on these two sets of dummy variables, \([A_{r \times p}, B_{c \times q}]\), can be accomplished by the solution of the determinantal equation,

\[ |R_{aa}^{-1}R_{ab}R_{ba}^{-1}R_{bb} - \rho^2 I| = 0 \]

where the \(R\) matrices contain the cross-variable and within-variable squared canonical correlations, and \(\rho^2_i\) are the \(i=1,\ldots,s\) squared canonical correlations. It is further known that a test of the hypothesis of no association can be accomplished by a \(X^2\) test on \((pq)\) degrees of freedom given in Equation 7.

With \(s = [\min(p,q)]\) eigenvalues in this problem the multivariate measure of strength of association for the data is \(\eta^2_v = \chi^2_i/s\). The numerical value of the Pearson \(X^2\) computed on the original frequencies of the cross classification table and the \(X^2\) computed from the canonical analysis is identical. The general proof of this equivalence is given by Kshirsagar (1972), Kendall and Stuart (1973), and Williams (1952). It is also useful to note that the value of \(\eta^2_v\) is also identical to Cramer’s \(V^2\) (Cramer, 1946) computed as a function of the \(X^2\) in the contingency table analysis: \(V^2 = \chi^2/n\), a fact also noted by Cohen (1988) and Zwick & Cramer (1986).

The close connection between the measures of association computed from \(X^2\) contingency table analysis and those computed from the canonical correlation analysis suggest that the measures of association derived from either method will yield biased estimates of the population parameter, and are therefore subject to overfitting at the level of the sample statistic. This point has been proven by Cohen and Nee (1984) with respect to \(\eta^2_v\) in the typical canonical case. Since many measures of association frequently cited in the literature of contingency table analysis are \(X^2\) based statistics, and since there is such a close relationship between \(X^2\) on contingency tables and canonical analysis, it follows that some degree of bias could be expected to be present in each of these several popular measures. In what follows we develop measures of correction for the bias in these measures for four measures of categorical association.
Unbiased $\chi^2$ Based Measures of Association

Cohen and Nee (1984) introduced a relatively direct alternative proof to Wherry’s (1931) correction for bias in the values of sample $R^2$ and applied this method to developing nearly unbiased estimates of two multivariate estimators of strength of association (Wilks’ $\Lambda$ and Pillai’s $V$). Our development in this section follows the method of Cohen and Nee (1984).

Let $\Omega^2$ represent a population measure of strength of association between the aggregate variance of a multicategorical response variable ($r \geq 3$) and a multicategorical explanatory variable ($c \geq 3$). Dummy coding of the categories for both outcome and explanatory variables results in a $(r-1) \times (c-1)$ vector representation of the $r \times c$ contingency table whose entries are frequency counts. The canonical solution to the problem (see for example, Kshirsagar, 1972) yields an $(r-1) \times (r-1)$ matrix of correlations among outcome vectors, a $(c-1) \times (c-1)$ matrix of intercorrelations among the explanatory variables, and an $(r-1) \times (c-1)$ matrix of intercorrelations between the explanatory and outcome vectors. The solution to the determinantal equation \([8]\) above yields the $s$ characteristic roots, $\eta^2$, whose sum defines Pillai’s Trace criterion $V$, and for which a proper measure of multivariate strength of association is $\eta^2 V = \chi^2/\text{ns}$. The $\chi^2$ on $(r-1) \times (c-1)$ degrees of freedom is defined by $\chi^2 = Vn$ (Kshirsagar, 1972). It can be shown that $\eta^2 V$ is precisely Cramer’s (1946) $V^2$ in this context. Since $\eta^2 V$ is fundamentally a least squares measure of association, it can be expected to overestimate the population parameter $\Omega^2$. An unbiased estimate of $\Omega^2$ may be developed by recognizing that the relation between the sample measure $\eta^2$ and the population parameter $\Omega^2$ is monotonically linear (asymptotically) and the relation between the two can be defined by selecting two points on the line describing the relationship (Cohen & Nee, 1984). The model describing the relation would then be written as,

\[ \Omega^2 = a + (b)E(\eta^2) \]

Cohen and Nee (1984) argue that the linear relation of $\eta^2$ to $\Omega^2$ can be accomplished by defining two points. Choosing as the upper bound the point such that $\Omega^2 = 1$ when $\eta^2 = 1$, and a second point $[\Omega^2 = 0$ when $\eta^2 = E(\eta^2)]$ such that $\Omega^2$ will be zero when $\eta^2$ equals its null expected value. With these points and using the fact that the $E(\chi^2) = df$ (asymptotically) we may write a pair of simultaneous equations capturing the upper and lower bounds of the linear relation between $\eta^2$ and $\Omega^2$ such that,

\[ 1 = a + b(1) \]
\[ 0 = a + (b)E(\eta^2) \]

Solving the equations simultaneously leads to the values of $a$ and $b$ in \([10]\) as,

\[ a = 1/(1 - E(\eta^2)) \]
\[ b = -E(\eta^2)/(1 - E(\eta^2)) \]

Hence the linear relation between $\eta^2$ and $\Omega^2$ may be expressed as,

\[ \Omega^2 = \left[1/(E(\eta^2))\right] - \left[E(\eta^2)/(1-E(\eta^2))\right][\eta^2] \]

Substitution of the appropriate definition of the expected values of $\eta^2$ for each of the four measures of categorical association addressed here into \([12]\), and solving the equation should provide the desired shrunk estimate of $\Omega^2$.

Pearson’s $\chi^2$ and Cramer’s $V^2$

Since it can be shown that Pearson’s $\chi^2$ is a special case of Cramer’s $V^2$ in the $2 \times 2$ table, we derive only the estimator for $V^2$. From the previous definition we may write $\eta^2 V = V^2$ as a function of $\chi^2$, $V^2 = \chi^2/\text{ns}$ and since $E(\chi^2) = df$, then

\[ E(\eta^2 V) = E(\chi^2)/\text{ns} = df/\text{ns} \]

where $df = pq = (r-1)(c-1)$: $p$ being the dummy coded vectors of the $r-1$ leading rows, and $q$ being the $c-1$ leading columns of the contingency table.

Substituting equation \([13]\) into equation \([12]\) and solving we obtain the definition of the estimator for both $\Omega^2$ and $V^2$,

\[ \Omega^2 = \frac{nSV^2 - df}{\text{ns} - df} \]

1323
which is the desired quantity for correcting for the bias in the sample measure of $\eta^2$.

Tschupperow's $T^2$

A frequently cited measure of association in contingency tables due to Tschupperow (Bishop, Feinberg & Holland, 1975) that is designed for use in nonsquare contingency tables may be written as,

$$T^2 = \chi^2 / [n(df)^{1/2}]$$

whose expected value may be obtained as a function of $E(\chi^2)$,

$$E(T^2) = E(\chi^2) / [n(df)^{1/2}] = \frac{df}{n(df)^{1/2}} = \frac{(df)^{1/2}}{n}$$

Substitution into equation [12] and solution leads to the estimator,

$$\Omega^2 = \frac{nT^2 - (df)^{1/2}}{n - (df)^{1/2}}$$

which should yield a shrunken estimate of the population measure of association based on $T^2$.

Pearson's Contingency Coefficient, $C^2$

Pearson's coefficient of contingency, $C^2$, has been routinely presented to students of behavioral statistics as a measure of association appropriate for use with $\chi^2$ tests of significance, despite its obvious interpretive problems (Bishop, Feinberg & Holland, 1975). The definition of Pearson's $C^2$ is,

$$C^2 = \frac{\chi^2}{\chi^2 + n}$$

and the expected value of $C^2$ may be written as,

$$E(C^2) = E(\chi^2) / [E(\chi^2) + n] = \frac{df}{df + n}$$

Substitution into equation [12] and solving we obtain the estimator $\Omega^2$,

$$\Omega^2 = \frac{(df + n)C^2 - df}{n}$$

which is the desired estimator for the shrunken value of $C^2$ from sample data in any contingency table.

Light and Margolin's Measure of Explained Variation

In conjunction with their development of methods for analysis of variance solutions to dichotomous dependent variable data Light and Margolin (1971) present a measure of asymmetric strength of association (asymmetric in the sense that one classification variable is designated as dependent and the second classification variable is designated as a predictor). The Light and Margolin index, denoted here as $LM^2$, is written as,

$$LM^2 = \frac{\chi^2}{[(n - 1)(c - 1)]}$$

in which $\chi^2$ on $(r-1)(c-1)$ degrees of freedom yields an index of association that is equivalent to Fisher's $\eta^2$ defined in ANOVA models as the ratio of the between groups-to-total sums of squares ($SS_e/SS_t$). The $c$ columns of the contingency table represent the categorical explanatory factor, while the rows of the contingency table contain 1 or 0, depending on the presence or absence of the value of the response variable for all subjects. The expected value of $LM^2$, based on its relationship to $\chi^2$, may be written as,

$$E(LM^2) = E(\chi^2) / [(n-1)(c-1)] = \frac{df}{(n-1)(c-1)}$$

Substitution of $E(LM^2)$ into equation [12] and solving leads to a definition of an unbiased estimator of $\Omega^2$,

$$\Omega^2 = \frac{(n-1)(c-1)LM^2 - df}{(n-1)(c-1) - df}$$

Monte Carlo Studies

Equations 14, 17, 20, and 22 may be used respectively to obtain "shrunken" values of $\eta^2$ (using $\eta^2$ to denote the general case subsuming $\psi^2$, $V^2$, $T^2$, $C^2$, and $LM^2$) Each equation should produce an unbiased estimate of population strength of association for the measures reviewed here. In order to test the accuracy of the formulae, several Monte Carlo experiments were conducted. A BASIC program was written to generate 2000 samples under the null hypothesis for each experiment from the multinomial distribution with fixed marginals for
the explanatory variable. The experiment was repeated for contingency tables of size 2 x 2, 3 x 3, 4 x 4, and 5 x 5, for samples of 20, 30, 60, 100, 200 and 400 (approximately). For each sample a value of $\chi^2$ and the measures of association ($V^2$, $T^2$, $C^2$, and $LM^2$) were computed, as were the shrunken values of each statistic based on equations 14, 17, 20, and 22. If the estimators of these equations are useful, then the average empirical value obtained from them should be near zero.

The results of these Monte Carlo experiments reveals that the derived formulas to correct for overfitting in each of the tested measures of association is quite efficacious, producing nearly unbiased estimates of the population parameter $\Omega^2$. Discrepancies between expected values and obtained values, as well as the average $\chi^2$ for each experiment also verify that the Monte Carlo process behaved according to theoretical expectations. The computed value of $\chi^2$ never differed from the expected value by more than $\pm$ .22, and the average discrepancy between expected and computed values of the measures of association was -.0003 across the 51 experiments. The estimators of each of the measures of association should, of course, produce shrunken values that are zero. In virtually all cases the means of the Monte Carlo empirical shrunken values are exceptionally close to zero (some values fall below zero and would be considered to be zero in practice). The average shrunken empirical value of the estimators across the 51 experiments using the formulas presented here was .0002. The average shrunken values for the 2x2, 3x3, 4x4, and 5x5 designs for sample sizes of 20-400 are summarized in Table 1.

Using the shrunken values from each experiment as a dependent variable, an analysis of variance was performed to ascertain if the size of the design, sample size, and N by design size interaction unduely influenced the magnitude of the shrunken estimates. The results reveal that sample sizes of n < 30 lead to the most biased estimates. Shrunken values for the 2x2 and 3x3 tables with n = 20 and n = 27, respectively, yielded shrunken estimates of .003 and .002, which are considerably different from the bulk of the results (.0002). As the sample size increases, the quality of the estimator increases concomitantly. For the entire data set the ANOVA revealed that sample size (N) had a significant influence on the magnitude of the estimated value of $\Omega^2$ ($F_{(3.43)} = 6.51, p < .02$). Similarly the size of the contingency table had a similar effect ($F_{(3.43)} = 3.04, p < .04$), largely produced by the four 2x2 and 3x3 experiments with N < 30. Dropping these outliers from the data, the ANOVA results are altered markedly. Without the very small sample size experiments, neither N nor size of contingency table had a significant effect on the quality of the estimators $\Omega^2$. The average value of $\Omega^2$ does not differ significantly from zero when the small sample size (n < 30) are not included in the analysis ($\chi^2 = - .18; p > .90$).

The analyst is cautioned to be circumspect about the degree of bias remaining in the estimators for designs in which sample sizes drop below 30. For larger sample sizes and any size contingency table tested here the estimators appear to have very desirable properties and should prove useful in analysis of categorical data where an unbiased measure of strength of association is desired. Such measures as studied here appear to be subject to the same "rules" of shrinkage as are other measures of magnitude of effect derived under the least squares model. The shrinkage of sample measures of association based on $\chi^2$ for categorical data appears to behave in a fashion similar to shrinkage in $R^2$ in linear regression analysis.

References


Table 1. Mean empirical shrunken value of the estimators averaged across all measures of $V^2$, $C^2$, and $LM^2$

<table>
<thead>
<tr>
<th>N</th>
<th>2x2</th>
<th>3x3</th>
<th>4x4</th>
<th>5x5</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 30</td>
<td>.003</td>
<td>.002</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>30-75</td>
<td>.001</td>
<td>.000</td>
<td>-.000</td>
<td>.000</td>
</tr>
<tr>
<td>80-200</td>
<td>-.001</td>
<td>-.000</td>
<td>-.000</td>
<td>-.000</td>
</tr>
<tr>
<td>200-400</td>
<td>-.000</td>
<td>-.000</td>
<td>-.000</td>
<td>-.001</td>
</tr>
</tbody>
</table>