Canonical correlation analysis for studying the relationship between egg production traits and body weight, egg weight and age at sexual maturity in layers

Y. Akbaş, Ç. Takma

Department of Biometry and Genetics, Faculty of Agriculture, Ege University, Izmir, Turkey

ABSTRACT: In this study, canonical correlation analysis was applied to layer data to estimate the relationships of egg production with age at sexual maturity, body weight and egg weight. For this purpose, it was designed to evaluate the relationship between two sets of variables of laying hens: egg numbers at three different periods as the first set of variables (**Y**) and age at sexual maturity, body weight, egg weight as the second set of variables (**X**) by using canonical correlation analysis. Estimated canonical correlations between the first and the second pair of canonical variates were significant ($P < 0.01$). Canonical weights and loadings from canonical correlation analysis indicated that age at sexual maturity had the largest contribution as compared with body weight and egg weight to variation of the number of egg productions at three different periods.

Keywords: canonical correlation; egg production; body weight; egg weight; age at sexual maturity

Canonical correlation is a measure of the interrelationships between sets of multiple dependent variables and multiple independent variables. Canonical correlation analysis is a generalization of multiple regression analysis with more than one set of dependent variables. Therefore this multivariate statistical technique is designed to assist the researcher in studying the complex interactions of data from two sets of variables. It is also concerned with two sets of variables related to each other and with how much variance of one set is common with or predictable from the other set (Weiss, 1972).

Since only a few canonical variates are needed to represent the association between the two sets of variables, canonical correlation analysis is a data reduction technique (Sharma, 1996).

Hotelling was the first to introduce the canonical correlation technique in 1935 (Wood and Erskine, 1976). He showed that the number of pairs of weighted linear functions was equal to the number of variables in the smallest set. Then he called the relationship function within these pairs as a canonical correlation. Afterwards, some researchers developed generalized solutions to the simultaneous relationships obtained between more than two data sets.

This powerful multivariate technique has gained acceptance in many fields such as psychology, social science, political science, ecology, education, sociology-communication and marketing (Jaiswal *et al.*, 1995).

Recently, application of this technique began to increase with the availability of related computer packages. In poultry science and even in animal science, however, there are a few studies (Gürbüz, 1989; Jaiswal *et al.*, 1995) in which canonical correlation analysis was applied.

In this study, canonical correlation analysis was used to estimate the relationships of egg production with age at sexual maturity, body weight and egg weight in layer.

MATERIAL AND METHODS

Material

Data, provided by a commercial breeding company, consisted of age at sexual maturity (ASM), number of produced eggs at 28 wk, 36 wk and 40 wk of age, average daily egg weight during 38–40 wk (EW) of 921 brown layers.

Egg numbers were recorded daily from onset of lay (22 wk) to 40 wk of age. All birds were weighed at sexual maturity age (BW) which was obtained individually according to the laying of the first egg.

Since egg production is a main trait in layer, egg productions from three different periods were included in the first variable set (Y_i) while the second set of variables (X_i) consisted of other traits as follows:

Y_1 = number of eggs up to 28 wk of age (EN_1)
Y_2 = number of eggs up to 36 wk of age (EN_2)
Y_3 = number of eggs up to 40 wk of age (EN_3)
X_1 = body weight at sexual maturity (BW)
X_2 = average egg weight during 38–40 wk (EW)
X_3 = age at sexual maturity (ASM)

Method

Canonical correlation analysis was used to examine the relationships between two sets of the traits by using PROC CANCORR procedure of SAS (1988).

Canonical correlation analysis focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. Linear combinations of variables are useful for predictive or comparative purposes (Johnson and Wichern, 1986). Therefore the canonical variates representing the optimal linear combinations of dependent and independent variables and the canonical correlation showing the relationship between them are results of interest (Hair *et al.*, 1998). Considering the below equations it can be defined that W_m and V_m are canonical variates.

$$W_m = a_{m1}X_1 + a_{m2}X_2 + \dots + a_{mp}X_p$$

$$V_m = b_{m1}Y_1 + b_{m2}Y_2 + \dots + b_{mq}Y_q$$

The correlation between W_m and V_m can be called canonical correlation (C_m). Squared canonical correlation (canonical roots or eigenvalues) represents the amount of variance in one canonical variate accounted for by the other canonical variate (Hair *et al.*, 1998).

The standardized coefficients are similar to the standardized regression coefficients in multiple regressions that can be used as an indication of relative importance of the independent variables in determining the value of dependent variable. Therefore the aim of canonical correlation analysis is to estimate canonical coefficients ($a_{m1}, a_{m2}, \dots a_{mp}$ and $b_{m1}, b_{m2}, \dots b_{mq}$) when the canonical correlation is maximum.

A serial process can explain the maximization technique as follows. Let the first group of p variables be represented by the random vector, $X_{(px1)}$, and let the second group of q variables be represented by the random vector, $Y_{(qx1)}$. For the random vectors X and Y, population mean and (co)variances would be as follows:

$$E(X) = \mu \qquad E(Y) = \mu$$

$$Cov(X) = \Sigma_{11} \quad Cov(Y) = \Sigma_{22} \quad Cov(X, Y) = \Sigma_{12} = \Sigma'_{21}$$

Furthermore, X and Y random vectors and (co)variance matrices can be written as follows:

$$\begin{bmatrix} X_{px1} \\ Y_{qx1} \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \\ Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma 11_{pxp} & \Sigma 12_{pxq} \\ \Sigma 21_{qxp} & \Sigma 22_{qxq} \end{bmatrix}$$

So that the linear combination of the components of X and the components of Y would be $W = a'X$ and $V = b'Y$, respectively. Then, W and V have the expectation of zero and (co)variances:

$$\text{var}(W) = a'Cov(X)a = a'\Sigma_{11}a$$

$$\text{var}(V) = b'Cov(Y)b = b'\Sigma_{22}b$$

$$\text{cov}(W,V) = a'Cov(X, Y)b = a'\Sigma_{12}b$$

The correlation coefficient between W and V is therefore

$$C = \frac{a'\Sigma_{12}b}{\left[(a'\Sigma_{11}a)(b'\Sigma_{22}b)\right]^{1/2}}$$

Furthermore, the null and alternative hypotheses for assessing the statistical significance of the canonical correlations are:

$$H_0 : C_1 = C_2 = \dots = C_m = 0$$

$H_a : C_1 \neq C_2 \neq \ldots \neq C_m \neq 0$

For testing the above hypotheses, the most widely used test statistic is Wilks' lambda, given by $\Lambda = \prod_{i=1}^{m}(1-C_i^2)$. Bartlett showed that under the null hypothesis (the sets X and Y are linearly unrelated) a particular function of Λ would be distributed approximately as a chi-squared variate (Levine, 1977). Therefore the statistical significance of Wilk's Λ requires the calculation of the following statistic:

$$\chi^2 = -[N - 0.5\ (p + q + 1)]\ \ln\Lambda$$

Where N is the number of cases, ln denotes the natural logarithm function, p is the number of variables in one set and q is the number of variables in the other set.

Large canonical correlation does not always mean that there is a powerful relationship between the two sets of the traits because canonical correlation maximizes the correlation between linear combinations of variables in two groups but does not maximize the amount of variances accounted for in one set of variables by the other set of variables.

Therefore it is suggested to calculate the redundancy measure for each canonical correlation to determine how much of the variance in one set of variables is accounted for by the other set of variables (Sharma, 1996).

Redundancy measure can be formulated as below

$$RM_{V_i|W_i} = AV(Y \mid V_i)xC_i^2 \qquad AV(Y \mid V_i) = \frac{\Sigma_{j=1}^{q} LY_{ij}^2}{q}$$

$AV(Y\backslash V_i)$ = the averaged variance in Y variables that is accounted for by the canonical variate V_i

LY_{ij}^2 = the loading of the jth Y variable on the ith canonical variate

q = the number of traits in canonical variates mentioned

C_i^2 = the shared variance between V_i and W_i

RESULTS

The cross-product correlation between the traits considered is presented in Table 1. These correlations show that the relationships of egg numbers at different periods with ASM and BW were negative and higher with ASM than BW. Instead of interpreting many correlations given above, only three correlations need to be interpreted in this study because the number of canonical correlations that needs to be interpreted is minimum number of traits within X or Y set. Estimated canonical correlations between the pairs of canonical variates were found to be 0.81, 0.15 and 0.003 and their probabilities of significance from the likelihood ratio test were 0.0001, 0.0003 and 0.9280, respectively (Table 2).

The canonical correlations between the first and second pair of canonical variates were found to be significant ($P < 0.01$) from the likelihood ratio test. The remaining canonical correlation is not statistically significant ($P > 0.05$). Significance of likelihood ratio test is also equal to the significance of Wilks' Λ. However the redundancy measure of 0.23 for the first canonical variate suggests that about 23% of the variance in the Y variables is accounted for by the X variables while it was only about 1% for the second canonical variate.

The coefficients of canonical variates from the raw data are given in Table 3. These coefficients of canonical equations are not unique. So the coefficients should be scaled that the resulting canonical variates had the mean of zero and variance of one. Standardized canonical coefficients or canonical weights for the X and Y variables are given in Table 4. Magnitudes of these weights represent their relative contributions to the related variate.

Egg production at the third period (EN_3) and the age at sexual maturity (ASM) have positive and high

Table 1. The correlation matrix between traits

	BW	EW	EN_1	EN_2	EN_3
ASM	0.18	0.004	–0.58	–0.67	–0.46
BW		0.31	–0.04	–0.15	–0.07
EW			0.09	–0.05	0.02
EN_1				0.52	0.73
EN_2					0.80

Table 2. Canonical correlations between two sets of variables, eigenvalues, likelihood ratios and their probabilities

	Canonical correlation	Squared canonical correlation	Degree of freedom	Eigen values	Likelihood ratio	Probability Pr > F
1	0.813	0.661	9	1.95	0.33	0.0001
2	0.152	0.023	4	0.02	0.98	0.0003
3	0.003	0.000009	1	0.00	0.99	0.9280

Table 3. Canonical coefficients of variates

	V_1	V_2	V_3		W_1	W_2	W_3
EN_1	–0.089	0.113	–0.078	ASM	0.135	–0.009	0.024
EN_2	–0.108	–0.092	–0.060	BW	5.311	0.003	–0.008
EN_3	0.076	0.013	0.149	EW	–0.020	2.829	2.439

Table 4. Standardized canonical coefficients of variates

	V_1	V_2	V_3		W_1	W_2	W_3
EN_1	–0.799	1.027	–0.711	ASM	0.999	–0.071	0.178
EN_2	–1.169	–0.999	–0.657	BW	0.0006	0.406	–0.991
EN_3	0.951	0.168	1.859	EW	–0.005	0.798	0.688

Table 5. Correlations between the variables and related canonical variates (canonical loadings)

	V_1	V_2	V_3		W_1	W_2	W_3
EN_1	–0.712	0.632	0.307	ASM	1.000	0.005	0.003
EN_2	–0.828	–0.330	0.453	BW	0.179	0.643	–0.745
EN_3	–0.565	0.125	0.816	EW	–0.001	0.925	0.379

Table 6. Correlations between the variables and the other set of canonical variates (canonical cross loadings)

	W_1	W_2	W_3		V_1	V_2	V_3
EN_1	–0.579	0.096	0.0009	ASM	0.813	0.0007	0.000
EN_2	–0.673	–0.050	0.001	BW	0.145	0.100	–0.002
EN_3	–0.460	0.019	0.002	EW	–0.001	0.140	0.001

coefficients for the canonical variate V_1 and W_1, respectively. On the other hand, EN_1 and EN_2 for the canonical variate V_1 have negative coefficients while coefficient of EN_2 shows the highest contribution to the canonical variate in absolute value (Table 4). Since the canonical coefficients can be unstable due to small sample size or presence of multicolinearity in the data, the loadings were also considered to provide substantive meaning of each variable for the canonical variate.

The loadings are shown in Table 5. The loadings for the ASM, BW and EW suggest that ASM is the most influential variable in forming W_1 compared to BW and EW. On the other hand, for the second pair of canonical variate W_2, EW and BW are about equally influential and more influential than ASM in forming W_2. The loadings for EN_1, EN_2 and EN_3 are about equally influential in forming V_1 (Table 5).

There are high canonical cross loadings for EN_2 and EN_1 with the canonical variate W_1 and for ASM with the canonical variate V_1 (Table 6). These reflect the high-shared variance between the variables EN_1 and EN_2. By squaring these figures (0.34 and 0.45), it can be concluded that 34% of the variance in EN_1 and 45% of the variance in EN_2 are explained by the variate W_1. For the other set of variables, 66% of the variance in ASM, 2% of the variance in BW and no variance in EW are explained by the canonical variate V_1.

High cross loadings correspond to the high canonical loadings. EN_1, EN_2 and EN_3 have the negative cross loadings and inverse relationships in the variate W_1. However cross loadings of ASM and BW are positive and have direct relationships with V_1 while EW has an inverse relationship in the variate V_1.

DISCUSSION

The cross-product moment correlation gives information only on the relationship between two variables without considering simultaneously other variables that related with each other. Canonical correlation, however, gives us the chance to estimate the correlation between two sets of variables including more than one trait in each at the same time.

Since canonical correlations between the first and second pair of canonical variables were found to be significant, only two pairs of canonical variates are considered. Estimated canonical correlation was the highest (0.81) for the first pair of canonical variates (V_1 and W_1) but small (0.152) for the second pair of variates (V_2 and W_2) while the third pair shows no association. From the first pair of canonical variates we can say that EN_1, EN_2 and EN_3 are highly correlated with ASM, BW and EW.

The signs of the standardized coefficients reflect the effects of ASM, BW and EW on EN_1, EN_2 and EN_3. From this source of information we can say that ASM and BW have a positive impact on the number of produced eggs at three different periods while EW has a negative impact on it.

Moreover, a redundancy measure of 0.333 for the first canonical correlation suggests that 33.3% of the variance in the *Y* variables (EN_1, EN_2 and EN_3) is accounted by the *X* variables (ASM, BW and EW). This value revealed that the first canonical correlation has a high practical significance.

When we considered the first pair of canonical variates and their coefficients, EN_1 and EN_2 contrast with EN_3 in V_1 and ASM is more effective than BW and EW in terms of the contribution to canonical variates W_1.

The sign of the loading for EN_3 does not agree with the sign of its canonical coefficient in the variate V_1. This can be explained by small sample size or multicolinearity in the data.

While the loadings ignore the presence of the other variables, canonical coefficients give the contribution of each variable in the presence of all the other variables. Therefore canonical coefficients are important to determine the importance of each variable in canonical variates. However, loadings provide substantive meaning for the canonical variates.

When canonical cross loadings are examined, egg production traits shared similar variance in W_1 as expected, because they are different measurements of the same trait. On the other hand, ASM shared in the variance in V_1. It can be concluded that selection for the age at sexual maturity (ASM) will affect the improvement of the number of produced egg when the aim is to increase egg production.

REFERENCES

Gürbüz F. (1989): Değişken takımları arasındaki ilişkilerin kanonik korelasyon yöntemi ile araştırılması. Ankara Üniversitesi Ziraat Fakültesi Yayınları, 1162. Ankara.

Hair J.F., Anderson R.E., Tatham R.H., William C.B. (1998): Multivariate Data Analysis. 5th Edition. Prentice Hall, Inc., 442–462.

Jaiswal U.C., Poonia J.S., Kumar J. (1995): Canonical correlation analysis for studying relationship among several traits: An example of calculation and interpretation. Indian J. Anim. Sci., *65*, 765–769.

Johnson R.A., Wichern D.W. (1986): Applied multivariate statistical analysis. Prentice Hall, Upper Saddle River, New Jersey. 460–485.

Levine M.S. (1977): Canonical Analysis and Factor Comparison. Sage Publications. Beverly Hills, London. 11–26.

SAS (1988): SAS User's Guide. SAS Inst. Inc., Cary, NC.
Sharma S. (1996): Applied Multivariate Techniques. John Wiley and Sons, Inc., Canada. 391–404.
Weiss D.J. (1972): Canonical correlation analysis in counselling psychology research. J. Counsel. Psychol., *19*, 241–252.
Wood D.A., Erskine J.A. (1976): Strategies in canonical correlation with application to behavioural data. Educ. Psychol. Meas., *36*, 861–878.

Received: 04–05–17
Accepted after corrections: 04–11–05

Corresponding Author

Prof. Dr. Yavuz Akbas, Ege University, Agriculture Faculty, Department of Biometry and Genetics,
35100 Bornova-Izmir, Turkey
Tel. +90 232 388 40 00 (ext-2917), fax +90 232 388 18 67, e-mail: yavuz@ziraat.ege.edu.tr