Short scientific report *Journal of Plantation Crops, 2008, 36 (2): 132-136*

Application of nonparametric covariance analysis in field trial with reference to YLD management trial in arecanut

(Manuscript Received: 25-09-07, Revised: 29-12-07, Accepted: 28-05-08)

Keywords: Covariance analysis, disease management trial, nonparametric regression, partial linear regression, smoothing technique

Linear covariance analysis (ANCOVA) is generally used to control the experimental error due to the variations in pretreatment values in disease or pest management trials. In linear covariance analysis, it is assumed that the covariate (pre treatment value) is linearly related to the response variable. But in many situations this assumption may not satisfy. In this paper we propose nonparametric covariance analysis (NPANCOVA) which does not require much assumption about the functional relationship between the response variable and the covariate. The only assumption about the relationship is that it is a smooth function. The method is also extended to analyze the data under the presence of treatment x covariate interaction effect. The performance of the proposed method is verified through simulation studies. The method is applied to the data on Yellow Leaf Disease (YLD) management trial in arecanut. The comparison of mean square errors (MSE) indicated that the performance of the proposed method is better than the traditional ANCOVA technique.

The statistical model for the simple analysis of covariance (ANCOVA) is of the form

$$Y = X\beta + U\gamma + \varepsilon \qquad (1)$$

where Y is the response variable, X is the design matrix, β is the vector of treatment effect, U is the covariate vector and ε is the error term. It is assumed that the relationship between $(Y - X\beta)$ and the covariate U is linear. But in many situations this assumption may not be true. A less rigid assumption is that the relationship between $(Y - X\beta)$ and the covariate U is a smooth function. In that case a simple nonparametric analysis of covariance (NPANCOVA) model of the following form can be considered.

$$Y = X\beta + \phi(u) + \varepsilon \qquad (2)$$

where, Y, X and β are as defined in model (1), $\phi(u)$ is the nonparametric function representing the relationship between $(Y - X\beta)$ and the covariate U and ε is the error term assumed to be independently and identically distributed random variable with mean zero and variance σ^2. Backfiting algorithm (Buja *et al*, 1989) is used to estimate the treatment and covariate effect in the regression model and it is given by

$$\hat{\beta} = \left[X^T(I-S)X\right]^T X^T(I-S)Y \text{ and } \hat{\phi} = S\left(Y - X\hat{\beta}\right)$$

where, S is the smoothing matrix derived using local linear regression (Ruppert and Wand, 1994). An estimate of σ^2 is given by

$$\hat{\sigma}_1^{\,2} = \frac{1}{\left(n - p - 1 - trace(S)\right)}\left[Y - X\hat{\beta} - \hat{\phi}\right]^T\left[Y - X\hat{\beta} - \hat{\phi}\right]$$

The variance of $\hat{\beta}$ is estimated by

$$V(\hat{\beta}) = PP^T\hat{\sigma}_1^{\,2}$$

where, $P = (X^T(I\text{-}S)X)^{-1}X^T(I\text{-}S)$. The significance of the covariate effect ϕ is tested using the lack of fit statistic by comparing parametric and nonparametric models (Hart, 1997). Under the null hypothesis that the covariate effect $\phi = 0$, the mean residual sum of squares obtained by fitting the model (1) is given by

$$= (Y - X\hat{\beta})^T[(I - X(X^TX)^{-1}X]^T[(I - X(X^TX)^{-1}X](Y - X\hat{\beta})/(n - p - 1)$$

The lack of fit test statistic is given by

$$R_0 = \frac{\hat{\sigma}_0^{\,2}}{\hat{\sigma}_1^{\,2}}$$

The test statistic R_0, asymptotically follows an F distribution with (n–p–1), (n–p–trace(S)) degrees of freedom and it can be used for testing lack of fitness of ANCOVA model.

Analysis of data in the presence of treatment x covariate interaction effect:

In the semiparametric regression model (2), it is assumed that the treatment x covariate interaction effect is absent and in the presence of this effect, the model (2) can be modified as follows

$$y_{ij} = f_i(u_{ij}) + \varepsilon_{ij},\ i = 1, 2,..., p; j = 1, 2,.... n_i;\ \Sigma n_i = n \quad (3)$$

where, y_{ij} is the observed value and u_{ij} is the covariate value corresponding to the j^{th} replicate of the i^{th} treatment and ε_{ij} is the iid random error with mean zero and constant variance σ^2. It is assumed that f_i, $i = 1,\ldots,p$ are smooth functions corresponding to the i^{th} treatment.

Let $Y_i = [y_{i1} \ldots\ldots y_{ini}]^T$ and $F_i = [f_i(u_{i1}) \ldots\ldots f_i(u_{ini})]^T$, $i = 1,\ldots, p$. Then the model (3) can be written as

$$Y^* = F + \varepsilon^*$$

where, $Y^* = [Y_1 \ldots\ldots Y_p]^T$, $F = [F_1 \ldots\ldots F_p]^T$ and ε^* is the error component corresponding to Y^*. The solution to the above regression problem is obtained as

$$\hat{F} = S^* Y^*$$

where, the smoothing matrix S^* is given by

$$S^* = \begin{bmatrix} S_1 & 0 & \ldots & 0 \\ 0 & S_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & S_p \end{bmatrix}$$

and S_i is the $n_i \times n_i$ smoother matrix for the observations u_{ij}, $j = 1,2, \ldots. n_i$. An estimate of σ^2 is given by

$$\hat{\sigma}_2^{\,2} = \frac{1}{(n - trace(S^*))}\left[Y^* - \hat{F}\right]^T\left[Y^* - \hat{F}\right]$$

The significance of the treatment x covariate interaction effect can be tested by comparing the fitted models of the equations (2) and (3) using the lack of fit test statistic

$$R_1 = \frac{\hat{\sigma}_1^{\,2}}{\hat{\sigma}_2^{\,2}}$$

The statistic R_1 asymptotically follows an F distribution with (n-p-trace(S)), (n-trace(S*)) degrees of freedom and it can be used for testing the significance of the treatment x covariate interaction effect. An approximate α-level point wise confidence band around the estimated function f is given by

$$\hat{f}_i(u_{ij}) \pm z_{\alpha/2}\hat{\sigma}_2\sqrt{\left[S_i S_i^{\,T}\right]_{jj}} \quad \text{for } j = 1, \ldots, n_i;\ i = 1,\ldots,p$$

where, $\left[S_i S_i^T\right]_{jj}$ represents the element in the jj^{th} position of the matrix $\left[S_i S_i^T\right]$.

Simulated data is used to see the performance of the proposed procedure. The data were generated based on the models such as ANCOVA, NPANCOVA and the NPANCOVA with interaction effect.

With regard to the ANCOVA, model (1) with $\beta' = [2, 2\sqrt{2}, 2\sqrt{2}, 3\sqrt{3}]$, $\gamma = 2\sqrt{2}$, $U \in [0, 1]$ and the random error $\varepsilon \sim N(0, 0.25)$ is considered. Based on the above 200 observations were generated by randomly allotting the value of the covariate and the treatment.

In the case of NPANCOVA, model (2) with $\phi(u) = 2(1+sin\ 3u)$, and all other values were retained as defined for the ANCOVA. Based on the above 200 observations were generated by randomly allotting the value of the covariate and the treatment.

The NPANCOVA model with the presence of treatment x covariate interaction effect considered for the simulation study is of the form

$$y_{ij} = f_i(u_{ij}) + \varepsilon_{ij},\ i = 1,2,3,4; j = 1, 2,....n_i;\ \Sigma n_i = 200 \quad (4)$$

where, y_{ij} is the observed value and $u_{ij} \in [0, 1]$ is the value of the covariate corresponding to the j^{th} replicate of the i^{th} treatment $f_1(u) = 2, f_2(u) = 2(\sqrt{2} + sin3u), f_3(u) = 2(\sqrt{2} + sin3u), f_4(u) = 3\sqrt{3} + 2(1 + sin3u)$ and the random error ε_{ij} follows $N(0, 0.25)$. Based on the above 200 observations were generated by randomly allotting the value of the covariate and the treatment.

The proposed method is applied to the YLD management trial data conducted at 5 locations in Karnataka during the year 1999 to 2002 with 8 treatments applied randomly to a total of 809 palms. The yellow leaf disease index (George *et al*, 1980) during the year 1999 is taken as the covariate and the disease index during the year 2002 is taken as the response variable. For the present study, the data was pooled on all the locations to compare the effect of various treatments.

The estimated values of the Mean Square Errors (MSE) and the lack of fit statistics R_0 and R_1 of the simulated data corresponding to the models ANCOVA, NPANCOVA and the NPANCOVA with interaction effect are given in Table 1. It can be seen that the estimated values of $\hat{\sigma}_0^2, \hat{\sigma}_1^2, \hat{\sigma}_2^2$, corresponding to the ANCOVA model is almost same as that of the true value σ^2 and also the lack of fit statistics R_0 and R_1 are not significant. This indicate that if the data follows an ANCOVA model or in other words, the relationship between the covariate and the response variable is linear, the ANCOVA technique is sufficient and there is no advantage for using NPANCOVA technique. The analysis of the simulated data corresponding to NPANCOVA model where the relationship between the covariate and the response variable is not linear, the MSE obtained by fitting the ANCOVA model ($\hat{\sigma}_0^2$) is much higher than the true value (σ^2), whereas the MSE obtained by fitting the NPANCOVA model ($\hat{\sigma}_1^2$) is almost same as the true value. The value of the lack fit test statistics R_0 also shows that the NPANCOVA technique is significantly better than the ANCOVA technique. The analysis of the simulated data corresponding to the model (4), where the relationship between the covariate and the response variable is not linear and the treatment x covariate interaction effect is present, the MSE obtained by fitting the ANCOVA model ($\hat{\sigma}_0^2$) and NPANCOVA model ($\hat{\sigma}_1^2$) are much higher than the true value (σ^2), whereas the MSE obtained by fitting the NPANCOVA model with interaction effect ($\hat{\sigma}_2^2$) is almost same as the true value. The value of the lack fit test statistics R_1 shows that the NPANCOVA (with interaction effect) technique is significantly better than the ANCOVA and NPANCOVA technique for analyzing the data obtained based on model (4). The above simulation study clearly indicate that whenever the relationship between the covariate and the response variable is not linear NPANCOVA technique is much superior than the usual ANCOVA technique and when the relationship is not linear and under the presence of treatment x covariate interaction effect, NPANCOVA (with interaction effect) technique is better. The lack of fit test statistic can be used to identify the proper model for the data analysis.

Table 1. The values of the Mean Square Errors (MSE) and the lack of fit statistics R_0 and R_1 of the simulated data

Models	σ^2	$\hat{\sigma}_0^2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	R_0	R_1
ANCOVA	0.25	0.239	0.246	0.246	0.97	1.00
NPANCOVA	0.25	0.986	0.248	0.245	3.98**	1.01
NPANCOVA with interaction	0.25	0.492	0.469	0.267	1.05	1.76**

** significant at P=0.01

The estimated treatment means with standard errors for the YLD management trial data obtained by using the simple ANOVA, ANCOVA and NPANCOVA techniques are given in Table 2. The MSE of the ANCOVA and NPANCOVA method are significantly less than that of the simple ANOVA. It indicates that for the present data, the covariance analysis is more suitable than the simple ANOVA. The estimated values of the treatment means are almost same for both ANCOVA and NPANCOVA techniques but the standard errors of the treatment means are less in the case of NPANCOVA than the ANCOVA technique. In all types of analysis, the treatment T_2 has the minimum mean value.

Table 2. Estimated treatment means and standard errors under different models

Treat	n_i	Pre-treat Mean	Simple ANOVA		ANCOVA		NPANCOVA	
			b	se	b	se	b	se
1	103	10.66	12.90	1.14	12.75	1.09	12.65	1.02
2	90	8.17	10.21	1.22	10.65	1.17	10.73	1.08
3	105	10.74	12.60	1.13	12.42	1.08	12.49	1.01
4	103	9.78	10.95	1.14	11.00	1.09	11.06	1.02
5	92	9.93	12.91	1.21	12.93	1.15	13.07	1.07
6	107	10.41	11.44	1.12	11.34	1.07	11.33	1.00
7	109	9.50	11.96	1.11	12.08	1.06	12.02	0.99
8	100	10.60	11.56	1.16	11.42	1.11	11.29	1.03
MSE			134.28		122.09		119.28	

The functional relationship between the covariate and $(Y - X\beta)$ based on the ANCOVA and NPANCOVA techniques are given in Figure 1. It can be noted that, between the disease index (covariate) value of 3 and 45, the estimated values of the covariate function in the case of ANCOVA and NPANCOVA is almost same. For higher values (>45) of the covariate (disease index) as well as for very low values (<3), the estimated covariate function based on NPANCOVA is higher than that of ANCOVA. Note that when the functional relationship is linear both the methods give almost the same result but if it is not linear, the NPANCOVA is more suitable than that of the ANCOVA technique.

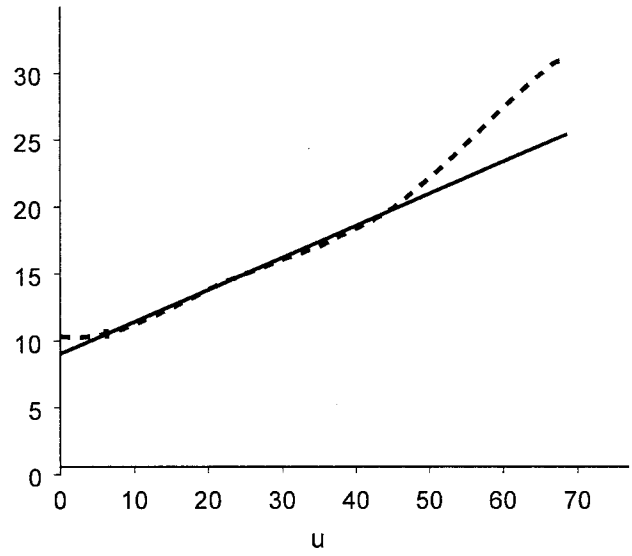

Fig.1. Covariate function of ANCOVA model (solid line) and NPANCOVA model (dotted line)

The value of the mean regression function corresponding to the covariate value (pre-treatment index) 0 is near to 10 (Figure 1). For very low values of the covariate (≤10), the estimated value of the mean regression function is more than the covariate value and for higher values (>10), the estimated value of the mean regression function is less than the covariate value. This is reflected in the frequency distribution table of pre and post treatment disease index (Table 3). There were 592 palms having the disease index less than or equal to 10 in the pre-treatment observation (covariate), but in the post treatment observation there were only 448 palms in this category. Note that the number of apparently healthy palms (disease index 0) in the pre-treatment observation was 294, which was reduced to 173 in the post treatment observation. Similarly there were 52 palms having the disease index greater than 40 in the pre-treatment observation, but in the post treatment observation there were only 20 palms in this category. The above observations indicate that even though the disease is spreading, the intensity of the disease can be controlled by the proper management of the garden.

Table 3. Frequency table of covariate (U) x response variable (Y)

U	Y						
0	0-10	10-20	20-30	30-40	>40	Total	
0	89	85	72	24	16	8	294
0-10	70	129	55	29	14	1	298
10-20	7	39	29	8	2	2	87
20-30	5	10	12	7	6	2	42
30-40	2	5	22	4	3	0	36
>40	0	7	17	17	4	7	52
Total	173	275	207	89	45	20	809

The model (3) is fitted to see the covariate x treatment interaction effect. The MSE of 116.51 obtained by fitting the model (3) is slightly less than that of ANCOVA and NPANCOVA techniques. The graphical representation of the fitted functions corresponding to different treatments indicates the relationship between the covariate and the response variable (Figure 2). Note that all the functions show more or less similar trend. This means the covariate x treatment interaction effect is not significant. The comparison of model (2) and model (3) by using the lack of fit test also indicates that the interaction effect is not significant. The graphical representation of the estimated covariate functions of different treatments of model (3) indicates that the treatment T_2 is comparatively better than the other treatments in controlling the disease. The model (3) is more adequate when the covariate x treatment interaction effect is significant.

In disease management trials we use ANCOVA technique for comparing different treatments with pre-treatment value as the covariate. The relationship between the response variable and the covariate is taken as linear in the usual ANCOVA technique for field experiments. But in many situations this assumption may not satisfy and subsequently the experimental error will become very high. NPANCOVA technique, which does not require much assumptions about the functional relationship between the response variable and the covariate is tried to overcome this problem. In NPANCOVA technique, the only assumption about the functional relationship between the response variable and the covariate is that it is smooth. Therefore, the NPANCOVA technique is more robust against the usual assumptions in the linear covariance analysis. The simulation study clearly shows that whenever the relationship between the covariate and the response variable is not linear NPANCOVA technique is much superior than the usual ANCOVA technique and when the relationship is not linear and under the presence of treatment x covariate interaction effect, NPANCOVA (with interaction effect) technique is better. The analysis of YLD management trial data in arecanut using different techniques indicates that the NPANCOVA technique performs better than the usual linear ANCOVA technique. This paper shows that whenever the relationship between the response variable and the covariate is unknown we have to fit ANCOVA, NPANCOVA and NPANCOVA with interaction effect models to the data and the proper model can be identified by testing the lack of fit statistics R_0 and R_1. The analysis of data using the proper model will reduce the experimental error considerably.

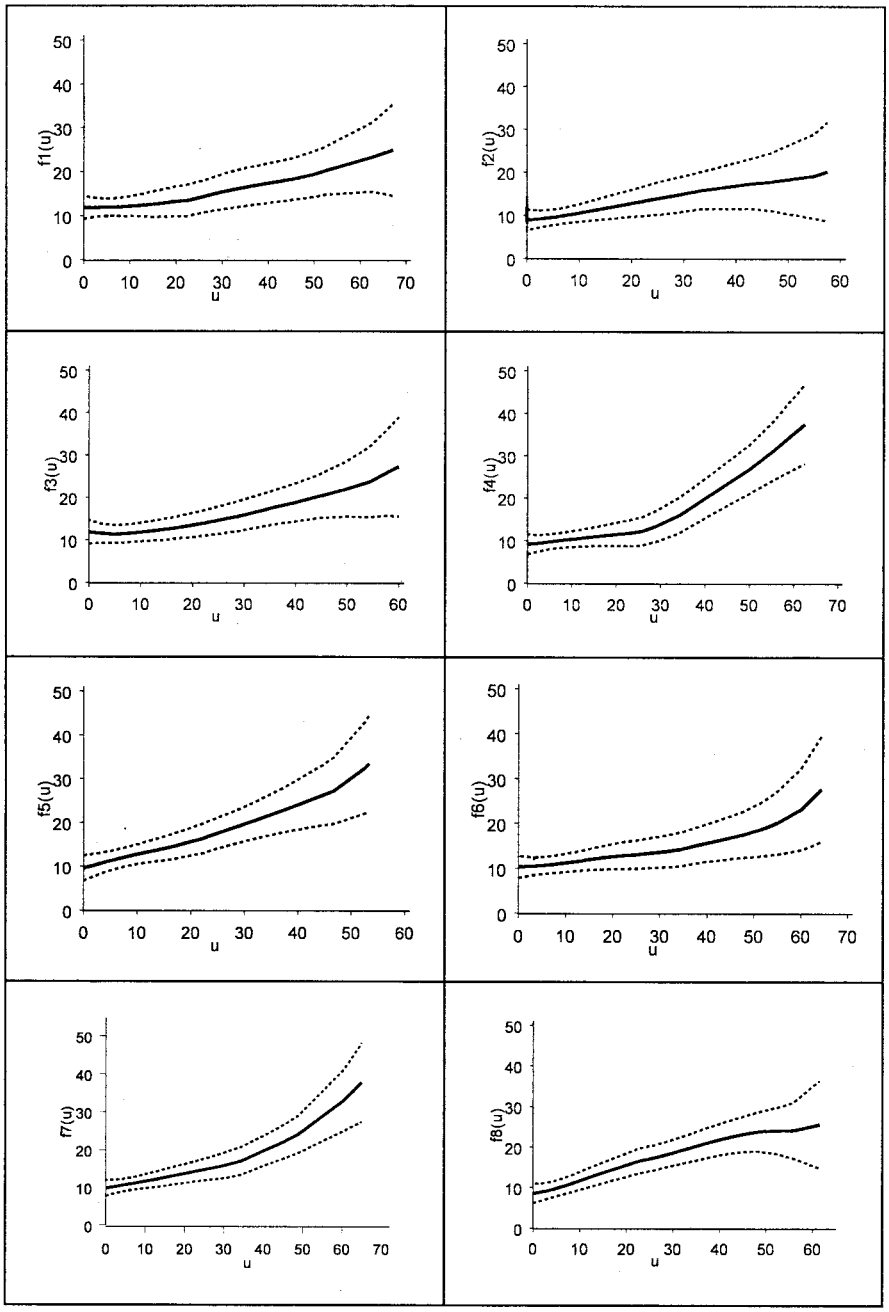

Fig.2. Estimated functions (solid line) with confidence interval (dotted line) of model (3)

References

Buja, A. Hastie, T.J and Tibshirani, R.J. 1989. Linear smoothers and additive models. *Annals of Statistics* **17**: 453-555.

George, M.V., Mathew, J. and Nagaraja, B. 1980. Indexing the yellow leaf disease of arecanut. *J. of Plantn. Crops* **8**: 82-85.

Hart, J.D. 1997. *Nonparametric smoothing and lack-of-fit tests.* Springer Verlag, New York, 287 p.

Ruppert, D. and Wand, M.P. (1994). Multivariate locally weighted least squares regression. *Annals of Statistics* **22**: 1346-70.

Central Plantation Crops Research Institute,
Regional Station, Vittal-574 243, Karnataka,
INDIA, Email: ctjos@yahoo.com

C.T. Jose[1],
N. Saraswathy

[1]*Corresponding author*