© South African Society for Animal Science

The influence of outliers on a model for the estimation of crossbreeding parameters for weaning weight in a beef cattle herd

M.A. Aziz[2], S.J. Schoeman[#1] and G.F. Jordaan[1]
[1]Department of Animal Sciences, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa
[2]Department of Animal Production, Alexandria University, Egypt

Abstract

Data on 17348 weaning weight records from a beef cattle crossbreeding operation were used to determine the effect of outliers on regression coefficients. Different criteria were used for detecting potential influential points. Eliminating a small number (932 or 5.4%) influential points resulted in the improvement of the model fitted. The R^2 values increased from 41% to 49% while the mean square error was reduced from 672.9 to 500.4. The use of diagnostic statistics for detecting influential observations is recommended before any analysis is performed.

Keywords: Beef cattle, crossbreeding, genetic effects, outliers
[#]Corresponding author. E-mail: sjsc@sun.ac.za

Introduction

An evaluation of breeds of cattle with respect to their direct, maternal and non-additive gene action is necessary to determine which breeds or combinations to use in a crossbreeding system. To estimate these genetic effects, various crosses among a number of breeds must be made and evaluated.

Several techniques of separating genetic effects were described. These techniques were based mainly on linear functions of the various cross means or multiple regression procedures (Alenda *et al*., 1980; Dillard *et al*., 1980; Robison *et al*., 1981; Cunningham & Magee, 1988; Schoeman *et al*., 1993; Skrypzeck *et al*., 2000). Generally, the importance of additive, maternal and non-additive genetic effects was determined and used for prediction of the performance of crosses that have not actually been tested. However, no attention has been paid to the problem areas of least square analysis relating to the failure of the basic assumptions, i.e. normality, common variance and independence of the errors. One of the most serious problems that violate these assumptions is the problem of outliers. Inferences based on ordinary least squares regression can be influenced by one or a few animals represented in the data. Hence, the fitted model may reflect unusual features of those animals instead of the overall relationships between variables.

A data point may be an outlier or a potentially influential point because of errors in recording or data entry or because the data point is from a different population. The latter could result from management changes that take the system out of the realm of interest or the occurrence of atypical environmental conditions. According to Rawlings (1988), a single point far from the other data points could have as much influence on the regression results as all other points combined. Little confidence could be placed on regression results that have been dominated by a few observations, regardless of the total size of the data.

The objectives of this investigation were to determine additive, maternal and non-additive gene action of different breed groups, to investigate the relative influence of individual animals on the inferential process and to determine the genetic components after deleting the influential points in question.

Materials and Methods

Data of this study were derived from the two farms of the Johannesburg Metropolitan Council. More details regarding feeding and management regimes and replacement and selection procedures are found in Paterson (1978; 1981), Paterson *et al*. (1980), MacGregor (1997) and Skrypzeck *et al*. (2000).

Data consisted of 17348 calf weaning weight records collected from 1968 to 1992. All those having weaning weights which deviated from the mean by more than three times the standard deviation, were omitted from the data before analysis.

Five breeds of cattle, namely Afrikaner (AF), Hereford (H), Aberdeen Angus (AA), Simmentaler (ST) and Charolais (CH), were mated to produce 129 different breed groups of calves. No distinction was made between Hereford (H) and Aberdeen Angus (AA), as earlier studies (Fredeen *et al*., 1982; Tosh *et al*., 1999) did not detect important differences between them. Therefore, both breeds were pooled and considered as one breed (HA). Data were classified into sex (male and female), age of dam (ranged from 2 to 9 years), herd-year-season (HYS) and breed groups. Some of the HYS subclasses contained small

© South African Society for Animal Science

numbers of observations. These were pooled with the next HYS subclass having the same season in the following year within the same farm. Season of birth was recorded as Winter (June to September) or Summer (December to March).

Least Squares Analysis of Variance was conducted for the trait using the GLM procedure of SAS (2000). The initial model fitted included the fixed effects of breed groups, HYS, sex of calf and age of dam and all possible one-way interactions. Interactions with no effect ($P > 0.05$) were excluded. The final model was:

$$Y_{ijkl} = \mu + G_i + HYS_j + S_k + D_l + Age_m + e_{ijklm}$$

where:

Y_{ijkl} is the observation,
μ is the overall mean,
G_i is the effect of the i^{th} breed group,
HYS_j is the effect of the j^{th} HYS contemporary group
S_k is the effect of k^{th} sex of calf,
D_l is the effect of l^{th} age of dam,
Age_m is the age of the calf at weaning (covariable), and
e_{ijkl} is a random error assumed to be randomly and independently distributed with mean 0 and variance σ^2_e.

Subsequently, the data were adjusted for the significant effects ($P \leq 0.05$), except for the effect of breed group. Multiplicative adjustment factors were used to adjust for the effects of sex of calf, age of dam and HYS, assuming that variances among factor subclasses were heterogeneous. The base of comparison was a male calf born to a 7-year old cow in the winter of 1968 in herd1.

After adjustment, a multiple regression analysis was conducted, assuming independency of the variables. The analysis was performed using the ARC computer package of Cook & Weisberg (1999). The model fitted to the adjusted data included the genetic effects, namely the breed additive, breed maternal, average individual heterotic and average maternal heterotic effects. The coefficients used for the genetic effects were the proportions of genes contributed by each breed which were considered as continuous variables. The mating plan did not allow for estimation of individual and maternal heterosis, due to a lack of observations in specific crosses. Only average individual and maternal heterosis were thus estimated. The regression model used was:

$$Y = \beta_0 + \beta_1 DAF + \beta_2 DCH + \beta_3 DST + \beta_4 DHA + \beta_5 MAF + \beta_6 MCH + \beta_7 MST + \beta_8 MHA + \beta_9 HI + \beta_{10} HM + e$$

where:

β_0 is constant (the intercept),
β_1, β_2, β_3, β_4 are the regression coefficients of breed additive effects,
DAF, DCH, DST, DHA are the percentages of genes contributed by AF, CH, ST and HA, respectively,
β_5, β_6, β_7, β_8 are the regression coefficients of breed maternal effects,
MAF, MCH, MST, MHA are the percentages of genes contributed by dams of AF, CH, ST and HA, respectively,
β_9 is the regression coefficient of the average individual heterosis, due to the interaction of two alleles at the same locus, with alleles being from different breeds,
HI is the average individual heterosis,
β_{10} is the regression coefficient of the average maternal heterosis, due to the interaction of two alleles from different breeds in the dam,
HM is the percentage of loci in the dam with one gene from one breed and the other from a different breed,
e is the error term.

Subsequently, several criteria were used to detect outliers. These criteria included Cook's distance, studentized residual, leverage as well as graphical procedures proposed by Weisberg (1985), Rawlings (1988) and Cook & Weisberg (1999).

© South African Society for Animal Science

Cook's distance (D_i) measures the distance from the regression coefficient before and after deleting the influential points, in terms of the joint confidence ellipsoids about the regression coefficient before deletion. It can also be interpreted as the Euclidean distance between the fitted value before and after deletion and hence, measure the shift in the fitted value caused by deleting the influential observations. The equation of Cook's distance is:

$$D_i = \frac{r_i^2}{p'}\left(\frac{v_{ii}}{1 - v_{ii}}\right)$$

where:

r_i is the standardized residual,

p' is the number of the parameters in the model,

v_{ii} is the diagonal element of the H matrix.

Studentized residual: Belsley *et al*. (1980) suggested standardizing each residual with an estimate of standard deviation that is independent of the residual. The result is the studentized residual denoted by (r_i^*) with:

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1 - v_{ii}}}$$

where:

e_i is the residual

$s_{(i)}$ is the square root of the residual mean square from analysis where that observation has been omitted.

v_{ii} as previously defined

Studentized residual is distributed as Student's t with ($n - p' - 1$) degrees of freedom when normality of the residual holds.

Potentially influential points are those with high leverages. Belsley *et al.* (1980) suggested using $v_{ii} > 2\frac{p'}{n}$ to identify potentially influential points. A leverage value is generally considered to be large if it is substantially greater than most of the other leverage values or if it is greater than twice the average leverage value.

Employing these procedures resulted in the elimination of 932 (or 5.4%) of the initial observations.

Results and Discussion

Regression coefficients and their standard errors for the direct and maternal effects of each breed and average individual and average maternal heterosis, both before and after deleting the influential points, are presented in Table 1. Re-analysing the data after deleting the influential points, resulted in a large change in point estimates (e.g. the intercept changed from 78.1 tot 170.7) of all effects, except the coefficients of the heterotic effects. Neither the direct breed nor maternal breed effects had any influence on the estimates ($P > 0.05$). They were characterised by large standard errors, which slightly decreased after deleting the influential points. Likewise, the R^2 value had slightly increased, and the mean square error had slightly decreased, indicating a slightly better fit.

In several other studies significant ($P \leq 0.05$) direct and maternal effects involving Charolais and Hereford were obtained (Peacock *et al*., 1981; Franke *et al*., 2001). In these studies, as well as others (Dillard *et al*., 1980; Alenda & Martin, 1981; Schoeman *et al*., 1993; Skrypzeck, *et al*., 2000) appreciably lower standard errors for these estimates were obtained. This together with the change in the b_i estimates, could be caused by extreme multi-collinearity amongst the independent (assumed) variables.

A lack-of-fit test of the functional form of the regression model before and after removing the influential points are presented in Table 2. Before deleting the influential cases, a lack-of-fit of the model was evident ($P < 0.01$). After removing the influential points, the model fit improved ($P > 0.05$). The pure

© South African Society for Animal Science

error term, which measures the variability among observations treated alike, greatly reduced after eliminating the influential points, reflecting a decrease in the degree of dispersion among the observations.

Table 1 Regression coefficients (±s.e.) for direct, maternal, individual and maternal heterotic effects for weaning weight before and after deleting influential points

Genetic effect	Before deleting	After deleting
Constant (intercept)	78.1 (110)	170.7 (103.5)
DAF*	0.1 (57.4)	-46.3 (51.8)
DCH	32.3 (57.4)	-11.6 (51.8)
DST	26.3 (57.4)	-18.9 (51.9)
DHA	2.3 (57.6)	-43.6 (52.9)
MAF	-41.1 (87.0)	-90.1 (83.0)
MCH	-44.2 (86.9)	-93.6 (82.9)
MST	-34.9 (86.9)	-82.9 (81.7)
MHA	-47.4 (87.3)	-96.3 (83.2)
HI	5.3 (0.82)**	7.9 (0.74)**
HM	2.2 (0.62)**	2.3 (0.56)**
Number of observations	17348	16416
R^2 (%)	41.3	49.4
Mean square error	672.9	500.4

* Abbreviations: DAF, DCH, DST and DHA – direct effects for Afrikaner, Charolais, Simmentaler and Hereford-Angus, respectively MAF, MCH, MST and MHA – maternal effects for Afrikaner, Charolais, Simmentaler and Hereford-Angus, respectively HI – individual heterosis, MI – maternal heterosis.

** $P \leq 0.01$

Table 2 Lack-of-fit test for weaning weight of the models before and after deleting the influential points

	Before deletion		After deletion	
	d.f.	Mean squares	d.f.	Mean squares
Lack-of-fit	4706	805.4**	4393	496.3
Pure error	12630	623.5	12011	501.9

**$P \leq 0.01$

A scatter plot of the residuals on the fitted values of weaning weight before deleting the influential points is presented in Figure 1. If the fitted model is correct and the assumptions are met, the residuals should appear as random variation around zero (Cook & Weisberg, 1999). Although most points were within the band of e±100, the dispersion increased with larger fitted values, indicating a heterogenous variance problem amongst breed groups. The scatter plot, after deleting the influential points, is presented in Figure 2. Removal of the influential points resulted in an improvement of the model, as was indicated by the improved dispersion of the residuals around zero.

A normal probability plot of the ordered residuals on the normal order statistics, which are the expected values of the ordered observations from the normal distribution with zero mean and unit variance (Galpin & Hawkins, 1984) for weaning weight before deleting the influential points, is presented in Figure 3. Some points clearly depart from the expected straight line, clearly suggesting a violation of the normality assumptions in multiple regression. The normal probability plot after eliminating the influential points is presented in Figure 4. These observations markedly superimposed the straight line, indicating only a slight departure from normality as compared to the situation in Figure 3.

© South African Society for Animal Science

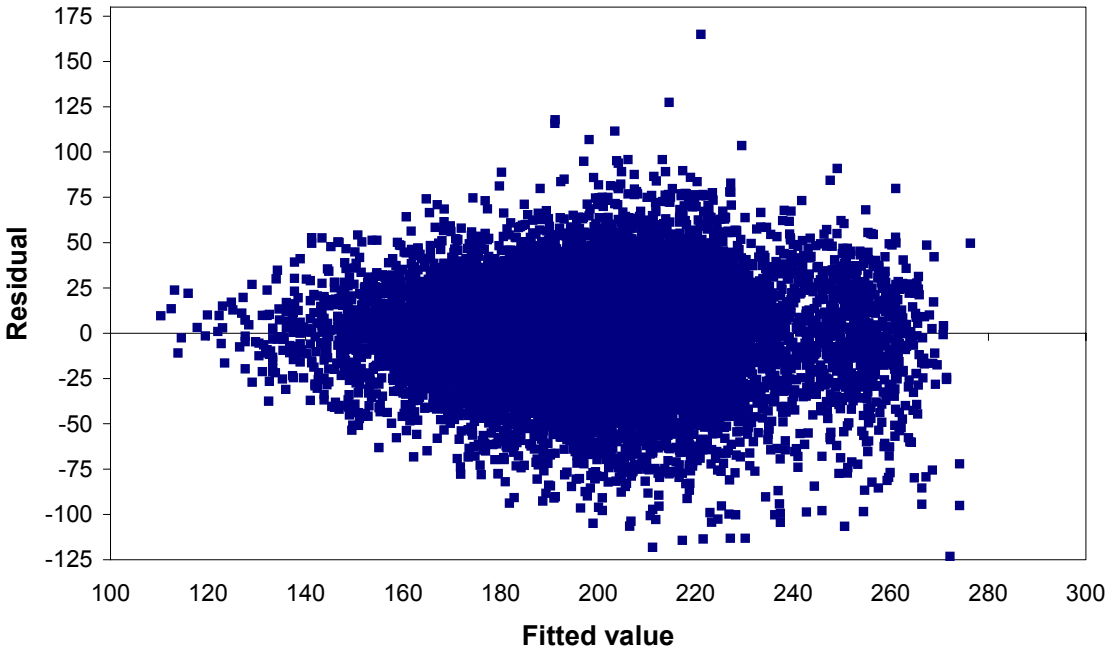

Figure 1 Scatter plot of the residuals on the fitted values of weaning weight before deleting the influential points

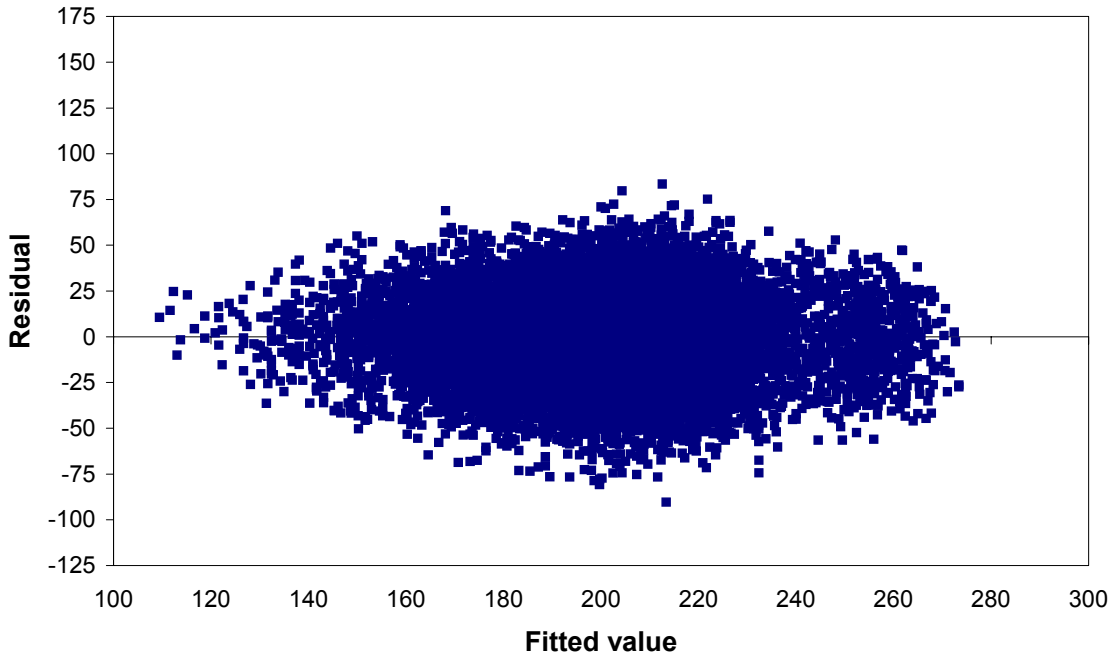

Figure 2 Scatter plot of the residuals on the fitted values of weaning weight after deleting the influential points

© South African Society for Animal Science

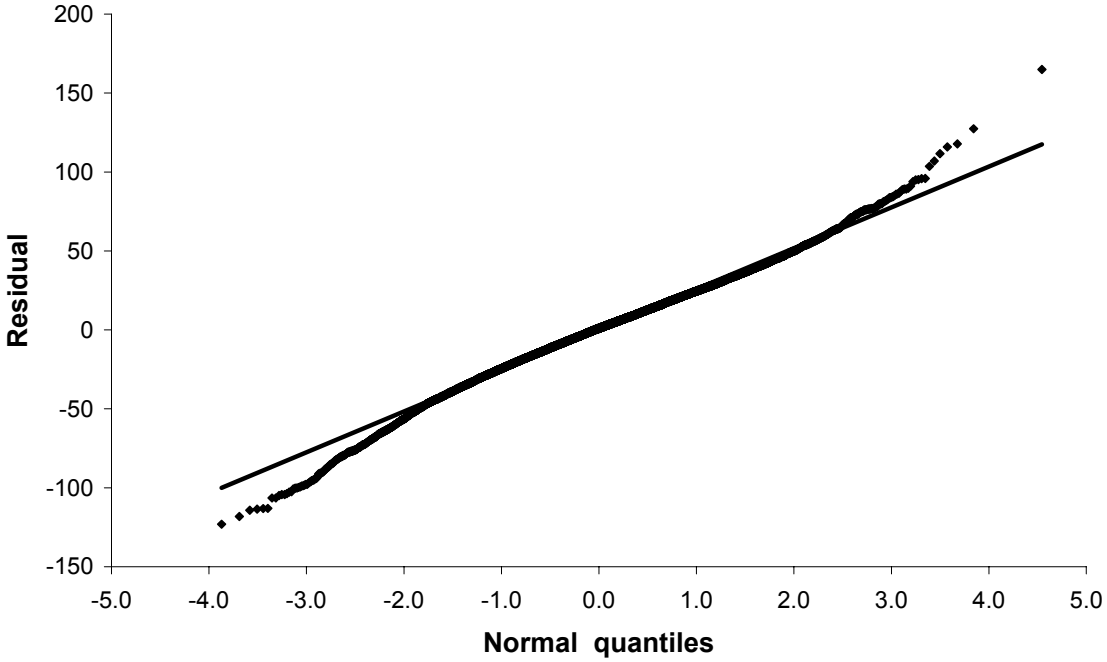

Figure 3 Normal probability plot of the residuals of weaning weight on the normal quantities before deleting the influential points

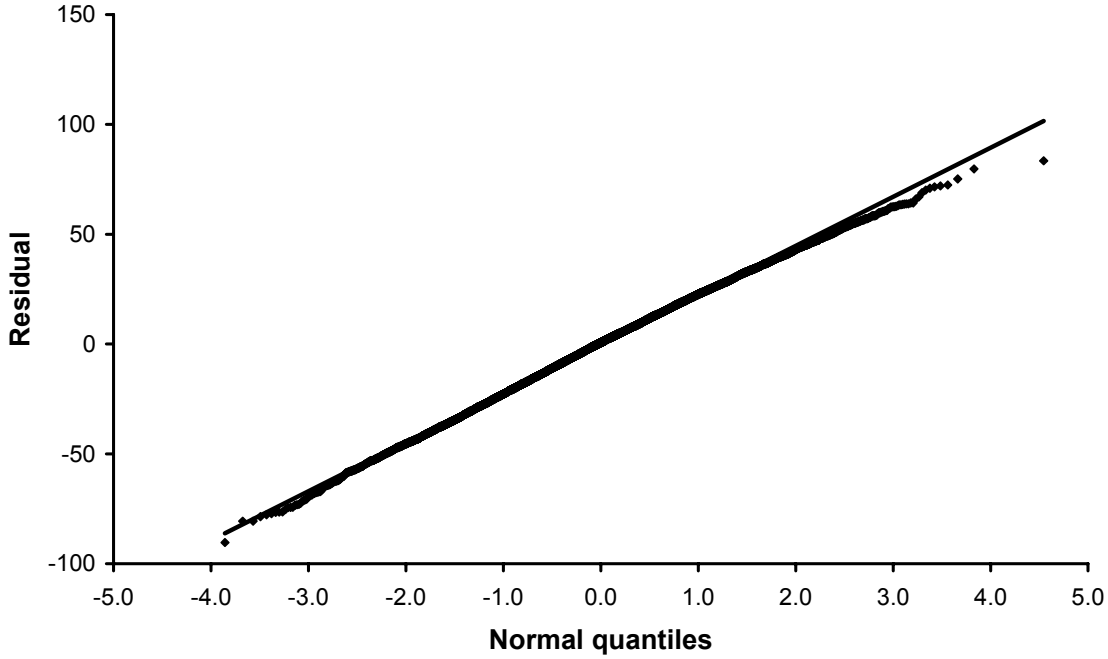

Figure 4 Normal probability plot of the residuals of weaning weight on the normal quantities after deleting the influential points

Conclusions

Multiple regression is a standard procedure in the estimation of crossbreeding parameters. In this, the method of ordinary least squares gives equal weight to every observation. The method is, furthermore, based on the assumptions that the errors are additive and are normally distributed independent random variables with a common variance. When these assumptions hold, least squares estimators have the desired properties of being the best. However, every observation does not have equal impact on the least squares

© South African Society for Animal Science

results, with some observations having the property of violating the underlying assumptions. In this study, after initially eliminating those observations which deviate by more than three times the standard deviation from the mean, an additional only about 5% of the data, which were detected to be outliers, violated the assumptions of the model. The application of diagnostic statistics for identifying outliers and possible influential points when using least squares procedures in animal breeding data is suggested.

References

Alenda, R., Martin, T.G., Lasley, J.F. & Ellersieck, M.R., 1980. Estimation of genetic and maternal effects in crossbred cattle of Angus, Charolais and Hereford parentage. Birth and weaning weights. J. Anim. Sci. 50, 226-234.

Alenda, R. & Martin, T.G., 1981. Estimation of genetic and maternal effects in crossbred cattle of Angus, Charolais and Hereford parentage. III. Optimum breed composition of crossbred. J. Anim. Sci. 53, 347-353.

Belsley, D.A., Kuh, E. & Welsh, R.E., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Wiley & Sons Inc., New York.

Cook, R.D. & Weisberg, S., 1999. Applied Regression Including Computing and Graphics. John Wiley and Sons, Inc., New York. 449 pp.

Cunningham, B.E. & Magee, W.T., 1988. Breed direct, breed-maternal and nonadditive genetic effects for preweaning traits in crossbred calves. Can. J. Anim. Sci. 68, 83-92.

Dillard, E.U., Rodrigues, O. & Robison, O.W., 1980. Estimation of additive and nonadditive direct and maternal genetic effects from crossbreeding beef cattle. J. Anim. Sci. 50, 653-663.

Franke, D.E., Habet, O., Tawah, L.C., Williams, A.R. & De Rouen, S.M., 2001. Direct and maternal genetic effects on birth and weaning traits in multibreed cattle data and producted performance of breed crosses. J. Anim. Sci. 79, 1713-1722.

Fredeen, H.T., Weiss, G.M., Rahnefeld, G.W., Lawson, J.E. & Newman, J.A., 1982. Environmental and genetic effects on preweaning performance of calves from first-cross cows. II. Growth traits. Can. J. Anim. Sci. 62, 51-67.

Galpin, J.S. & Hawkins, D.M., 1984. The use of recursive residuals in checking the model fit in linear regression. American Statistician 38, 94-105.

MacGregor, R.G., 1997. Evaluation of methods of measuring reproduction and production in beef cows. PhD thesis, University of Pretoria, South Africa.

Paterson, A.G., 1978. Statistical analyses of factors affecting preweaning growth of beef cattle under intensive pasture conditions. MSc (Agric) thesis, University of Pretoria, South Africa.

Paterson, A.G., Venter, H.A.W. & Harwin, G.O., 1980. Pre-weaning growth of British, *Bos indicus*, Charolais and dual purpose type cattle under intensive pasture conditions. S. Afr. J. Anim. Sci. 10, 125-133.

Paterson, A.G., 1981. Factors affecting post-weaning growth and reproduction of crossbred cattle under an intensive production system. DSc (Agric) thesis, University of Pretoria, South Africa.

Peacock, F.M., Koger, M., Olson, T.A. & Crockett, J.R., 1981. Additive genetic and heterosis effects in crosses among cattle breeds of British, European and Zebu origin. J. Anim. Sci. 52, 1007-1013.

Rawlings, J.O., 1988. Applied Regression Analysis. A Research Tool. Wadworth & Brooks, Pacific Grove, California. 553 pp.

Robison, O.W., McDaniel, B.T. & Rincon, E.J., 1981. Estimation of direct and maternal additive and heterotic effects from crossbreeding experiments in animals. J. Anim. Sci. 52, 44-50.

SAS, 2000. Statistical Analysis Systems user's guide, (Version 8). SAS Institute Inc., Cary, North Carolina.

Schoeman, S.J., Van Zyl, J.G.E. & De Wet, R., 1993. Direct and maternal additive and heterotic effects in crossbreeding Hereford, Simmentaler and Afrikaner cattle. S. Afr. J. Anim. Sci. 23, 61-66.

Skrypzeck, H., Schoeman, S.J., Jordaan, G.F. & Neser, F.W.C., 2000. Estimates of crossbreeding parameters in a multibreed beef cattle crossbreeding project. S. Afr. J. Anim. Sci. 30, 193-200.

Tosh, J.J., Kemp, R.A. & Ward, D.R., 1999. Estimates of direct and maternal genetic parameters for weight traits and backfat thickness in a multibreed population of beef cattle. Can. J. Anim. Sci. 79, 433-439.

Weisberg, S., 1985. Applied Linear Regression. 2nd Edition. Johan Wiley and Sons, Inc., New York.