Example 2 (referred to in module 4)

Fixed effects least squares analysis – an example in quantitative methods

John Rowlands

International Livestock Research Institute, P.O. Box 30709, Nairobi, Kenya

Background

Helminths constitute one of the most important constraints to livestock production in the tropics. Widespread infection with internal parasites in grazing animals, associated production losses, costs of anthelmintics and death of infected animals are some of the major problems. Current control methods focus on reducing contamination of pastures through anthelmintics treatment and/or controlled grazing. In the tropics, these methods are limited by the high costs of anthelmintics, their uncertain availability and increasing frequency of drug resistance. In this situation an attractive, sustainable solution is breeding for disease resistance. Indeed, there is a large and diverse range of indigenous breeds of sheep and goats in the tropics, some of which appear to have the genetic ability to resist or tolerate helminthiasis.

The data used in this example come from a study carried out at Diani Estate of the Baobab Farms, 20 km south of Mombasa in the sub-humid coastal region of Kenya between 1991 and 1997. The purpose of the experiment was to compare the genetic resistance to helminthiasis of two sheep breeds – Dorper and Red Maasai. Throughout the six years Dorper (D), Red Maasai (R) and FI (RXD) ewes were mated to Red Maasai and Dorper rams to produce a number of different lamb genotypes. For the purposes of this example, only the following four offspring genotypes are considered:-

D x D, D x R, R x D and R x R. (For shorthand we shall use DD, DR, RD and RR with the first letter referring to the breed of the sire and the second to the breed of the dam, respectively). Eight hundred and eighty two lambs within these genotypes were born to 74 rams and 367 ewes. Thus, each ewe gave birth on average to approximately two to three lambs, one each in a different year.

Measurements of lamb body weight were made periodically over a period of about a year. In addition, the age at weaning, the lamb's sex, the age of its dam and the identity of both sire and dam were recorded. In this example we shall consider weaning weight as the response variable and determine the effect of breed and other covariates on it.

Descriptive statistics

Before undertaking a full least squares analysis it is useful to first explore relationships between weaning weight and certain covariate fixed effects to see how best to_characterise them. In this example it is hypothesised that, in addition to year of birth and sex, both age of dam and age at weaning may also influence weaning weight. But they influence age at weaning? The following section shows how we might do this.

Firstly, let us look at some of the patterns in the data. The following table describes the distribution of the data on weaning weight by breed.

Table 1: *Summary of the distribution of weaning weight of Diani lambs*

SIRE_BRD	D					
	Count	Nobservd	Mean	Minimum	Maximum	Median
DAM_BRD						
D	310	220	11.84	5.300	19.10	11.60
R	123	101	10.45	4.500	16.10	10.50
SIRE_BRD R						
DAM_BRD						
D	234	198	11.81	3.800	18.20	11.75
R	215	181	9.79	4.100	15.20	10.10

The difference between 'count' and 'Nobserved' gives the number of missing observations in each breed category. These are mostly the numbers of animals that died before weaning together with a few whose weights at weaning were not recorded.

Note that the largest number of losses is for the lambs born of Dorper rams and ewes.

The data analysis that follows disregards the cases for which the response variable, namely weaning weight, is not recorded. The 2-way tables below by breed give counts of the numbers of lambs recorded with weaning weights each year and for each age of dam.

Table 2a: *Numbers of lambs with recorded weaning weight (WEANWT), by year of birth and breed*

BREED	YEAR						Count
	91	92	93	94	95	96	
DD	71	49	49	15	23	13	220
RD	73	47	54	9	9	6	198
RR	0	7	31	40	53	50	181
DR	0	6	34	15	22	24	101
Count	144	109	168	79	107	93	700

Table 2b: *Numbers of lambs with recorded weaning weight (WEANWT) by age of dam and breed*

BREED	DAMAGE										Count
	1	2	3	4	5	6	7	8	9	10	
DD	0	24	49	38	47	32	22	8	0	0	220
RD	0	16	28	47	61	19	19	6	1	1	198
RR	1	17	41	51	40	31	0	0	0	0	181
DR	0	15	40	21	14	11	0	0	0	0	101
Count	1	72	158	157	162	93	41	14	1	1	700

The numbers of lambs born to Dorper ewes were greater during the first three years. In contrast, mating to the Red Maasai ewes did not start until 1992, and more lambs were born during the later years of the study. The total number of DR lambs weaned is approximately half the number of DD lambs. These observations reveal an imbalance in the data. In particular, there were no RR and DR lambs in 1991 and the RD lambs were few in number during the last three years of the study. Since there were different numbers of lambs born for the different breeds in the different years, it is important to take year of birth into account in the analysis, since the effect of breed on weaning weight is partially confounded with year.

The number of lambs characterised by age of dam also reveals an imbalance. The oldest Red Maasai ewes were aged 6 years whereas some Dorper ewes were older. From the numbers of lambs for each age category it can be seen that dams between the ages of 2 to 6 years were most common. Extreme age classes of 1, 9 and 10 years had only one lamb each. Since age of dam is a factor to be considered in the analysis of weaning weight of lambs, it would not be sensible to keep these classes separate. Thus, one could either omit these three records or pool them with existing ones. We have chosen to put age 1 year and 2 years together to form one class (2 years and below) and to put ages 9 and 10 years together into the age 8 year category to form an '8 years and above' class. When fitting a classification factor in a model it is always important to check that there are reasonable numbers of observations within each level; attempts to fit parameter terms to sparse data often leads to spurious estimates.

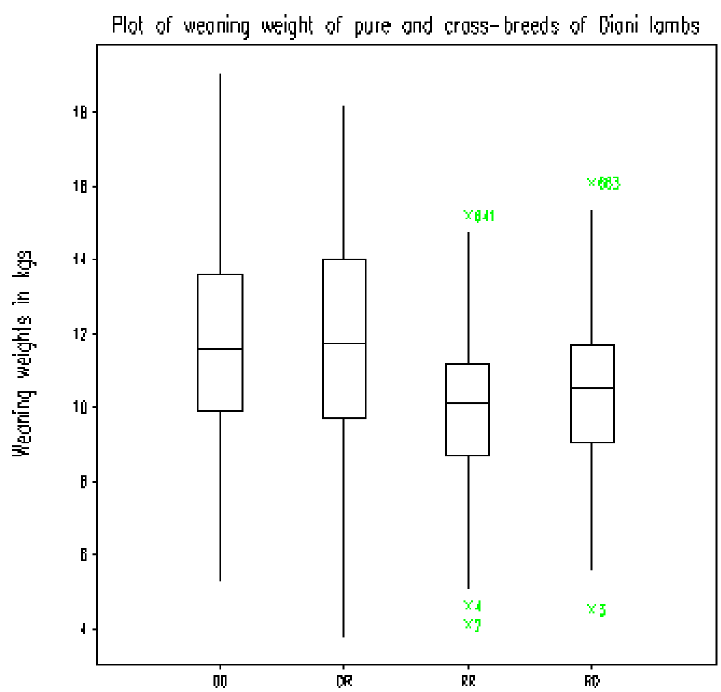


Figure 1: *Boxplot of weaning weight by breed.*

The figure above (known as a boxplot) reveals that the lambs sired by Dorper rams (the first two boxes) have a generally higher weaning weight than those sired by Red Maasai rams (the second two boxes). However, within breed groups, weaning weights would appear to be fairly normally distributed as indicated by the relative positions of the medians within the respective boxes, which contain half the data. The numbers indicated beyond the extremities of the lines that contain most of the additional data are potential outliers. One could investigate these weights further and decide on whether to exclude them from the analysis or not. We have decided to retain them for the analysis since they are not far from the ends of the lines. Thus, the assumption of normality required for analysis of variance appears to be valid.

The following boxplot, by age of dam, illustrates the association between weaning weight and age of dam. The plot shows that weaning weight increases with age of dam from 2 to 5 years of age and decreases from 6 years onwards. We can fit age as a factor with 7 levels. Alternatively, we may be able to represent the relationship, either by a polynomial curve, possibly up to order 3 (cubic), or by using fewer discrete subclasses, amalgamating some of the ages (e.g. 2, 3-4, 5-6, 7-8 years). These alternatives are considered later.

The distributions of weaning weight within each group of age of dam are also fairly normal as revealed by these plots. The spread of the weights is similar for all age of dam groups except possibly for lambs born to dams aged 6 years; which, apart from some ‘outliers’, shows a narrower distribution.

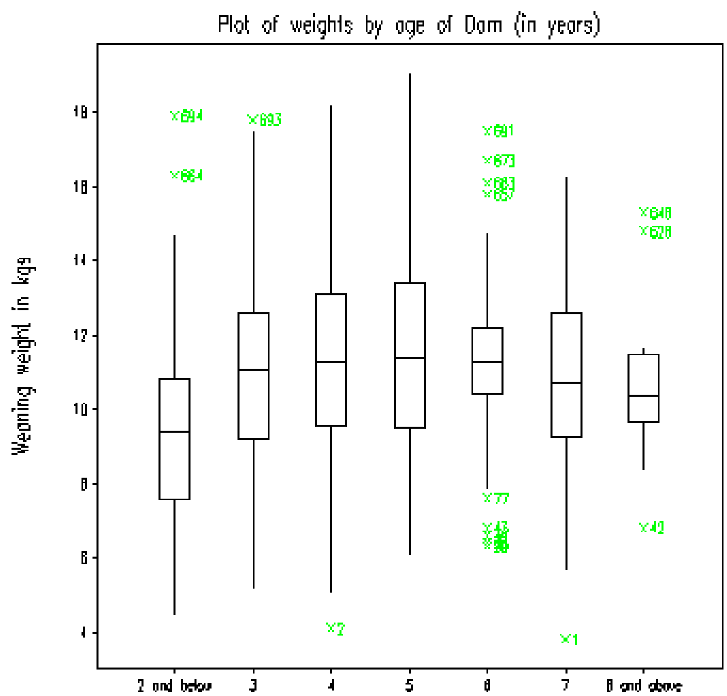
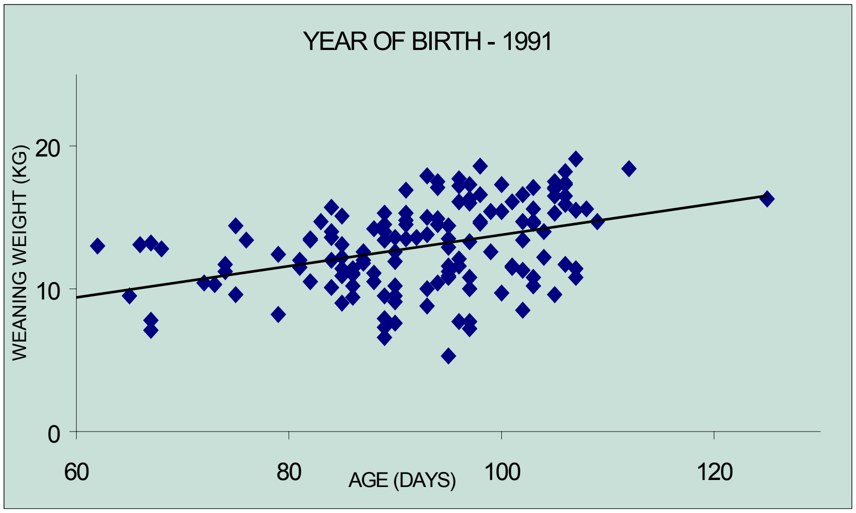


Figure 2: *Boxplot of weaning weights grouped by age of dam.*

Weaning weights recorded during the first two years are plotted below against age at weaning. There is a general pattern indicating a linear relationship with age. Similar patterns occurred for the other four years. A trend line added to each plot confirms this slight positive correlation. Age at weaning is therefore proposed for inclusion in the model as a continuous covariate in order to correct for its effect on weaning weight.



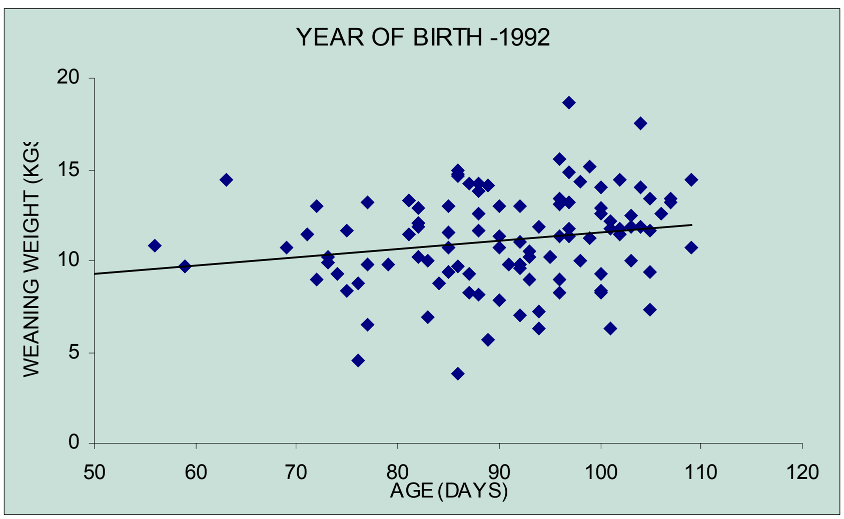


Figure 3: *Relationship between age of lamb and weaning weight.*

Least squares analysis of variance

Following our exploratory analysis a full least squares analysis is now undertaken to investigate the influence of each of the fixed effects on weaning weight. The model fitted includes term for:

BREED	(DD, RD, DR, RR)
YEAR	(1991, …., 1996)
SEX	(female, male)
WEANAGE	(linear regression)
DAMAGE	(2, 3, …..8)

Part of the output is shown below. It includes parameter estimates adjusted for other effects in the model and an accumulated analysis of variance indicating the sums of squares accounted for by each term as it is added. Thus, each sum of squares in the analysis of variance is adjusted for preceding terms. This does not apply, however, to parameter estimates. For each factor, parameter estimates are adjusted for all other factors regardless of the order in which they are included in the model.

Regression Analysis

Response variate: WEANWT
Fitted terms: Constant + BREED + YEAR + SEX + AGEWEAN + DAMAGE

Estimates of parameters

	estimate	s.e.	t(683)	t pr.
Constant	4.327	0.883	4.90	<.001
BREED RD	-0.408	0.222	-1.84	0.066
BREED RR	-1.008	0.272	-3.71	<.001
BREED DR	-0.493	0.306	-1.61	0.107
YEAR 92	-1.551	0.308	-5.03	<.001
YEAR 93	-1.228	0.291	-4.22	<.001
YEAR 94	-2.983	0.388	-7.69	<.001
YEAR 95	-3.258	0.346	-9.40	<.001
YEAR 96	-2.333	0.423	-5.51	<.001
SEX M	0.482	0.170	2.84	0.005
AGEWEAN	0.07058	0.00886	7.97	<.001
DAMAGE 3	1.833	0.319	5.75	<.001
DAMAGE 4	2.741	0.331	8.28	<.001
DAMAGE 5	2.742	0.322	8.52	<.001
DAMAGE 6	2.322	0.382	6.07	<.001
DAMAGE 7	1.754	0.462	3.79	<.001
DAMAGE 8 or more	1.405	0.647	2.17	0.030

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ BREED	3	570.427	190.142	38.68	<.001
+ YEAR	5	735.646	147.129	29.93	<.001
+ SEX	1	59.013	59.013	12.00	<.001
+ AGEWEAN	1	336.792	336.792	68.51	<.001
+ DAMAGE	6	445.076	74.179	15.09	<.001
Residual	683	3357.495	4.916		
Total	699	5504.450	7.875		

Notice first that the first level for each factor (i.e. those with discrete levels) is omitted. Each parameter estimate represents the deviation of the level of the factor it represents from the first. Thus, breed RD lambs have an average weaning weight 0.408kg less than breed DD lambs, when adjusted for other fixed effects in the model. This difference in weaning weight has a standard error of 0.222 kg. Looking at the corresponding t-value and t probability value it can be seen that this difference is not significant (P = 0.066). The interpretation of the WEANAGE term is different from that for discrete classification factors. This is a continuous covariate and so the value of 0.07058 kg/day represents the slope of the linear regression of WEANWT on WEANAGE adjusted for all other factors.

We shall now explore, a little more closely, different representations of the effect of age of dam on weaning weight. First, let us replace the 7 factor levels for age of dam by a quadratic function. With the codes DL and DQ representing the linear and quadratic terms for age of dam, respectively, the following output is obtained.

Regression Analysis

Response variate: WEANWT
Fitted terms: Constant + BREED + YEAR + SEX + AGEWEAN +DL + DQ

Estimates of parameters

	Estimat	es.e.	t(687)	t pr.
Constant	2.702	0.929	2.91	0.004
BREED RD	-0.389	0.220	-1.77	0.078
BREED RR	-1.040	0.271	-3.84	<.001
BREED DR	-0.511	0.304	-1.68	0.093
YEAR 92	-1.565	0.293	-5.34	<.001
YEAR 93	-1.099	0.276	-3.99	<.001
YEAR 94	-2.820	0.359	-7.85	<.001
YEAR 95	-3.215	0.346	-9.30	<.001
YEAR 96	-2.342	0.391	-5.99	<.001
SEX M	0.476	0.170	2.81	0.005
AGEWEAN	0.07026	0.00886	7.93	<.001
DL	2.188	0.249	8.79	<.001
DQ	-0.2688	0.0340	-7.90	<.001

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ BREED	3	570.427	190.142	38.50	<.001
+ YEAR	5	735.646	147.129	29.79	<.001
+ SEX	1	59.013	59.013	11.95	<.001
+ AGEWEAN	1	336.792	336.792	68.19	<.001
+ DL	1	101.581	101.581	20.57	<.001
+ DQ	1	308.044	308.044	62.37	<.001
Residual	687	3392.947	4.939		
Total	699	5504.450	7.875		

We see that the residual mean square is slightly increased from the value of 4.916 in the previous model to 4.939 kg^2 here, implying a slightly poorer fit. However, let us look at the analysis a little more closely.

From the results of the two analyses we can break down the sum of squares of 445.076 for DAMAGE in the first analysis and present it together with the residual line as follows:

	d.f	S.S.	M.S	F
DAMAGE	6	445.076	74.179	
DL	1	101.581	101.581	
DQ	1	308.044	308.044	
Remainder	4	35.451	8.863	1.80
Residual	683	3357.495	4.916	

The 'Remainder' term, which represents the DAMAGE variation not accounted for by the quadratic function, is not significant (F = 1.80). This implies that the quadratic fit is a good one, and possibly as good as can be expected considering that the size of the remaining variation is not statistically significant. Had the curve reduced the remainder mean square to a value below that of the residual mean square then it is likely that the curve may have been an overfit. Thus, we can argue that it is not necessary to add a cubic term. We decide not do so.

An alternative approach to the analysis is to try the same model with the age of dam grouped into fewer discrete categories. This has been done with 4 instead of the original 7 different age categories. The 4 levels chosen for the new groups are as suggested earlier, namely 2, 3-4, 5-6, 7-8 years.

The least squares analysis now becomes:

Estimates of parameters

	estimate	s.e.	t(686)	t pr.
Constant	4.159	0.888	4.68	<.001
BREED RD	-0.286	0.221	-1.29	0.197
BREED RR	-1.069	0.273	-3.92	<.001
BREED DR	-0.651	0.305	-2.14	0.033
YEARB 92	-1.591	0.304	-5.23	<.001
YEARB 93	-1.392	0.284	-4.91	<.001
YEARB 94	-2.763	0.366	-7.56	<.001
YEARB 95	-3.092	0.347	-8.92	<.001
YEARB 96	-2.278	0.416	-5.47	<.001
SEX 2	0.499	0.171	2.92	0.004
AGEWEAN	0.07260	0.00892	8.14	<.001
DAMAGE 2	2.401	0.286	8.40	<.001
DAMAGE 3	2.276	0.383	5.94	<.001
DAMAGE 4	1.619	0.427	3.79	<.001

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ BREED	3	570.427	190.142	37.99	<.001
+ YEAR	5	735.646	147.129	29.40	<.001
+ SEX	1	59.013	59.013	11.79	<.001
+ AGEWEAN	1	336.792	336.792	67.30	<.001
+ DAMAGE	3	369.467	123.156	24.61	<.001
Residual	686	3433.105	5.005		
Total	699	5504.450	7.875		

This analysis gave a slightly higher residual mean square than that for the quadratic function i.e. an increase from 4.939 to 5.005 kg^2. Note that the sum of squares accounted for by each of the other factors is, however, unchanged as they occupy the same positions in this analysis as they did in the previous one and are fitted before DAMAGE.

Applying the same steps as in the case of the quadratic function we obtain:

	d.f	S.S.	M.S.	F
DAMAGE	6	445.076	74.179	
Grouped	3	369.467	123.156	
Remainder	3	75.609	25.203	5.127
Residual	683	3357.495	4.916	

Here the remainder mean square is significantly greater than the residual mean square (P<0.01) so, compared with the quadratic, the reduced number of categories is not such a good representation of the association with age of dam.

We shall decide to use the quadratic relationship for DAMAGE in our final analysis. We now change the order in which the effects are fitted so that BREED is added last. For the purposes of this output GENSTAT has also provided the least squares means for breed together with their standard errors.

Regression Analysis

Response variate: WEANWT
Fitted terms: Constant + YEAR + SEX + AGEWEAN + DL + DQ +BREED

Estimates of parameters

	estimate	s.e.	t(687)	t pr.
Constant	2.702	0.929	2.91	0.004
YEAR 92	-1.565	0.293	-5.34	<.001
YEAR 93	-1.099	0.276	-3.99	<.001
YEAR 94	-2.820	0.359	-7.85	<.001
YEAR 95	-3.215	0.346	-9.30	<.001
YEAR 96	-2.342	0.391	-5.99	<.001
SEX M	0.476	0.170	2.81	0.005
AGEWEAN	0.07026	0.00886	7.93	<.001
DL	2.188	0.249	8.79	<.001
DQ	-0.2688	0.0340	-7.90	<.001
BREED RD	-0.389	0.220	-1.77	0.078
BREED RR	-1.040	0.271	-3.84	<.001
BREED DR	-0.511	0.304	-1.68	0.093

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ YEAR	5	1208.149	241.630	48.92	<.001
+ SEX	1	55.983	55.983	11.34	<.001
+ AGEWEAN	1	344.206	344.206	69.69	<.001
+ DL	1	151.513	151.513	30.68	<.001
+ DQ	1	275.795	275.795	55.84	<.001
+ BREED	3	75.857	25.286	5.12	0.002
Residual	687	3392.947	4.939		
Total	699	5504.450	7.875		

Response variate: WEANWT

BREED	Prediction	S.e.
DD	11.552	0.159
RD	11.163	0.176
RR	10.512	0.193
DR	11.041	0.240

Note that the parameter estimates remain the same regardless of the order in which the terms are added to the model. Comparing this output with the one given earlier in which the term BREED was fitted first, the mean square accounted for by BREED, after correcting for all the other terms in the model, is reduced from 190.142 to 25.286 kg^2. Breed differences are nevertheless still significant (P<0.01). No parameter estimate is shown for breed DD, which is used as a reference level against which the estimate for each of the other breeds is compared. The least squares 'predicted' means, however, are calculated for all four breeds. They indicate that the pure breed RR lambs had the lowest mean weaning weight of 10.512 kg, whilst the pure breed DD lambs had the highest mean weaning weight of 11.552 kg.

From the corresponding parameter estimates one can see that the RR parameter estimate of –1.04 is highly significant (P<0.001). In other words, the mean weaning weight of RR lambs was significantly lower, (by 1.04 kg), than that of DD lambs (P<0.001). The parameter estimates also suggest that the cross breed RD and DR lambs have mean weaning weights in between those for the pure breeds.

Study questions

1. One thing has been overlooked in this analysis. Consider again the design of the study and the way the data are structured, taking into account the lamb's parentage, and determine whether there is more than one level of experimental unit and, if so, which ones are appropriate for the different fixed effects in the model. What is the likely impact of this on the sizes of the standard errors for comparing fixed effects?

2. Adding the cubic term (DC) for DAMAGE to the model results in the following output.

Regression Analysis

Response variate: WEANWT
Fitted terms: Constant + BREED + YEAR + SEX + AGEWEAN + DL + DQ + DC

Estimates of parameters

	estimate	s.e.	t(686)	t pr.
Constant	0.99	1.13	0.88	0.378
BREED RD	-0.400	0.219	-1.82	0.069
BREED RR	-1.012	0.270	-3.76	<.001
BREED DR	-0.503	0.303	-1.66	0.097
YEAR 92	-1.555	0.292	-5.33	<.001
YEAR 93	-1.233	0.279	-4.42	<.001
YEAR 94	-2.954	0.361	-8.18	<.001
YEAR 95	-3.251	0.344	-9.44	<.001
YEAR 96	-2.325	0.389	-5.98	<.001
SEX M	0.482	0.169	2.85	0.004
AGEWEAN	0.07072	0.00882	8.01	<.001
DL	4.096	0.757	5.41	<.001
DQ	-0.853	0.222	-3.85	<.001
DC	0.0519	0.0194	2.67	0.008

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ BREED	3	570.427	190.142	38.84	<.001
+ YEAR	5	735.646	147.129	30.06	<.001
+ SEX	1	59.013	59.013	12.06	<.001
+ AGEWEAN	1	336.792	336.792	68.80	<.001
+ DL	1	101.581	101.581	20.75	<.001
+ DQ	1	308.044	308.044	62.93	<.001
+ DC	1	34.820	34.820	7.11	0.008
Residual	686	3358.126	4.895		
Total	699	5504.450	7.875		

Following the steps in the notes above, calculate the remainder sum of squares and compare it with the residual term. Comment on your findings. Would you include a cubic term or not?

3. The analysis has considered breed as 4 distinct entities or genotypes. These genotypes are results of a factorial mating of rams and ewes. Reparameterise the model to take into account this factorial structure and the interaction between parent breeds. Describe how the new effects will be shown in the output.

4. Write in two sentences a summary of the results given in this case study describing the difference in weaning weight between genotypes. Is there any more information not provided in the output from GENSTAT that you would ideally like to have?

5. The data analysis has been done on lambs that survived to weaning. You will note from an earlier table that survival rate was greater in the Red Maasai than the Dorper lambs. What implication do you think that this might have on the interpretation of the weaning weight results?

6. What do you know about traits in general in Dorper and Red Maasai breeds and how do such traits compare with those of sheep and goat breeds that are indigenous to, or raised in your own country?