Case study 3: Exploratory analysis of factors affecting weaning weight of Dorper and Red Maasai lambs

Damaris Yobera[a], James Audho[b] and Eric Aduda[b]
[a]Crop Science Department, University of Nairobi, P.O. Box 30197, Nairobi, Kenya.
[b]International Livestock Research Institute, P.O. Box 30709, Nairobi, Kenya.

Contents

Summary

This case study explores the patterns in relationships of various factors with weaning weights of Red Maasai and Dorper lambs and their crosses, such as year of birth, sex, age at weaning and age of dam. In particular, different formulations of the relationship between age of dam and weaning weight are compared. Ways of subdividing the sums of squares in an analysis of variance to evaluate alternative parameterisations of the statistical model are described. Parameters describing the patterns with these various factors are then included in a model to compare the weaning weights of the different genotypes when adjusted for these factors. Finally, alternative forms of presentation of results from general least squares analyses of variance are discussed.

This case study is part of a larger investigation to compare the performance of these different genotypes when exposed to helminthiasis. Case Study 4 evaluates the genetic components of variation associated with a lamb's dam or sire.

Glossary

A number of 'scientific' terms used in this case study, perhaps not familiar to the reader, are listed here.

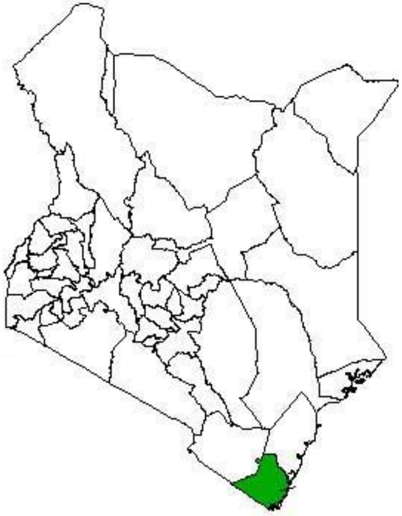
Anthelmintic treatment:	a chemical treatment used to treat gastro-intestinal parasites.
Dam:	the female parent of an animal, especially of domestic livestock.
Diallel design:	a two-way factorial design involving sire breed and dam breed in an animal experiment.
Faecal egg count:	a count of the number of parasite eggs found in a sample of faeces from an animal.
Genotype:	the genetic constitution of an animal.
Helminth:	a parasite found in the intestines of livestock.
Helminthiasis:	the disease caused by the presence of helminths.
Offspring:	the immediate descendant of an animal.
Packed cell volume:	the percentage of blood cells in an animal's blood (measurement obtained by spinning a sample in a centrifuge to separate blood cells from serum - the watery liquid component).
Sire:	the male parent of an animal, especially of domestic livestock.

Background

Helminths (parasites that reside in an animal's intestines) constitute one of the most important constraints to small ruminant livestock production in the tropics resulting in widespread infection in grazing animals, associated production losses, high costs of treatment and death. Current control methods in the tropics focus on reducing contamination of pastures through anthelmintic treatment of animals and/or controlled grazing. But there are problems with increasing frequencies of drug resistance.

An attractive, alternative and sustainable solution is the breeding for disease resistance. Indeed, anecdotal evidence suggests that, among the large and diverse range of indigenous breeds of sheep and goats in the tropics, there are some that appear to have the genetic ability to resist or tolerate helminthiasis. One of these is the Red Maasai breed found in East Africa and perceived to be resistant to helminthiasis. The Red Maasai is a fat-tailed sheep associated with the Maasai tribe found in northern Tanzania and south-central Kenya.

In 1990 ILRI decided to investigate the degree of resistance exhibited by this breed and initiated a study at Diani Estate of the Baobab Farms, 20 km south of Mombasa in the sub-humid coastal region of Kenya., ,

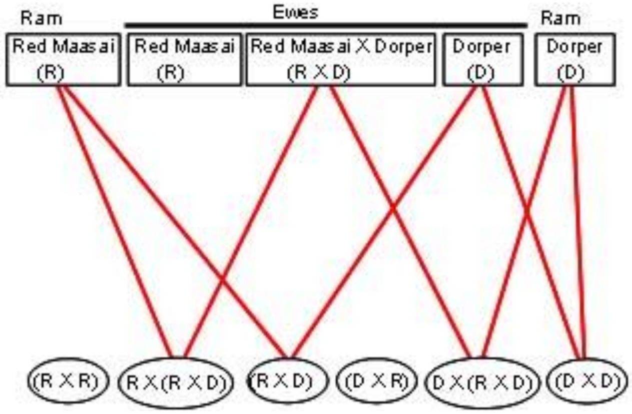


To do so, a susceptible breed, the Dorper, originally from South Africa, was chosen to provide a direct comparison with the Red Maasai. The Dorper breed was developed in South Africa in the 1940s by interbreeding the Dorset Horn and Black Head Persian breeds. The Dorper is particularly well adapted to harsh, arid conditions and was imported into Kenya in the 1960s. This breed is also larger than the Red Maasai, and this makes these sheep attractive to farmers.

Study design

The purpose of the experiment was to compare the genetic resistance to helminthiasis of the two indigenous breeds of sheep - Dorper and Red Maasai, together with their crosses. Throughout six years from 1991 to 1996 Dorper (D), Red Maasai (R) and Red Maasai - Dorper crossed ewes were mated to Red Maasai and Dorper rams to produce a number of different lamb genotypes.

In the first year 12 Dorper and 12 Red Maasai rams were mated to Dorper and Red Maasai x Dorper ewes (Red Maasai genes coming from the sire (ram) and Dorper genes coming from the dam (ewe)). Red Maasai ewes were not available in this year and the study formed what is known in the discipline of genetics as a partial diallel design.



The numbers of offspring, when summed over years, were estimated as being sufficient to determine genetic parameters with the required precision, and provided the number of lambs that could be accommodated and raised at the farm each year. By replicating over six years a range in different weather patterns was covered.

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 the abbreviation DD, DR, RD and RR, respectively, with the first letter referring to the breed of the sire and the second to the breed of the dam. Eight hundred and eighty two lambs within these genotypes were born during the six years 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, before being replaced. A few twins were born but these were not included in the experiment.

Lambs were weighed and blood and faecal samples were taken periodically over a period of about a year for each of the six batches of lambs born annually during the study. Measurements were made of blood packed cell volume (PCV), which reduces in volume when an animal becomes anaemic due to disease, and faecal egg count (FEC) which estimates the numbers of helminths present in the intestines. These measurements were made monthly up to weaning at about three months of age, and on average every two to three months thereafter to about 12 months of age. The periods from birth to weaning and from weaning to 12 months are distinct periods of growth and thus required separate analysis. Further details of the experimental design are given in Baker et al. (1999) and Baker et al. (2003).

Objectives

The objectives of the study were primarily:

- to compare the performance of the Red Maasai and Dorper breeds and their various crosses in terms of productivity under high disease risk
- to study genetic sources of variation among lambs within the two breeds and their crosses

In this example we shall consider the first objective but just look at weaning weight as the performance variable. We shall determine the effect of breed and other factors or covariates on weaning weight. Weaning of lambs took place on a single day each year when the average age of the lambs was 3 months. However, dates of birth varied and so weaning weight will, to some extent, have been influenced by the age of the lamb on that day.

Source: Isaac Kosgey

In **Case Study 4** we shall study the genetic variation expressed among offspring from different rams and ewes.

Questions to be addressed

The questions to be addressed herein are:

- Are there differences in weaning weight among the four genotypes?
- Are there other factors or covariates that can be included in the analysis to improve the precision with which weaning weight comparisons can be made among genotypes?
- One such covariate might be age of dam (ewe). How best can the association between age of dam and weaning weight be parameterised? Another possible covariate is age at weaning? How can this also be best incorporated in the model?
  - In answering these questions we shall first produce some summary statistics and graphs to understand how the data are distributed in relation to the different covariates.
  - Having decided how best to define our statistical model we shall, using weaning weight as the response variable, fit a least squares analysis of variance incorporating the various factors and covariates to reflect the patterns we have observed.

Source material

The complete data set used in this example is stored in the Excel file CS3Data. The fields are described in the associated word file CS3Doc but the same information is also included at the top of the data file. These cover both data originally collected in the study and others derived for the statistical analysis. A number of variables (both original and derived) have already been defined as factors.

Data management

The CS3Data data file has already been read by GenStat and a number of factors and levels defined.

Notice how, after exporting this file to Excel, the uppermost rows in the Excel file have been used to describe the variables that are stored below. This provides a useful method for documenting the data. This would normally be done once all statistical analysis had been completed and the data are ready for archiving.

Of course, the documented file may need to be read again. When opening the file in GenStat, the user must request that the first four non-empty rows to be skipped and state that the column names must appear in row 5 (i.e. click 'Next', then 'Specified Range' and type 5 in box alongside).

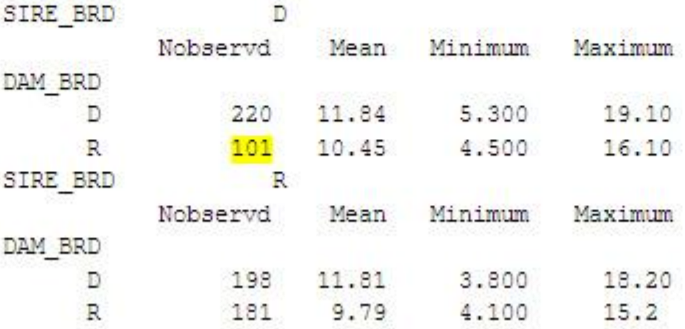
Exploration & description

Before undertaking a final statistical analysis it is useful to first explore the relationships between weaning weight and its covariates to see how best to define how these relationships might be included in the statistical model. In this example we hypothesise that, in addition to year of birth and sex, age of dam and age of lamb at weaning may also influence weaning weight. But how? The following pages show how we might do this.

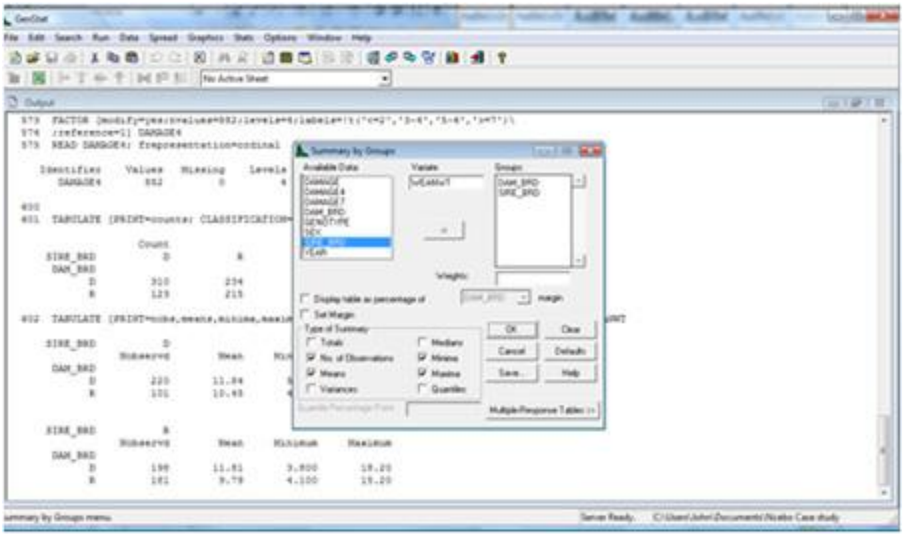
Firstly, let us look at some of the patterns in the data. The tabular output results from the steps Stats → Summary Statistics → Summaries by Groups (Tabulation...) and the production of the dialog box shown alongside.

The output includes the frequency distribution of recorded weaning weight by genotype. Animals that died before weaning or a few whose weaning weights were not recorded are excluded from 'Nobserved'. The total number of lambs born can be obtained withStats → Summary → Statistics Frequency Tables... Note that there are half the numbers of lambs at weaning for the RD genotype compared with the others.

Note that there are half the numbers of lambs at weaning for the RD genotype compared with the others; also that the largest proportion of losses ((310-220)/310)=0.29 is for pure Dorper lambs compared with ((215-181)/215)=0.16 for pure Red Maasai lambs.



```
SIRE_BRD               D
         Nobservd    Mean   Minimum   Maximum
DAM_BRD
       D      220   11.84     5.300     19.10
       R      101   10.45     4.500     16.10
SIRE_BRD               R
         Nobservd    Mean   Minimum   Maximum
DAM_BRD
       D      198   11.81     3.800     18.20
       R      181    9.79     4.100     15.2
```

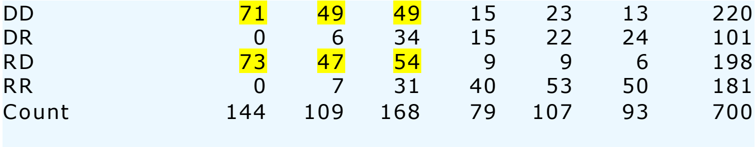


```
573  FACTOR [modify=yes;nvalues=852;levels=4;labels=!t('<=2','3-4','5-6','>=7')\
574  ;reference=1] DAMAGE4
575  READ DAMAGE4; frepresentation=ordinal

   Identifier   Values   Missing    Levels
     DAMAGE4       852         0         4

430
431  TABULATE [PRINT=counts; CLASSIFICATION=

                 Count
  SIRE_BRD           D          R
   DAM_BRD
         D         310        234
         R         123        215

432  TABULATE [PRINT=nobs,means,minima,maxima

  SIRE_BRD           D
              Nobservd       Mean
   DAM_BRD
         D         220      11.84
         R         101      10.45

  SIRE_BRD           R
              Nobservd       Mean    Minimum    Maximum
   DAM_BRD
         D         198      11.81      3.800      18.20
         R         181       9.79      4.100      15.20
```

The data exploration that follows disregards the cases for which the response variable weaning weight was not recorded. This can be achieved by first using the GenStat Spread → Restrict/Filter command to exclude missing values for weaning weight. Two-way frequency tables can then be produced by Stats → Summary Statistics → Frequency Tables … and then completing the dialog box that appears to give counts of the numbers of lambs recorded with weaning weights each year and for each lamb genotype.

The numbers of lambs (DD and RD) born to Dorper ewes were more during the former three than the latter three years.

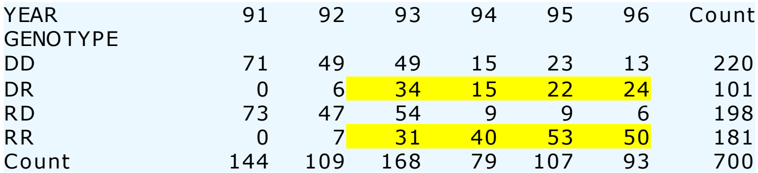
```
YEAR              91     92     93     94     95     96    Count
GENOTYPE
```



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

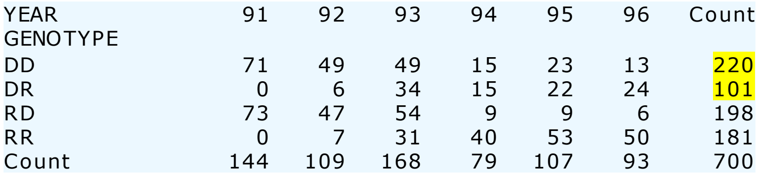
The data exploration that follows disregards the cases for which the response variable weaning weight was not recorded. This can be achieved by first using the GenStat Spread → Restrict/Filter command to exclude missing values for weaning weight. Two-way frequency tables can then be produced by Stats → Summary Statistics → Frequency Tables… and then completing the dialog box that appears to give counts of the numbers of lambs recorded with weaning weights each year and for each lamb genotype.

In contrast, mating to the Red Maasai ewes did not start until 1992, and more lambs were born to Red Maasai than Dorper ewes during the latter years of the study.



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

By weaning there were approximately half the number of DR lambs in the study compared with the 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 it is clear that the effect of genotype on weaning weight is partially confounded with year.



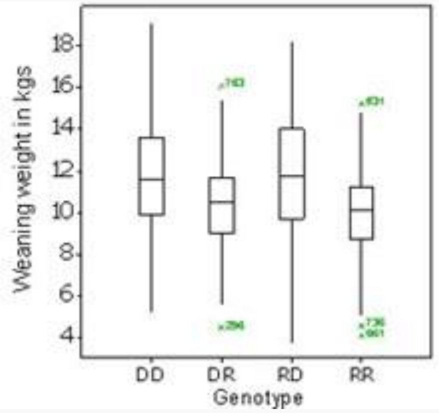
```
YEAR            91    92    93    94    95    96   Count
GENOTYPE
DD              71    49    49    15    23    13     220
DR               0     6    34    15    22    24     101
RD              73    47    54     9     9     6     198
RR               0     7    31    40    53    50     181
Count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

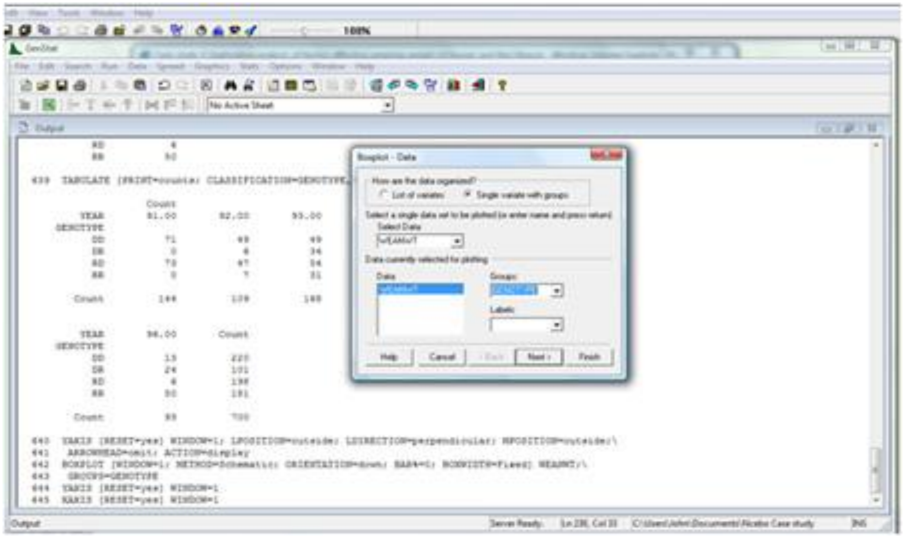
The number of lambs characterised by age of dam also reveals a frequency imbalance. The oldest Red Maasai ewes were aged 6 years whereas one Dorper ewe was as old as 10 years. From the numbers of lambs for each age category it can be seen that dams between the ages of 2 and 6 years were most common. 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. One can 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. The column DAMAGE7 inCS3Data has been created using the GenStat command Spread → Factor → Change levels… to put the extreme values into the neighbouring categories.

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

When fitting a classification factor in a statistical model it is always important to check that there are reasonable numbers of observations within each category level; attempts to fit parameter terms to sparse data often leads to spurious estimates.

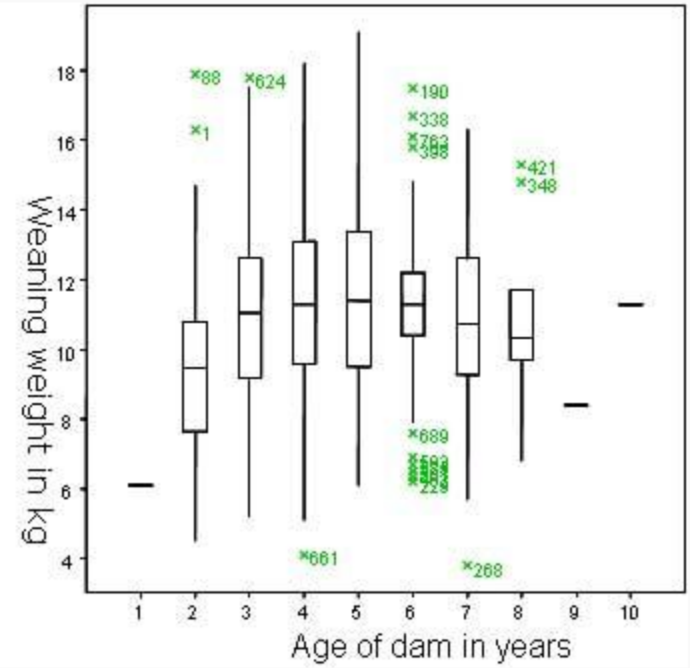
A box plot by genotype (Graphics → Boxplot...) and completing the dialog box reveals that the lambs born to Dorper dams (the first and third boxes) appear to have a generally higher weaning weight than those born to Red Maasai dams (the second and fourth boxes). However, weaning weights within genotypes appear to be fairly normally distributed, as indicated by the relative positions of the medians within the respective boxes that contain half the data. The numbers 763, 531, 296, 661 and 736 indicated beyond the extremities of the vertical lines point to the record numbers in CS3Data that are 'outliers'.




GenStat
File Edit Search Run Data Spread Graphics Stats Options Window Help
No Active Sheet
Output
Boxplot - Data
How are the data organized?
○ List of variates ● Single variate with groups
Select a single data set to be plotted (or enter name and press return)
Select Data
WEANWT
Data currently selected for plotting
Data
WEANWT
Groups:
GENOTYPE
Labels:
Help
Cancel
< Back
Next >
Finish
Server Ready
Ln 238, Col 33

```
         RD          4
         RR         50

439  TABULATE [PRINT=counts; CLASSIFICATION=GENOTYPE,

                 Count
        YEAR     81.00       82.00       83.00
    GENOTYPE
          DD        71          88          48
          DR         0          43          34
          RD        72          68          54
          RR         0           7          31

       Count       144         178         168


        YEAR     84.00       Count
    GENOTYPE
          DD        13         220
          DR        24         101
          RD         4         198
          RR        50         181

       Count        89         700

440  YAXIS [RESET=yes] WINDOW=1; LPOSITION=outside; LDIRECTION=perpendicular; NPOSITION=outside;\
441   ARROWHEAD=omit; ACTION=display
442  BOXPLOT [WINDOW=1; METHOD=Schematic; ORIENTATION=down; BAR=1; BOXWIDTH=fixed] WEANWT;\
443   GROUPS=GENOTYPE
444  YAXIS [RESET=yes] WINDOW=1
445  XAXIS [RESET=yes] WINDOW=1
```

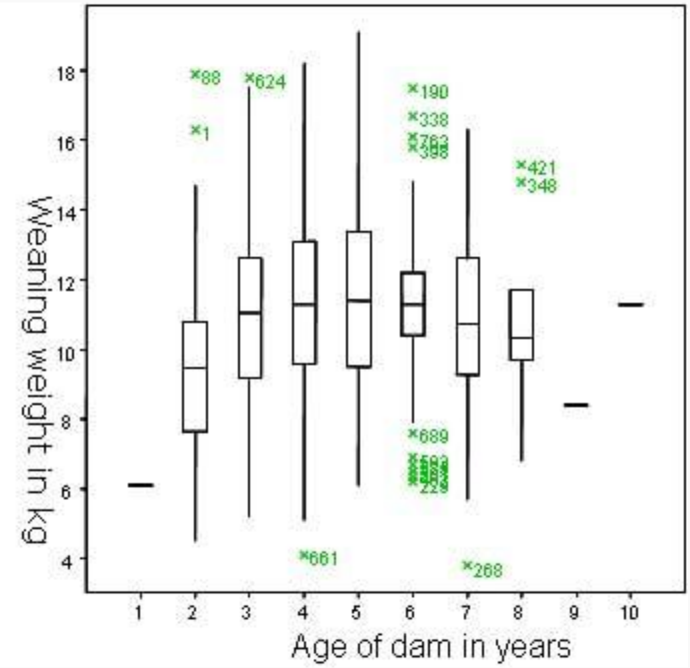
This box plot by age of dam illustrates the association between weaning weight and the age of a lamb's dam. There are more 'outliers' shown in this diagram than the one for genotype. This is probably because the variation among genotypes is not accounted for in this series of boxplots. The plot shows that an offspring's weaning weight appears to increase as a dam increases in age from 2 to 5 years and to decrease from 6 years onwards. We can fit age as a factor with seven levels.



This box plot illustrates the association between weaning weight and the age of a lamb's dam. There are more `outliers' shown in this diagram than the one for genotype. This is probably because the variation among genotypes is not accounted for in this series of boxplots. The plot shows that an offspring's weaning weight appears to increase as a dam increases in age from 2 to 5 years and to decrease from 6 years onwards. We can fit age as a factor with seven levels.

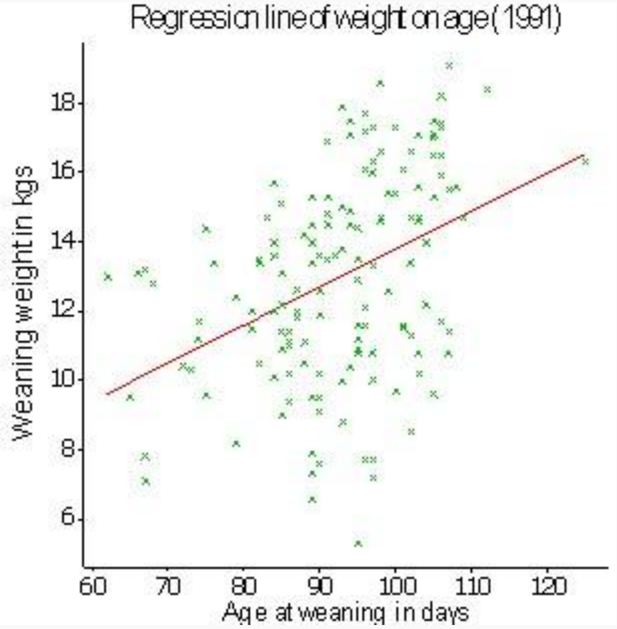
Alternatively, we may be able to represent the relationship, either by a polynomial curve, possibly up to order 3 (cubic), or by amalgamating some of the ages by using fewer discrete subclasses (e.g. 2, 3-4, 5-6, 7-8 years). These alternatives are considered later.

For each age the distributions of lamb



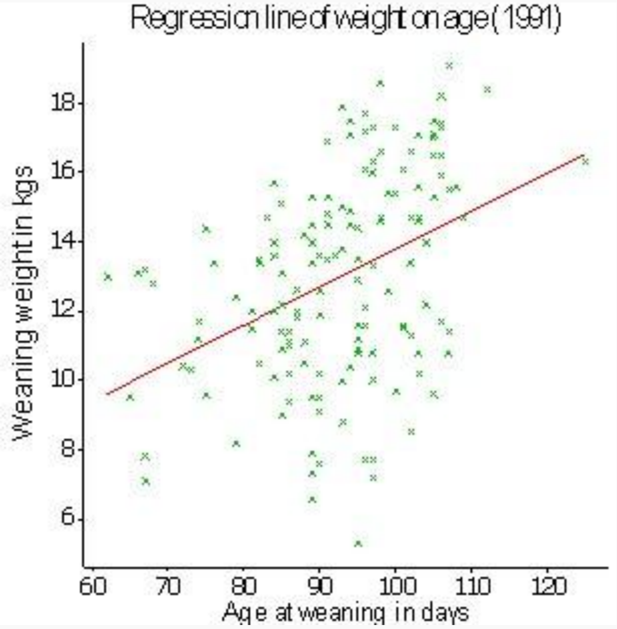
weaning weights are also fairly normal as revealed by the box plot. The spread of the weights is similar for all age of a dam except possibly that for lambs born to dams aged 6 years.

By restricting the data to 1991 and then fitting a regression line **Stats** → **Regression** → **Analysis** → **Linear Models...** and completing the dialog box, a regression analysis of weaning weight on weaning age can be obtained. By clicking **Further Output**, then, clicking **Fitted Model** in the next dialog box, the scatter plot plus the fitted regression line, as shown for weaning weight versus age at weaning is produced.



Lambs were born over a short period spanning a few weeks but were all weaned together on the same day. There is a general pattern indicating a linear relationship of weaning weight with age.

By similarly restricting the data to 1992 a second scatter plot is produced. This can be repeated for 1993, and so on. It can similarly be shown that there are similar patterns for the other four years. 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.



In summary, we can deduce that the statistical model to be fitted needs to include terms for year, age of dam (either as fixed effects or as a polynomial regression) and a linear term for age at weaning. We have not compared weaning weights of male and female lambs but it is well known that male lambs grow faster than females. Thus, sex will invariably be included in a model such as this.

Statistical modelling

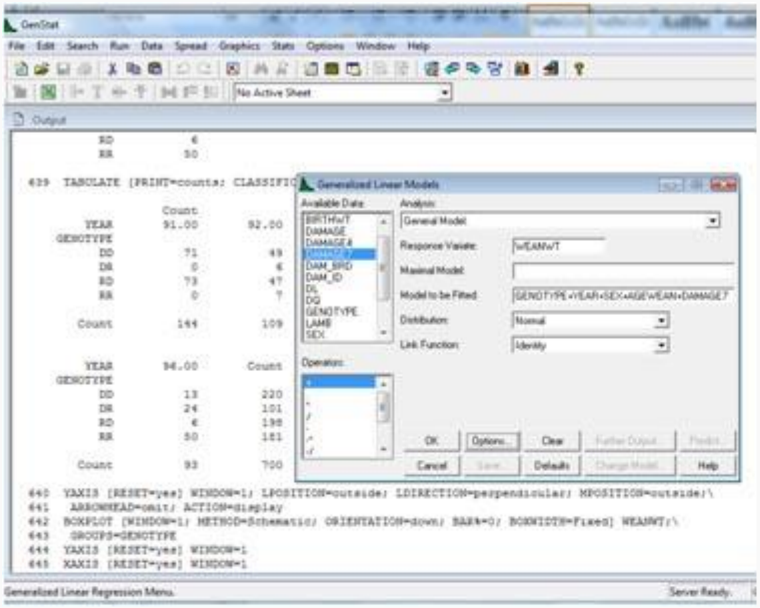
Least squares analysis of variance

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

Factor	Levels
GENOTYPE	DD, RD, DR, RR
YEAR	1991, ..., 1996
SEX	female, male
WEANAGE	linear regression
DAMAGE7	≤ 2, 3, 4, 5, 6, 7, ≥ 8

We can fit the model by using the dialog box shown below obtained by Stats → Regression Analysis → Generalized Linear Models... Then by clicking the Optionsbutton, then ticking Accumulated, an analysis of variance is shown which gives the sums of squares accounted for by each parameter in the model.

```
***** Regression Analysis *****
  Response variate: WEANWT
      Fitted terms: Constant + GENOTYPE + YEAR +
                 SEX + AGEWEAN + DAMAGE7
*** Estimates of parameters ***


Parameter              estimate       s.e.          t
Constant                  4.327     0.883        4.90
GENOTYPE DR              -0.493     0.306       -1.61
GENOTYPE RD              -0.408     0.222       -1.84
GENOTYPE RR              -1.008     0.272       -3.71
YEAR 92                  -1.551     0.308       -5.03
YEAR 93                  -1.228     0.291       -4.22
YEAR 94                  -2.983     0.388       -7.69
YEAR 95                  -3.258     0.346       -9.40
YEAR 96                  -2.333     0.423       -5.51
SEX M                     0.482     0.170        2.84
AGEWEAN                 0.07058   0.00886        7.97
DAMAGE7 3                 1.833     0.319        5.75
DAMAGE7 4                 2.741     0.331        8.28
DAMAGE7 5                 2.742     0.322        8.52
DAMAGE7 6                 2.322     0.382        6.07
DAMAGE7 7                 1.754     0.462        3.79
DAMAGE7 >=8               1.405     0.647        2.17


*** Accumulated analysis of variance ***


Change               d.f.         s.s.        m.s.      v.r.
+GENOTYPE               3      570.427     190.142     38.68
```



```
          RD          6
          RR         50

 639  TABULATE [PRINT=counts; CLASSIFIC

                   Count
      YEAR         91.00        92.00
  GENOTYPE
          DD          71           49
          DR           0            6
          RD          73           47
          RR           0            7

     Count           144          109


      YEAR         96.00        Count
  GENOTYPE
          DD          13          220
          DR          24          101
          RD           6          198
          RR          50          181

     Count            93          700

 640  YAXIS [RESET=yes] WINDOW=1; LPOSITION=outside; LDIRECTION=perpendicular; MPOSITION=outside;\
 641   ARROWHEAD=omit; ACTION=display
 642  BOXPLOT [WINDOW=1; METHOD=Schematic; ORIENTATION=down; BAR%=0; BOXWIDTH=Fixed] WEANWT;\
 643   GROUPS=GENOTYPE
 644  YAXIS [RESET=yes] WINDOW=1
 645  XAXIS [RESET=yes] WINDOW=1
```

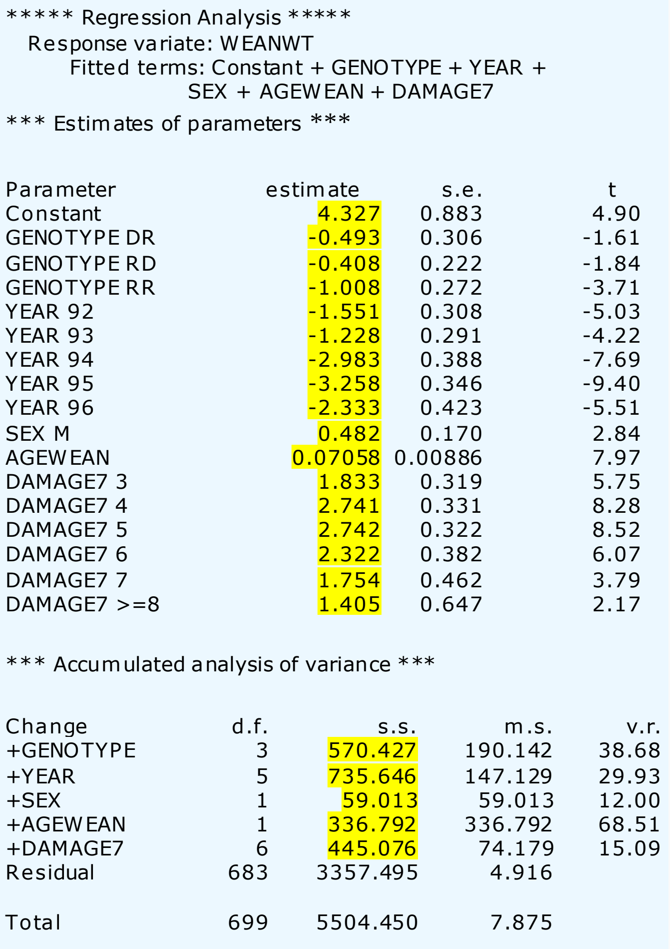
Analysis: General Model
Response Variate: WEANWT
Maximal Model:
Model to be Fitted: GENOTYPE+YEAR+SEX+AGEWEAN+DAMAGE7
Distribution: Normal
Link Function: Identity

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 DR lambs have an average weaning weight 0.493 kg less than breed DD lambs, when adjusted for other fixed effects in the model. This difference in weaning weight has a standard error of 0.306 kg. The absolute Student's t-value of 1.61 is less than 2, the approximate value that t needs to exceed to be significant (P<0.05). In contrast the parameter estimate for RR lambs is statistically significant (absolute t ≥ 2).

```
+YEAR              5     735.646     147.129     29.93
+SEX               1      59.013      59.013     12.00
+AGEWEAN           1     336.792     336.792     68.51
+DAMAGE7           6     445.076      74.179     15.09
Residual         683    3357.495       4.916

Total            699    5504.450       7.875
```

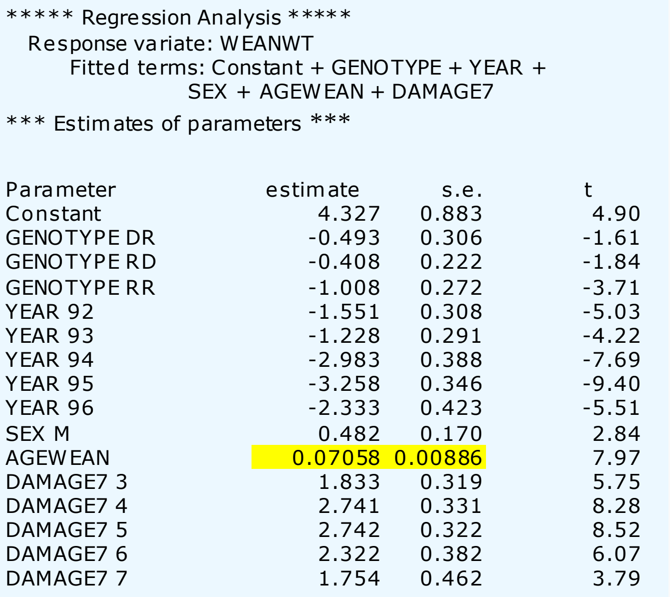
```
***** Regression Analysis *****
  Response variate: WEANWT
      Fitted terms: Constant + GENOTYPE + YEAR +
                  SEX + AGEWEAN + DAMAGE7
*** Estimates of parameters ***

Parameter                estimate       s.e.          t
Constant                    4.327      0.883       4.90
GENOTYPE DR                -0.493      0.306      -1.61
GENOTYPE RD                -0.408      0.222      -1.84
GENOTYPE RR                -1.008      0.272      -3.71
YEAR 92                    -1.551      0.308      -5.03
YEAR 93                    -1.228      0.291      -4.22
YEAR 94                    -2.983      0.388      -7.69
YEAR 95                    -3.258      0.346      -9.40
YEAR 96                    -2.333      0.423      -5.51
SEX M                       0.482      0.170       2.84
AGEWEAN                   0.07058    0.00886       7.97
DAMAGE7 3                   1.833      0.319       5.75
DAMAGE7 4                   2.741      0.331       8.28
DAMAGE7 5                   2.742      0.322       8.52
DAMAGE7 6                   2.322      0.382       6.07
DAMAGE7 7                   1.754      0.462       3.79
DAMAGE7 >=8                 1.405      0.647       2.17


*** Accumulated analysis of variance ***


Change              d.f.        s.s.        m.s.      v.r.
+GENOTYPE              3     570.427     190.142     38.68
+YEAR                  5     735.646     147.129     29.93
+SEX                   1      59.013      59.013     12.00
+AGEWEAN               1     336.792     336.792     68.51
+DAMAGE7               6     445.076      74.179     15.09
Residual             683    3357.495       4.916

Total                699    5504.450       7.875
```

In addition, the output contains parameter estimates adjusted for other effects in the model, and an accumulated analysis of variance indicating the sums of squares accounted for as each term 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.

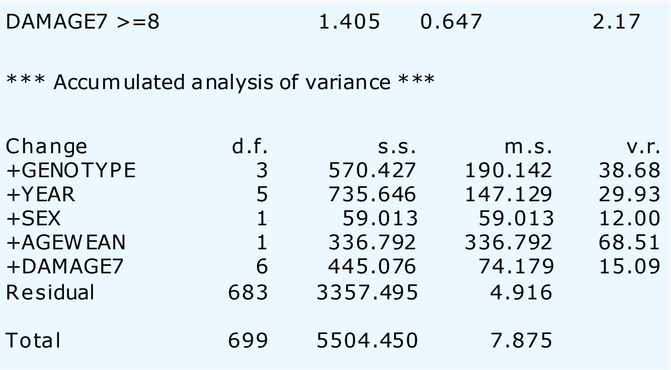


```
***** Regression Analysis *****
  Response variate: WEANWT
      Fitted terms: Constant + GENOTYPE + YEAR +
                 SEX + AGEWEAN + DAMAGE7
*** Estimates of parameters ***

Parameter                   estimate       s.e.           t
Constant                       4.327      0.883        4.90
GENOTYPE DR                   -0.493      0.306       -1.61
GENOTYPE RD                   -0.408      0.222       -1.84
GENOTYPE RR                   -1.008      0.272       -3.71
YEAR 92                       -1.551      0.308       -5.03
YEAR 93                       -1.228      0.291       -4.22
YEAR 94                       -2.983      0.388       -7.69
YEAR 95                       -3.258      0.346       -9.40
YEAR 96                       -2.333      0.423       -5.51
SEX M                          0.482      0.170        2.84
AGEWEAN                      0.07058    0.00886        7.97
DAMAGE7 3                      1.833      0.319        5.75
DAMAGE7 4                      2.741      0.331        8.28
DAMAGE7 5                      2.742      0.322        8.52
DAMAGE7 6                      2.322      0.382        6.07
DAMAGE7 7                      1.754      0.462        3.79
DAMAGE7 >=8                    1.405      0.647        2.17

*** Accumulated analysis of variance ***

Change              d.f.         s.s.        m.s.      v.r.
+GENOTYPE              3      570.427     190.142     38.68
+YEAR                  5      735.646     147.129     29.93
+SEX                   1       59.013      59.013     12.00
+AGEWEAN               1      336.792     336.792     68.51
+DAMAGE7               6      445.076      74.179     15.09
Residual             683     3357.495       4.916

Total                699     5504.450       7.875
```

The interpretation of the AGEWEAN term is simpler than that for a factor with discrete levels such as DAMAGE7. This is a continuous covariate and so the value of 0.07058 (± 0.00886) kg/day represents the slope (± s.e.) of the linear regression of WEANWT on AGEWEAN adjusted for all other factors.



```
***** Regression Analysis *****
  Response variate: WEANWT
      Fitted terms: Constant + GENOTYPE + YEAR +
                 SEX + AGEWEAN + DAMAGE7
*** Estimates of parameters ***

Parameter                   estimate       s.e.           t
Constant                       4.327      0.883        4.90
GENOTYPE DR                   -0.493      0.306       -1.61
GENOTYPE RD                   -0.408      0.222       -1.84
GENOTYPE RR                   -1.008      0.272       -3.71
YEAR 92                       -1.551      0.308       -5.03
YEAR 93                       -1.228      0.291       -4.22
YEAR 94                       -2.983      0.388       -7.69
YEAR 95                       -3.258      0.346       -9.40
YEAR 96                       -2.333      0.423       -5.51
SEX M                          0.482      0.170        2.84
AGEWEAN                      0.07058    0.00886        7.97
DAMAGE7 3                      1.833      0.319        5.75
DAMAGE7 4                      2.741      0.331        8.28
DAMAGE7 5                      2.742      0.322        8.52
DAMAGE7 6                      2.322      0.382        6.07
DAMAGE7 7                      1.754      0.462        3.79
```



```
DAMAGE7 >=8              1.405    0.647          2.17

*** Accumulated analysis of variance ***

Change              d.f.        s.s.        m.s.      v.r.
+GENOTYPE              3     570.427     190.142     38.68
+YEAR                  5     735.646     147.129     29.93
+SEX                   1      59.013      59.013     12.00
+AGEWEAN               1     336.792     336.792     68.51
+DAMAGE7               6     445.076      74.179     15.09
Residual             683    3357.495       4.916

Total                699    5504.450       7.875
```

We shall now explore, a little more closely, different representations of the effect of age of dam on weaning weight. The DAMAGE and residual lines in the analysis of variance (see previous screen view) read:

Source of variation	d.f	s.s.	m.s.	v.r.
+DAMAGE7	6	445.076	74.179	15.09
Residual	683	3357.495	4.916	

First, let us replace the seven factor levels for DAMAGE7 by a quadratic function. We can do this by inserting the following two lines into the GenStat Input Window obtained via File → New... and submitting them for execution.

```
CALCULATE DL=DAMAGE7
CALCULATE DQ=DL*DL
```

Alternatively we can use Spread → Calculate → Column... and insert the appropriate calculations there.

With the codes DL and DQ representing the linear and quadratic terms for age of dam, respectively, the output, as shown, can be obtained.

We see that the residual mean square is slightly increased from the value of 4.916 kg2 in the previous model to 4.939 kg2 here, implying a slightly poorer fit.

Note, however, that the sum of squares accounted for by each of the other factors is unchanged as they occupy the same positions in this analysis as they did in the previous one and are fitted before DL and DQ.

However, let us look at the analysis a little more closely.

```
***** Regression Analysis *****
   Response variate: WEANWT
        Fitted terms: Constant + GENOTYPE + YEAR +
                                SEX + AGEWEAN + DL + DQ
*** Estimates of parameters ***

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

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

From the results of the two analyses we can break down the sum of squares of 445.076 for DAMAGE7 in the first analysis into components for DL and DQ (101.581 and 308.044 in the second analysis - see previous screen) and a remainder (445.076 - 101.581 - 308.044 = 35.451). Presenting these values together with the residual line we get:

Source of variation	d.f	s.s	m.s.	v.r.
DAMAGE7	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 DAMAGE7 variation not accounted for by the quadratic function, is not significant (VR = 1.80). Since the size of this remaining variation is not statistically significant it can be deduced that the quadratic fit is a good one. We can also argue that it is not necessary to add a cubic term to the polynomial equation and decide not to do so.

An alternative approach to the statistical analysis is to try the same model again but with the age of dam grouped into fewer discrete categories. DAMAGE7 has been recoded as DAMAGE4

```
***** Regression Analysis *****
   Response variate: WEANWT
        Fitted terms: Constant + GENOTYPE + YEAR +
                        SEX + AGEWEAN + DAMAGE4
```

with four instead of seven age categories. The four levels chosen for the new categories are, as suggested earlier, namely 2, 3-4, 5-6 and 7-8 years.

This analysis gives a slightly higher residual mean square than that originally obtained, i.e. an increase from 4.916 to 4.991 kg^2.

```
*** Estimates of parameters ***

                              Parameter         s.e.      t(686)
Constant                          4.170        0.887        4.70
GENOTYPE RD                      -0.297        0.221       -1.34
GENOTYPE RR                      -1.022        0.270       -3.79
GENOTYPE DR                      -0.573        0.302       -1.90
YEAR 92                          -1.606        0.296       -5.42
YEAR 93                          -1.363        0.281       -4.85
YEAR 94                          -2.734        0.363       -7.54
YEAR 95                          -3.155        0.345       -9.14
YEAR 96                          -2.415        0.378       -6.39
SEX M                             0.493        0.171        2.89
AGEWEAN                         0.07242      0.00890        8.13
DAMAGE4 3-4                       2.260        0.297        7.62
DAMAGE4 5-6                       2.549        0.305        8.36
DAMAGE4 >=7                       1.642        0.425        3.86

*** Accumulated analysis of variance ***

Change                     d.f.        s.s.       m.s.          v.r.
+ GENOTYPE                    3     570.427    190.142         38.10
+ YEAR                        5     735.646    147.129         29.48
+ SEX                         1      59.013     59.013         11.82
+ AGEWEAN                     1     336.792    336.792         67.48
+ DAMAGE4                     3     378.859    126.286         25.30
Residual                    686    3423.713      4.991

Total                       699    5504.450      7.875
```

Doing the same steps as for fitting the quadratic function we obtain:

Source of variation	d.f	s.s	m.s.	v.r.
DAMAGE7	6	445.076	74.179	
DAMAGE4	3	378.859	126.286	
Remainder	3	66.217	22.072	4.49
Residual	683	3357.495	4.916	

Here the remainder mean square is significantly greater than the residual mean square ($F_{3,683}$ = 4.49; P<0.01), so, the reduced number of categories is not as good a representation of the association with age as that with seven categories.

We decide to use the quadratic relationship with DL and DQ in our final analysis.

We now change the order in which the effects are fitted so that GENOTYPE is added last. For the purposes of this output GenStat has also provided the predicted least squares means for each genotype (obtained by clicking the Predict button

```
Response variate: WEANWT
   Fitted terms: Constant + YEAR + SEX + AGEWEAN + DL
                       + DQ + GENOTYPE
*** Estimates of parameters ***
                              estimate          s.e.      t(687)
Constant                          0.25          1.07        0.23
YEAR 92                         -1.565         0.293       -5.34
YEAR 93                         -1.099         0.276       -3.99
YEAR 94                         -2.820         0.359       -7.85
YEAR 95                         -3.215         0.346       -9.30
YEAR 96                         -2.342         0.391       -5.99
SEX M                            0.476         0.170        2.81
AGEWEAN                        0.07026       0.00886        7.93
DL                               2.725         0.315        8.65
```

that appears after the model has been fitted) together with their standard errors.

Note that the parameter estimates remain the same regardless of the order in which the terms are added to the model. However, compairing this output with the one given earlier in which the term GENOTYPE was fitted first, the mean square accounted for by GENOTYPE, after correcting for all the other terms in the model, is reduced from 190.142 to 25.286 kg2. This demonstrates the impact of the partial confounding of other factors, particularly year, on the variation that can be attributed uniquely to genotype. Genotype differences, nevertheless, remain significant (P < 0.01)

```
DQ                          -0.269        0.034      -7.90
GENOTYPE RD                 -0.389        0.220      -1.77
GENOTYPE RR                 -1.040        0.271      -3.84
GENOTYPE DR                 -0.511        0.304      -1.68
*** Accumulated analysis of variance ***
Change                 d.f.        s.s.         m.s.        v.r.
+YEAR                     5    1208.149      241.630       48.92
+SEX                      1      55.983       55.983       11.34
+AGEWEAN                  1     344.206      344.206       69.69
+DL                       1     151.513      151.513       30.68
+DQ                       1     275.795      275.795       55.84
+GENOTYPE                 3      75.857       25.286        5.12
Residual                687    3392.947        4.939
Total                   699    5504.450        7.875
GENOTYPE          Prediction       s.e.
DD                    11.552      0.159
DR                    11.041      0.240
RD                    11.163      0.176
RR                    10.512      0.193
```

```
Response variate: WEANWT
   Fitted terms: Constant + YEAR + SEX + AGEWEAN + DL
                  + DQ + GENOTYPE
*** Estimates of parameters ***
                          estimate         s.e.     t(687)
Constant                      0.25         1.07       0.23
YEAR 92                     -1.565        0.293      -5.34
YEAR 93                     -1.099        0.276      -3.99
YEAR 94                     -2.820        0.359      -7.85
YEAR 95                     -3.215        0.346      -9.30
YEAR 96                     -2.342        0.391      -5.99
SEX M                        0.476        0.170       2.81
AGEWEAN                    0.07026      0.00886       7.93
DL                           2.725        0.315       8.65
DQ                          -0.269        0.034      -7.90
GENOTYPE RD                 -0.389        0.220      -1.77
GENOTYPE RR                 -1.040        0.271      -3.84
GENOTYPE DR                 -0.511        0.304      -1.68
*** Accumulated analysis of variance ***
Change                 d.f.        s.s.         m.s.        v.r.
+YEAR                     5    1208.149      241.630       48.92
+SEX                      1      55.983       55.983       11.34
+AGEWEAN                  1     344.206      344.206       69.69
+DL                       1     151.513      151.513       30.68
+DQ                       1     275.795      275.795       55.84
```

No parameter estimate is shown for the pure breed DD, which is used as a reference level against which the estimate for each of the other genotypes is compared. The least squares 'predicted' means, however, are calculated for all four genotypes. They indicate that the purebred RR lambs had the lowest mean weaning weight of 10.512 ± 0.193 kg, whilst the purebred DD lambs had the highest mean weaning weight of 11.552 ± 0.159 kg.

```
+GENOTYPE                3        75.857          25.286          5.12
Residual               687      3392.947           4.939
Total                  699      5504.450           7.875
GENOTYPE          Prediction        s.e.
DD                11.552           0.159
DR                11.041           0.240
RD                11.163           0.176
RR                10.512           0.193
```

rom the corresponding parameter estimates it can be seen that the RR parameter estimate of -1.04 is highly significant (t=-3.84, 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 crossbred RD and DR lambs have a similar mean weaning weight lying between those for the pure breeds.

```
Response variate: WEANWT
   Fitted terms: Constant + YEAR + SEX + AGEWEAN + DL
                       + DQ + GENOTYPE
*** Estimates of parameters ***
                              estimate            s.e.      t(687)
Constant                         2.702           0.929        2.91
YEAR 92                         -1.565           0.293       -5.34
YEAR 93                         -1.099           0.276       -3.99
YEAR 94                         -2.820           0.359       -7.85
YEAR 95                         -3.215           0.346       -9.30
YEAR 96                         -2.342           0.391       -5.99
SEX M                            0.476           0.170        2.81
AGEWEAN                        0.07026         0.00886        7.93
DL                               2.725           0.315        8.65
DQ                              -0.269           0.034       -7.90
GENOTYPE RD                     -0.389           0.220       -1.77
GENOTYPE RR                     -1.040           0.271       -3.84
GENOTYPE DR                     -0.511           0.304       -1.68
*** Accumulated analysis of variance ***
Change                      d.f.          s.s.         m.s.          v.r.
+YEAR                          5      1208.149      241.630         48.92
+SEX                           1        55.983       55.983         11.34
+AGEWEAN                       1       344.206      344.206         69.69
+DL                            1       151.513      151.513         30.68
+DQ                            1       275.795      275.795         55.84
+GENOTYPE                      3        75.857       25.286          5.12
Residual                     687      3392.947        4.939
Total                        699      5504.450        7.875
GENOTYPE          Prediction        s.e.
DD                11.552           0.159
DR                11.041           0.240
RD                11.163           0.176
RR                10.512           0.193
```

Findings, implications and lessons learned

1. This case study has shown methods for evaluating the contributions of different explanatory variables to a statistical model. Different representations of one of the explanatory variables, namely the age of the dam, are investigated to determine the most suitable way to express the relationship. The appropriate formulations of the terms for inclusion in the model are determined by first exploring the patterns of the associations of weaning weight with these factors and covariates.
2. There were major variations in mean weaning weights among years. Because of the imbalance across years in the distribution of lambs belonging to the different genotypes, it would have been clearly wrong to ignore year of birth when making comparisons across genotypes. It is thus important to make sure that all potentially important factors and covariates are accommodated in the model.

3. The example also shows how to calculate the sums of squares remaining when fewer degrees of freedom are used to represent an alternative parameterisation for a variable in the model. It was shown, for example, that age of dam was best fitted using a quadratic relationship term and that this accounted for most of the variation among the individual age categories.
4. Sometimes reparameterisation results in the remainder mean square falling to a value below that of the residual mean square. Had it happened here (it did not) then it is possible that the curve might have been over fitting the data and that the quadratic term was probably not necessary. To find out whether this might have been the case the DL term could have been tried on its own.
5. A common mistake (when individual values are known, as here) is to fit a regression model to mean values and then to calculate standard errors and draw conclusions based on the residual variation among the means alone. By doing so the precision with which the mean values have themselves been calculated is ignored. The correct approach is the one described here.

Reporting

Here we illustrate how the results of this statistical analysis can be presented in a table.

Tables demonstrating parameter estimates and their standard errors can be presented, either as parameter estimates with one parameter level for each parameter set to zero and often referred to as the 'reference' or 'baseline' level, or as least squares mean estimates.

Consider first a suitable table based on the parameter estimates provided in the GenStat output. Note that a suitable number of decimal places has been used that allows a reasonable number of significant figures to be presented in a way that enhances readability but at the same time allows estimates to be

Parameter Genotype	Estimate	s.e
Dorper (D)	reference	-
Red Maasai (R)	-1.04	0.27
R x D	-0.39	0.22
D x R	-0.51	0.30

compared. Thus, the numbers of decimal places printed in the output have been reduced in most cases to two. An extra decimal place has been included for the regression coefficients for weaning age and the quadratic component for dam age to allow two significant figures to be presented.

Year		
1991	reference	-
1992	-1.56	0.29
1993	-1.10	0.28
1994	-2.82	0.36
1995	-3.22	0.35
1996	-2.34	0.39
Sex		
Female	Reference	-
Male	0.48	0.17
Age at weaning (day)	0.070	0.009
Age of dam (year)		
Linear	2.19	0.25
Quadratic	-0.269	0.034

Note that neither t-values nor probability levels are given. The approximate statistical significance of a parameter estimate can be easily derived by eye by dividing the estimate by its standard error and comparing the result with t-values of 2, 2.6 or 3.3. These are the approximate t-values for P = 0.05, P = 0.01 and P = 0.001 levels of significance, respectively. Divide each parameter estimate in this table by its standard error and see what you conclude.

Parameter Genotype	Estimate	s.e.
Dorper (D)	reference	-
Red Maasai (R)	-1.04	0.27
R x D	-0.39	0.22
D x R	-0.51	0.30
Year		
1991	reference	-
1992	-1.56	0.29
1993	-1.10	0.28
1994	-2.82	0.36
1995	-3.22	0.35
1996	-2.34	0.39
Sex		
Female	Reference	-
Male	0.48	0.17
Age at weaning (day)	0.070	0.009
Age of dam (year)		
Linear	2.19	0.25
Quadratic	-0.269	0.034

Another way to present the results is to use the GenStat **Predict** command to produce least squares estimates of means for each factor. This command becomes available within the Genstat dialog box once the model has been fitted. Note that this needs to be done for each factor one at a time; when all three factors are predicted simultaneously a 3-way table of least squares means is produced.

It is helpful to include the numbers of observations for each level of each parameter, both in this table and sometimes in the previous one too. These can be produced using the **Stats → Summary Statistics → Frequency Tables...** command. It can be seen from the table that the number of observations influences the magnitudes of the standard error, the larger the number of observations generally the smaller the standard errors. Regression coefficients have also been included at the foot of the table. This is optional.

Parameter Genotype	No.	Mean	S.E.
Dorper (D)	220	11.56	0.16
Red Maasai (R)	181	10.51	0.19
R x D	198	11.16	0.18
D x R	101	11.04	0.24
Year			
1991	144	12.73	0.22
1992	109	11.16	0.22
1993	168	11.63	0.18
1994	79	9.91	0.26
1995	107	9.53	0.24
1996	93	10.39	0.28
Sex			
Female	323	10.84	0.12
Male	377	11.32	0.12
		Regression coefficient	S.E.
Age at weaning (day)		0.070	0.010
Age of dam (year)			
Linear		2.19	0.25
Quadratic		-0.269	0.034

It is generally more helpful to replace the individual standard errors for each factor level by a single value that represents the approximate standard error of the differences between pairs of means. The validity of doing so depends on the distribution of numbers of observations across the different levels of a parameter. There is some disparity in numbers of lambs across the four levels for genotype and the six levels for year in the previous table. Even with the variation in frequencies here, however, it is reasonable to calculate average standard errors. It is best to calculate the average standard error as the square root of the average of the sum of the squares of the individual standard errors. The average S.E.D. is then this figure multiplied by the square root of 2.

Thus, for genotype (see last GenStat output)

average standard error of difference (S.E.D) = square root $[2(0.16^2 + 0.18^2 + 0.19^2 + 0.24^2)/4] = 0.28$

In general this can be considered to be a fairly satisfactory estimate provided there is no covariance between the different levels of the parameter. To get the exact average value one would, for genotype for instance, need to rerun the model three times, changing the reference level each time. One gets a series of standard errors of differences from the reference levels. By picking out the six values that correspond to the standard errors of the differences between the six different pairs of means, and calculating the average, one gets the correct average. In practice this is usually unnecessary.

By applying this formula to each factor the table can be written as shown. Here it is particularly important to give the numbers of lambs for each factor level so as to be able to gauge how approximate the S.E.D. is when applied to each level of a factor.

Same table of means but with standard errors of differences between means

Parameter	No.	Mean
Genotype		
Dorper (D)	220	11.56
Red Maasai (R)	181	10.51
RXD	198	11.16
DXR	101	11.04
Average S.E.D		0.28
Year		
1991	144	12.73
1992	109	11.16
1993	168	11.63
1994	79	9.91
1995	107	9.53
1996	93	10.39
Average S.E.D.		0.33
Sex		
Female	323	10.84
Male	377	11.32
Average S.E.D.		0.17

Study questions

1. Write a summary of the results given in this case study in about 100 -150 words describing the difference in weaning weight among genotypes and the effects of the various covariates on weaning weight.
2. Fit the final statistical model again and save the residuals. Produce the boxplots by genotype and age of dam, as done earlier in this case study. How many outliers are now shown? What will you do about them?
3. Fit an additional cubic term (DC) for DAMAGE7 to the model and examine its statistical significance. Based on what you find would you include a cubic term or not?
4. We have observed the partial confounding between genotype and year. Run the model again without year and comment on the parameter estimates and their standard errors that you now find for genotype. Calculate the new value for the average standard error of difference between pairs of genotype means and compare with the earlier value.
5. Include BIRTHWT as a covariate in the model. What effect does this have on the analysis? Why? Using Genstat to calculate an analysis of variance to compare birth weights among genotypes may help to explain what has happened. Based on this do you

think it appropriate to include BIRTHWT in the model and, if so, what interpretation do you put on the results?

6. We have assumed that regression lines between weaning weight and age at weaning are parallel for each year. Fit a statistical model to include an overall regression line and an interaction with year and comment on your findings.
7. 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. On the basis of an average of 100 Dorper and 100 Red Maasai lambs that might be born in a flock, and assuming the growth rates and survival rates given in this study, calculate the total weight of the lambs that reach weaning for each breed. Compare the results with those given for weaning weight in this case study. Comment.
8. A few weights for lambs that were still alive at weaning were missing in the data sheet at weaning. Is there any way that these weights might be estimated from other weights recorded for the lambs? If so, explain how you might do so. Comment on whether this would be necessary for this data set.
9. Use the GenStat Spread → Sheet → Title… facility as an alternative method of documenting the data. Comment on the advantages/disadvantages of this method compared with that used in the case study.

Related reading

Baker, R.L., Mwamachi, D.M., Audho, J.O., Aduda, E.O. and Thorpe W. 1999. Genetic resistance to gastrointestinal nematode parasites in Red Maasai , Dorper and Red Maasai x Dorper ewes in the sub-humid-tropics. *Animal Science* **69**:335-344. Abstract

Baker, R.L., Nagda, S., Rodriguez-Zas, S.L., Southey, B.R., Audho, J.O., Aduda, E.O. and Thorpe, W. 2003. Resistance and resilience to gastro-intestinal nematode parasites and relationships with productivity of Red Maasai, Dorper and Red Maasai x Dorper crossbred lambs in the sub-humid tropics. *Animal Science* **76**: 119-136. Abstract

Rege, J.E.O., Tembely, S., Mukasa-Mugerwa, E., Sovani, S., Anindo, D., Lahlou-Kassi, A. Nagda, S. and Baker, R.L. 2002. Effect of breed and season on production and response to infections with gastro-intestinal nematode parasites in sheep in the highlands of Ethiopia. *Livestock Production Science* **78**: 159-174. Abstract

Acknowledgements

We are grateful to Dr. Leyden Baker, International Livestock Research Institute, who directed the research, for permission to use data collected during the experiment for this case study. Teachers and other users of this case study are welcome to use the data for further analysis but not to publish any results without permission from the International Livestock Research Institute. We also thank Dr. John Rowlands for guidance in developing this case study.

Unless otherwise indicated, photographs were taken by Dave Elseworth.