## Case study 1: A survey of cattle keeping practices of the Orma tribe in eastern Kenya and levels of milk production

Thomas Achia [a], Damaris Yobera [b], Patrick Irungu [c]
[a] Department of Mathematics, University of Nairobi, P. O. Box 30197, Nairobi, Kenya
[b] Crop Science Department, University of Nairobi, P.O. Box 29053, Nairobi, Kenya
[c] Kenya Trypanosomiasis Research Institute, P.O. Box 362, Kikuyu, Kenya

### Contents

### Summary

This case study is taken from a pilot household survey investigating the cattle breeding practices of the Orma people in eastern Kenya. The case study combines questionnaire and quantitative measurement approaches to assess levels of milk production by village Orma Boran cattle and the extent to which these are surplus to the consumption needs of the household. It demonstrates the application of methods of regression analysis to compare associations in milk offtake with age of calf of cows kept in two separate village locations.

Having explored graphically the nature of the relationships the case study shows how to fit separate regression lines for each location, firstly in parallel and secondly with different slopes. Reporting methods for the presentation of results of regression analysis are also illustrated. The results are then used together with data collected on average daily household consumption to assess the levels of surplus milk offtake available for sale by a household.

The case study then goes on to consider the suitability of the survey approach in answering the different aims of the project. One aim was to determine avenues for further research. Important information was gained about the extent of trypanosomosis in the cattle and the methods that farmers were using to control the disease. This led to further investigations using participatory methods to gain more information from farmers on their knowledge of disease and to seek recommendations from them on the best options for future disease control.

Source: Bernard Sacher

Data analysis is illustrated by both R and GenStat software.

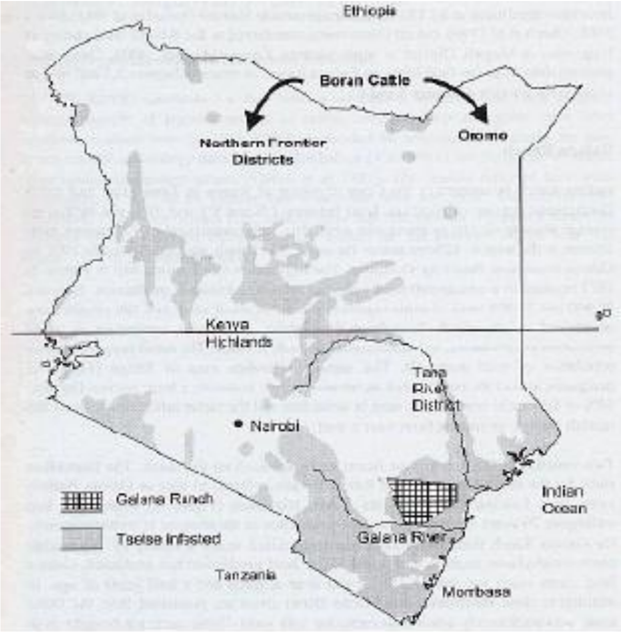
## Background

Trypanosomosis, a parasitic disease transmitted by the tsetse fly, is widespread throughout the humid and sub-humid areas of sub-Saharan Africa. Certain indigenous cattle breeds have evolved in these areas and have developed varying degrees of tolerance to the disease. One of these breeds is the Orma Boran cattle that belong to the Orma people.

Source: Rob Hutchinson

Source: John Rowlands

The Orma people are descendants of the Oromo, who originated in Borana Province in Ethiopia and brought their Boran cattle south. These nomadic pastoralists finally settled in the tsetse infested lands of the Tana River district in Kenya.



Source: Rosemary Dolan Source: John Rowlands

Studies at Galana Ranch, situated nearby, have shown that Orma Boran cattle do better when exposed to high tsetse challenge than improved Kenya Boran cattle that have been bred in areas of the Kenya highlands where trypanosomosis is absent. Infection and mortality rates due to trypanosomosis in the Orma Boran were approximately half of those observed in their counterparts (Dolan, 1998). However, these results were obtained under ranch conditions and little was known on how Orma Boran cattle fared under village management, especially in terms of milk production. A pilot household survey was therefore carried out among the Orma people to obtain information on their cattle keeping practices, to estimate levels of daily milk production and to obtain indicators of what further research was needed to combat the effects of trypanososmosis.

Source: Bernard Sacher

## Research strategy

The area of the Tana Delta where the Orma people live is in a fairly remote area of Kenya. There is little in the way of documentation of the villages or of the households contained within them, and little was known of the sizes of either the human or the cattle population. Access is also difficult. Thus, it was impossible to develop a sampling frame from which villages and households within villages could be randomly selected.

It was concluded therefore that a pilot survey was needed in order to explore the distribution and accessibility of villages in the region, make contact with the community and collect some basic data on cattle keeping practices, disease prevalence and milk production.

Source: John Rowlands

Source: John Rowlands

As well as using a questionnaire approach opportunities would be taken to make quantitative measurements of milk offtakes in a sample of cows.

One of the objectives of the survey was to help to determine the types of research studies that might be possible in the future. The results from this initial survey indicated that the Orma people had good knowledge of the different diseases that affected their cattle and the treatment of them. The research strategy thus developed into a more thorough investigation which used participatory methods to determine the views of the Orma people on the importance of trypanosomosis and its control. This latter study is summarised briefly at the end of this case study.

### Study design

A survey of households from various villages was planned to provide information about the Orma pastoralists and their cattle management practices. The study also aimed to gather data on levels of milk production of their cows. Twelve villages from two locations were visited during the course of the study, 48 households from different manyattas (villages) were selected and the household heads interviewed. At the same time milk offtakes were measured in selected cows in lactation at the household.

The households were identified using purposive sampling

g with the help of the chief of the village. On the day of interview, milk offtake (i.e. the amount of milk collected for human consumption apart from that consumed by the calf) for both morning and evening milking was determined using a calibrated plastic measuring jar to an accuracy of 50 ml. Milk offtakes from a total of 164 cows were collected and the respective ages of calves recorded, as reported by the owner.

## Study design

The survey was conducted in the Tana Delta area itself; referred to as Bilisa, and in Assa location, a more arid region to the west, in order to investigate differences in management practices between the two regions and their impact on milk offtake.

A series of questions were asked of the household head (Irungu, 2000); those related to family size CS1Quest1 and milk sales and milk consumption in the households CS1Quest2 are included here. As will be seen during the course of this case study the form that the questioning took proved not to be ideal. For a preliminary study such as this the range of questions was too wide. This meant that the length of the interview was a little long with the result that, as will be seen later, information was incomplete.

## Objectives

The objectives of the study were to undertake a preliminary household survey among Orma pastoralists in the Tana River district in south-eastern Kenya in order to collect data on:

- General demographic data on the Orma people
- Their cattle keeping practices
- The important disease affecting their cattle
- Average levels of milk production (milk offtake) in their cows
- Average amounts of consumption of milk in the home and of sales of milk

An additional important objective in carrying out this preliminary study was to use results from the survey to determine future avenues for investigation and research.

## Questions to be addressed

The specific questions to be addressed herein relate to milk production and are:

- What is the general level of milk offtake in the Bilisa and Assa locations and do mean milk offtakes differ between them?
- How does milk offtake vary with age of calf (in other words with month of lactation) and do the patterns differ between Bilisa and Assa locations?
- To what extent does milk offtake meet the needs of the household both in terms of

human consumption and marketing?

We shall also evaluate the suitability of the study design in terms of both the questionnaire that was used and the quantitative milk recording component.

GenStat is used within the case study. The same analysis using R is demonstrated by Nagda (2009).

- In response to these questions we shall first summarise milk offtake by location and use graphical methods to explore the overall variation in milk offtake and to find out how average milk offtake differs between the two locations.
- By treating milk offtake as the response variable and the age of calf (synonymous with stage of lactation) as the explanatory variable, we shall use regression analysis to investigate the nature of the relationship between the two variables and the extent to which it differed between the two locations. We shall achieve this by including an additional parameter to describe location. We shall firstly use analysis of variance to fit parallel regression lines for the locations, and then fit two regression lines with separate slopes.
- These results will then be used together with questionnaire information on average milk consumption by members of the household to estimate surplus levels of milk production available for marketing.

## Source material

The data sets used for this case study are in CS1Data1 and CS1Data2. Files CS1Doc1 and CS1Doc2 describe the variables contained in the two data files, respectively.

The former file contains recordings of daily milk offtakes measured in 164 cows during the course of the survey.

The latter contains details of information provided on family size, daily milk consumption in the household, milk given to friends and milk sold. Parts of the original questionnaire that provided the details contained in this file have been put together and stored in CS1Quest1 and CS1Quest2.

Source: Bernard Sacher

## Data management

The data file CS1Data1 produced from the original source data contains data on recorded milk offtake for each cow and the reported age of her calf. These offtakes were measured in cows residing in a number of households (sometimes more than one cow per household) in various villages in Bilisa and Assa locations, respectively. These locations have been coded as 1 and 2, respectively.

TOTALM, the sum of morning and afternoon recordings, is also included in the file. Although easy to calculate at the time of recording, and perhaps of interest to do so, only the individual morning and afternoon values need to be entered initially into the data file. Variables that can be calculated from other variables are best done by computer. This saves unnecessary work and reduces mistakes during data collection and entry.

The data file CS1Data2 contains a number of variables extracted and derived from the questionnaires CS1Quest1 and CS1Quest2. The data file is divided into four spreadsheets for the purpose of analysis. The second spread sheet contains the original data and an edited version (see later) is contained in the first spreadsheet. Separate work sheets are then prepared for Bilisa and Assa.

## Data management

If one compares closely the details of the questionnaire and the data extracted and stored in the data file one can see that a fair amount of work was needed to code the data in a form suitable for analysis. For example, milk consumption needed to be first transformed into litres before data entry.

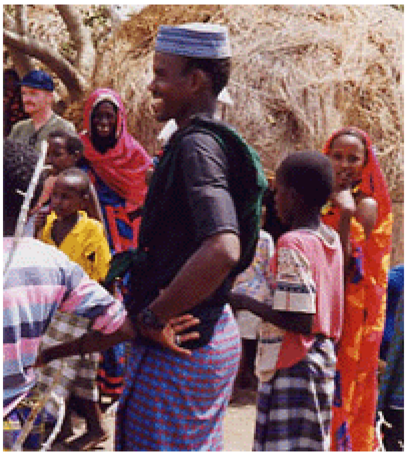
It was decided to ignore ages of members of the household, as this information proved difficult to collect, and to just enter the total amount consumed. Sour milk was not drunk and hence this column was ignored. The total number of children in the household was also calculated from the numbers given per wife.

Files CS1Doc1 and CS1Doc2 describe the variables contained in CS1Data1 and CS1Data2, respectively. Documentation is an important component of any investigation. Investigators are often dilatory in documenting their data but, if this is not done carefully, then this limits the possibilities for further exploitation of the data by another researcher at a later date. The documentation files provided here describe the data but have omitted descriptions of the sources of the data - when they were collected and for what purpose. Such information will need to be included when the data are archived at the completion of a project.

This pilot survey was planned to gather preliminary information over a number of households. Looking back it can now be seen that, whilst valuable information was obtained that enabled further research studies to be planned, attempts were made to collect too much detail. Indeed, it is now appreciated that it would have been better to have simplified the questionnaire and designed it in such a way that the data could have been extracted and entered directly from the questionnaire forms into the computer.

We hope that by including this questionnaire in the form that it was used will be instructive in alerting others to the types of problems that can be encountered. One of the questions at the end of this case study is to redesign the questionnaire in a form that will provide answers to the important questions and allow easy data entry without further manipulation.

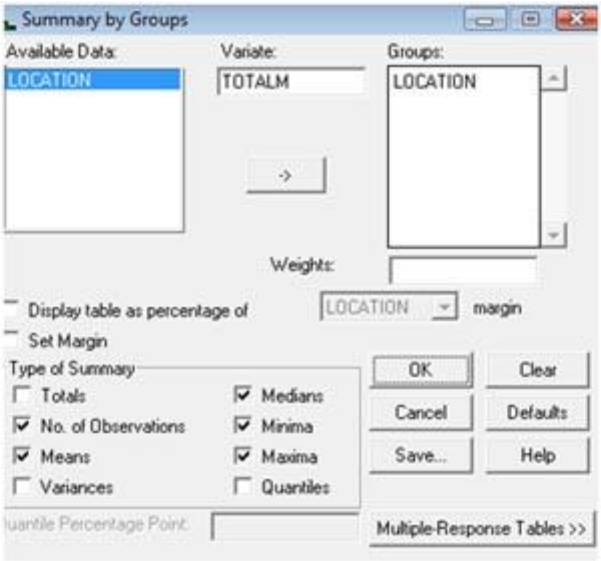
This case study provides a valuable lesson into the need for careful questionnaire design in relation to the way that information is collected and handled.



## Exploration & description

### Milk offtake

Descriptive and graphical methods provide a useful starting point for the analysis of the milk offtake data contained in CS1Data1. They help to reveal differences in the patterns of milk offtake between locations and the nature of associations between milk offtake and age of calf. Having converted LOCATION to a factor, general descriptive statistics can be produced by via **Stats → Summary Statistics → Summaries of Groups (Tabulation)**.



Summary by Groups

Available Data: LOCATION

Variate: TOTALM

Groups: LOCATION

Weights:

Display table as percentage of LOCATION margin

Set Margin

Type of Summary

□ Totals ☑ Medians

☑ No. of Observations ☑ Minima

☑ Means ☑ Maxima

□ Variances □ Quantiles

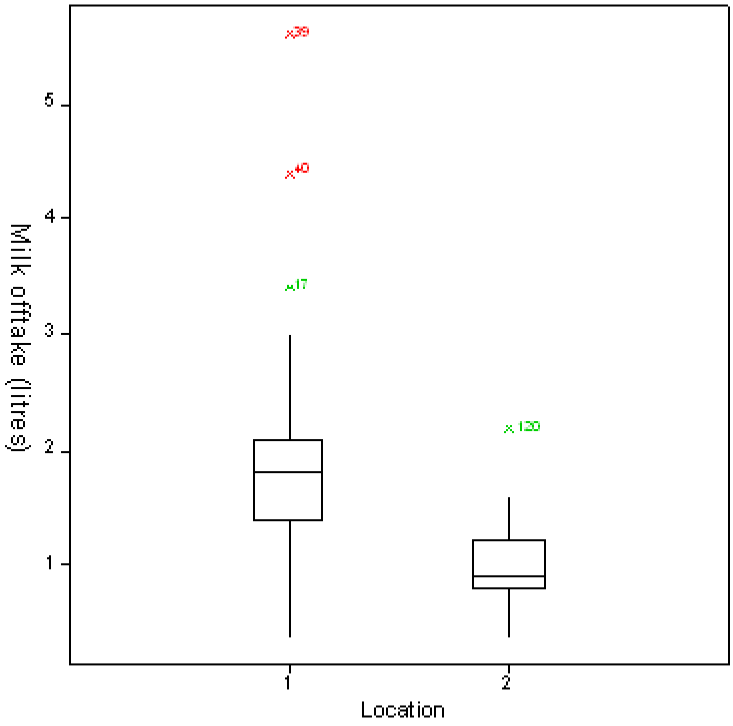
uantile Percentage Point:

The means and medians in both locations are comparatively close indicating generally symmetric distributions. The range in milk offtake in Bilisa (LOCATION 1), however, is 5.2 litres per day compared with 1.8 litres per day in Assa (LOCATION 2).

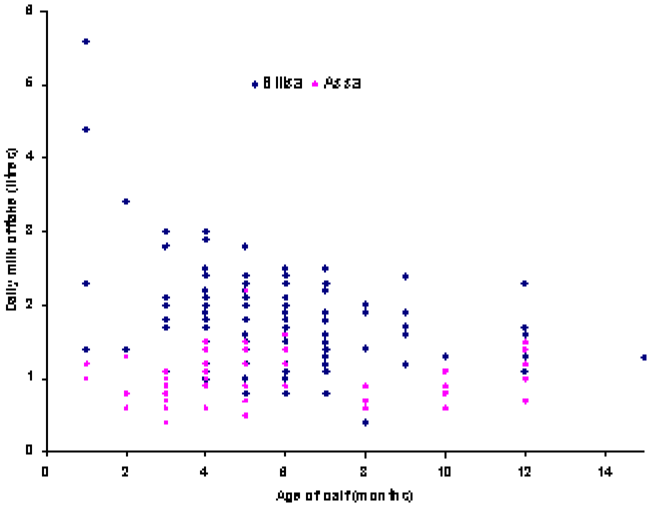
|  | Nobservd | Mean | Minimum | Maximum | Median |
|---|---|---|---|---|---|
| LOCATION |  |  |  |  |  |
| 1 | 111 | 1.843 | 0.4000 | 5.600 | 1.800 |
| 2 | 53 | 1.025 | 0.4000 | 2.200 | 0.900 |

## Milk offtake

A box plot (produced by **Graphics** ➔ **Boxplot** and using the **Tools editors** to revise the title and legends) illustrates further the differences in variation in milk offtakes in Bilisa and Assa and indicates three ´ outliers ´ for Bilisa and one for Assa.

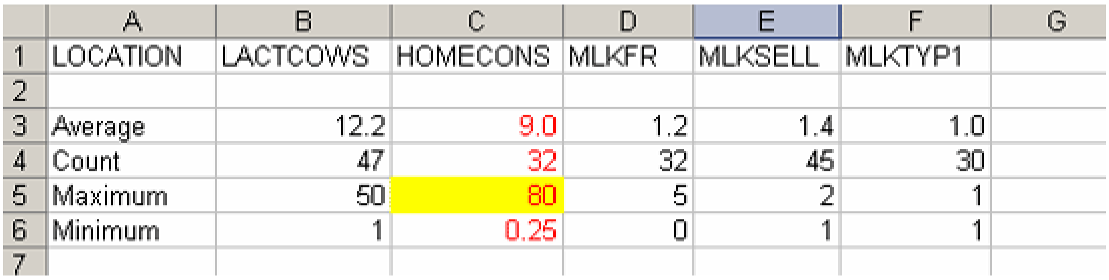


One can pick out the four outliers in the scatter plot (**Graphics** → **Create Graph...** → **2D Scatter Plot** together with use of the **Tools editors)**. Furthermore, the distributions of the other points in the body of the figure suggest different patterns between milk offtake and calf age in the two locations. These patterns support the use of a multiple regression analysis including a term to describe different intercepts on the y-axis for the two locations.



## Milk consumption & marketing

Summary statistics are included at the bottoms of the columns in CS1Data2. The range in home consumption values is from 0.25 to 80 litres with a mean of 9 litres (printed in red in the ´original´ worksheet and also summarised below). The value of 80 litres (shaded) for Household 36 is clearly wrong and has been changed to 8 litres in the´edited ´ sheet. This shows the importance of checking data before plunging into the statistical analysis. Examination of the data file shows many missing data items. This is going to make it difficult to derive some useful information.



|   | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| 1 | LOCATION | LACTCOWS | HOMECONS | MLKFR | MLKSELL | MLKTYP1 |  |
| 2 |  |  |  |  |  |  |  |
| 3 | Average | 12.2 | 9.0 | 1.2 | 1.4 | 1.0 |  |
| 4 | Count | 47 | 32 | 32 | 45 | 30 |  |
| 5 | Maximum | 50 | 80 | 5 | 2 | 1 |  |
| 6 | Minimum | 1 | 0.25 | 0 | 1 | 1 |  |
| 7 |  |  |  |  |  |  |  |

## Statistical modelling

### Milk offtake

The first step in the analysis of the milk offtake data is to try and fit a regression equation with a term to describe a common slope for the pattern and a term to allow separate intercepts on the y-axis for the two locations. We can do this by adding a parameter for LOCATION to that for AGEC. The model can be written algebraically in the form:

yi = a + bxi + Lj + ei

where yi and xi are the milk offtake and the age of calf, respectively, for cow i (i =1, ..., 164), where Lj (j=1, 2) is a parameter that describes the location where the cow resides with reference to a constant a, and where ei is the residual term.

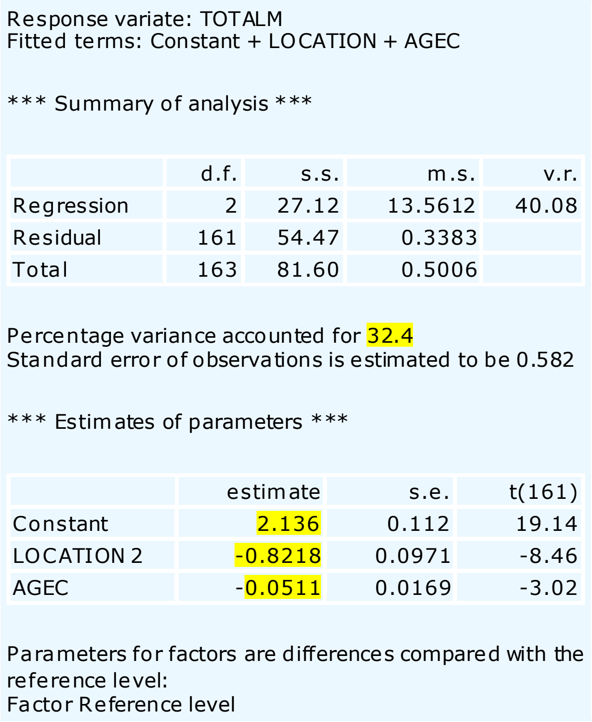
The location parameter Li signifies that the regression lines for the two locations cross the y-axis at a + L1 and a + L2, respectively.

- With the algebraic constraint L1 + L2 = 0, this reduces to a + L1 and a − L1, respectively.
- With the algebraic constraint L1 = 0, this reduces to a and a + L2, respectively.

GenStat uses the second constraint.

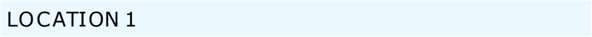
Using CS1Data1 the model to be fitted (LOCATION+AGEC) can be done using Stats → Regression Analysis → Generalized Linear Models.... . The output describes the analysis of variance and shows evidence of a highly statistically significant association accounting for 32.4% of the variation.

A table of parameter estimates follows, the constant giving the intercept on the y-axis for Bilisa (the baseline or reference location (code 1)), LOCATION 2 representing the average difference between Bilisa and Assa, and AGEC representing the regression coefficient or slope for the two parallel regression lines. The standard error (s.e.) gives the precision with which the



```
Response variate: TOTALM
Fitted terms: Constant + LOCATION + AGEC


*** Summary of analysis ***


                 d.f.      s.s.       m.s.     v.r.
Regression          2     27.12    13.5612    40.08
Residual          161     54.47     0.3383
Total             163     81.60     0.5006


Percentage variance accounted for 32.4
Standard error of observations is estimated to be 0.582


*** Estimates of parameters ***


                    estimate       s.e.     t(161)
Constant               2.136      0.112      19.14
LOCATION 2           -0.8218     0.0971      -8.46
AGEC                 -0.0511     0.0169      -3.02


Parameters for factors are differences compared with the
reference level:
Factor Reference level
```

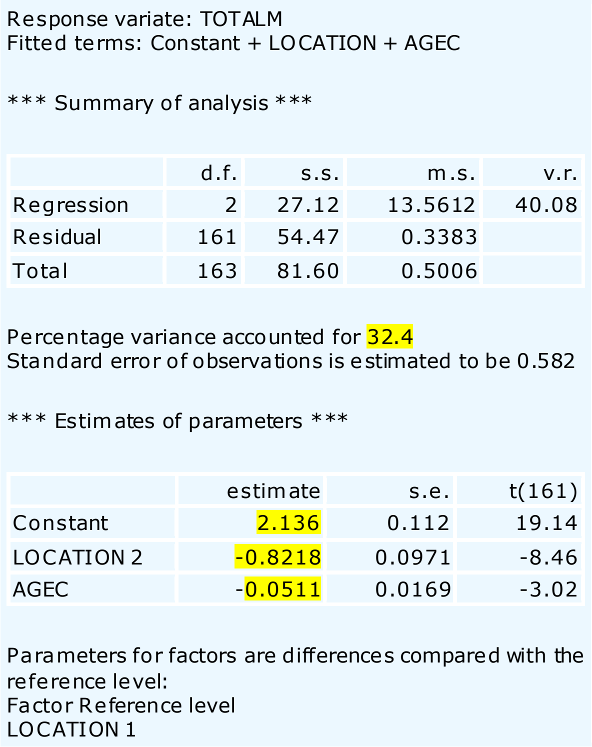
parameter estimates are determined.



```
LOCATION 1
```

Using CS1Data1 the model to be fitted (LOCATION+AGEC) can be done using Stats → Regression Analysis → Generalized Linear Models.... . The output describes the analysis of variance and shows evidence of a highly statistically significant association accounting for 32.4% of the variation.

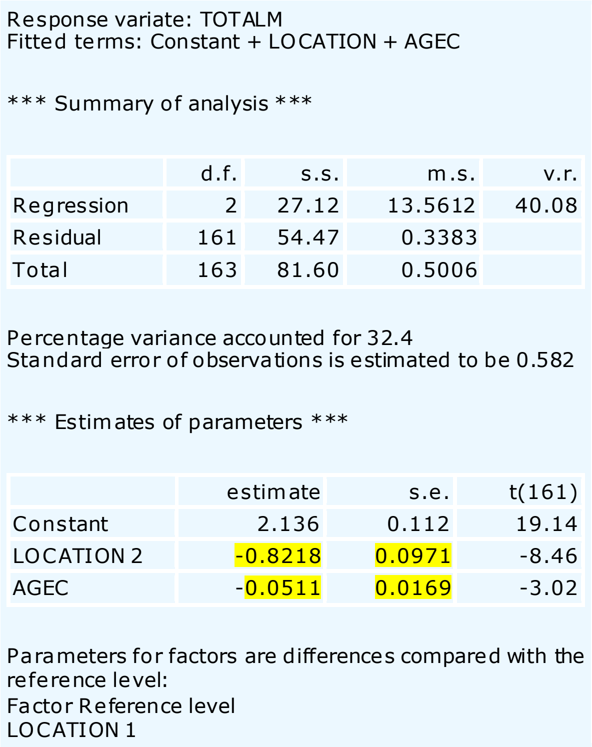
A table of parameter estimates follows, the constant giving the intercept on the y-axis for Bilisa (the baseline or reference location (code 1)), LOCATION 2 representing the average difference between Bilisa and Assa, and AGEC representing the regression coefficient or slope for the two parallel regression lines. The standard error (s.e.) gives the precision with which the parameter estimates are determined.



```
Response variate: TOTALM
Fitted terms: Constant + LOCATION + AGEC


*** Summary of analysis ***

                 d.f.      s.s.       m.s.      v.r.
Regression          2     27.12    13.5612     40.08
Residual          161     54.47     0.3383
Total             163     81.60     0.5006

Percentage variance accounted for 32.4
Standard error of observations is estimated to be 0.582


*** Estimates of parameters ***

                  estimate       s.e.    t(161)
Constant             2.136      0.112     19.14
LOCATION 2         -0.8218     0.0971     -8.46
AGEC               -0.0511     0.0169     -3.02

Parameters for factors are differences compared with the
reference level:
Factor Reference level
LOCATION 1
```

Multiplying the s.e. by 2 and adding to and subtracting, respectively, from the mean, gives an approximate estimate of the 95% confidence intervals (namely, -1.0160 to -0.6276 for LOCATION and -0.0849 to -0.0173 for AGEC) within which the true difference between locations and the true slope are expected to lie.

We can see that neither of the pairs of limits contain the value zero, confirming that the data can be represented by separate lines with a slope that is significantly different from zero.



```
Response variate: TOTALM
Fitted terms: Constant + LOCATION + AGEC


*** Summary of analysis ***

                 d.f.      s.s.       m.s.      v.r.
Regression          2     27.12    13.5612     40.08
Residual          161     54.47     0.3383
Total             163     81.60     0.5006

Percentage variance accounted for 32.4
Standard error of observations is estimated to be 0.582


*** Estimates of parameters ***

                  estimate       s.e.    t(161)
Constant             2.136      0.112     19.14
LOCATION 2         -0.8218     0.0971     -8.46
AGEC               -0.0511     0.0169     -3.02

Parameters for factors are differences compared with the
reference level:
Factor Reference level
LOCATION 1
```

From the parameter estimates the fitted equation can be written as:

TOTALMi = 2.136(±0.112) - 0.8218(±0.0971)L2 - 0.0511(±0.0169)AGECi

where the coefficient for L2 refers to the difference in intercepts for Assa from Bilisa. It is customary to include standard errors in parentheses. The results show that the intercept is 0.8218 litres lower for Assa than for Bilisa.

Thus, separate regression lines for the two locations can be written:

For Bilisa: yi = 2.136 (±0.112) - 0.0511(±0.0169) xi
For Assa: yi = 1.314 (±0.125) - 0.0511(±0.0169) xi

The constant term for Assa is determined by subtracting the value 0.8218 from the constant, namely 2.136 - 0.8218 = 1.314. However, the standard error for the constant term for Assa is not so easily obtained. One way is to run GenStat again but first to reorder the LOCATION code so that level 2 is recognised as the first level, i.e. making Assa the reference location.

Changing the LOCATION code can be done using the factor reordering facility (**Spread → Factor → Reorder levels...**). The factor levels themselves do not change in the spread sheet but they are considered to occur in a different order.

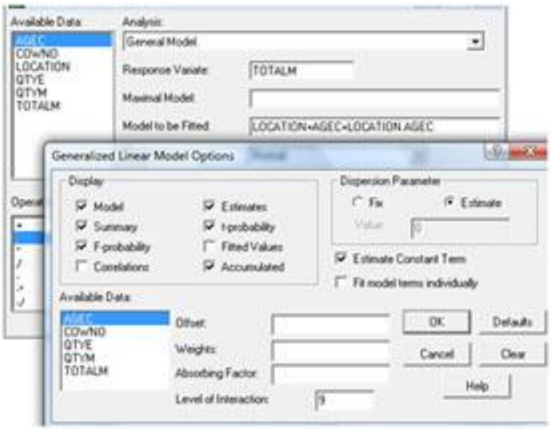
The parameter value for LOCATION 1 for Bilisa is now shown (the same value as before but with the sign reversed). The constant term value of 1.314 now refers to Assa and its standard error can be seen to be the value 0.125 included in the equation on the previous page.

```
***** Regression Analysis *****


Response variate: TOTALM
Fitted terms: Constant + LOCATION + AGEC


*** Summary of analysis ***


                 d.f.      s.s.       m.s.      v.r.
Regression          2     27.12    13.5612     40.08
Residual          161     54.47     0.3383
Total             163     81.60     0.5006


*** Estimates of parameters ***


                   estimate       s.e.     t(161)
Constant              1.314      0.125      10.52
LOCATION 1           0.8218     0.0971       8.46
AGEC                -0.0511     0.0169      -3.02


Parameters for factors are differences compared with the
reference level:
Factor Reference level
LOCATION 2
```

The next step is to investigate whether non-parallel lines better represent the data. This is achieved by fitting an interaction term in the model.Here we use the **Options** button to allow **Accumulated** to be ticked (see below). This allows an accumulated analysis of variance to be included in the output which shows the sums of squares accounted for by each term as it is added to the model.

```
Response variate: TOTALM
Fitted terms:
Constant+LOCATION+AGEC+AGEC.LOCATION
*** Summary of analysis ***
                 d.f.      s.s.       m.s.      v.r.
Regression          3     30.23    10.0766     31.39
Residual          160     51.37     0.3210
Total             163     81.60     0.5006


Percentage variance accounted for 35.9


*** Estimates of parameters ***


                        estimate       s.e.     t(160)
Constant                   0.983      0.161       6.09
```

The output shows that the interaction term is significant (P<0.01). The percentage variance accounted for increases from 32.4% in the previous analysis to 35.9% here.



Available Data: AGEC, COWNO, LOCATION, QTYE, QTYM, TOTALM

Analysis: General Model

Response Variate: TOTALM

Maximal Model:

Model to be Fitted: LOCATION+AGEC+LOCATION.AGEC

Generalized Linear Model Options

Display:
☑ Model ☑ Estimates
☑ Summary ☑ t-probability
☑ F-probability □ Fitted Values
□ Correlations ☑ Accumulated

Dispersion Parameter: ○ Fix ● Estimate

Value: 1

☑ Estimate Constant Term

□ Fit model terms individually

Available Data: AGEC, COWNO, QTYE, QTYM, TOTALM

Offset:

Weights:

Absorbing Factor:

Level of Interaction: 3

```
LOCATION 1              1.411      0.212      6.67
AGEC                   0.0073     0.0250      0.29
AGEC.LOCATION 1       -0.1036     0.0333     -3.11

Parameters for factors are differences compared with the
reference level:
Factor Reference level: LOCATION 2

*** Accumulated analysis of variance ***

Change               d.f.      s.s.      m.s.    v.r.
+ LOCATION              1   24.0448   24.0448   74.90
+ AGEC                  1    3.0777    3.0777    9.59
+ AGEC.LOCATION         1    3.1073    3.1073    9.68
Residual              160   51.3656    0.3210
Total                 163   81.5953    0.5006
```

The output also shows how care must be taken in interpreting the parameter estimates. Each is corrected for the others in the model with the t-value measuring the significance of the parameter when included in addition to all other parameters in the model.

The accumulated analysis of variance, on the other hand, shows the additional sum of squares accounted for as each variable is added in turn. The order in which the terms are included to the model is important. Each sum of squares is corrected for variables already included in the model but not for those to be added later. Therefore the F-value has a different interpretation from the t-value.

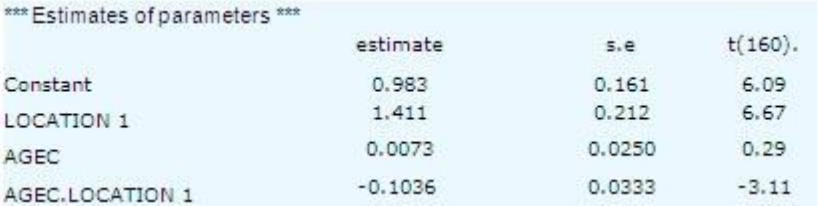
```
Response variate: TOTALM
Fitted terms:
Constant+LOCATION+AGEC+AGEC.LOCATION
*** Summary of analysis ***
                   d.f.      s.s.       m.s.     v.r.
Regression            3     30.23    10.0766    31.39
Residual            160     51.37     0.3210
Total               163     81.60     0.5006

Percentage variance accounted for 35.9

*** Estimates of parameters ***

                     estimate       s.e.    t(160)
Constant                0.983      0.161      6.09
LOCATION 1              1.411      0.212      6.67
AGEC                   0.0073     0.0250      0.29
AGEC.LOCATION 1       -0.1036     0.0333     -3.11

Parameters for factors are differences compared with the
reference level:
Factor Reference level: LOCATION 2

*** Accumulated analysis of variance ***

Change               d.f.      s.s.      m.s.    v.r.
+ LOCATION              1   24.0448   24.0448   74.90
+ AGEC                  1    3.0777    3.0777    9.59
+ AGEC.LOCATION         1    3.1073    3.1073    9.68
Residual              160   51.3656    0.3210
Total                 163   81.5953    0.5006
```

The fitted regression lines for the two locations can now be calculated as:

for Assa: $y_i = 0.983\ (\pm 0.161) - 0.0073(\pm 0.025)\ x_i$
for Bilisa: $y_i = 2.394\ (\pm 0.137) - 0.0963(\pm 0.022)\ x_i$

with parameter estimates for Bilisa calculated as 2.394 = 0.983 + 1.411
and - 0.0963 = 0.0073 - 0.1036 (see values in table below)

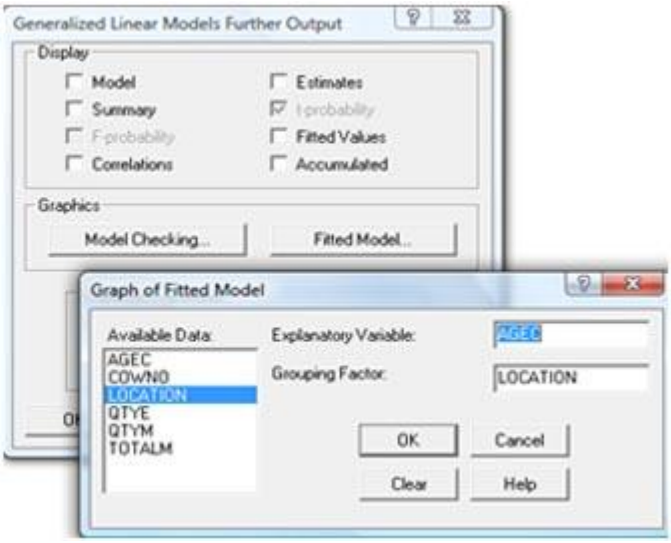
and corresponding standard errors calculated by rerunning GenStat with LOCATION codes changed back to their original order.



```
*** Estimates of parameters ***
                      estimate        s.e     t(160).
Constant                 0.983      0.161       6.09
LOCATION 1               1.411      0.212       6.67
AGEC                    0.0073     0.0250       0.29
AGEC.LOCATION 1        -0.1036     0.0333      -3.11
```

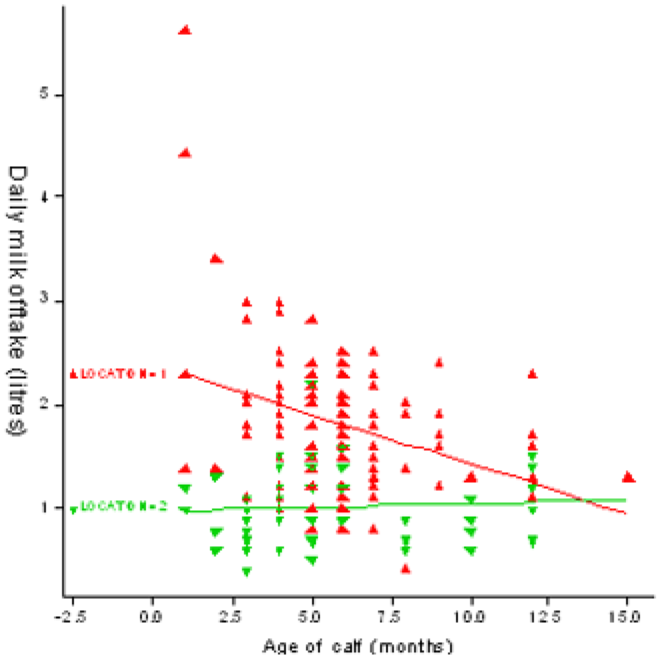
Note that parameter estimate given for the level of a factor is the difference in value from that of the reference level. Here LOCATION 2 (namely Assa) is being used as the reference level for the LOCATION factor.

The scatter plot with the two fitted regression lines is produced by clicking the **Further output …** button in the dialog box shown earlier, then the Graphics **Fitted Model …** button and finally completing the dialog box below. The graph shows how milk offtake decreases with stage of lactation of cows residing in Bilisa but not at Assa where milk offtakes remain low throughout.

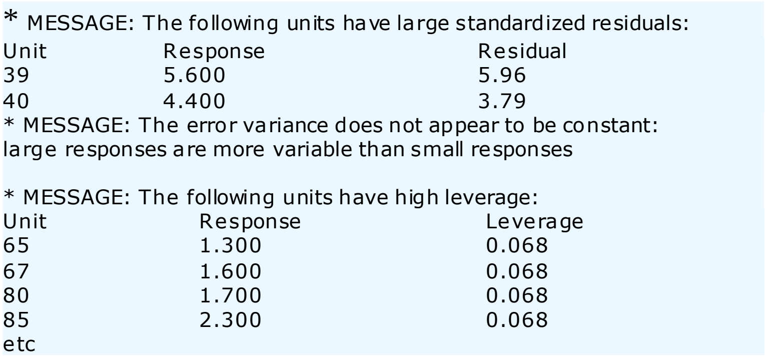


Explanatory Variable: AGEC

Grouping Factor: LOCATION



GenStat also produces warning messages as appropriate during the analysis. The messages shown here, which should be read in conjunction with the scatter diagram on the previous page, were produced when non-parallel lines were fitted.



```
* MESSAGE: The following units have large standardized residuals:
Unit            Response                        Residual
39              5.600                           5.96
40              4.400                           3.79
* MESSAGE: The error variance does not appear to be constant:
large responses are more variable than small responses

* MESSAGE: The following units have high leverage:
Unit                Response                     Leverage
65                  1.300                        0.068
67                  1.600                        0.068
80                  1.700                        0.068
85                  2.300                        0.068
etc
```

Message 1. The two units have standardised residuals (calculated as the deviation of an observation from its fitted value divided by the overall residual standard deviation) meaning that they fall some distance away from the fitted line and are ´outliers´.

Message 2. This suggests that the assumption the *y*-variable has a constant variance may not be tenable.

Message 3. Units with high leverage are those points that have a strong influence on the direction of the regression line. These points in this example are those that lie to the extreme right for calves aged 12 months and beyond.

## Milk consumption & marketing

A summary of the results of the questionnaire survey on how milk offtakes were utilised is given in the table. The table collects together results of calculations done within the Excel file itself. These can be seen by opening the Bilisa and Assa spreadsheets in CS1Data2 and comparing the coloured sections.

| | | | Milk offtake (litres per day) | | | | |
|---|---|---|---|---|---|---|---|
| Location | Family size | No. of lactating cows | Total milk offtake | Consumed at home | Given to friend | Sold | Unaccounted fo |
| Bilisa | 16.2 ± 1.0 | 11.5 ± 1.8 | 21.2 ± 0.6 | 4.8 ± 0.9 | 1.2 ± 0.2 | 2.4 ± 0.3 | 12.8 |
| | (38) | (38) | (111) | (24) | (27) | (36) | |
| | | | | | | | & |
| Assa | 15.4 ± 1.9 | 15.2 ± 2.5 | 15.6 ± 1.2 | 12.8 ± 2.2 | 1.1 ± 0.3 | 0.2 ± 0.2 | 1.5 |
| | (9) | (9) | (53) | (8) | (5) | (9) | |

The total milk offtake values per household were calculated as:

Average no. of lactating cows (see above) x (offtake mean per cow calculated earlier from the milk production survey – see mean values below)
= 11.5 x 1.843 = 21.2 litres for Billisa; = 15.4 x 1.025 = 15.6 litres for Assa.

```
               Nobservd     Mean     Minimum       Maximum     Median
LOCATION
1                   111    1.843      0.4000         5.600      1.800
2                    53    1.025      0.4000         2.200      0.900
```

A summary of the results of the questionnaire survey on how milk offtakes were utilised is given in the table. The table collects together results of calculations done within the Excel file itself. These can be seen by opening the Bilisa and Assa spreadsheets in CS1Data2 and comparing the coloured sections.

| | | | Milk offtake (litres per day) | | | | |
|---|---|---|---|---|---|---|---|
| Location | Family size | No. of lactating cows | Total milk offtake | Consumed at home | Given to friend | Sold | Unaccounted fo |
| Bilisa | 16.2 ± 1.0 | 11.5 ± 1.8 | 21.2 ± 0.6 | 4.8 ± 0.9 | 1.2 ± 0.2 | 2.4 ± 0.3 | 12.8 |
| | (38) | (38) | (111) | (24) | (27) | (36) | |
| | | | | | | | & |
| Assa | 15.4 ± 1.9 | 15.2 ± 2.5 | 15.6 ± 1.2 | 12.8 ± 2.2 | 1.1 ± 0.3 | 0.2 ± 0.2 | 1.5 |
| | (9) | (9) | (53) | (8) | (5) | (9) | |

The total milk offtake values per household were calculated as:

Average no. of lactating cows (see above) x (offtake mean per cow calculated earlier from

the milk production survey – see mean values below)
= 11.5 x 1.843 = 21.2 litres for Billisa; = 15.4 x 1.025 = 15.6 litres for Assa.

```
                          Nobservd     Mean      Minimum        Maximum     Median
LOCATION
1                              111    1.843       0.4000          5.600      1.800
2                               53    1.025       0.4000          2.200      0.900
```

The standard errors for total milk offtake (0.6 and 1.2, respectively, in the table on the previous page) were estimated by using the residual mean square 0.3210 for milk offtake in the final analysis of variance table in the GenStat regression analysis output, dividing by the number of cows used in the analysis, and multiplying by the average number of lactating cows per household. Thus:

11.5 x (0.3210/111) = 0.6 for Bilisa
15.4 x (0.3210/53) = 1.2 for Assa.

```
*** Summary of analysis ***
                          d.f.        s.s.        m.s.
Regression                   3       30.23     10.0766
Residual                   160       51.37      0.3210
Total                      163       81.60      0.5006
```

Although family sizes were similar in Bilisa and Assa there was a large difference in recorded milk consumption in the home (an apparent 4.8 litres per day in Bilisa and 12.8 litres per day in Assa). Dividing the average quantity consumed by the average size of the family size gives about 0.3 litres per household member in Bilisa and about 0.8 litres in Assa. The difference is surprising.

But the quantities unaccounted for in the table are:

for Bilisa: 21.2 - 4.8 -1.2 - 2.2 = 12.8 litres
for Assa: 15.6 -12.8 -1.1 - 0.2 = 1.5 litres

Taking into account the sizes of the standard errors, and the different ways in which the data were collected, the results for Assa look reasonable, but those for Bilisa not. One thus needs to consider carefully the quality of data collected and possible sources of error.

The large proportion of households that did not provide information on quantities of milk consumed in the households in Bilisa places a question mark on the justification of using a mean from the remaining households. By dividing amounts of milk consumed per household by family size in the CS1Data2 edited spreadsheet it can be seen that there is a wide range in individual proportions and thus a poor correlation. One reason, both for this poor correlation and the lack of response in some cases, was that the head of the household often said that he could not give an answer because milking and milk management largely fell into the hands of his wives.

In view of the possible inaccuracies in the collection of these data it will be unwise to publish the results shown in the table for wider circulation.

The limited conclusions that one might draw are that households in Assa had very little milk to spare. For households in Bilisa, however, where average milk offtakes were higher, there are opportunities for marketing, but the precise amounts of milk available for this purpose are not clear.

Findings, implications and lessons learned

Regression analysis

- There is a difference in mean milk offtake between cows sampled in Bilisia and Assa.
- Milk offtake declined with age of calf (or stage of lactation) at Bilisa but not Assa.
- Residual patterns in the data suggest that certain observations may have been influential in determining the fitted patterns and that there were others that possibly did not belong to the overall pattern.
- The last point has implications on the suitability of the study design that will be investigated further in the study questions. It illustrates some of the difficulties in interpreting results from small studies.

We have also seen how by switching of the order in which the factor LOCATION is coded, in other words redefining the reference level, we can calculate standard errors for both levels.

## Findings, implications and lessons learned

### Survey on milk consumption & marketing

- In retrospect there were shortcomings in the design of the questionnaire survey in relation to milk consumption and marketing. Too many questions were asked. It is likely as a result that respondents became tired and may not have understood some of the questions. The interviews were conducted in the local tribal language and there may have been some translation difficulties.

- The structure of the questionnaire could have taken better account of how the data to be collected would be stored in a computer. This resulted in additional manipulation of the data in order to get them into a form in which they could be used.

- It is important, when planning any type of survey, to be clear of the objectives and focus the questions accordingly. Thus, a pilot survey should be as simple as possible with a few questions that can be easily answered and can give the broad picture.

### Preliminary or ´pilot´ studies

Another lesson to be taken from this case study is that the research process is often an iterative one built on a series of studies, one following another. Preliminary or ´pilot´ studies can often be undertaken to test an idea or investigate some fact before proceeding with the next step. In order to minimise costs the researcher will wish to use as few experimental or sampling units as possible. The danger is that if studies are too small the data will not render themselves suitable for statistical analysis and hence the results will be difficult to interpret.

The design of preliminary investigations is as important as the design of main studies and it is necessary to ensure that sample sizes can allow conclusions to be made that can justify decisions taken for the next phase of the research. The biometrician often finds himself/herself advising on the design of pilot studies and needs to ensure that he understands the goals that the researcher has in mind and how the results from a current study will lead to the next phase of the research. Sometimes it is possible to plan a study which in itself may be too small to merit analysis on its own but, if the results look promising, the study can be replicated and the two studies analysed and reported together.

## Reporting

Here we represent the basis of a brief report on the regression analysis of milk offtake that can be suitably conveyed to other researchers.

We first present a suitable presentation for a summary table (Table 1). Note that no levels of significance are quoted in the table (these are quoted in the text), that the heading for the table is self explanatory (i.e. a reader can understand the contents of the table without necessarily referring to the text), and that numbers are presented in the table with a precision that is both merited by the data and makes the table easily interpretable and readable.

The table illustrates both the use of standard deviations (to give a measure of the spread of the data) and standard errors (to give a measure of the precision with which the regression coefficients are estimated). Standard errors are more commonly reported than standard deviations, since it is usually the precision with which parameter estimates are determined that is of primary interest.

Table 1. Mean milk offtake and regression coefficients with age of calf in Orma Boran cows sampled in Bilisa and Assa locations in Tana River district, Kenya.

| Location | Number of cows | Mean ± s.d. (litres) | Regression coefficient ± s.e. (age in months) |
|---|---|---|---|
| Bilisa | 111 | 1.8 ± 0.06 | -0.096 ± 0.022 |
| Assa | 53 | 1.0 ± 0.05 | 0.007 ± 0.025 |

The statistical analysis can be summarised by a simple statement as follows. One does not necessarily need to describe each step in the process of finding the appropriate model.

## Statistical analysis

Milk offtake was analysed by analysis of variance fitting separate regression lines to the data from Bilisa and Assa.

The results of the statistical analysis can then be put simply as follows:

## Results

Average milk offtake was higher in cows sampled in villages in Bilisa than in villages in Assa, a drier, more arid region (Table 1), especially during early lactation (Fig. 1). For cows in Bilisa milk offtake declined linearly by an average of 0.096 litres per month with increasing age of the cow´s calf (P<0.001). However, no trend was evident for cows in Assa where milk offtake remained low at about 1 litre per day throughout lactation. , ,

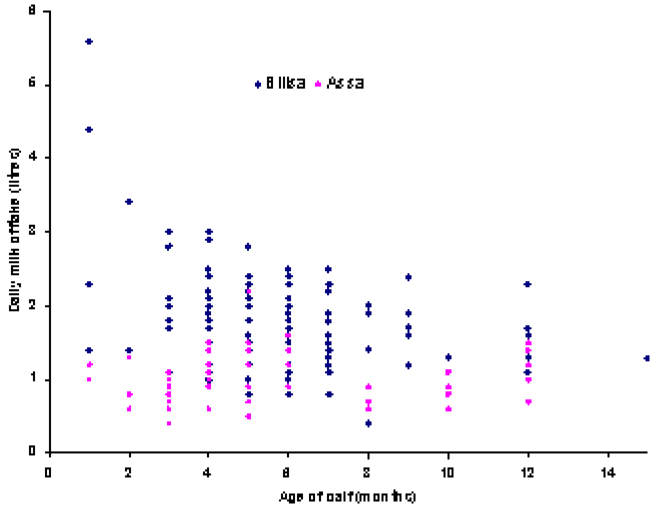


*Fig1.Association between milk offtake and age of calf in Orma Boran cows sampled in Bilisa and Assa locations in Tana River district, Kenya*

## Participatory study

The full report of the survey is given in Irungu (2000). One of the important preliminary findings from the survey was that the households interviewed put trypansomosis as the most important disease affecting their cattle and they appeared to be able to recognise different forms of the disease. Trypanocidal drugs provided the main method of disease control but other indigenous methods were also used.

The researchers proposed in their report that some form of community-based method of tsetse control was needed to reduce the impact of the diseases. Before doing so, however, it was decided to investigate more thoroughly the Orma people´s knowledge of the disease and to seek their views on the way forward for improving its control. A study using participatory methods of ´matrix scoring´ and ´proportional piling´ was therefore instigated (Catley et al., 2002) in order to understand local perceptions of incidence of different diseases, their clinical signs and causes, and preferences for indigenous and modern disease control methods.

The participatory methods involved villagers placing stones in squares traced on the ground which described the clinical signs that they associated most with a particular disease. Drawings and objects to describe the different diseases and their possible clinical signs were put on the ground in the shape of a matrix with diseases in one direction and signs in another.

The table, which illustrates some of the results of this exercise, shows five diseases (trypansomosis (in two forms as described by Orma herdsmen − *gandi* and *buku*), foot and mouth disease, pneumonia and rinderpest) and five clinical signs (weight loss, animal seeking shade, diarrhoea, coughing, reduced appetite). The villagers were asked to score the clinical signs for each disease by distributing a pile of 20 stones amongst them. The more important they felt a particular sign the more stones they put in the pile.

The method was replicated with three groups of individuals within each of four villages. The table shows the average numbers of stones (from a total of 20) placed alongside each disease.

| | Disease | | | | |
|---|---|---|---|---|---|
| | Gandi | Buku | Foot and mouth | Pneumonia | Rinderpest |
| Weight loss | O O O O O | | O O | O O O O O O O O O O O O O | |
| Animal seeks shade | O | O O | O O O O O O O O O O O O O O O O | O | |
| Diarrhoea | O O | O O O O O | | | O O O O O O O O O O O O O |
| Coughing | O O O O | O | | O O O O O O O O O O O O O O O | |
| Reduced appetite | O O O O O | O O O | O O O O O O O | O O O | O O |

The table, which illustrates some of the results of this exercise, shows five diseases (trypansomosis (in two forms as described by Orma herdsmen − *gandi* and *buku*), foot and mouth disease, pneumonia and rinderpest) and five clinical signs (weight loss, animal seeking shade, diarrhoea, coughing, reduced appetite). The villagers were asked to score the clinical signs for each disease by distributing a pile of 20 stones amongst them. The more important they felt a particular sign the more stones they put in the pile.

The method was replicated with three groups of individuals within each of four villages. The table shows the average numbers of stones (from a total of 20) placed alongside each disease.

|  | Disease |  |  |  |  |
|---|---|---|---|---|---|
|  | Gandi | Buku | Foot and mouth | Pneumonia | Rinderpest |
| Weight loss | O O O O O |  | O O | O O O O O O O O O O O O O |  |
| Animal seeks shade | O | O O | O O O O O O O O O O O O O O O O | O |  |
| Diarrhoea | O O | O O O O O |  |  | O O O O O O O O O O O O O |
| Coughing | O O O O | O |  | O O O O O O O O O O O O O O O |  |
| Reduced appetite | O O O O O | O O O | O O O O O O O | O O O | O O |

## Participatory study

The results of this investigation revealed potential constraints in implementing community-based tsetse control, an idea originally considered by the researchers. Instead the villagers recommended wider use of trypanocidal drugs as the best option and requested assistance in learning how best to apply them. The proposed research strategy was thus changed and new proposals prepared to investigate existing drug use, assess levels of possible drug resistance and design participative training courses on ´better use of trypanocides´.

Participatory methods can thus provide a useful alternative to the traditional questionnaire approach for obtaining information from farmers.

## Study questions

1. Change LOCATION from a factor to a variable and rerun the regression analysis. Are there any differences in the results? Explain why. Rewrite the algebraic expression for the model given in the case study with the term representing LOCATION now taking on the form of a covariate. Plot a scatter plot of milk offtake versus location and explain why the regression coefficient, which is equivalent to the slope of the line between the mid points for the two locations, is equivalent to the mean difference in offtakes.
2. What other factors or traits do you consider might have important effects on milk offtake? Write down a suitable statistical model that incorporates these factors. What implications do any of these factors have in study design?
3. Prepare a protocol for estimating average milk offtake and average human consumption in a group of smallholder farms possessing in the range of 1-4 cows. Explain how you would record the data and what approach you would use to analyse the data. Discuss any limitations of such a study.
4. In light of the results obtained in this Case Study redesign the part of the questionnaire given in CS1Quest1 so that it better achieves the objectives set and allows simpler and direct computer entry. Sketch out a simple computer screen to show how the data might be entered.
5. One of the proposals resulting from the participatory study was to conduct another

participatory study to quantify drug use. Using similar methods as described in the paper plan how you would find out which types of animals the farmers would select for prophylactic treatment and at what dose.

6. Write in your own way a brief report of the data analysis on milk offtake and the findings of the survey on consumption and marketing for an agricultural extension worker who may not be familiar with such statistical terminology as regression lines or standard errors. Discuss the adaptations that you have made to the report given in the case study.
7. Discuss the suitability of the cross-sectional approach used here for estimating milk offtake in relation to the objectives outlined earlier. Are there other types of study designs that you might recommend as being more suitable, either for this study or for any follow-up? Choose one alternative approach, describe, in general terms, how the study would be organised and discuss any advantages or disadvantages from that carried out here.
8. Exclude the two high milk offtakes in early lactation. Rerun the GenStat regression analysis and compare with the previous output. Which results do you think you should present? Discuss under what circumstances it is permissible to exclude outliers.
9. Exclude the two milk offtakes with the highest leverages and rerun GenStat. What do you deduce from this output? Should these points be omitted?
10. The general relationship between milk offtake and stage of lactation in cows is known to be curvilinear, increasing from calving to a peak value around 4 − 6 weeks and then decreasing. Include a quadratic term for total milk offtake in the model and rerun GenStat. Does the analysis suggest that a quadratic term should be included? Do you think that the type of data collected provides the best way for determining the shape of the lactation curve? If not, how would you design a study to achieve this aim?

## Related reading

Catley, A., Irungu, P., Simiyu, K., Dayde, J., Mwakio, W., Kiragu, J. and Nyamwaro, S.O. (2002). Participatory investigations of bovine trypanosomiasis in Tana River District, Kenya. Medical and Veterinary Epidemiology 16: 55-66. Abstract

Dolan, R.B. (1998). The Orma Boran, a trypanotolerant East African breed. Fifteen years of research on Galana Ranch in Kenya. Kenya Tryponomiasis Research Institute, Muguga, Kenya, 88pp. Full text

Irungu, P. (2000). Cattle keeping practices of the Orma people. Results of a household survey in Tana River District, Kenya. KETRI-ILRI Collaboratrive Study, Kenya Trypanosomiasis Research Institute, Muguga, 60pp. Full text

Nagda, Sonal (2009). Linear regression using R. Research Methods Group, ILRI, Nairobi 15pp. Full text

## Acknowledgements

We thank the Director of Kenya Trypanosomiasis Research Institute for allowing the data collected during the survey to be used in this Case Study, and also for permission to include for reference purposes the full report by Irungu (2000). The original study could not have been done without the support of the local District Veterinary Office in Tana River, the willingness of the Orma pastoralists to participate in the survey and the cattle that they owned.