Example 1 (referred to in module 4)

Regression analysis – an example in quantitative methods

John Rowlands

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

Background

Eighty seven percent of the country's landmass in Kenya is arid and semi-arid and, as such, not suitable for arable farming. These rangelands support over half of the country's national cattle herd and about 60% of these rangelands are tsetse infested (Irungu 2000). The ever-growing population in Kenya has given rise to an increasing demand for livestock products. The harsh climatic conditions and disease constraints are such that improved exotic breeds cannot be maintained in many areas. Increasing livestock production through the use of indigenous breeds is an important option. The Orma people, descendants of the Oromo, originated from the Borana Province in Ethiopia; a country to the north of Kenya, bringing their Boran cattle with them: the Orma Boran (Ensminger 1966). These nomadic pastoralists finally settled in the tsetse infested lands of the Tana River District. Studies at Galana Ranch (situated in Tana River and Kilifi Districts) have shown that Orma Boran cattle do better than improved Kenya Borans under high tsetse challenge with infection and mortality rates from trypanosomosis in the former being approximately half those observed in their counterparts (Dolan 1998).

In a study aimed at providing information on the Orma pastoralists and the management practices of Orma Boran cattle in their own environment, a household survey was conducted in the Tana River District of Kenya. A total of 48 households from different villages (manyattas) in various villages, were selected and the household heads interviewed. At the same time, data were collected on milk offtake of the cows in calf at the household. Through the help of the local administrative officer and purposive sampling, the households were identified. On the day of interview milk offtake for both morning and evening milkings was determined using a calibrated plastic measuring jar to an accuracy of 50ml. A total of 164 cases together with their respective age of calf, ascertained by the herd owner, were recorded. The survey was conducted in the Tana Delta itself; namely Bilisa location and in Assa location; a more arid region to the west.

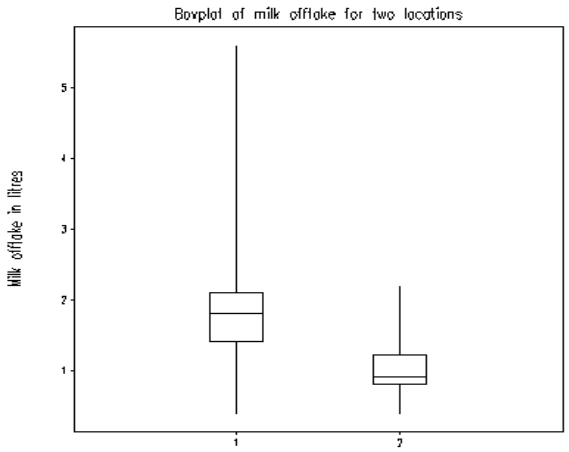
The data set used for this case study consists of recordings of daily milk offtake (the sum of morning and evening milkings), identified by location, village, sex and age of calf. For the analysis, the following codes will be used; 1 and 2 representing Bilisa and Assa locations respectively. The analyses were carried out in Genstat.

Descriptive statistics

The descriptive patterns in the data may be revealed by the following statistics in the Genstat output:

LOCATION	Nobservd	Mean	Minimum	Maximum	Median
ASSA	53	1.025	0.4000	2.200	0.900
BILISA	111	1.843	0.4000	5.600	1.800

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



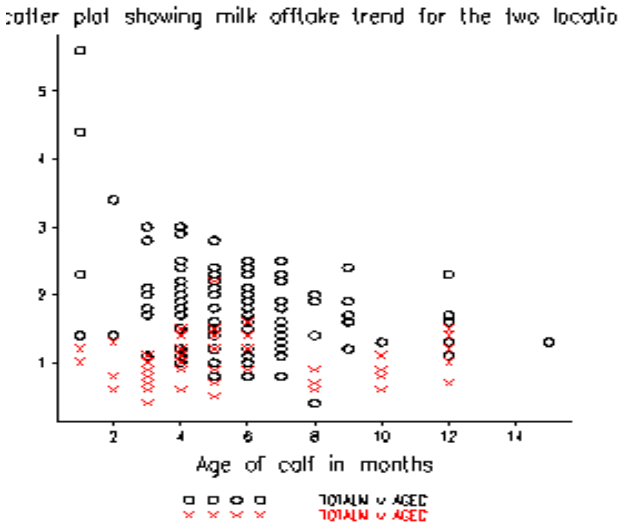


Fig.1a: Boxplot

Fig.1b: Scatter diagram

The boxplot shows that milk offtake in Bilisa is higher and more variable than that for Assa (Figure1a).

A scatter diagram of the two variables, milk offtake (TOTALM) against age of calf (AGEC), by location; symbol 'o' and 'x' for Bilisa and Assa, respectively, suggests the kind of relationship to expect between milk offtake and age of calf. The diagram also suggests that the expected relationship may be different in the two locations. A regression analysis is indicated which fits different lines to the data from the two locations.

Regression analysis

The first step is to try and fit a straight line of the form:

Y_i = constant + bX_i

where Y_i = milk offtake and X_i = age of calf for cow (i=1,2,...,164)

The Genstat output for the analysis follows:

Regression analysis

Response variate: TOTALM
Fitted terms: Constant, AGEC

Summary of analysis

	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	2.90	2.8989	5.97	0.016
Residual	162	78.70	0.4858		
Total	163	81.60	0.5006		

Percentage variance accounted for 3.0

Estimates of parameters

	estimate	s.e.	t(162)	t pr.
Constant	1.861	0.128	14.55	<.001
AGEC	-0.0496	0.0203	-2.44	0.016

The summary describes the analysis of variance and shows evidence of a statistically significant linear relationship between milk offtake and age of calf ($P<0.05$) although only 3% of the variation is accounted for. A table of parameter estimates follows this, the constant representing the intercept on the Y-axis and AGEC representing the regression coefficient or the slope of the line of milk offtake on age of calf. From the parameter estimates the fitted equation can be written as:

$$Y = 1.861(\pm 0.128) - 0.0496(\pm 0.0203)X$$

The second step is to include a parameter to describe location. The following output fits location and age of calf.

Regression Analysis

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

Summary of analysis

	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	27.12	13.5612	40.08	<.001
Residual	161	54.47	0.3383		
Total	163	81.60	0.5006		
Change	-1	-24.22	24.2236	71.60	<.001

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

The summary reveals that the relationship is improved with location included in the model. The residual mean square reduced to 0.3383 litres^2 from 0.4858 litres^2 in the previous analysis, and the percentage of variance accounted for increases from 3.0% to 32.4%.

Estimates of parameters

	estimate	s.e.	t(161)	t pr.
Constant	2.136	0.112	19.14	<.001
LOCATION 2	-0.8218	0.0971	-8.46	<.001
AGEC	-0.0511	0.0169	-3.02	0.003

With the parameter estimates given, the fitted equation may be written as:

$$Y = 2.136\ (\pm\ 0.112) - 0.8218(\pm 0.0971)L_2 - 0.0511(\pm 0.0169)X$$

where L_2 refers to Assa. The parameter estimate associated with it, means that the intercept on the Y-axis for Assa is 0.8218 litres lower than for Bilisa; which is given by the constant term.

Both regression coefficients for AGEC and LOCATION 2 are highly significant. The accumulated analysis of variance given below shows the reduction in sum of squares due to fitting AGEC having already accounted for LOCATION 2.

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ LOCATION	1	24.0448	24.0448	71. 60	<.001
+ AGEC	1	3.0777	3.0777	9.10	0.003
Residual	161	54.4729	0.3383		
Total	163	81.5953	0.5006		

Separate regression lines for the two locations can be written in the following way:

For Bilisa;

$$Y = 2.136(\pm 0.112) - 0.511(\pm 0.0169)X$$

For Assa;

$$Y = 1.314(\pm 0.125) - 0.0511(\pm 0.0169)X$$

The constant term for Assa is determined by adding the difference mentioned above to the constant term of Bilisa i.e.

$$2.136 - 0.8218 = 1.314$$

However, the standard error is not so easily obtained. One way is to run the program again but with the location code interchanged to give the following output:

Estimates of parameters

	estimate	s.e.	t(161)	t pr.
Constant	1.314	0.125	10.52	<.001
LOCATION 1	0.8218	0.0971	8.46	<.001
AGEC	-0.0511	0.0169	-3.02	0.003

Where, LOCATION 1 now stands for Bilisa.

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.

The next step is to investigate whether the data are better represented by non-parallel lines. Fitting an interaction term resulting in the output below does this.

Regression Analysis

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

Summary of analysis

	d.f.	s.s.	m.s.	v.r.	Fpr.
Regression	3	30.23	10.0766	31.39	<.001
Residual	160	51.37	0.3210		
Total	163	81.60	0.5006		
Change	-1	-3.11	3.1073	9.68	0.002

Percentage variance accounted for 35.9
Standard error of observations is estimated to be 0.567

Estimates of parameters

	estimate	s.e.	t(160)	t pr.
Constant	2.394	0.137	17.50	<.001
LOCATION 2	-1.411	0.212	-6.67	<.001
AGEC	-0.0963	0.0220	-4.38	<.001
AGEC.LOCATION 2	0.1036	0.0333	3.11	0.002

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

Accumulated analysis of variance

Change	d.f.	s.s.	m.s.	v.r.	F pr.
+ LOCATION	1	24.0448	24.0448	74.90	<.001
+ AGEC	1	3.0777	3.0777	9.59	0.002
+ AGEC.LOCATION	1	3.1073	3.1073	9.68	0.002
Residual	160	51.3656	0.3210		
Total	163	81.5953	0.5006		

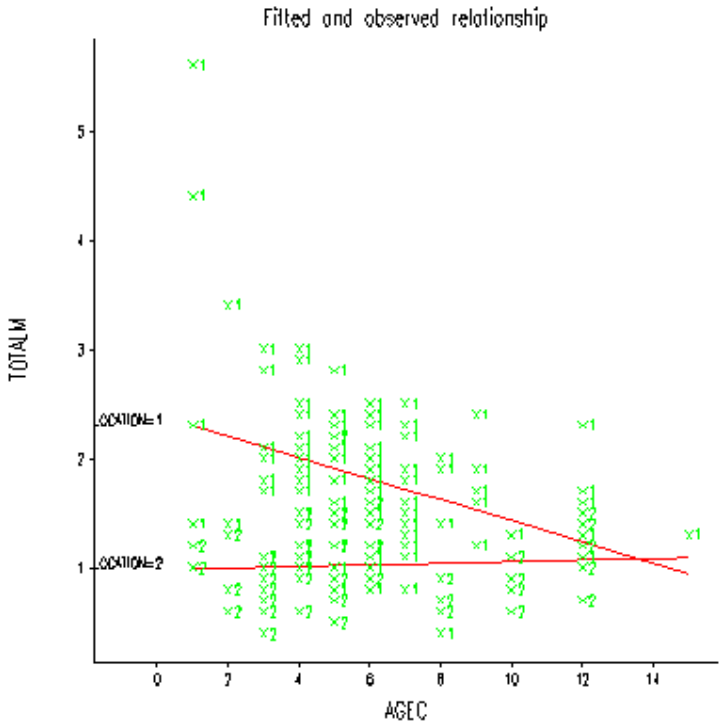


Fig. 2: The fitted and observed relationship between milk offtake and age of calf with 'X1' and 'X2' representing data for Bilisa and Assa respectively.

The fitted regression lines for the two locations are given below:

for Bilisa: $Y = 2.394(\pm 0.137) - 0.0963(\pm 0.022)X$
for Assa: $Y = 0.983(\pm 0.161) + 0.0073(\pm 0.025)X$

The regression coefficient of 0.0073 for Assa with its standard error is determined in a similar way to that described earlier, by interchanging the location code resulting in the following parameter estimates:

Estimates of parameters

	estimate	s.e.	t(160)	t pr.
Constant	0.983	0.161	6.09	<.001
LOCATION 1	1.411	0.212	6.67	<.001
AGEC	0.0073	0.0250	0.29	0.771
AGEC.LOCATION 1	-0.1036	0.0333	-3.11	0.002

Genstat also produces warning messages at every stage of the analysis. At this stage, in particular, the following messages were given.

* 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
86	1.100	0.068
100	1.300	0.139
117	1.500	0.097
121	1.200	0.097
122	1.400	0.097
139	0.700	0.097
153	1.000	0.097

These can be explained as follows:

- Units that have high standardised residuals (calculated as the deviation of an observation from its fitted value divided by the overall residual standard deviation) are those milk offtakes that fall some distance away from the fitted line. These may be considered 'outliers'.

- The message concerning the error variance is an indicator that the assumption of constant variance for the Y-variable may not be tenable.

- The units with high leverage are those milk offtakes that have a strong influence on the direction of the regression line. The sum of leverages of all the units in the sample in question is always equal to the number of parameters used in the regression model. In this case the number of units is 164 and of parameters in the model, 4. The message appears for those units with more than about twice the average influence on the model.

The two units with high-standardised residuals can be seen to lie to the top left of the scatter diagram. Further analysis may be done by fitting the same model with these two milk offtakes omitted. The points with high leverage are those for an age of calf 12 months and beyond. These observations could also be omitted to see the effect on the analysis.

Study questions

1. Excluding the two milk offtakes in early lactation resulted in the following parameter estimates:

Estimates of parameters

	estimate	s.e.	t(158)	t pr.
Constant	2.101	0.118	17.77	<.001
AGEC	-0.0543	0.0188	-2.88	0.004
LOCATION 2	-1.117	0.178	-6.27	<.001
AGEC.LOCATION 2	0.0616	0.0280	2.20	0.029

Compare this with the output with all the data.
What does this tell you?
When would you consider it permissible to exclude outliers?

2. The message concerning the error variance indicates vertical variation in milk offtake about the regression line increases with the fitted value. This can be seen on the scatter diagram as one moves from right to left, particularly for location Bilisa. Explain why you think the error variance is not constant.
   How might you overcome this on the analysis?
   Do you think this is necessary?

3. Excluding the milk offtakes with high leverage results in the following output:

Estimates of parameters

	estimate	s.e.	t(147)	t pr.
Constant	2.178	0.150	14.54	<.001
AGEC	-0.0699	0.0263	-2.66	0.009
LOCATION 2	-1.150	0.215	-5.34	<.001
AGEC.LOCATION 2	0.0665	0.0382	1.74	0.084

   What do you deduce from this output?
   Should these points be omitted from the data for analysis?

4. The association between milk offtake and stage of lactation in cows in general is curvilinear decreasing from a peak offtake at around 4 – 6 weeks. Including a quadratic term on the model to account for this resulted in the following:

Estimates of parameters

	estimate	s.e.	t(145)	t pr.
Constant	2.241	0.323	6.95	<.001
LOCATION 2	-1.657	0.459	-3.61	<.001
AGEC	-0.096	0.119	-0.80	0.423
AGEC.LOCATION 2	0.279	0.172	1.62	0.108
C2	0.0024	0.0106	0.22	0.824
C2.LOCATION 2	-0.0181	0.0147	-1.23	0.222

   where C2 is the quadratic term AGEC*AGEC

What conclusions do you draw?
Comment on the results of this analysis.

5. Discuss the suitability of this cross-sectional study for estimating milk offtake. Are here any ways that it could be improved? Describe how you would go about determining levels of milk offtake in a breed of cattle.

6. Are there other factors or covariates that you might consider important in field studies of assessment of levels of milk offtake?

7. Illustrate how you might categorise the factors into different levels. For example, age of cow might be important. Suggest how you might define the level of a factor that you could use to represent age in the model.

In: Animal Genetics Training Resource, version 2, 2006. Ojango, J.M., Malmfors, B. and Okeyo, A.M. (Eds).International Livestock Research Institute, Nairobi, Kenya, and Swedish University of Agricultural Sciences, Uppsala, Sweden.