Research Methods Group

ILRI

INTERNATIONAL
LIVESTOCK RESEARCH
INSTITUTE

World Agroforestry Centre
TRANSFORMING LIVES AND LANDSCAPES

LINEAR REGRESSION USING R

Using Case Study 1 from the
BIOMETRICS & RESEARCH METHODS TEACHING RESOURCE

BY

Sonal Nagda

www.ilri.org/rmg
P. O. Box 30709 - 00100, Nairobi, Kenya
Phone + 254 2 422 3000
Fax +254 2 422 3001

www.worldagroforestrycentre.org/rmg
P. O. Box 30677, Nairobi, Kenya
Phone + 254 20 722 4000
Fax + 254 2 722 4001

CONTENTS

1. Introduction

R is an open-source software which is free to use, distribute and modify under the open-source type licence. The newest version of R and its documentation can be downloaded from http://www.R-project.org

The data used come from a study carried out in the Tana Delta area of Kenya; referred to as Bilisa and Assa location. The study was conducted to investigate differences in the management practices between the two locations and their impact on milk offtake. For more information, refer to Case Study 1 in the Biometrics & Research Methods Teaching Resource.

The specific questions to be addressed regarding milk production in this example using the R software are:

- What is the general level of milk offtake in Bilisa and Assa locations.

- How does milk offtake vary with age of calf (i.e. with month of lactation) and do the patterns differ between Bilisa and Assa location.

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 also find out how average milk offtake differs between the two locations. Then, we will use regression analysis to investigate the nature of the relationship between age of calf and the two locations with the response variable milk offtake.

2. Description of the data

The data used in this example is stored in **CS1data1.xls**, on the Biometrics & Research methods Teaching Resource CD.

Field	**Description**
COWNO	Cow number Unique ID
VILLAGE	Village where homestead located
LOCATION	Location code 1 = Bilisa referring to general area in the Eastern and western half of the Delta, 2 = Assa, located in the western and drier area.
AGEC	Age of calf (months)
QTYM	Recorded morning milk offtake (litres)
QTYE	Recorded evening milk offtake (litres)
TOTALM	Total recorded milk offtake (litres) calculated as the sum of QTYM and QTYE.

Importing Data into R

Data can be stored in a variety of software programs (e.g ACCESS , EXCEL, GENSTAT etc). The best way is to export into an ASCII or Excel file which can be used in R.

From Excel, the data can be saved as ".csv" (comma separated values) format. The first row in Excel should be reading the variable names and then the data below. Any rows above the variable names should be deleted. Save the CS1Data1 file as a .csv file into a folder on your computer (e.g. c://CaseStudyData/CS1Data1.csv).

Commands in R to read the .csv file:

```
> data1<-read.csv("c://CaseStudyData/CS1Data1.csv",  header=TRUE,  sep=",")
> data1
```

Alternatively, to read EXCEL or ACCESS file directly, the **RODBC** package needs to be installed in R. Then use the commands:

```
data1<-" c://CaseStudyData/CS1Data1.xls"
connect2<-odbcConnectExcel(data1)
data1<-sqlFetch(channel=connect2, sqtable="Sheet1")
```

To display the names of variables, type in "**names(data1)**" to give the output:

```
> names(data1)
[1] "COWNO"    "VILLAGE"  "LOCATION" "AGEC"     "QTYM"     "QTYE"     "TOTALM"
```

To display the characteristic of the variables, type in “**str(data1)**” to give the output:

```
> str(data1)
'data.frame':   164 obs. of  7 variables:
 $ COWNO         :num  1 2 3 4 5 6 7 8 9 10 ...
 $ VILLAGE       :Factor w/ 12 levels "ASSA","CHIRA",..: 5 5 5 5 11 11 11 8 10 10 ...
 $ LOCATION      :num  1 1 1 1 1 1 1 1 1 1 ...
 $ AGEC          :num  5 4 4 1 3 3 5 7 5 8 ...
 $ QTYM          :num  1.2 0.9 1 1 0.7 1.3 0.8 1 0.3 1 ...
 $ QTYE          :num  1.2 0.6 1.2 1.3 1 1.5 1.3 1.2 0.5 1 ...
 $ TOTALM        :num  2.4 1.5 2.2 2.3 1.7 2.8 2.1 2.2 0.8 2 ...
```

If the variable is not numeric then R usually considers it to be a factor or categorical variables (e.g. Village above).

To transform numerical variables (e.g. “LOCATION”) into factor type use the command:

```
> data1$LOCATION=as.factor(data1$LOCATION)
```

Check again with “**str(data1)**” that it has been converted to a factor.

To simplify the commands in R, run the “**attach(data1)**” command, so that when specifying a variable in a function the “**data1$**” is no longer required. I.e. instead of “**data1$LOCATION**” above use “**LOCATION**”.

```
> attach(data1)
```

3. Data exploration

The first step, before undertaking any statistical analysis, is to explore the data.

To summarize the variables in **data1** type “**summary(data1)”** to obtain the following output:

```
> summary(data1)
     COWNO             VILLAGE        LOCATION        AGEC              QTYM
   Min.: 1.00            KONE: 42       1: 111     Min.: 1.000        Min.: 0.1000
1st Qu.: 41.75        KONKONA: 35       2: 53   1st Qu.: 4.000     1st Qu.: 0.5000
 Median: 82.50           TULU: 27               Median: 5.000      Median: 0.7000
   Mean: 82.50          SHELI: 20                 Mean: 5.701        Mean: 0.7488
3rd Qu.: 123.25         CHIRA: 10               3rd Qu.: 7.000     3rd Qu.: 1.0000
   Max.: 164.00        HAMESA: 9                   Max.: 15.000       Max.: 2.9000
                      (Other): 21
     QTYE              TOTALM
   Min.: 0.2000          Min.: 0.400
1st Qu.: 0.6000       1st Qu.: 1.100
 Median: 0.8000        Median: 1.500
   Mean: 0.8299          Mean: 1.579
3rd Qu.: 1.0000       3rd Qu.: 2.000
   Max.: 2.7000          Max.: 5.600
```

For continuous variables, the summary statistics are shown (min, median, mean, max, 1st & 3rd quartiles) and for factors a frequency tabulation is displayed.

To obtain the mean, standard deviation, minimum, maximum and median cross-tabulated by location type use:

```
>aggregate(data.frame(TOTALM=TOTALM),by=list(LOCATION),mean)
>aggregate(data.frame(TOTALM=TOTALM),by=list(LOCATION),sd)
>aggregate(data.frame(TOTALM=TOTALM),by=list(LOCATION),min)
>aggregate(data.frame(TOTALM=TOTALM),by=list(LOCATION),max)
>aggregate(data.frame(TOTALM=TOTALM),by=list(LOCATION),median)
```

Output:

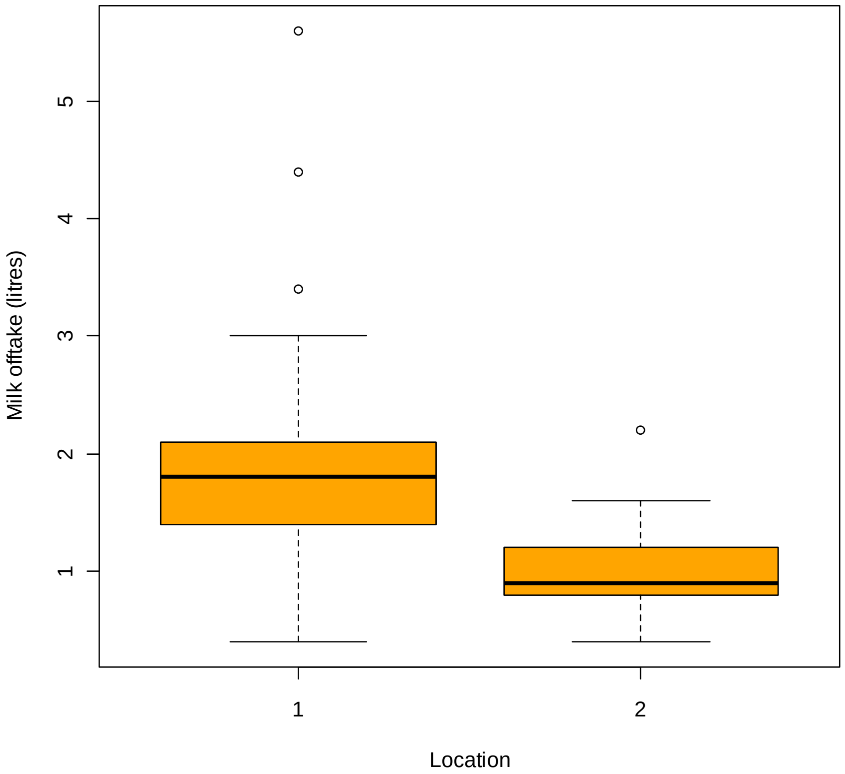
```
> aggregate(data.frame(TOTALM=TOTALM),by=list(LOCATION),mean)
  Group.1   TOTALM
1       1 1.843243
2       2 1.024528
> aggregate(data.frame(TOTALM=TOTALM),by=list(LOCATION),sd)
  Group.1    TOTALM
1       1 0.6840550
2       2 0.3418871
> aggregate(data.frame(TOTALM=TOTALM),by=list(LOCATION),min)
  Group.1 TOTALM
1       1    0.4
2       2    0.4
> aggregate(data.frame(TOTALM=TOTALM),by=list(LOCATION),max)
  Group.1 TOTALM
1       1    5.6
2       2    2.2
> aggregate(data.frame(TOTALM=TOTALM),by=list(LOCATION),median)
  Group.1 TOTALM
1       1    1.8
2       2    0.9
```

The means and medians in both locations are comparatively close indicating a symmetric distribution. However, the range of milk offtake in Location 1 (Bilisa) is 5.2 litres per day compared to 1.8 litres per day in Location 2 (Assa).

To view the differences in variation in milk offtakes use the commands below to produce a boxplot:

```
> boxplot(TOTALM~LOCATION, col="orange",xlab="Location", ylab="Milk offtake (litres)" )
```

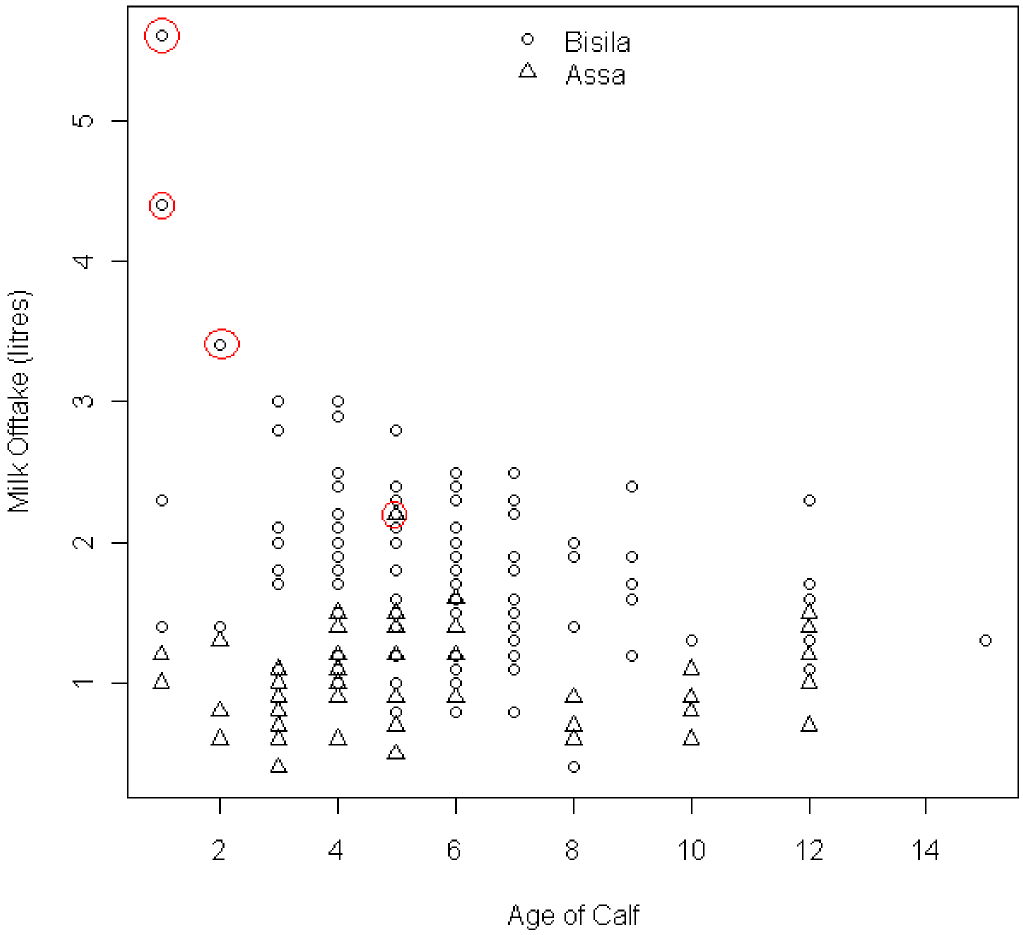
The boxplot shown is equivalent to a Schematic boxplot in GenStat showing the median (horizontal line in the box), the interquartile range (the box), the upper/lower fence (end of the lines – representing 1.5 x the interquartile range above/below) and outliers in the data (lying outside of the upper/lower fence).



A boxplot illustrates the differences in variation in milk offtakes in Bilisa and Asssa. It shows three outliers for Bilisa (Location 1) and one for Assa (Location 2).

One can see the four outliers (but not as clearly) in a scatter plot, produced using the following commands:

```
> psymb=as.numeric(LOCATION)
> plot(TOTALM~AGEC,pch=psymb,xlab="Age of calf", ylab="Milk offtake (litres)")
> legend("top","center",legend = c("Bisila","Assa"),pch=1:2,bty="n")
```



The graph shows the different patterns between the milk offtake and age of the calf for the two locations. These patterns support the use of a multiple regression analysis to describe different intercepts on the y-axis for the two locations and potential differing slopes.

4. Statistical modelling

The first step in the analysis of the milk offtake is to fit a regression with a term to describe a common slope for the pattern (AGEC) and a term to allow separate intercepts (LOCATION) on the y-axis for the location. We can do this by fitting these two parameters (AGEC, LOCATION) using the commands:

```
>fm2=lm(TOTALM~LOCATION+AGEC)
>fm2
>summary(fm2)
```

Output:

```
Call:
lm(formula = TOTALM ~ LOCATION + AGEC)

Coefficients:
(Intercept)   LOCATION2        AGEC
   2.13562     -0.82180     -0.05111

> summary(fm1)

Call:
lm(formula = TOTALM ~ LOCATION + AGEC)

Residuals:
    Min            1Q          Median          3Q            Max
-1.32675       -0.32897       -0.01918       0.27103        3.51549

Coefficients:
                         Estimate      Std. Error      t value          Pr(>|t|)
(Intercept)               2.13562       0.11156        19.143           < 2e-16 ***
LOCATION2               -0.82180        0.09712        -8.461           1.53e-14 ***
AGEC                    -0.05111        0.01695        -3.016           0.00298 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5817 on 161 degrees of freedom
Multiple R-squared: 0.3324,     Adjusted R-squared: 0.3241
F-statistic: 40.08 on 2 and 161 DF,  p-value: 7.473e-15
```

The percentage of variation accounted for by the model is 32.4% and both parameters are significant.

Multiplying the s.e by 2 and adding and subtracting respectively from the estimates of 95% confidence intervals gives us the interval (-1.0160 to -0.6276 for LOCATION and -0.0850 to -0.0172 for AGEC) within which the true difference between location and the true slope are expected to lie. We can see that neither of the pair of limits contain value zero, confirming that the data can be represented by separate lines with a slope that is significantly different from 0.

R gives the coefficient for LOCATION 2 (Assa) which refers to the difference between LOCATION 1 (Bilisa) and LOCATION 2 (Assa). In R LOCATION 1 (Bilisa) is the default reference level.

However, we can recode the data so that LOCATION 1 (Assa) is the reference level and rerun the model. Using the commands below we can create another variable called **LOCATIONB** where the reference level is “Assa”. Note that R auto-corrects LOCATIONB to be a factor (as LOCATION was a factor).

```
> LOCATIONB=(ifelse(LOCATION==1,"Bilisa","Assa"))
>fm3=lm(TOTALM~LOCATIONB+AGEC)
>fm3
>summary(fm3)
```

Output:

```
Call:
lm(formula = TOTALM ~ LOCATIONB + AGEC)

Coefficients:
  (Intercept)          LOCATIONBBilisa                    AGEC
     1.31382          0.82180                        -0.05111

> summary(fm3)

Call:
lm(formula = TOTALM ~ LOCATIONB + AGEC)

Residuals:
     Min         1Q          Median         3Q         Max
-1.32675     -0.32897     -0.01918     0.27103     3.51549

Coefficients:
                        Estimate           Std. Error    t value       Pr(>|t|)
(Intercept)              1.31382            0.12484      10.524        < 2e-16 ***
LOCATIONBBilisa          0.82180            0.09712       8.461        1.53e-14 ***
AGEC                    -0.05111            0.01695      -3.016        0.00298 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5817 on 161 degrees of freedom
Multiple R-squared: 0.3324,     Adjusted R-squared: 0.3241
F-statistic: 40.08 on 2 and 161 DF,  p-value: 7.473e-15
```

The fitted regression lines for the two locations can now be calculated using the two models (fm2 & fm3)

For **Bilisa** y_i = 2.136 (±0.112) - 0.051 (±0.017)AGEC

For **Assa** y_i = 1.314 (±0.125) - 0.051 (±0.017)AGEC

The next step is to investigate whether non-parallel lines would better represent the data. This is achieved by fitting an interaction in the model (LOCATIONB*AGEC). In this case we will also request for the accumulated analysis of variance (ANOVA) to be shown:

```
> fm4=lm(TOTALM~LOCATIONB+AGEC+LOCATIONB*AGEC)
> fm4
> summary(fm4)
> anova(fm4)
```

Output:

```
Call:
lm(formula = TOTALM ~ LOCATIONB + AGEC + LOCATIONB * AGEC)

Coefficients:
      (Intercept)          LOCATIONBBilisa              AGEC
         0.983316           1.410689                  0.007281

LOCATIONBBilisa:AGEC
        -0.103556

> summary(fm4)

Call:
lm(formula = TOTALM ~ LOCATIONB + AGEC + LOCATIONB * AGEC)

Residuals:
    Min      1Q   Median      3Q     Max
-1.22381 -0.31356 -0.01440  0.28737  3.30227

Coefficients:
                        Estimate    Std. Error    t value     Pr(>|t|)
(Intercept)             0.983316    0.161472      6.090       8.10e-09 ***
LOCATIONBBilisa         1.410689    0.211613      6.666       4.03e-10 ***
AGEC                    0.007281    0.024994      0.291       0.77120
LOCATIONBBilisa:AGEC   -0.103556    0.033286      -3.111      0.00221 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5666 on 160 degrees of freedom
```

```
Multiple R-squared: 0.3705,     Adjusted R-squared: 0.3587
F-statistic: 31.39 on 3 and 160 DF,  p-value: 5.194e-16

> anova(fm4)
Analysis of Variance Table

Response: TOTALM
                     Df      Sum Sq       Mean Sq      F value      Pr(>F)
LOCATIONB            1       24.045       24.0448      74.8976      5.025e-15 ***
AGEC                 1       3.078        3.0777       9.5867       0.002315 **
LOCATIONB:AGEC       1       3.107        3.1073       9.6789       0.002208 **
Residuals            160     51.366       0.3210
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The output shows that the interaction term is significant (p<0.01). The variance accounted for (R^2) increases from 32.4% to 35.9%.

The accumulated analysis of variance shows the additional sum of squares accounted for. Each sum of squares is corrected for variables already included in the model.

Similarly, we can run the model with Bisila (Location 1) as the reference level:

```
>fm5=lm(TOTALM~LOCATION+AGEC+LOCATION*AGEC)
>fm5
>summary(fm5)
```

Output:

```

                     Estimate     Std. Error    t value      Pr(>|t|)
(Intercept)            2.39400    0.13677       17.503       < 2e-16 ***
LOCATION2            -1.41069     0.21161       -6.666       4.03e-10 ***
AGEC                 -0.09627     0.02198       -4.380       2.14e-05 ***
LOCATION2:AGEC        0.10356     0.03329        3.111       0.00221 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

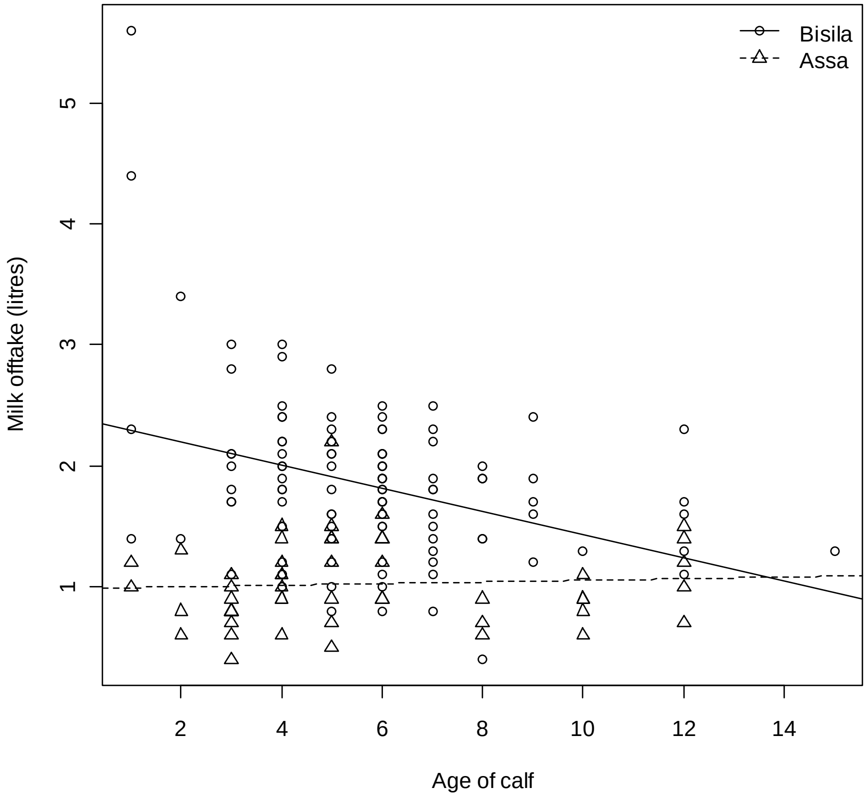
The fitted regression lines for the two locations can now be calculated as

For **Assa** y_i = 0.983 (±0.161) - 0.007 (±0.025)AGEC

For **Bilisa** y_i = 2.394 (±0.137) - 0.096 (±0.022)AGEC

To plot the fitted two regression lines use the commands:

```
>psymb=as.numeric(LOCATION)
>plot(TOTALM~AGEC, pch=psymb,xlab="Age of calf", ylab="Milk offtake (litres)")
>abline(lm(TOTALM~AGEC,subset=LOCATION==1))
>abline(lm(TOTALM~AGEC,subset=LOCATION==2), lty=2)
>legend("topright",legend = c("Bisila","Assa"),pch=1:2,lty=1:2,bty="n")
```



The output shows how milk offtake decreases with stage of lactation for cows residing in Bisila but not at Assa where milk offtakes remain low throughout the lactation.

5. Findings, implications and lessons learned

Regression analysis

- There is a difference in mean milk offtake between cows sampled in Bilisa and Assa.

- Milk offtake declined with age of calf (or stage of lactation) at Bilisa but not at Assa.

- Patterns in the data suggest that certain observations may have been influential in determining the fitted patterns and that there were others that there were others that possibly did not belong to the overall pattern.

- Switching the order in which the factor LOCATION is coded, i.e. redefining the reference level, allows the standard errors for both levels to be easily extracted.