New York Science Journal

Websites:
http://www.sciencepub.net/newyork
http://www.sciencepub.net

Emails:
newyorksci@gmail.com
editor@sciencepub.net

MARSLAND PRESS
Multidisciplinary Academic Journal Publisher

Review On Analysis Of Continuous Data By Using R Software

Umer Seid

Oda Bultum University, Collage of Agriculture, Department of Animal Science, Chiro, Ethiopia
Omerseid76@gmail.com

Abstract: Surveys primarily collect quantitative data, it can contain many kinds of questions; these questions are oft en called variables. There are some basic types of variables. It is important to understand the different types of varia bles, because the type of variable can lead to different kinds of data and guide your analysis. They are discrete data a nd continuous data. Discrete data can take at most countable number of values, whereas continuous data can take an y number of values. There is different test that used to analysis the continuous data these include: The *t*-test (single t -test, paired t-test, analysis of variance (ANOVA), linear regression (simple and multiple linear regression)). Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. Multiple regression models thus describe how a single response variable Y depends linearly on a number of predictor variables. A data set that constitutes body weight and linear body measurements recorded f or a random sample of 32 yearling Borana steers purchased from Yabello market was been used to analysis by multi ple regression. Because my data contains four explanatory variables multiple linear regressions was found to be the appropriate statistical analysis method. First model simplification was done on the data set to remove not significant variable then interpreted on simplify model. Then check the major assumption of multiple linear regression that incl udes: normality, homoscedasticity and linearity whether it is fulfill this assumption or not and the assessing individu al observation include outlier test, assessments of leverage values and influential observations
[Seid U**. Review on Analysis of Continuous Data by Using R Software**. *N Y Sci J* 2021;14(4):13-30]. ISSN 1554-0200 (print); ISSN 2375-723X (online). http://www.lifesciencesite.com. 2. doi:10.7537/marsnys140421.02.

Keywords: Review; Analysis; Continuous Data; Using; Software.

INTRODUCTION

Data is the most salient entity in statistics as it is necessarily the "study of the collection, organization, analysis, and interpretation of data". The numerical data used in statistics fall in to two main categories. They are discrete data and continuous data. Discrete data can take at most countable number of values, whereas continuous data can take any number of values. Discrete data usually occurs when data is collected by counting, but continuous data usually occurs when data is collected by taking measurements. Constructs or factors being studied are represented by "variables". Variables (also sometimes called "factors") have "values" or "levels". Variables summarize and reduce data, attempting to represent the "essential" information. Variables can be classified in various ways. A continuous variable takes on all values within its permissible range, so that for any two allowable values there are other allowable values in between. A continuous variable (sometimes called a "measurement variable") can be used in answer to the question "how much". Measurements such as weight, height, and blood pressure can, in principle, be represented by continuous variables and are frequently treated as such in statistical analysis. In practice, of course, the instruments used to measure these and other phenomena and the precision with which values are recorded allow only a finite number of values, but these can be regarded as points on a continuum(Ragland, 1992).

Mathematically, a discrete variable can take only certain values between its maximum and minimum values, even if there is no limit to the number of such values (e.g., the set of all rational numbers is countable though unlimited in number). Discrete variables that can take any of a large number of values are often treated as if they were continuous. If the values of a variable can be placed in order, then whether the analyst elects to treat it as discrete and/or continuous depends on the variable's distribution, the requirements of available analytic procedures, and the analyst's judgment about interpretability (Hertz-Picciotto, 1999).

Continuous variables has two type's interval and ratio. Interval is differences (intervals) between values are meaningful, but ratios of values are not. That is, if the variable takes on the values 11-88, with a mean of 40, it is meaningful to state that subject A's score of 60 is "twice as far from the mean" as subject B's score of 50. But it is not meaningful to say that subject A's score is "1.5 times the mean". The reason is that the zero point for the scale is arbitrary, so values of

the scores have meaning only in relation to each other. Without loss of information, the scale can be shifted: 11-88 could be translated into 0-77 by subtracting 11(Sun, 2011a).

Scale scores can also be multiplied by a constant. After either transformation, subject A's score is still twice as far from the mean as is subject B's, but subject A's score is no longer 1.5 times the mean score. Psychological scales (e.g., anxiety, depression) often have this level of measurement. An example from physics is temperature measured on the Fahrenheit or Celsius scale. Ratio is both differences and ratios are meaningful. There is a non-arbitrary zero point, so it is meaningful to characterize a value as "x" times the mean value. Any transformation other than multiplying by a constant (e.g., a change of units) will distort the relationships of the values of a variable measured on the ratio scale. Physiological parameters such as blood pressure or cholesterol are ratio measures. Kelvin or absolute temperature is a ratio scale measure (Zeger, 1991a).

Linear regression is suitable for modeling the outcome when it is measured on a continuous, or near-continuous scale. In regression analysis the relationship is asymmetric in that we think the value of one variable is caused by (or we wish to predict it by) the value or state of another variable.

The outcome variable is denoted as the dependent, or outcome, variable, whereas the 'causal' or 'predictor' variables are called the independent or predictor variables. We continue to refer to the predictor variable(s) of primary interest as the exposure variable(s). The predictor variables can be measured on a continuous, categorical or dichotomous scale(Dohoo et al., 2003).

Surveys primarily collect quantitative data, it can contain many kinds of questions; these questions are often called variables. There are some basic types of variables. It is important to understand the different types of variables, because the type of variable can lead to different kinds of data and guide your analysis (Woolf et al., 1990).

1. STATISTICAL TESTS TO ANALYZE CONTINUOUS DATA

The *t*-test and one-way analysis of variance (ANOVA) are basic tools for assessing the statistical significance of differences between the average values of a continuous outcome across two or more samples. Both the *t*-test and one-way ANOVA can be seen as methods for assessing the association of a categorical predictor – binary in the case of the *t*-test, with more than two levels in the case of one-way ANOVA – with a continuous outcome. Both are based in statistical theory for normally distributed outcomes, but work well for many other types of data; and both turn out to be special cases of linear regression models. The linear regression model with a continuous outcome and a single or more continuous predictor variables(Hertz-Picciotto, 1999)

1.1. Independent Sample t-test

An experiment with two groups can be either a **paired-samples design** or an **independent-samples design.** For either design, the appropriate statistical test is a ***t* test.** However, the two different designs require different formulas for calculating the *t* test, so you must decide what kind of design you have before you analyze the data. In an independent-samples design, there is no reason to pair up the scores in the two groups. You cannot tell the difference between the two designs just by knowing the independent variable and the dependent variable. And, after the data are analyzed, you cannot tell the difference from the *t*-test value or from the interpretation of the experiment. To tell the difference, you must know whether scores in one group are paired with scores in a second group. This is used where data are collected from groups which are unrelated, such as the length at one year of a group of infants who were breastfed, compared with a group who were not breastfed. While using *t*-test we assume that the population from which sample has been taken is normal or approximately normal, sample is a random sample, observations are independent(Spatz, 2007).

Test statistics

The test statistics is t and it is calculated as follows

$$t_{cal} = \frac{\overline{X}_2 - \overline{X}_1}{SE}$$

The test is simply an extension of a one sample hypothesis t test procedure. $\overline{X}_1$ and $\overline{X}_2$ are the means of the two groups of data. Also the SE in the denominator refers to standard error value. Since there are two standard error values associated with the respective group mean, the standard error in the denominator is the sum of the standard errors of the two group means. This standard error is called the standard error for the difference between two means. However as SE values are derived from variance and are not directly additive, they should be converted into variance of means and the latter are added(Yosef Tekle-Giorgis, 2017).

$$S_{\overline{X}1} = \sqrt{\frac{S_1^2}{n_1}} \quad \text{and} \quad S_{\overline{X}2} = \sqrt{\frac{S_2^2}{n_2}}$$

$$\text{Hence SE} = S_{\overline{X}1} + S_{\overline{X}2} = \sqrt{\frac{S_1^2}{n_1}} + \sqrt{\frac{S_2^2}{n_2}} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

1.2. Paired Sample t-test

The paired *t*-test is for use in settings where individuals or observations are linked across the two samples. Examples include measurements taken at two time points on the same individuals, or on other naturally linked pairs, as in a clinical trial where one eye is treated and the other serves as a control. In a paired-samples (or paired-scores) design, each dependent variable score in one treatment is matched or paired with a particular dependent-variable score in the other treatment. This pairing is based on some logical reason for matching up two scores and not on the size of the scores. The paired-samples design is a favourite of researchers if their materials permit. The logical pairing required for this design can be created three ways: natural pairs, matched pairs, and repeated measures. Fortunately, the arithmetic of calculating a *t*-test value is the same for all three. In a **natural pair's** investigation, the researcher does not assign the participants to one group or the other; the pairing occurs naturally, prior to the investigation. Matched pairs in some situations, the researcher has control over the ways pairs are formed and matches can be arranged. One method is for two participants to be paired on the basis of similar scores on a pretest that is related to the dependent variable. Repeated measures a third kind of paired-samples design is called a **repeated measures** design because more than one measure is taken on each participant. Assumption is the observation in each pair should be different and the set of differences for all pairs is approximately normally distributed even though the original observation in the groups may not be. This design may take the form of a before-and-after experiment. In this case the two data sets are generated from same or related individuals. Thus the two data sets are not independent (Spatz, 2007).

Test statistic

The test statistic is t and it is computed as follows:

$t_{cal} = \frac{\overline{d}}{S_{\overline{d}}}$, where

$\overline{d}$ = the mean paired difference = $\sum d_i/n$

d_i = is the paired difference obtained by deducting the paired data

$S_{\overline{d}}$ = the standard error of the mean paired difference and it is calculated as

$$S_{\overline{d}} = \sqrt{\frac{S_d^2}{n}} \quad S_d^2 = \frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n-1}$$

The value that we analyse for each pair is the ***difference*** *between the two measurements.*

$$t = \bar{x}/\text{s.e.}$$

where $\bar{x}$ = mean of the differences and s.e. = standard error of the differences.

$$\text{d.f.} = n - 1$$

where n = sample size.

$$\text{s.e.} = s/\sqrt{n}$$

where s = standard deviation of the differences and n = sample size.

1.3. Single Sample t-test

This test compares a sample mean with a population mean. Assumption the sample data is from normally distributed population of values and are representative of that population (random selection).

$$t = (\bar{x} - \mu)/\text{s.e.}$$

where $\bar{x}$ = sample mean, μ = population mean and s.e. = standard error of sample mean.

$$\text{d.f.} = n - 1$$

where n = sample size.

$$\text{s.e.} = s/\sqrt{n}$$

where s = standard deviation of sample mean and n = sample size.

Source: (Yosef Tekle-Giorgis, 2017)

1.4. Analysis of Variance (ANOVA)

Suppose that we need to compare sample averages across the arms of a clinical trial with multiple treatments, or more generally across more than two independent samples. For this purpose, one-way analysis of variance (ANOVA) and the *F*-test take the place of the *t*-test. The appropriate technique for analyzing continuous variables when there are three or more groups to be compared is the analysis of variance, commonly referred to as ANOVA. An example might be comparing the blood pressure reduction effects of the three drugs. The principals involved in the analysis of variance are the same as those in the t-test. Under the null hypothesis we would have the following situation: there would be one big population and if we picked samples of a given size from that population we would have a bunch of sample means that would vary due to chance around the grand mean of the whole population. If it turns out they vary around the grand mean more than we would expect just by chance alone, then perhaps something other than chance is operating. Perhaps they don't all come from the same population. Perhaps something distinguishes the groups we have picked. We would then reject the null hypothesis that all the means are equal and conclude the means are different from each other by more than just chance. Essentially, we want to know if the variability of all the groups' means is substantially greater than the variability within each of the groups around their own mean(Ahlbom, 1993).

We calculate a quantity known as the between-groups variance, which is the variability of the group means around the grand mean of all the data. We calculate another quantity called the within-groups variance, which is the variability of the scores within each group around its own mean. One of the assumptions of the analysis of variance is that the extent of the variability of individuals within groups is the same for each of the groups, so we can pool the estimates of the individual within group variances to obtain a more reliable estimate of overall within-groups variance. If there is as much variability of individuals within the groups as there is variability of means between the groups, the means probably come from the same population, which would be consistent with the hypothesis of no true difference among means, that is, we could not reject the null hypothesis of no difference among means. The ratio of the between-groups variance to the within-groups variance is known as the F ratio. Values of the F distribution appear in tables in many statistical texts and if the obtained value from our experiment is greater than the critical value that is tabled, we can then reject the hypothesis of no difference. There are different critical values of F depending on how many groups are compared and on how many scores there are in each group. To read the tables of F, one must know the two values of degrees of freedom (df). The df corresponding to the between-groups variance, which is the numerator of the F ratio, is equal to $k - 1$, where k is the number of groups. The df corresponding to the denominator of the F ratio, which is the within-groups variance, is equal to $k \times (n - 1)$, that is, the number of groups times the number of scores in each group minus one(Ahlbom, 1993)

1.5. Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable *causes* the other (for example, higher SAT scores do not *cause* higher college grades), but that there is some significant association between the two variables. A scatterplot can be a helpful tool in determining the strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatterplot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables(Zeger, 1991b).

Linear regression is the most basic type of regression and commonly used predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome variable? Is the model using the predictors accounting for the variability in the changes in the dependent variable? (2) Which variables in particular are significant predictors of the dependent variable? And in what way do they–indicated by the magnitude and sign of the beta estimates–impact the dependent variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. (3)

What is the regression equation that shows how the set of predictor variables can be used to predict the outcome? The simplest form of the equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent score, c = constant, b = regression coefficients, and x =independent variable(Woolf et al., 1990).

1.5.1. Simple Linear Regression

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables. The term simple linear regression 'model' is used to denote the formal statistical formula, or equation, that describes the relationship we believe exists between the predictor and the outcome(Sun, 2011b).

The regression equation that describes a simple linear type regression relationship in a population is expressed as:

$$\boldsymbol{Yi = \alpha + \beta\, Xi + \varepsilon i}$$

Where α is the intercept (the value of Y when X = 0) and β is the slope of the relationship.
εi is referred as residual or error term and it is the departure of an actual (measured) Y from the estimated Y using the above regression equation ($\hat{Y}$). The sample regression relationship is expressed as
$\boldsymbol{Yi = a + b\, Xi + \varepsilon i}$,
where the terms are as defined above.

For example, the model. is a statistical way of describing how the value of the outcome (variable *Y),* changes across population groups formed by the values of the predictor variable *Xi* More formally it says that the mean value of the outcome for any value of the predictor variable is determined using a starting point ***α,*** when *Xi* has the value O and, for each unit increase *in Xi,* the outcome changes by ***b*** units. ***α*** is usually referred to as the constant or the intercept term whereas ***b*** is usually referred to as the regression coefficient. The ***εi****:* component is called the error and reflects the fact that the relationship between *Xi* and *Y* is not exact. We will assume that these errors are normally and independently distributed, with zero mean and variance. We estimate these errors by residuals; these are the difference between the observed (actual) value of the observation and the value predicted by the model(Dohoo et al., 2003).

1.5.2. Multiple Linear Regression

So far, we have seen the concept of simple linear regression where a single predictor variable X was used to model the response variable Y. In many applications, there is more than one factor that influences the response. Multiple regression models thus describe how a single response variable Y depends linearly on a number of predictor variables. We move from the simple linear regression model with one predictor to the multiple linear regression model with two or more predictors. That is, we use the adjective "simple" to denote that our model has only predictor, and we use the adjective "multiple" to indicate that our model has at least two predictors. In the multiple regression setting, because of the potentially large number of predictors, it is more efficient to use matrices to define the regression model and the subsequent analyses. This lesson considers some of the more important multiple regression formulas in matrix form(Sun, 2011a)

The equation for multiple linear regression relationship is expressed as

$Yi = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots\ldots\ldots + \beta_k X_k + \varepsilon_i$, where
Yi = the dependent variable

$X_1 \ldots. X_k$ = independent (explanatory) variables considered to have influence on the Y variable
$\beta_1 \ldots. \beta_k$ = Partial regression slopes corresponding to the respective Xi
β_i is defined as the rate of change in Y for a unit change in Xi, while the effects of the other independent variables remain constant. ε_i is the residual variance in Y after taking into consideration the effects of the Xi variables included in the model. The parallel for a multiple regression equation based on sample data is given as

$Yi = \alpha + b_1 X_1 + b_2 X_2 + b_3 X_3 \ldots\ldots\ldots + b_k X_k$

Source: (Yosef Tekle-Giorgis, 2017)

1.6. Linear Correlation

Correlation analysis are about relationships among variables. Variables are said to be related when a change in the magnitude of one variable is associated with a change in the magnitude of the other variable. The change in magnitude for quantitative variables is an increase or decrease in the value of the variable. If an increase in the value of one of the variable is accompanied by an increase in the other variable, the relationship is referred as a positive relationship (Figure1a). If on the other hand an increase in the value of one of the variable is accompanied by a decrease in the other variable, the relationship is negative (Figure1b). Two or more variables are said to be unrelated if one of the variable is not responsive as the other variable changes. (Figure1c). For a nominal scale data, the positive relationship exists when the change in the categories is along the same direction. If not the relationship is negative(Yosef Tekle-Giorgis, 2017).

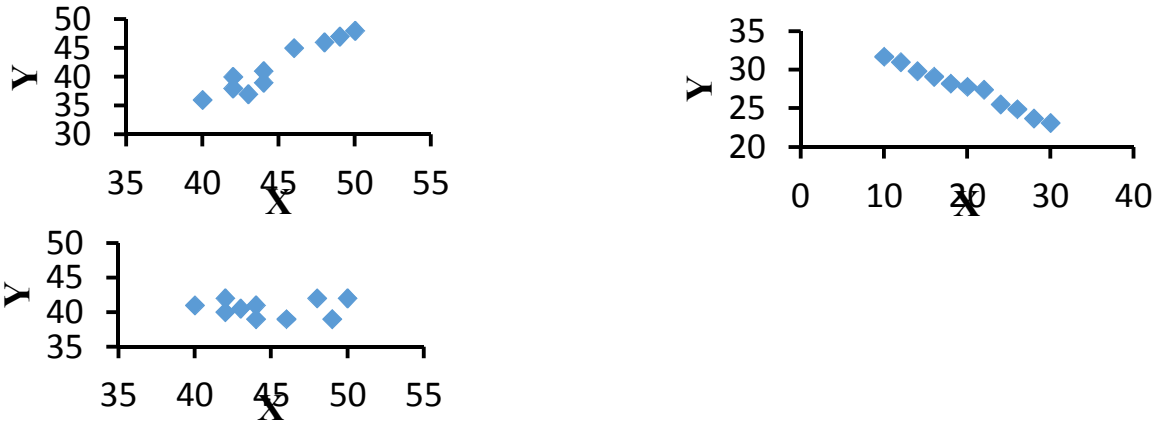

Figure 1. (Relationship b/n X and Y variables. a) Positive, b) negative, c) no relationship

2. ANALYSIS OF A CONTINUOUS DATA

2.1. Source of Data

The data was obtained from module called "Software Based Statistical Methods for Scientific Research Training Module" for PhD student of Haramaya University prepared by Yosef Tekle-Giorgis in 2017(Yosef Tekle-Giorgis, 2017). The data were slightly modified by I myself for getting some significant results for a better test of my understanding and fitting of the model at p-value=0.05 because most variables were non-significant.

2.2. Description of the Data

A data set that constitutes body weight and linear body measurements recorded for a random sample of 32 yearling Borana steers purchased from Yabello market. Body weight of animals is recorded for a variety of reasons but measuring the weight of large animals is quite a difficult task since the weighing scale is number some to carry from place to place. Especially when there is a need to take weight measurements of cattle from market or form households, the problem is severe as the balance is not easily portable. Besides, there is also a need to construct a chute to lead the animals to the weighing scale. On the other hand, body weight of animals (Kg) (cattle, small ruminants etc) can be estimated from linear body measurements like heart girth, wither height, crown height (cm) etc with a reasonable precision if the equation that relates the body weight of the animal with linear body measurements is developed. Table gives data on body weight and various linear body measurements of 32 randomly taken yearling Borana steers. Establish a multiple regression relationship between body weight and the linear measurements. Also select the linear measurements that are significantly related with body weight to be included in the regression model? (Yosef Tekle-Giorgis, 2017).

Table 1. Body weight (kg) and various linear body measurements recorded from 32 randomly taken yearling Borana steers.

ID	BWT	HG	WH	CH	Belly G
1	98	110	98.5	99	126.5
2	100	110	96.5	99.5	128.5
3	113	119	99.5	105	136
4	92	109	99.5	102	127
5	135	126	103	103.5	132
6	124	123	105.5	102.5	126
7	90	110	94.5	102	130
8	104	112	99	95.5	132
9	85.5	106	95	98.5	119.5

10	85	112	93	95	118.5
11	103	119	95	101	121
12	104	124	97	100	124
13	81	118	94	102	120
14	79	119	92	94	131
15	100	109	94	102	127
16	102	117	93	99	118
17	86	103	92	94	116
18	95	112	102	105	130
19	111	117	97.5	102.5	131
20	88	107	91.5	99	127.5
21	81	111	91	93	118
22	93	109	98	97	117
23	91	99.5	98.5	98	123
24	95	121	103	103	121
25	110	120	105	104	130
26	113.5	118.5	96	104	125
27	123	121	108	105	119
28	100.5	115	93.5	106	125
29	99	112	91.5	106	121
30	105.5	120	93.5	95.5	118.5
31	104.5	113.5	92.5	94.5	123
32	89	114	91	92	121

Source: (Yosef Tekle-Giorgis, 2017)

2.3. Types of Variables of the Data

Body Weight (BWT) into **dependent variable** (outcome variable).

Independent variables

HG	Heart Girth (cm)
WH	Wither Height (cm)
CH	Crown Height (cm)
Bell G	Belly Girth (cm)

2.4. Test for Analysis Data

2.4.1. Multiple linear regression

A multiple regression relationship gives a better prediction of Y than a simple linear regression relationship that takes each Xi separately. However not all Xi variables considered have significant influence on Y and it is important to select only those regressor variables that significantly influence Y and include them in the regression model equation. Popularly there are two step wise procedures employed to select the Xi variables for the model and these are Step up (Forward) selection and Step down (back ward illumination) procedures. In case of back ward illumination procedure, a regression is established with all the variables at hand and step by step those variables that are not significantly related with Y will be drooped. Whereas in case of forward selection Xi variables are added step by step one at a time by testing their significance. The steps for the backward illumination procedure is as follows: 1) Establish a regression relationship between Y and all Xi variables. Test for the significance of the relationship between Y and each Xi. This is done employing the t test procedure discussed earlier and testing each partial slope if different from 0 or not. 2) Drop the Xi variable that is most non-significantly related with Y and establish a regression relationship with the rest of Xi variables. 3) Again drop the Xi variable that is most non-significantly related with Y and reestablish the regression relationship with the rest. This procedure of dropping one Xi variable at a time will be continued until Xi variables that are significantly related with Y remain. Finally those that are significantly related with Y are considered to build the predictive regression equation(Zeger, 1991a).

2.4.2. Model Simplification

In our initial model "multipleintil" Belly G (Belly girth) and CH (crow height) are not significant. Thus, they can be dropped from the model. Thus, they can be dropped from the model.

```
multipleReduc<-lm(BWT~HG+WH)
> summary(multipleReduc)

Call:
lm(formula = BWT ~ HG + WH)

Residuals:
    Min      1Q  Median      3Q     Max
-18.990  -5.005   2.259   5.004  16.420

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -134.0936    37.6852  -3.558  0.00131 **
HG             0.9179     0.2795   3.284  0.00267 **
WH             1.3302     0.3777   3.522  0.00144 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.927 on 29 degrees of freedom
Multiple R-squared:  0.5716,
F-statistic: 19.35 on 2 and 29 DF,  p-value: 4.59e-06
```

```
Multiple_fit<-lm(BWT~HG+WH+`Belly G`+CH)
> summary(Multiple_fit)

Call:
lm(formula = BWT ~ HG + WH + `Belly G` + CH)

Residuals:
    Min      1Q  Median      3Q     Max
-18.659  -4.934   1.306   5.774  15.554

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -174.6567    49.1120  -3.556  0.00141 **
HG             0.8540     0.2851   2.995  0.00581 **
WH             1.0597     0.4326   2.449  0.02108 *
`Belly G`      0.2278     0.3307   0.689  0.49689
CH             0.4569     0.4848   0.942  0.35432
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now let's compare the two nested models (multipleintil and multipleReduc). In R we can compare nested models with the anova( )

```
anova(Multiple_fit,multipleReduc)
Analysis of Variance Table

Model 1: BWT ~ HG + WH + `Belly G` + CH
Model 2: BWT ~ HG + WH
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     27 2167.6
2     29 2310.9 -2   -143.33 0.8927 0.4213
```

Interpretation: the p-value = 0.4213 indicates that there is no significant difference between the two models. Thus, removing Belly G (Belly girth) and CH (crow height) from the model could not result in any information loss.

Model simplification using "*stepwise*" Alternatively, the "stepwise" function could be used to remove non-significant variables. You can perform stepwise selection (forward, backward, both) using the *stepAIC ( )* function from the 'MASS' package. *stepAIC ( )* performs stepwise model selection by exact AIC.

```
step<-stepAIC(multipleintil, direction = "both")
Start:  AIC=144.9
BWT ~ HG + WH + `Belly G` + CH

           Df Sum of Sq    RSS    AIC
- `Belly G`  1     38.08 2205.7 143.46
```

```
- CH        1    71.31 2238.9 143.94
<none>               2167.6 144.90
- WH        1   481.64 2649.2 149.32
- HG        1   720.22 2887.8 152.08

Step:  AIC=143.46
BWT ~ HG + WH + CH

          Df Sum of Sq    RSS    AIC
- CH        1   105.26 2310.9 142.95
<none>               2205.7 143.46
+ `Belly G`  1    38.08 2167.6 144.90
- WH        1   531.86 2737.6 148.37
- HG        1   732.46 2938.2 150.63

Step:  AIC=142.95
BWT ~ HG + WH

          Df Sum of Sq    RSS    AIC
<none>               2310.9 142.95
+ CH        1   105.26 2205.7 143.46
+ `Belly G`  1    72.02 2238.9 143.94
- HG        1   859.46 3170.4 151.07
- WH        1   988.64 3299.6 152.35
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
BWT ~ HG + WH + `Belly G` + CH

Final Model:
BWT ~ HG + WH


        Step Df  Deviance Resid. Df Resid. Dev      AIC
1                       27   2167.611 144.9006
2 - `Belly G`  1  38.07717        28   2205.689 143.4579
```

Initial Model:

BWT ~ HG + WH + CH + `Belly G` + BL

Final Model:
BWT ~ HG + WH

Interpretation: you can see the difference between the Initial and Final model, the Crown height, Belly girth has been dropped from the final model. The final model encompasses only HG (Heart girth) and WH (Wither height).

Let's develop a multiple regression model of -BHT- using HG (Heart girth) and WH (Wither height) as explanatory/predictor variable.

```
multifinal<-lm(BWT ~ HG +WH)
> summary(multifinal)

Call:
```

```
lm(formula = BWT ~ HG + WH)

Residuals:
    Min      1Q  Median      3Q     Max
-18.990  -5.005   2.259   5.004  16.420

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -134.0936    37.6852  -3.558  0.00131 **
HG             0.9179     0.2795   3.284  0.00267 **
WH             1.3302     0.3777   3.522  0.00144 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.927 on 29 degrees of freedom
Multiple R-squared:  0.5716,
F-statistic: 19.35 on 2 and 29 DF,  p-value: 4.59e-06
```

Interpretation:

When we inc1ude data on -Heart girth- and - Wither height -, the overall model is highly significant with an explained proportion of the total variation in the data (R2=SSM/SST) of 57.2%.

The coef. of HG(heart girth) (0.9179) indicate that when the Heart girth of the steers increases by one centimetre the body weight increases by 0.9179 Kg given that the other variables are held constant.

The coef. Of WH/ Wither height (1.3302) indicate that when the wither height of the steers increases by one centimetre the body weight increases by 1.3302 Kg given that the other variables are held constant. The positive value indicates the positive trend of the regression line.

Accordingly the final regression model used to estimate the body weight of the studied steers from linear body measurements is given as

Expected value (Body weight) (kg) = **-134.0936** + (**0.9179*** Heart girth (cm)) + (**1.3302*** Wither height (cm))
These expected values (i.e. regression line) can be calculated in R using the command

```
pred<-fitted(multifinal)
> pred
       1        2        3        4        5        6        7        8
 97.90685  95.24640 107.49854  98.31913 118.57990 119.15164  92.58596 100.40784
       9       10       11       12       13       14       15       16
 89.57930  92.42650 101.51253 108.76268  99.26437  97.52187  91.00290  97.01621
      17       18       19       20       21       22       23       24
 82.83481 104.39851 103.00221  85.84147  88.84812  96.32380  88.26847 113.99020
      25       26       27       28       29       30       31       32
115.73270 102.38379 120.64131  95.84544  90.43117 100.43514  93.13830  91.60194
```

These are the predicted values for the all observation that determine by the final model.

```
anova(multifinal)
Analysis of Variance Table

Response: BWT
          Df  Sum Sq Mean Sq F value    Pr(>F)
HG         1 2094.78 2094.78  26.287 1.781e-05 ***
WH         1  988.64  988.64  12.406  0.001438 **
Residuals 29 2310.94   79.69
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1
```

In R you can use coefficients () command to visualize the regression coefficients

```
coef(multifinal)
 (Intercept)          HG          WH
-134.0935837   0.9179408   1.3302228
confint(multifinal, level = 0.95)
                2.5 %     97.5 %
(Intercept) -211.1685040 -57.018663
HG             0.3462771   1.489605
WH             0.5578210   2.102625
```

2.4.3. Model diagnosis

Assessing the major assumptions

Assessing normality: The residuals of the model should also follow a normal distribution, for all values of the explanatory variables. Most commonly, the distribution of the residuals is evaluated in a histogram or in a Q-Q (quantile-quantile) plot.

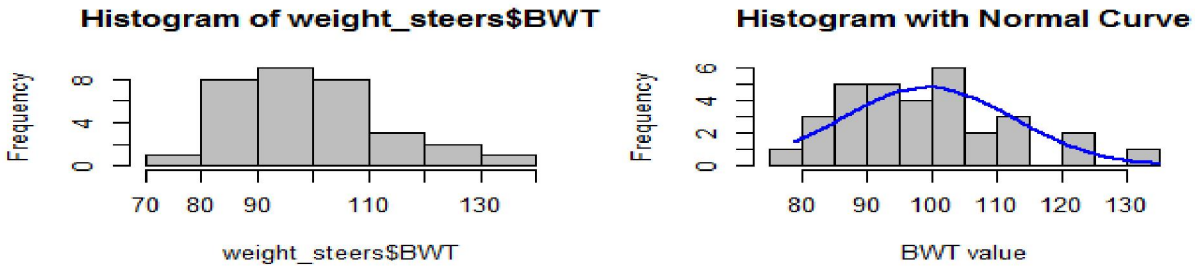

Figure 2. The left Histogram of BWT and the right Histogram of BWT with normal curve

The Q-Q plot displays the quantiles of the residuals versus the quantiles of the normal probability distribution.

```
qqnorm(res)
> qqline(res)
```

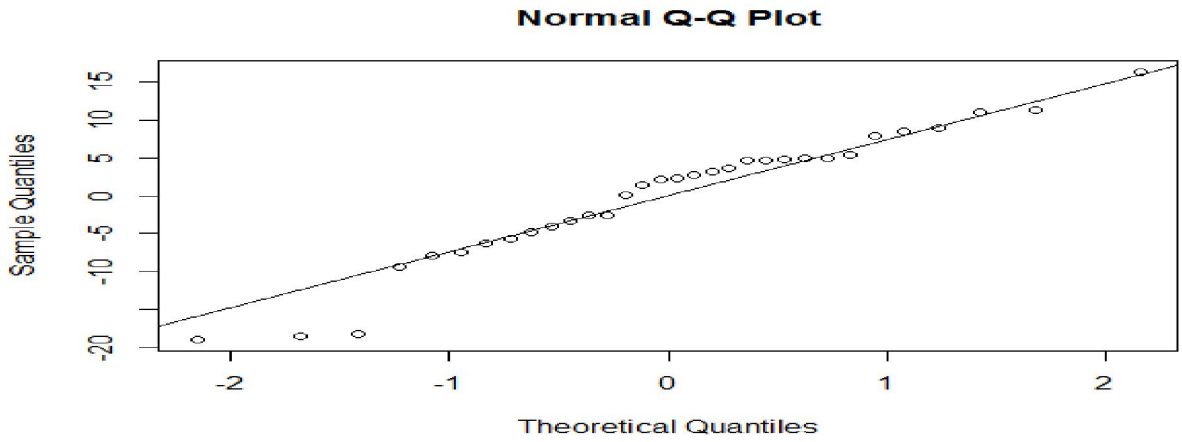

Figure 3. Q-Q plot (test of normality)

Interpretation: The resulting plot will be (approximately) a straight line above *45°* to the horizontal so the residuals of the model follow a normal distribution, for all values of the explanatory variables.
The official test for normality is the Shapiro-Wilk W test for normality.
shapiro.test(res)

Shapiro-Wilk normality test

data: res

W = 0.94949, p-value = 0.1394

Interpretation*:* as we can see from the Shapiro-Wilk normality test result, the residuals of this model follow a **normal distribution**. The Shapiro-Wilk's statistic, which for this example gives a value of W=0.94949 (small values are critical for Ho: normal distribution) and P> 0.1394 since the p-value is greater than 0.05, we can accept our null hypothesis which states that the data is normally distributed.

Assessing homoscedasticity: The variance of the outcome is constant at all levels of the explanatory variables and within all combinations of the explanatory variables. One can examine the homoscedasticity assumption, by plotting the standardised residuals against the predicted values.

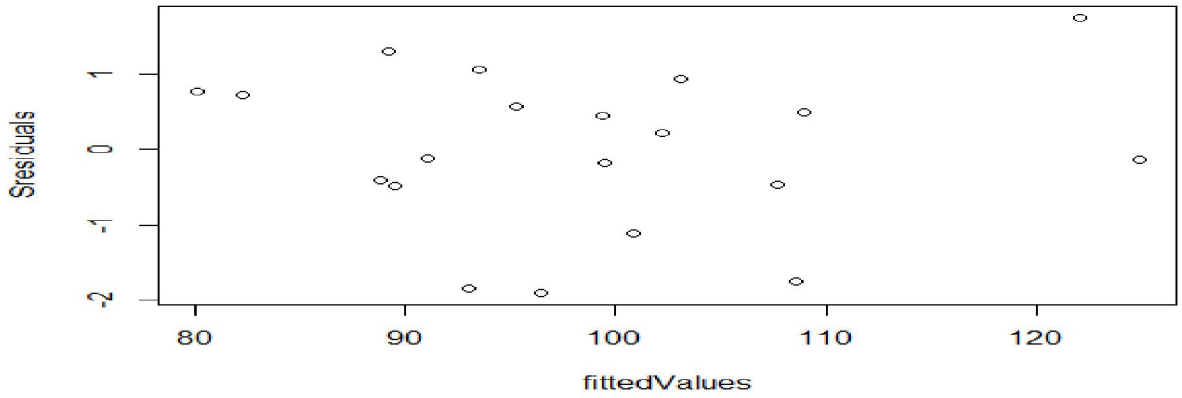

Figure 4. Plot of the standardised residuals against predicted values (homoscedasticity test)

Interpretation: the variance seems to remain relatively constant for all values:
the distribution creates a band around "0", without evident funnel shape. A formal test of equal variances (homoscedasticity) is the CookWeisberg test for heteroscedasticity (ie the null hypothesis is Ho: data is homoscedastic).
Homoscedasticity could be officially tested using ncvTest command in package "car"

ncvTest(multifinal)

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.175292 Df = 1 p = 0.1402429

Interpretation: The p-value is 0. 1402429 and it supports the null hypothesis (i.e. the variance of the outcome is constant at all levels of the explanatory variables and within all combinations of the explanatory variables). Thus, the assumption of homoscedasticity or constant variance is fulfilled.

Assessing linearity: regression models assume a linear relationship between the response and continuous explanatory variables. In our example dataset HG is a continuous explanatory variable. Thus, we expect a linear relationship between -BWT- and -HG-.

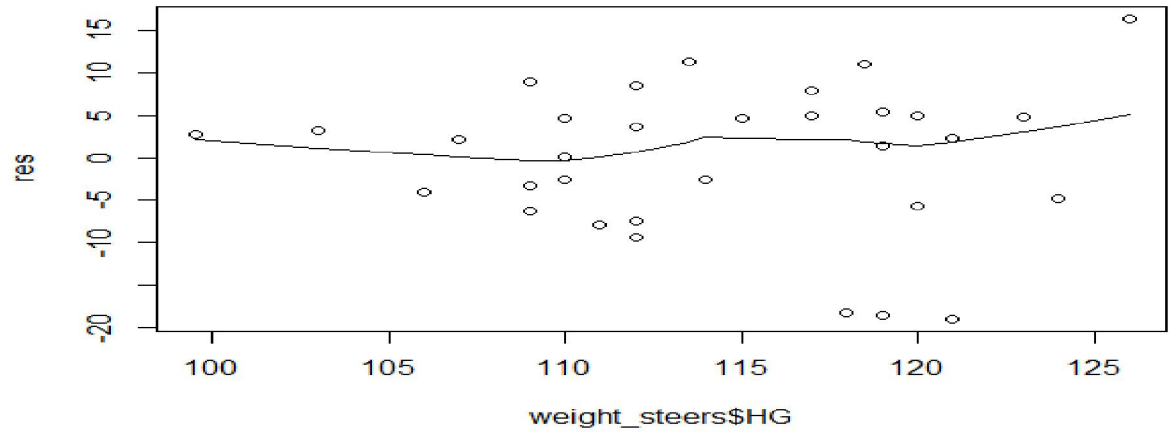

Figure 5. Plot of the residuals against HG (test of linearity)
Interpretation: in this example the relationship is not totally linear, but it is sufficiently close for our modelling purposes.

In our example dataset WH is a continuous explanatory variable. Thus, we expect a linear relationship between -BWT- and -WH-.

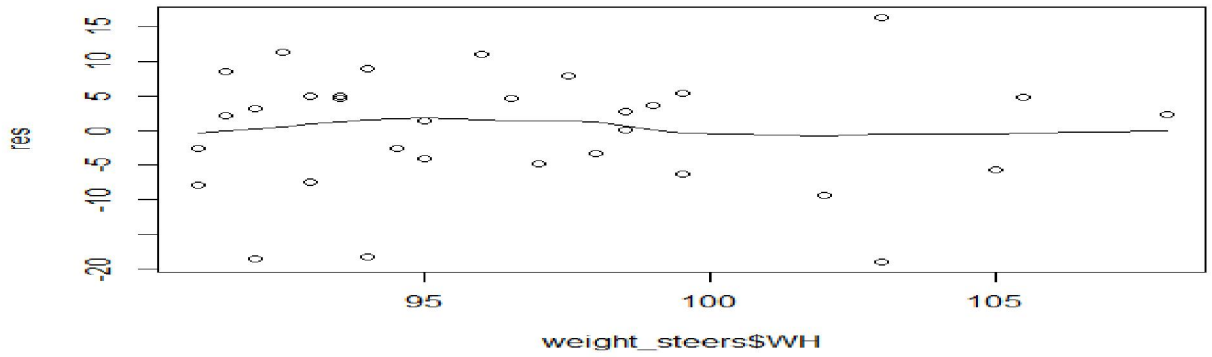

Figure 16. Plot of the residuals against WH (test of linearity)
Interpretation: in this example, the relationship is not totally linear, but it is sufficiently close for our modelling purposes.
plot(multifinal)

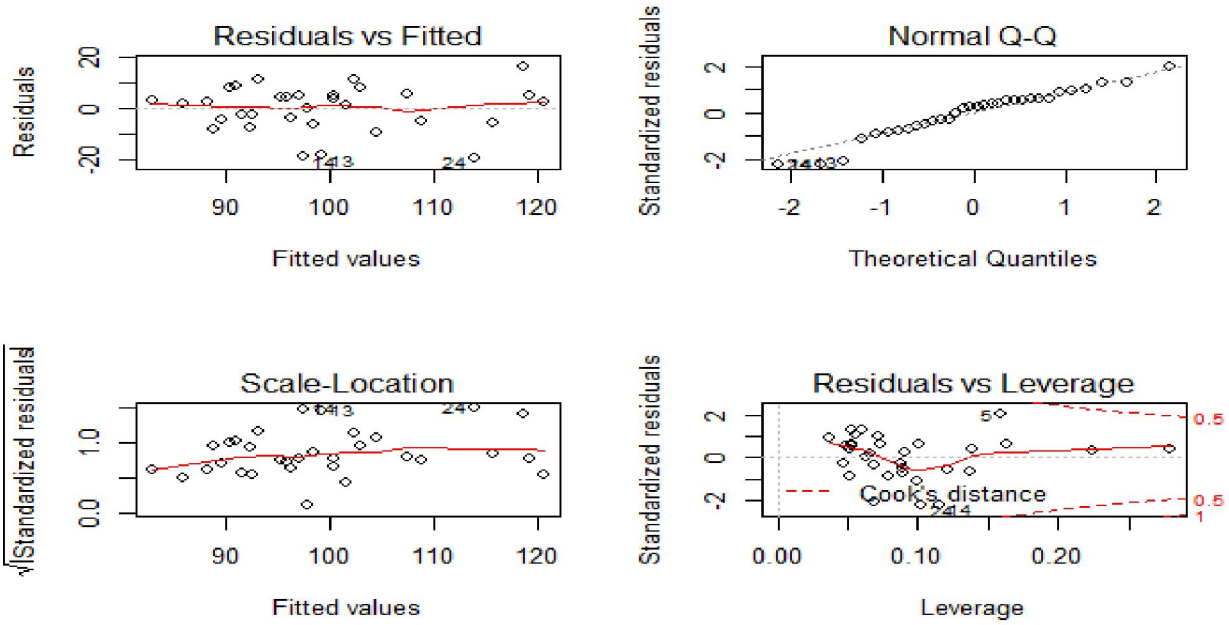

Figure 6. Plot graph to Assessing the major assumptions the at the same time
Plot of the standardised residuals against the predicted values (*homoscedasticity test*) on top left, the approximately equal-width band of points suggests our model likely meets the assumption of equal variances. Q-Q plot (test of normality) on the top of the right. Bottom right is leverage value.

2.4.4. Assessment of individual observations

Here we assess the fit of the model on an observation by observation basis.
Assessing outliers: Identify observations with large standardized residuals

outlierTest(multifinal)

No Studentized residuals with Bonferonni p < 0.05

Largest |rstudent|:

	rstudent	unadjusted p-value	Bonferonni p
24	**-2.426935**	**0.021914**	**0.70126**

Interpretation: as shown in the analysis there is no observation with standardised value greater than 3 and Bonferonni p-value less than 0.05, observation number 24 has standardised value= -2.427 which is still not greater than 3.
You can also graphically visualize outliers using a command "qqPlot"

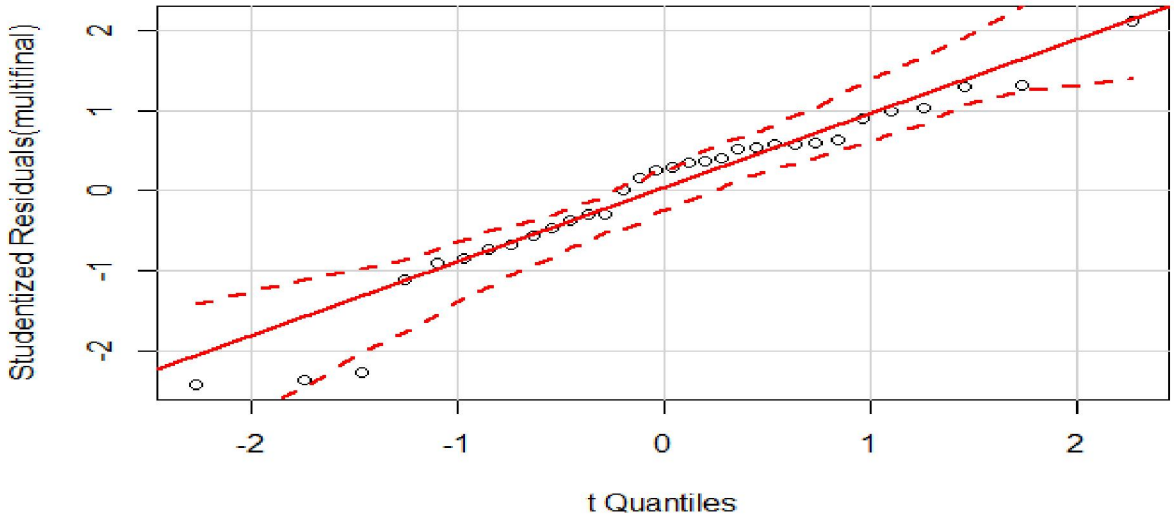

Figure 7: Plot of outliers
Interpretation: we can see that no observation appears to be significantly outlier

Assessing leverage values

A leverage effect occurs when one or a few excessively high X values are observed as continuous explanatory variable.

A common rule is to examine observations that have leverage values $>2(k+1)/n$, where k is the number of predictors in the model (or the number of regression parameters, excluding the intercept) and n is the number of observation. In our example k=2 and n =20. Thus, values > 2*(2+1)/32 = 0.1875 deserve attention.

lev[lev>.1875]

23	**27**
0.2796646	**0.2243369**

Observation with lev value greater than 0.1875 in lev dataset.
The leverage values could also be plotted

```
plot(hatvalues(multifinal))
> abline(h=0.1875)
```

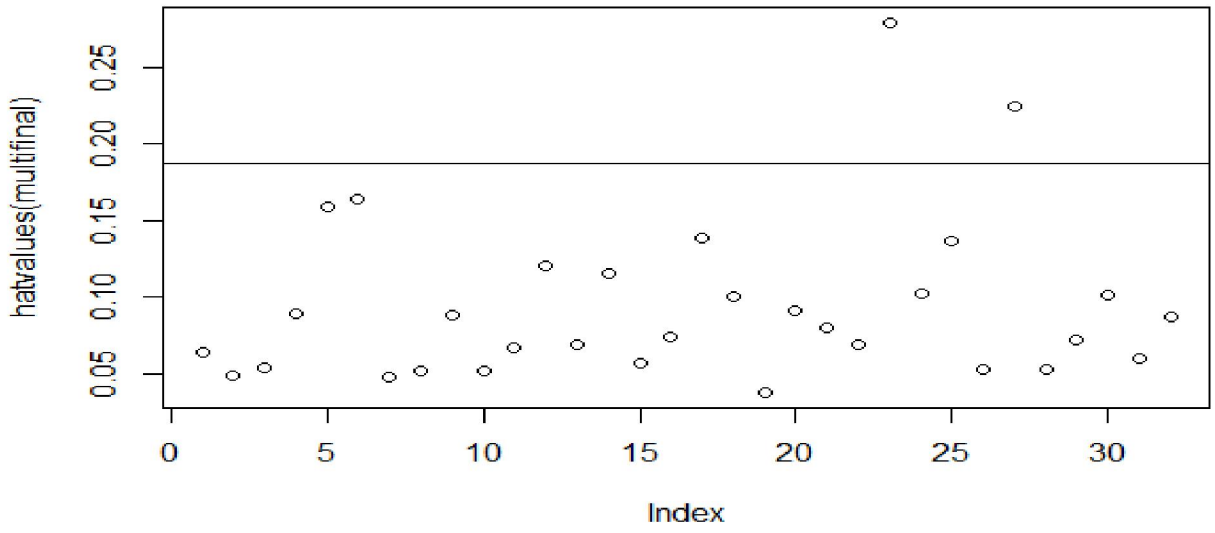

Figure 8: Plot of leverages (hatvalues)
Interpretation: the observations above the cut of line deserve attention.

Assessing influential observations: Cook's distance
We can estimate the Cook's value using the "cook.distance" command.

```
cook<-cooks.distance(multifinal)
```

```
> influencePlot(multifinal)
```

	StudRes	Hat	CookD
5	2.1237257	0.1590213	0.25358538
23	0.3550574	0.2796646	0.01682161
24	-2.4269352	0.1021062	0.19105068

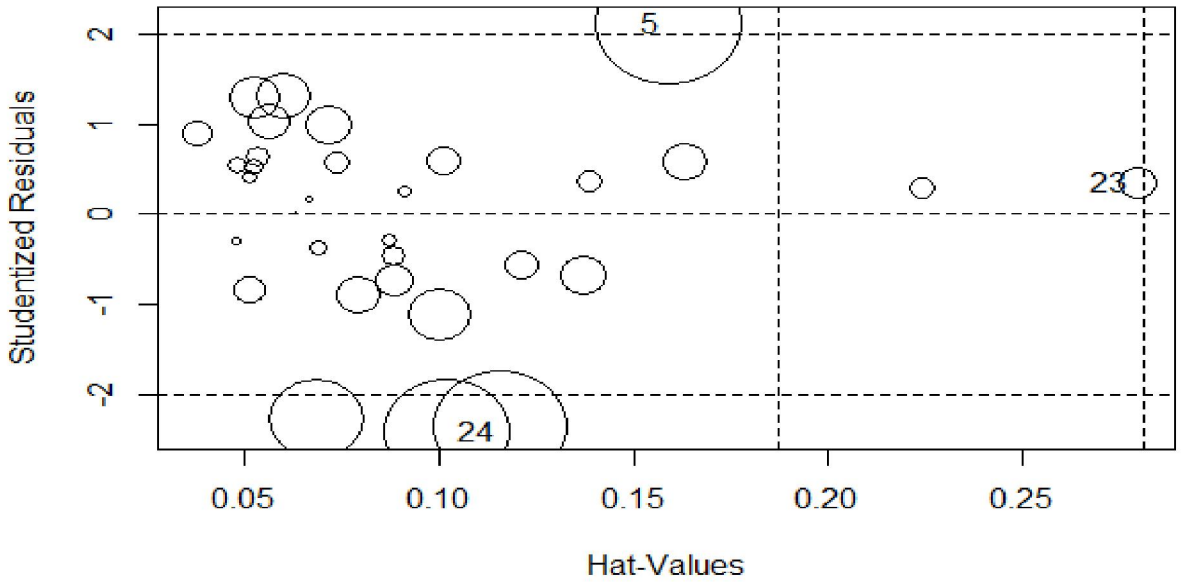

Figure 9. Plot of the residuals vs the hat values (influence plot)

Interpretation: the graph displays studentized residuals, hat-values and Cook's D on a single plot. The horizontal axis represents the hat-values; the vertical axis represents the studentized residuals; circles for each observation represent the relative size of the Cook's D. The radius is proportional to the square root of Cook's D, and thus the areas are proportional to the Cook's D. As indicated in the influential plot observations number 5, 23 and 24 deserves attention.

3. CONCLUSSION

The analyzed data using multiple linear regression in the "R" software was revealing that the model was fit. Multiple linear regression was also the best statistical analysis of continuous data analysis to determine the measures of the proportion of the variation in response variables "Y" that is explained by the variation in explanatory variable "X" using regression coefficients. A data set that constitutes body weight and linear body measurements recorded for a random sample of 32 yearling Borana steers purchased from Yabello market was been used to analysis by multiple regression. Because my data contains four explanatory variables multiple linear regressions was found to be the appropriate statistical analysis method. Frist model simplification was done on the data set to remove not significant variable then interpreted on simplify model. Then check the major assumption of multiple linear regression that includes: normality, homoscedasticity and linearity whether it is fulfill this assumption or not and the assessing individual observation include outlier test, assessments of leverage values and influential observations.

4. REFERENCE

[1]Ahlbom, A., 1993. Biostatistics for epidemiologists. CRC Press.

[2]Dohoo, I., Martin, W., Stryhn, H., 2003. Veterinary Epidemiologic Research. AVC Inc., Prince Edward Island, Canada. ISBN 0-919013-41-44.

[3]Hertz-Picciotto, I., 1999. What you should have learned about epidemiologic data analysis. Epidemiology 778-783.

[4]Ragland, D.R., 1992. Dichotomizing continuous outcome variables: dependence of the magnitude of association and statistical power on the cutpoint. Epidemiology 3, 434-440.

[5]Spatz, C., 2007. Basic statistics: Tales of distributions. Cengage Learning.

[6]Sun, B., 2011a. A study about the prediction of university library lending based on multiple regression analysis. Advances in Automation and Robotics 1, 525-532.

[7]Sun, B., 2011b. A study about the prediction of university library lending based on multiple regression analysis. Advances in Automation and Robotics 1, 525-532.

[8]Woolf, S.H., Battista, R.N., Anderson, G.M., Logan, A.G., Wang, E., on the Periodic, C.T.F., 1990. Assessing the clinical effectiveness of preventive maneuvers: Analytic principles and systematic methods in reviewing evidence and developing clinical practice recommendations A report by the Canadian task force on the periodic health examination. Journal of clinical epidemiology 43, 891-905.

[9]Yosef Tekle-Giorgis, 2017. **Software Based Statistical Methods for Scientific Research Training Module**.

[10]Zeger, S.L., 1991a. Statistical reasoning in epidemiology. American Journal of Epidemiology 134, 1062-1066.

[11]Zeger, S.L., 1991b. Statistical reasoning in epidemiology. American Journal of Epidemiology 134, 1062-1066.

4/25/2021