DATA ANALYTICS & VISUALIZATION
(UCS-633)
PRACTICAL FILE
ON
FERTILITY PREDICTION
BY
LINEAR REGRESSION

COMPUTER SCIENCE AND ENGINEERING
DEPARTMENT
THAPAR UNIVERSITY
PATIALA

SUBMITTED BY: Team no. 40

Rohan Goel (101403157)
Rohit Semwal (101403158)
Sachet Dhar (101403159)
Sahil Mahajan (101403160)

Abstract

The fertility potential analysis represents a research topic of great interest, and could help us to understand all the factors that difficult to have high fertility rates, which is quite relevant to be able to propose solutions and obtain an increase in the fertility levels especially for men. In this project,we implemented the linear regression on the dataset given, so we can conclude that our method proposal represents a tool to support decision making for patient analysis with fertility problems. To detect the relation between attributes and the outputs various methods were applied like chi-square test and correlation between the attributes and the outputs.

The regression line was plotted which shows deviation from the actual points. To check how accurate the regression line is, the root mean square error was obtained for each attribute of the given sample.

Table of Contents

Introduction

The World Health Organization (WHO) defines infertility as a disease. It’s a reproductive system disease that manifest as the inability of obtain clinical pregnancy after 12 months of having unprotected sexual relationships. Infertility causes several effects in different types of personal health: physical, mental, emotional, psychological, social and even religious, in the couples that suffer from it. It’s one of the most important causes of depression, and its social, psychological and cultural consequences have been catalogued in six levels of severity, ranging from guilt, fear, depression to violent death or suicide. This behavior is also confirmed, and clearly infertility can affect negatively generating frustration, and personality weakness.

The fertility potential analysis represents a research topic of great interest, and could help us to understand all the factors that leads to difficulty in having high fertility rates, which is quite relevant to be able to propose solutions and obtain an increase in the fertility levels especially for men. Some authors, say that during the last three decades, several reports have suggested that the semen quality in regular men has decreased, and talks about a trend in decrease of sperm count and seminal fluid volume in the last fifty years.

It has been mentioned that fertility rates have decreased drastically in the last two decades, especially in men, due to environmental issues and lifestyle that can affect the quality of semen. Several artificial intelligence techniques have become an emergent technology for decision support systems in medicine to patient identification with fertility problems.

A research was conducted to study semen volume, sperm concentration, progressive motility, vigor and percentage of normal forms and multiple anomalies. The semen volume didn’t decrease, but an important decrease in total sperm count was found (443,2 million in 1976 to 300.2 million in 2009), also in motility (64% in 1976 to 49% in 2009) and vigor (88% to 80%).

Then it was considered that semen analysis is standard for routine diagnosis of infertile couples, studies through the sperm count, and it´s strongly related to male infertility, and they also express the importance of sperm concentration in male infertility, since it is a relevant factor in diagnosis of this disease.

Based on previous studies in existing literature, we can infer that the proposed model would become an excellent tool to identify patterns and predict behaviors in fertility analysis, and also suggest its implementation in other areas of healthcare.

Methodology

Study Population

We use the data collected and shared in 2013 by the Department of Biotechnology of University of Alicante. 100 young healthy volunteers among students who were between 18 and 36 years old provided. A semen sample was taken for analysis as well as their socio-demographic data, environmental factors, health status, and life habits. Students with previous known reproductive alterations were excluded from the statistical analysis.

Database Description

Attributes	Description
Season in which the analysis was performed	• Winter • Spring • Summer • Fall. (-1, -0.33, 0.33, 1)
Age at the time of analysis	• 18-36 (0, 1)
Childish diseases (i.e. chicken pox, measles, mumps, polio)	• Yes • No (0, 1)
Accident or serious trauma	• Yes • No (0, 1)
Surgical intervention	• Yes • No (0, 1)
High fevers in the last year	• Less than three months • More than three months ago • No (-1,0,1)
Frequency of alcohol consumption	• Several times a day • Every day • Several times a week, • Once a week, • Hardly ever or never (0, 1)
Smoking habit	• Never • Occasional • Daily (-1,0,1)
Number of hours spent sitting per day ene-16	• 0 • 1
Output: Diagnosis	• Normal (N) • Altered (O)

Table No. 1 Database Description

Linear Regression

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or moreexplanatory variables (or independent variables) denoted X. The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression.

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimatedfrom the data. Such models are called linear models. Most commonly, the conditional mean of y given the value of X is assumed to be an affine function of X; less commonly, the median or some other quantile of the conditional distribution of y given X is expressed as a linear function of X. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of given X, rather than on the joint probability distribution of y and X, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

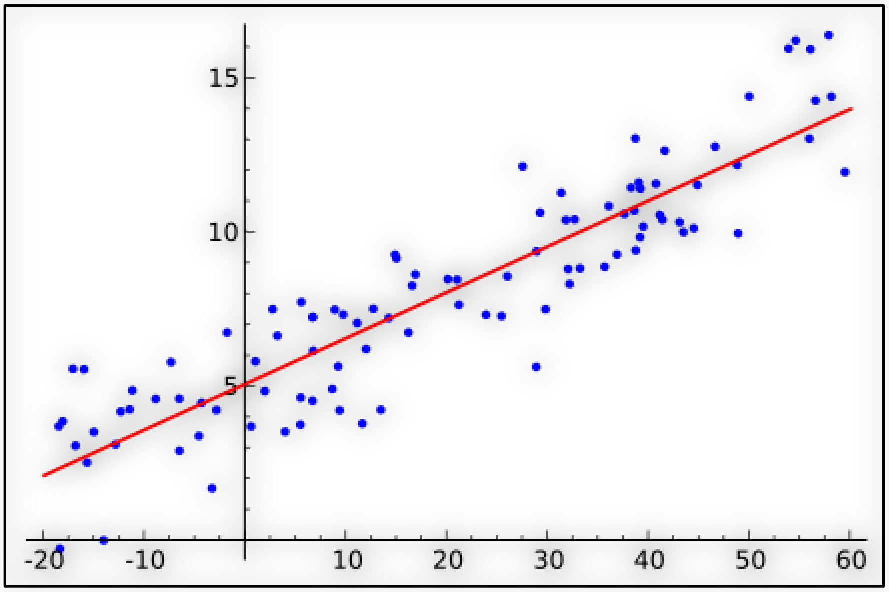


Fig No. 1 Linear Regression

Experimentation

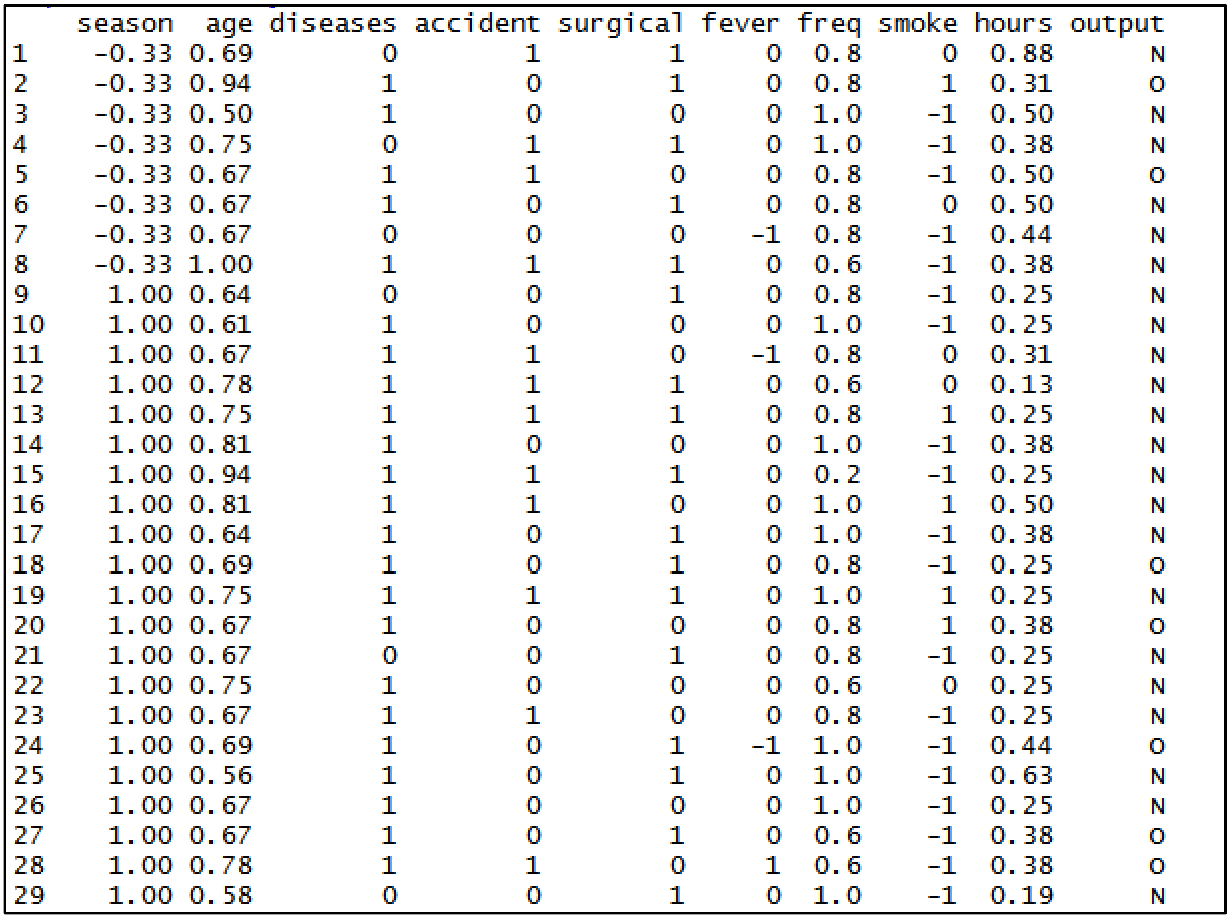
For this project, the “Fertility Dataset” was used, which can be found in the Machine Learning Repository UCI. The purpose of this research is to predict the fertility according to linear regression. Next, we describe the methodology used.

First, dataset was downloaded and prepared to start training and evaluation of the method created. Then the testing data was selected from the given data and the training data was selected by removing the testing data from the data. The various steps taken to obtain the result are as follows:

Data Modification

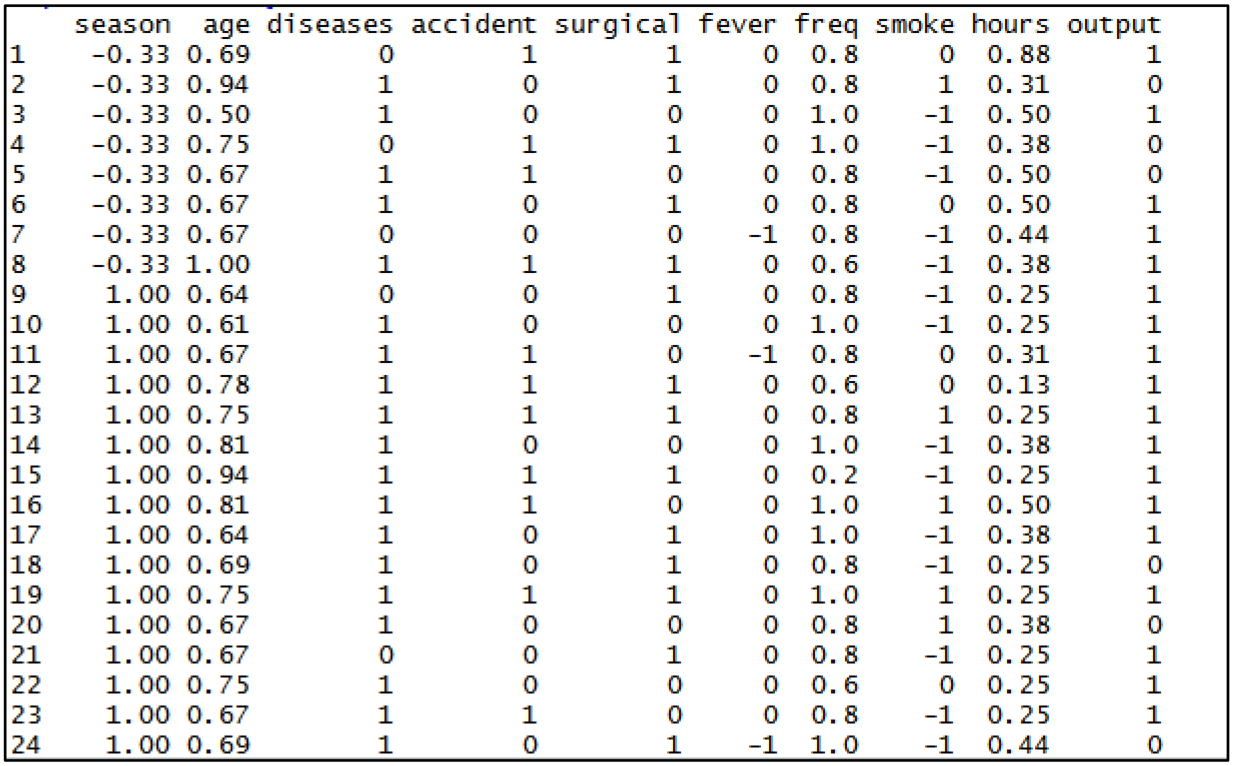
In the process of data preparation, it was necessary to run a balancing process for the data, because the output variable data was out of scale, and this situation has a negative impact on the proposed model, since the learning process must be made under one situation only.

```
## converting nominal to ordinal data


for (i in 1:n)
  if (fertility[i,10]=="N")
    fertility[i,10]=1.0
if(fertility[i,10]=="O")
  fertility[i,10]=0
print(fertility)
```



```
   season  age diseases accident surgical fever freq smoke hours output
1   -0.33 0.69        0        1        1     0  0.8     0  0.88      N
2   -0.33 0.94        1        0        1     0  0.8     1  0.31      O
3   -0.33 0.50        1        0        0     0  1.0    -1  0.50      N
4   -0.33 0.75        0        1        1     0  1.0    -1  0.38      N
5   -0.33 0.67        1        1        0     0  0.8    -1  0.50      O
6   -0.33 0.67        1        0        1     0  0.8     0  0.50      N
7   -0.33 0.67        0        0        0    -1  0.8    -1  0.44      N
8   -0.33 1.00        1        1        1     0  0.6    -1  0.38      N
9    1.00 0.64        0        0        1     0  0.8    -1  0.25      N
10   1.00 0.61        1        0        0     0  1.0    -1  0.25      N
11   1.00 0.67        1        1        0    -1  0.8     0  0.31      N
12   1.00 0.78        1        1        1     0  0.6     0  0.13      N
13   1.00 0.75        1        1        1     0  0.8     1  0.25      N
14   1.00 0.81        1        0        0     0  1.0    -1  0.38      N
15   1.00 0.94        1        1        1     0  0.2    -1  0.25      N
16   1.00 0.81        1        1        0     0  1.0     1  0.50      N
17   1.00 0.64        1        0        1     0  1.0    -1  0.38      N
18   1.00 0.69        1        0        1     0  0.8    -1  0.25      O
19   1.00 0.75        1        1        1     0  1.0     1  0.25      N
20   1.00 0.67        1        0        0     0  0.8     1  0.38      O
21   1.00 0.67        0        0        1     0  0.8    -1  0.25      N
22   1.00 0.75        1        0        0     0  0.6     0  0.25      N
23   1.00 0.67        1        1        0     0  0.8    -1  0.25      N
24   1.00 0.69        1        0        1    -1  1.0    -1  0.44      O
25   1.00 0.56        1        0        1     0  1.0    -1  0.63      N
26   1.00 0.67        1        0        0     0  1.0    -1  0.25      N
27   1.00 0.67        1        0        1     0  0.6    -1  0.38      O
28   1.00 0.78        1        1        0     1  0.6    -1  0.38      O
29   1.00 0.58        0        0        1     0  1.0    -1  0.19      N
```

Fig no. 2 Original Dataset

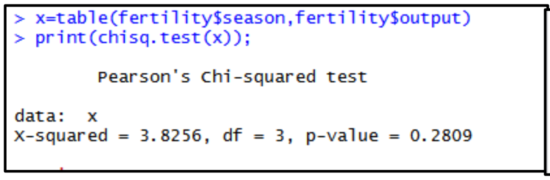


```
    season  age diseases accident surgical fever freq smoke hours output
1    -0.33 0.69        0        1        1     0  0.8     0  0.88      1
2    -0.33 0.94        1        0        1     0  0.8     1  0.31      0
3    -0.33 0.50        1        0        0     0  1.0    -1  0.50      1
4    -0.33 0.75        0        1        1     0  1.0    -1  0.38      0
5    -0.33 0.67        1        1        0     0  0.8    -1  0.50      0
6    -0.33 0.67        1        0        1     0  0.8     0  0.50      1
7    -0.33 0.67        0        0        0    -1  0.8    -1  0.44      1
8    -0.33 1.00        1        1        1     0  0.6    -1  0.38      1
9     1.00 0.64        0        0        1     0  0.8    -1  0.25      1
10    1.00 0.61        1        0        0     0  1.0    -1  0.25      1
11    1.00 0.67        1        1        0    -1  0.8     0  0.31      1
12    1.00 0.78        1        1        1     0  0.6     0  0.13      1
13    1.00 0.75        1        1        1     0  0.8     1  0.25      1
14    1.00 0.81        1        0        0     0  1.0    -1  0.38      1
15    1.00 0.94        1        1        1     0  0.2    -1  0.25      1
16    1.00 0.81        1        1        0     0  1.0     1  0.50      1
17    1.00 0.64        1        0        1     0  1.0    -1  0.38      1
18    1.00 0.69        1        0        1     0  0.8    -1  0.25      0
19    1.00 0.75        1        1        1     0  1.0     1  0.25      1
20    1.00 0.67        1        0        0     0  0.8     1  0.38      0
21    1.00 0.67        0        0        1     0  0.8    -1  0.25      1
22    1.00 0.75        1        0        0     0  0.6     0  0.25      1
23    1.00 0.67        1        1        0     0  0.8    -1  0.25      1
24    1.00 0.69        1        0        1    -1  1.0    -1  0.44      0
```

Fig no. 3 Modified Dataset

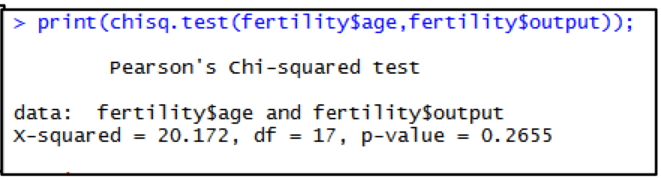
Chi-Square Test

The chi-square test was performed on each attribute to detect if there is any interdependence between any attribute and the result.



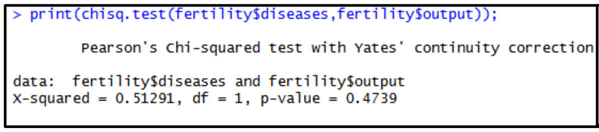
```
> x=table(fertility$season,fertility$output)
> print(chisq.test(x));

        Pearson's Chi-squared test

data:  x
X-squared = 3.8256, df = 3, p-value = 0.2809
```

Fig No. 4



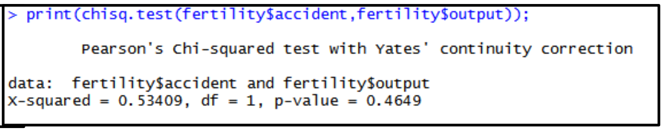
```
> print(chisq.test(fertility$age,fertility$output));

        Pearson's Chi-squared test

data:  fertility$age and fertility$output
X-squared = 20.172, df = 17, p-value = 0.2655
```

Fig No. 5



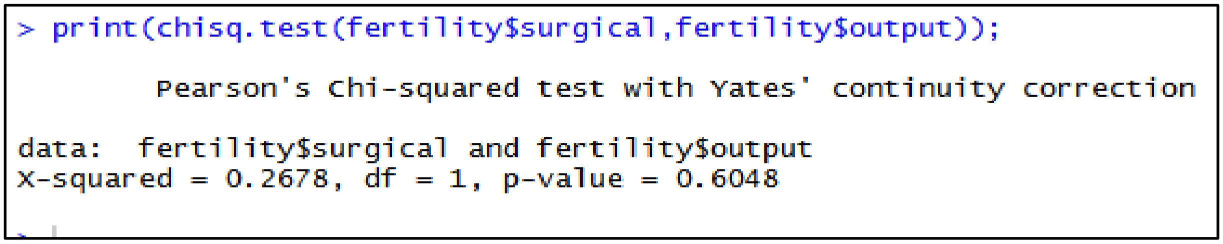
```
> print(chisq.test(fertility$diseases,fertility$output));

        Pearson's Chi-squared test with Yates' continuity correction

data:  fertility$diseases and fertility$output
X-squared = 0.51291, df = 1, p-value = 0.4739
```

Fig No. 6



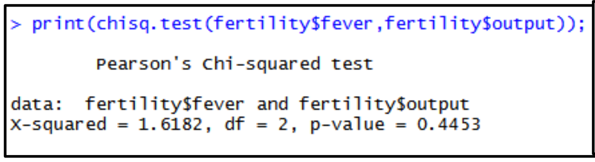
```
> print(chisq.test(fertility$accident,fertility$output));

        Pearson's Chi-squared test with Yates' continuity correction

data:  fertility$accident and fertility$output
X-squared = 0.53409, df = 1, p-value = 0.4649
```

Fig No. 7



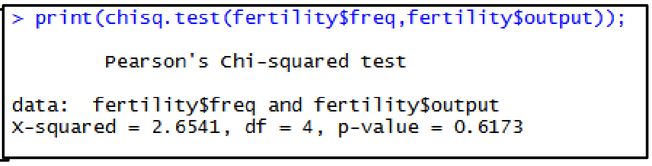
```
> print(chisq.test(fertility$surgical,fertility$output));

        Pearson's Chi-squared test with Yates' continuity correction

data:  fertility$surgical and fertility$output
X-squared = 0.2678, df = 1, p-value = 0.6048
```

Fig No. 8



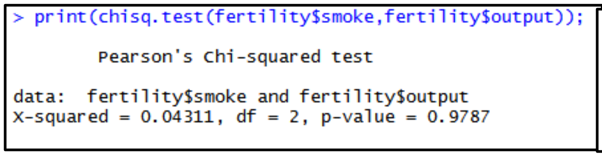
```
> print(chisq.test(fertility$fever,fertility$output));

        Pearson's Chi-squared test

data:  fertility$fever and fertility$output
X-squared = 1.6182, df = 2, p-value = 0.4453
```

Fig No. 9



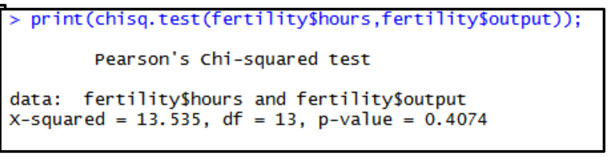
```
> print(chisq.test(fertility$freq,fertility$output));

        Pearson's Chi-squared test

data:  fertility$freq and fertility$output
X-squared = 2.6541, df = 4, p-value = 0.6173
```

Fig No. 10



```
> print(chisq.test(fertility$smoke,fertility$output));

        Pearson's Chi-squared test

data:  fertility$smoke and fertility$output
X-squared = 0.04311, df = 2, p-value = 0.9787
```

Fig No. 11



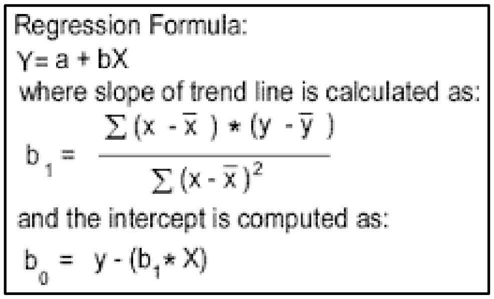
```
> print(chisq.test(fertility$hours,fertility$output));

        Pearson's Chi-squared test

data:  fertility$hours and fertility$output
X-squared = 13.535, df = 13, p-value = 0.4074
```

Fig No. 12

From the results,we have found that there does not exist any clear interdependence b/w the attributes and the output.

Plotting the Regression Lines

In a cause and effect relationship, the independent variable is the cause, and the dependent variable is the effect. Least squares linear regression is a method for predicting the value of a dependent variable *Y*, based on the value of an independent variable *X*. Using formula for linear regression as:



Regression Formula:

Y= a + bX

where slope of trend line is calculated as:

$$b_1 = \frac{\sum (x - \bar{x}) * (y - \bar{y})}{\sum (x - \bar{x})^2}$$

and the intercept is computed as:

$$b_0 = y - (b_1 * X)$$

Calculation of linear regression lines for every attributes with output.

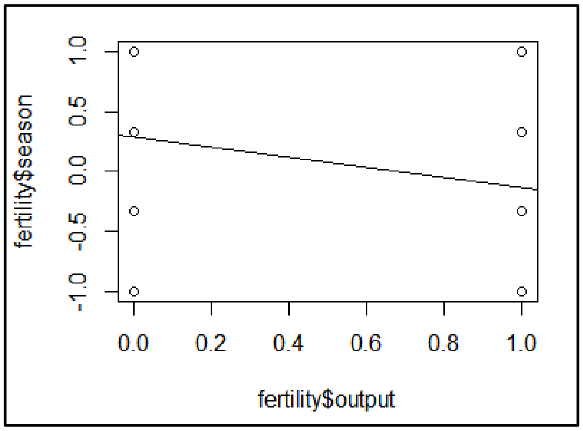


Fig No. 13

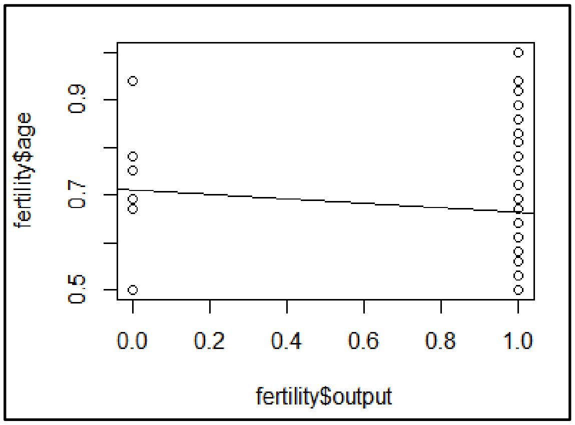


Fig No. 14

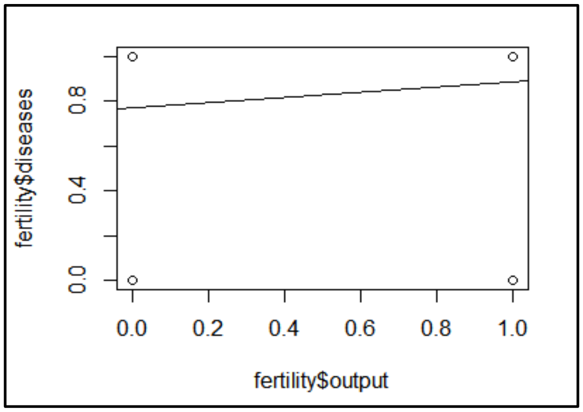


Fig No. 15

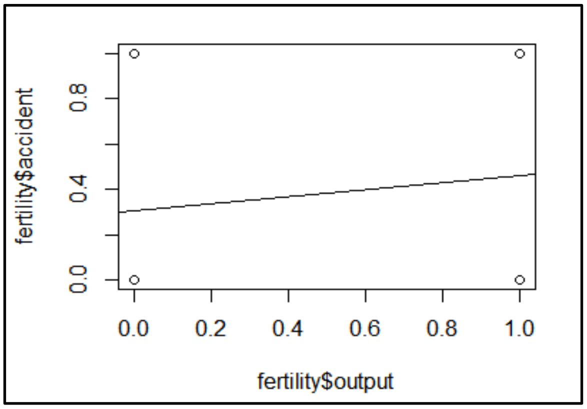


Fig No. 16

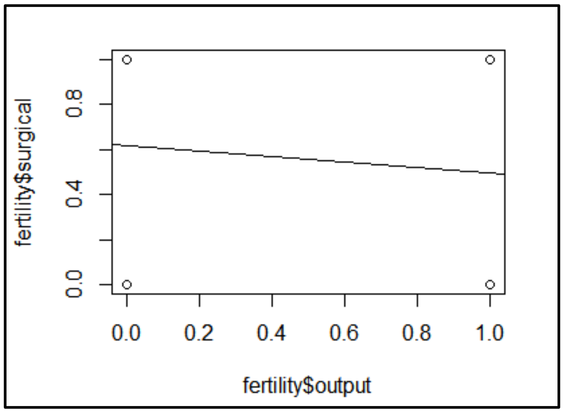


Fig No. 17

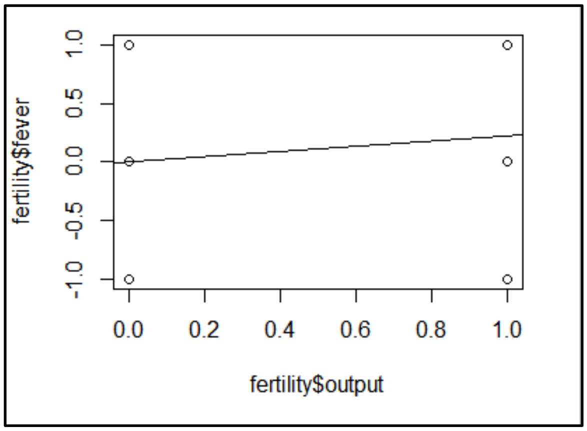


Fig No. 18

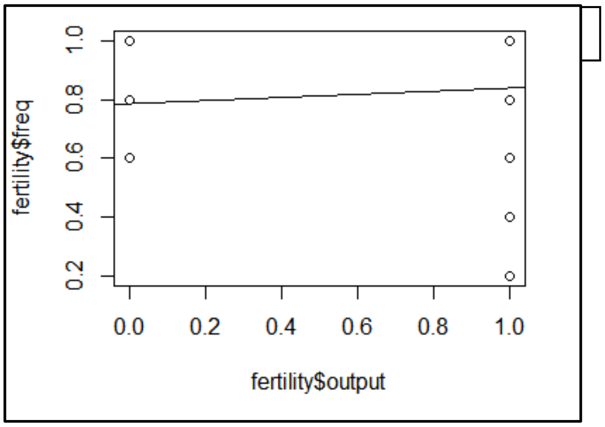


Fig No. 19

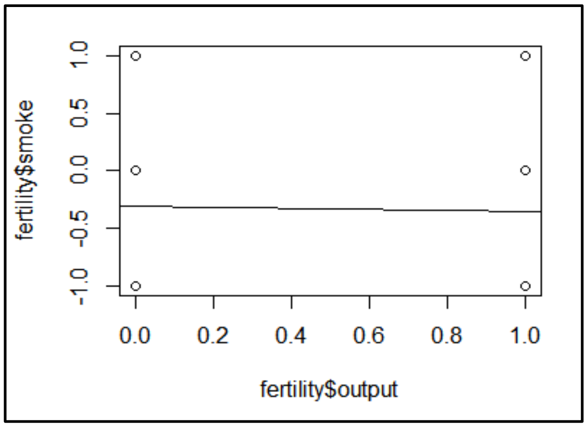


Fig No. 20

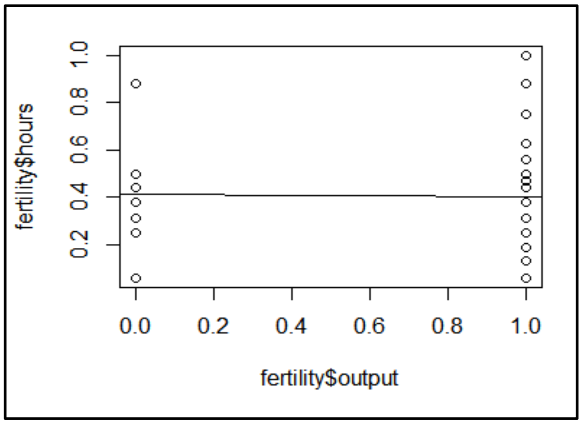


Fig No. 21

The above diagram shows linear regression plot of all 10 attributes with the output attributes.

Correlation

The need for correlation checking was due to:

- We might care more about the overall shape of expression profiles rather than the actual magnitudes
- That is, we might want to consider genes similar when they are "up" and "down" together

Then the correlation between the attributes and the output was checked. The correlation for any attribute does not show gives any light to any kind of relation between the output and attributes.

Here correlation coefficient (r) is given by:

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i}^{n} y_i$$

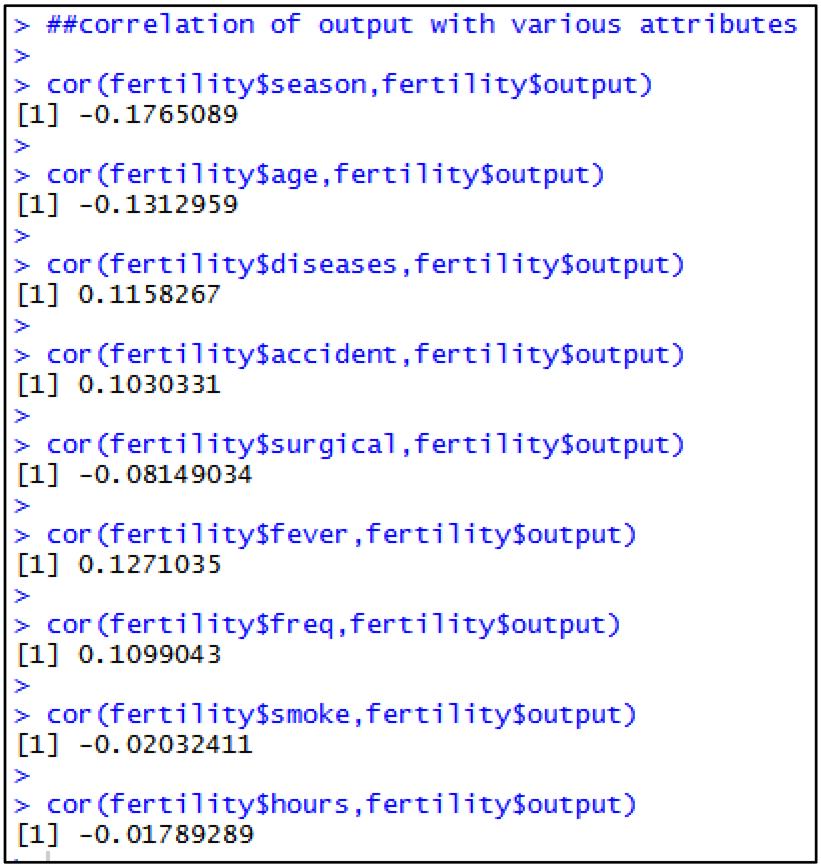
(Mean of x)

(Mean of y)

The **coefficient of correlation (r)** is a numerical measure of the strength of the linear relationship between 2 variables.

Values of *r* are always between -1 & 1; i.e., between 0 and 1 in absolute value.

r = 0 means no correlation; r = +-1 means perfect correlation; both rare.



```
> ##correlation of output with various attributes
>
> cor(fertility$season,fertility$output)
[1] -0.1765089
>
> cor(fertility$age,fertility$output)
[1] -0.1312959
>
> cor(fertility$diseases,fertility$output)
[1] 0.1158267
>
> cor(fertility$accident,fertility$output)
[1] 0.1030331
>
> cor(fertility$surgical,fertility$output)
[1] -0.08149034
>
> cor(fertility$fever,fertility$output)
[1] 0.1271035
>
> cor(fertility$freq,fertility$output)
[1] 0.1099043
>
> cor(fertility$smoke,fertility$output)
[1] -0.02032411
>
> cor(fertility$hours,fertility$output)
[1] -0.01789289
```

Fig No. 22

Since it is very clear from the above calculations that correlation coefficients for any of the two attributes lies in the range [-1,1]. So we can safely say that no two attributes are strongly positively or negatively dependent.

Root Mean Square Error(RMSD)

The root mean square was then found to check the difference……………….

The RMSD represents the sample standard deviation of the differences between predicted values and observed values.

The RMSE serves to aggregate the magnitudes of the errors in predictions into a single measure of predictive power.

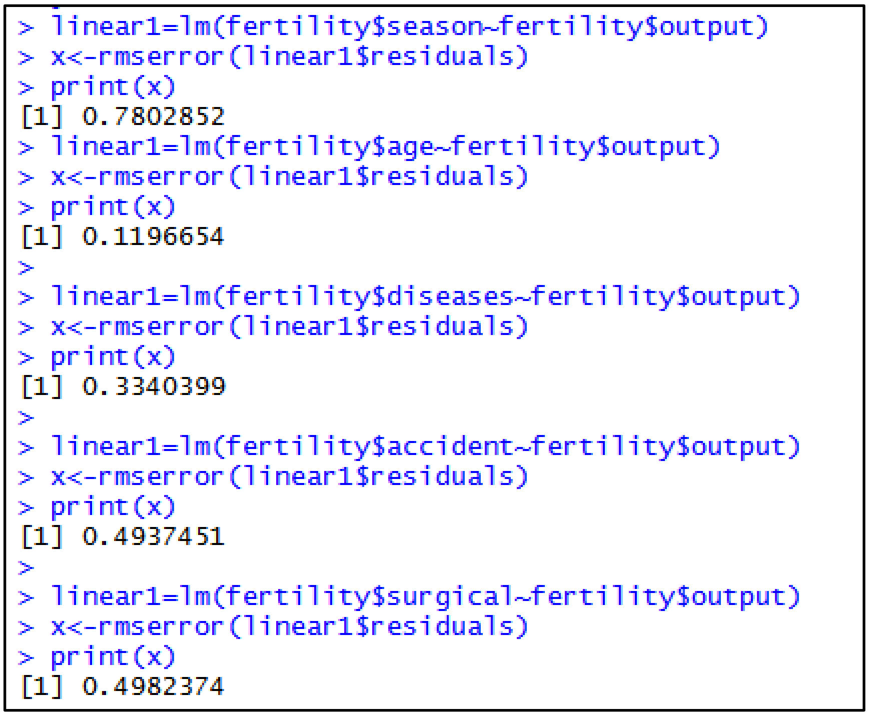
RMSE is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent.

Each RMSE can be interpreted as the average prediction error within the same scale (unit).

FORMULA :

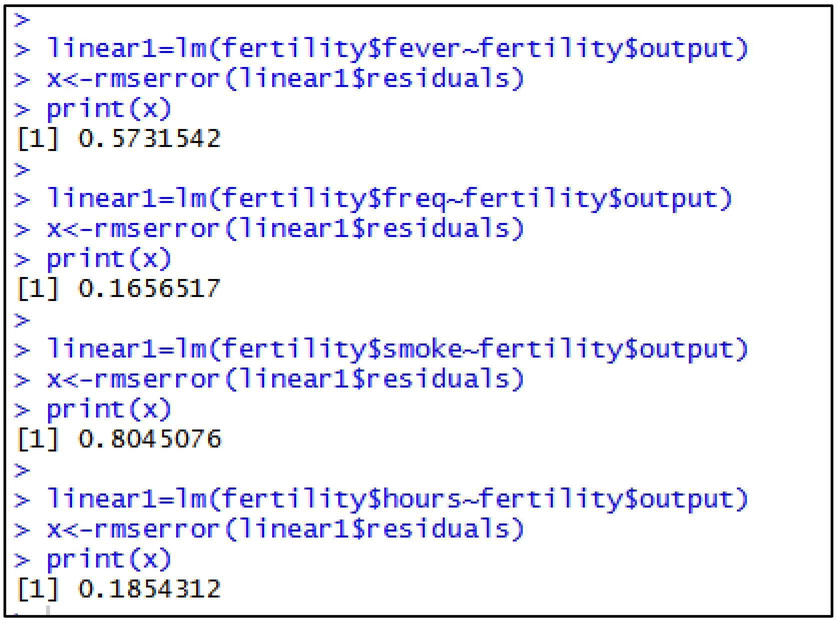
The RMSE is the square root of the average value of the square of the residual (actual - predicted)

$$\boxed{\text{Root mean squared error (RMSE}|\text{RMSD)} = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N}}}$$



```
> linear1=lm(fertility$season~fertility$output)
> x<-rmserror(linear1$residuals)
> print(x)
[1] 0.7802852
> linear1=lm(fertility$age~fertility$output)
> x<-rmserror(linear1$residuals)
> print(x)
[1] 0.1196654
>
> linear1=lm(fertility$diseases~fertility$output)
> x<-rmserror(linear1$residuals)
> print(x)
[1] 0.3340399
>
> linear1=lm(fertility$accident~fertility$output)
> x<-rmserror(linear1$residuals)
> print(x)
[1] 0.4937451
>
> linear1=lm(fertility$surgical~fertility$output)
> x<-rmserror(linear1$residuals)
> print(x)
[1] 0.4982374
```

Fig No. 23



```
>
> linear1=lm(fertility$fever~fertility$output)
> x<-rmserror(linear1$residuals)
> print(x)
[1] 0.5731542
>
> linear1=lm(fertility$freq~fertility$output)
> x<-rmserror(linear1$residuals)
> print(x)
[1] 0.1656517
>
> linear1=lm(fertility$smoke~fertility$output)
> x<-rmserror(linear1$residuals)
> print(x)
[1] 0.8045076
>
> linear1=lm(fertility$hours~fertility$output)
> x<-rmserror(linear1$residuals)
> print(x)
[1] 0.1854312
```

Fig No. 24