International Journal of Statistics and Applications 2016, 6(6): 386-390
DOI: 10.5923/j.statistics.20160606.07

A Chi-Squared approach to Obtaining Missing Values on Egg Production

Ogoke U. P.[1,*], Nduka E. C.[1], Soyinka A. T.[2]

[1]Department of Mathematics and Statistics, University of Port Harcourt, Nigeria
[2]Department of Research and Training, Federal Neuro-Psychiatric Hospital Aro, Nigeria

Abstract In this paper, we estimate the missing values in the outcome of an experiment that involved four different types of poultry birds fed over a period of eighteen months in four feed weight replicates per month. The outcome which is the number of eggs produced monthly by the birds is stochastically dependent on two inputs (feed weight and months). The study assumed a cubic polynomial interaction that follows a binomial series expansion between the inputs. A regression equation was used to link the outcome and the input interaction. The regression coefficients of the link function was determined by the method of ordinary least square via multiple imputation technique while replacing the missing values with the available outcomes in turn. So for each imputation, there is a corresponding likelihood value obtained as estimate of the missing data from the regression equation. The probability of the chi-square goodness of fit was then used to determine the maximum likelihood value of the missing data by plotting the graph of the obtained chi-square density against the available outcomes.

Keywords Missing values, Cubic interactions, Multiple imputations, Chi-Square density

1. Introduction

Missing data is an important aspect of statistical science that involves the unavoidable absence of an information as a result of factors which includes no response from the respondent due to acclaimed privacy violation, dropout of respondent in the middle of a research procedure and so making further information not available and finally as a result of initial improper capturing of critical statistics thus making such data inappropriate as a secondary source. However due to the fact that missing data reduce the representativeness of the sample and can therefore distort inferences about the population because majority of the statistical tools requires a complete data set for analysis; then there is need to prevent missing data or to properly estimate missing data before proceeding for further analysis. In situation where missing data are likely unavoidable, the researcher is advised to use statistical methods like statistical design and sampling techniques that allow for robustness in case of missing values. Hence the ultimate aim is to develop properly modeled statistical methods to estimate the missing values so that research work will not die prematurely. There are two major methods of estimating missing data in statistical science which are multiple imputations (MI) and maximum likelihood (ML) methods. Recently, modification of multiple imputations via local least square imputation (LLS) and Bayesian principal component analysis (BPCA) have been commonly used along with some form of neural network approach. Though MI and ML has been extensively used over years for estimating missing data, the precision and accuracy of the obtained estimate via sufficient statistics and standard error computation is still sketchy and difficult. Hence in this study, we bridged the advantages of MI and ML to obtain likelihood estimates of a missing data and confirm the most appropriate maximum likelihood estimate via chi square density approach (Aguilar (2003) [1], Allison (2006) [2], Stoop et al (2010) [3]).

* Corresponding author:
uchedubem@yahoo.com (Ogoke U. P.)
Published online at http://journal.sapub.org/statistics
Copyright © 2016 Scientific & Academic Publishing. All Rights Reserved

2. Model Assumption

Let the number of eggs produced y_i be dependent on the input factors feed weight s and months t. Then y_i is a function of the interaction between feed weight s and months t that is

$$y_i = f(s \times t). \quad (1)$$

Assuming that both the feed weight s and months t has different levels of cubic, quadratic, linear and constant interaction effect on the birds then by binomial expansion

$$y_i = (st+1)^3 = (st)^3 + 3(st)^2 + 3(st)^1 + (st)^0 \quad (2)$$

we obtain the general stochastic relationship

$$\begin{bmatrix} s^3t^3 & s^2t^3 & st^3 & t^3 \\ s^3t^2 & s^2t^2 & st^2 & t^2 \\ s^3t & s^2t & st & t \\ s^3 & s^2 & s & 1 \end{bmatrix} + 3\begin{bmatrix} s^2t^2 & st^2 & t^2 \\ s^2t & st & t \\ s^2 & s & 1 \end{bmatrix} + 3\begin{bmatrix} st & t \\ s & 1 \end{bmatrix} + 1 \tag{3}$$

which result into the general regression equation

$$\begin{aligned} y_i = {} & \beta_{15}s^3t^3 + \beta_{14}s^3t^2 + \beta_{13}s^3t^1 + \beta_{12}s^3 + \beta_{11}s^2t^3 + 4\beta_{10}s^2t^2 + 4\beta_9 s^2t^1 + 4\beta_8 s^2 \\ & + \beta_7 s^1t^3 + 4\beta_6 s^1t^2 + 7\beta_5 s^1t^1 + 7\beta_4 s^1 + \beta_3 t^3 + 4\beta_2 t^2 + 7\beta_1 t^1 + 8\beta_0 \end{aligned} \tag{4}$$

The matrix of the regression coefficients from (4) is

$$\beta^T = [\beta_{15} \quad \beta_{14} \quad \beta_{13} \quad \beta_{12} \quad \beta_{11} \quad \beta_{10} \quad \beta_9 \quad \beta_8 \quad \beta_7 \quad \beta_6 \quad \beta_5 \quad \beta_4 \quad \beta_3 \quad \beta_2 \quad \beta_1 \quad \beta_0] \tag{5}$$

For the period of eighteen months, the regression model is thus

$$y_{i[18\times1]} = X^T_{[18\times16]} \times \beta^T_{[16\times1]} + \varepsilon_{[18\times1]} \tag{6}$$

The design matrix for the feed weights (s) and months (t) interaction is given by the rectangular matrix (18 x 16) equation (7).

$$X^T = \begin{bmatrix} 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \\ 1&1&1&1&1&4&4&4&1&4&7&7&1&4&7&8 \end{bmatrix} \begin{bmatrix} s^3t^3 \\ s^3t^2 \\ s^3t \\ s^3 \\ s^2t^3 \\ s^2t^2 \\ s^2t \\ s^2 \\ st^3 \\ st^2 \\ st \\ s \\ t^3 \\ t^2 \\ t \\ 1 \end{bmatrix} \tag{7}$$

3. Application

The matrix of the egg produced for each month by the different categories of birds along with its missing spots is given below in equation (8)

$$y_{yaffa(5.0)} = \begin{bmatrix} 53 \\ 53 \\ 56 \\ 57 \\ 61 \\ 62 \\ 62 \\ 56 \\ 60 \\ 55 \\ 54 \\ - \\ 44 \\ - \\ 45 \\ 45 \\ 45 \\ 47 \end{bmatrix}; y_{niger(5.0)} = \begin{bmatrix} 53 \\ 54 \\ 56 \\ 59 \\ 61 \\ 60 \\ 61 \\ 54 \\ 59 \\ 54 \\ 55 \\ 52 \\ 45 \\ 45 \\ 45 \\ 48 \\ - \\ 47 \end{bmatrix}; y_{harco(5.0)} = \begin{bmatrix} 58 \\ 57 \\ 57 \\ 62 \\ 59 \\ 60 \\ 60 \\ 54 \\ 62 \\ - \\ 53 \\ 54 \\ 46 \\ 46 \\ 48 \\ 47 \\ 47 \\ 46 \end{bmatrix}; y_{harco(5.1)} = \begin{bmatrix} - \\ 48 \\ 45 \\ 46 \\ 46 \\ 48 \\ 62 \\ 52 \\ 57 \\ 54 \\ 57 \\ 51 \\ - \\ 48 \\ 45 \\ 46 \\ 46 \\ 48 \end{bmatrix}; y_{harco(5.2)} = \begin{bmatrix} - \\ 60 \\ 60 \\ 60 \\ 60 \\ 64 \\ 60 \\ 54 \\ 54 \\ 59 \\ 51 \\ 50 \\ 48 \\ 47 \\ 46 \\ 45 \\ 47 \\ 50 \end{bmatrix}; y_{blackp(4.9)} = \begin{bmatrix} 50 \\ 54 \\ 51 \\ 60 \\ 67 \\ 63 \\ 63 \\ 57 \\ - \\ 55 \\ 52 \\ 52 \\ 45 \\ 49 \\ 47 \\ 45 \\ 45 \\ 45 \end{bmatrix}; y_{blackp(5.0)} = \begin{bmatrix} 56 \\ 58 \\ 59 \\ 65 \\ - \\ 60 \\ 62 \\ 53 \\ 60 \\ 56 \\ 53 \\ 52 \\ 46 \\ 46 \\ 45 \\ 46 \\ 48 \\ 46 \end{bmatrix} \tag{8}$$

4. Methodology

Missing values Estimation steps

a. Substitutes the missing values in turn that is one after the other with each of the available outcome y_i .

b. At each substitution obtain regression coefficients β^T and estimate the missing value E_i.

c. For each available outcomes $y_1, y_2, y_3, y_5, y_8, y_9, y_{10}, y_{11}, y_{12}, y_{15}$ denoted as O_i there is corresponding equivalent value estimated for the missing values (-) denoted as E_i.

d. Recall the chi square goodness of fit.

Then evaluate $\chi^2_{cal} = \sum_{i=1}^{n}\left[\frac{(O-E_i)^2}{E_i}\right] = \sum_{i=1}^{n}\left[\frac{(O-E_i)}{\sqrt{E_i}}\right]^2$ based on the null hypothesis $O = E_i$. That is we obtain the observed value O that has the maximum chi-square probability of not been significantly different from the missing value estimates.

e. Then the Maximum Likelihood Estimate for the missing value is located at the point where the graph of χ^2_{cal} against O_i and E_i is minimum or at the point where the density of the chi square calculated against O_i and E_i at specified degree of freedom k is maximum.

5. Data Analysis and Result

1. Estimating the missing value in the column y_{blackp} with the feed weight of $s = 4.9$ (column 6) in equation (8).

Missing spot

> y_{blackp} <-c(50,54,51,60,67,63,63,57, - ,55,52,52,45,49,47,45,45,45)

Table 1.

O_i	45	47	49	50	51	52	54	55	57	60	63	67
E_i	52.5	52.611	52.72	52.78	52.83	52.89	53	53.06	53.17	53.33	53.5	53.72
χ^2	14.542	8.1965	3.6626	2.075	0.94	0.257	0.251	0.927	3.6375	11.099	22.64	44.36
P-value	0.2675	0.7696	0.9888	0.999	0.999	1	1	0.999	0.989	0.5205	0.031	0.00001

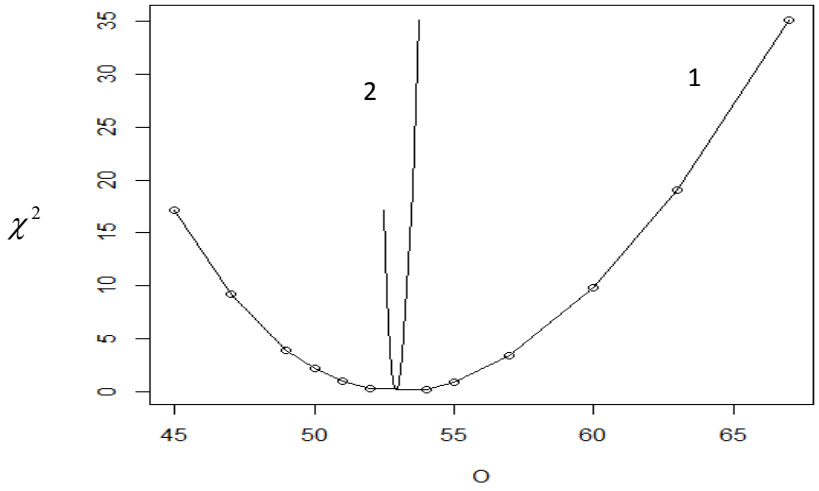

Figure 1. Graph of Chi-square χ^2 value against the observed O {1} and the expected value E_i {2}

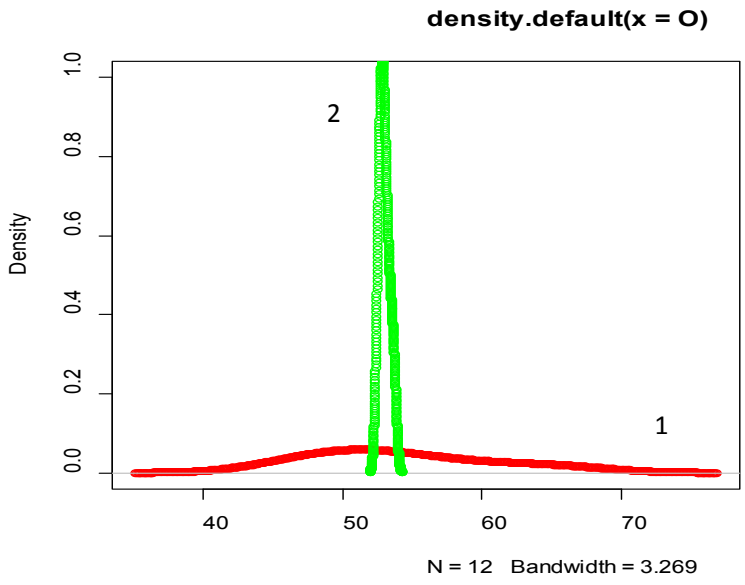

Figure 2. Graph of χ^2 probability density against observed O {1} and the expected values E_i {2}

The value of the missing spot is 52.

2. Estimating the missing value in the column y_{yaffa} with the feed weight of $s = 5.0$ (column 1) in equation (8).

> y_{yaffa} <-c(53,53,56,57,61,62,62,56,60,55,54, --- , 44, -- , 45, 45, 45, 47)

Missing Spots

Table 2.

O_i	44	45	47	53	54	55	56	57	60	61	62
E_i	52.39	52.5	52.72	53.39	53.5	53.61	53.72	53.83	54.17	54.28	54.39
χ^2	18.62	14.92	8.7576	0.1412	0.1447	0.5594	1.3853	2.6226	8.8021	11.68	14.98
P-value	0.068	0.1862	0.6443	1	1	0.999	0.999	0.9948	0.6402	0.3878	0.1835

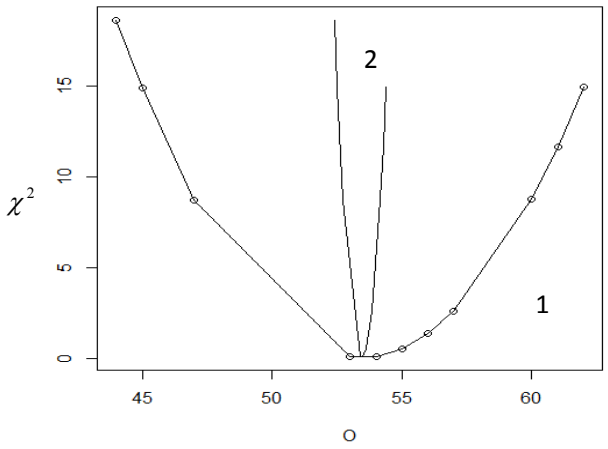

Figure 3. Graph of Chi-square value against the observed O {1} and the expected value E_i {2}

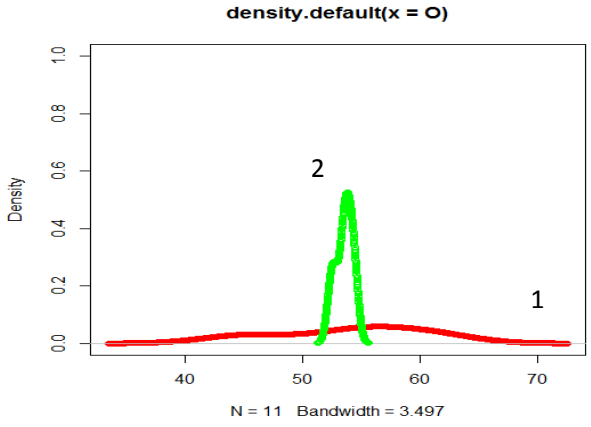

Figure 4. Graph of χ^2 probability density against observed O{1} and the expected values E_i {2}

The value of the missing spots is 54 a piece. Extending the procedure to other columns in equation 8 the estimate of the missing values are $y_{niger}(5.0) = 54$; $y_{harco}(5.0) = 54$; $y_{harco}(5.1) = 51$; $y_{harco}(5.2) = 54$; $y_{blackp}(5.0) = 53$. See appendix for the r program.

6. Conclusions and Recommendations

The chi square approach to estimating likelihood estimate of missing value(s) in data sets via multiple imputations and obtaining its maximum likelihood estimate via area under a chi square density is an approach that is easy to apply in practice. The approach is applicable to any factorial experiment where levels of each factors as well as its interactive factor levels are well defined. The missing data can then be estimated from other available outcomes with similar factor-level interaction. The approach can also be used to estimate missing data as a result of subject that drops out from a particular or joint intervention(s) in medical research.

REFERENCES

[1] R.S. Aguilar, "Missing value estimation methods for DNA microarrays," Statistics and Genomics Seminar and Reading group, 2003.

[2] P.D. Allison, "Modern Methods for Missing Data Statistical," Horizons LLC, 2006.

[3] I. Stoop, J. Billiet, A. Koch, and R. Fitzgerald, "Improving survey response lessons learned from European social survey," John wiley ISBN 0-470-51669-0, 2010.

[4] E. Craig, "Applied Missing Data Analysis," Guilford Press New York, 2010.

[5] J.L. Roderick and D. Rubin, "Statistical Analysis with Missing Data" John Wiley & Sons, Inc: Hoboken 2002.

[6] D. J. Fogarty, 2006. "Multiple imputation as a missing data approach to reject inference on consumer credit scoring," Interstat. URL, 2006.

[7] A. Dempster, N. Liard and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm (with discussion)" Journal of Royal Statistcal. Society B39, 1977, pp 1-38.

[8] K. Lakshminarayan, S. Harp, and T. Samad, "Imputation of missing data in industrial databases," Applied Intelligence 11 (3), 1999, 259 -275.

[9] D.B. Rubin, "Multiple imputation for Nonresponse in Surveys" New York: John Wiley and Sons, 1987.

[10] J.L. Schafer and J.W. Graham, "Missing Data: Our View of the State of the Art" Psychological Methods, 2002.