Treatment of missing and atypical values in the application of the Statistical Model of Impact Measuring in a study of 90 dairy farms in Pastaza province, Ecuador.

E.O. Segura[1] and Verena Torres[2]

[1]*Universidad Estatal Amazónica, km 2½, Napo, Puyo, Pastaza, Ecuador*

[2]*Instituto de Ciencia Animal, Apartado Postal 24, San José de las Lajas, Mayabeque, Cuba*

Email: edisonsegurachavez@gmail.com

The treatment of missing and atypical values is presented in the gathered information of a study of 90 dairy farms from Pastaza province, Ecuador, before applying the Statistical Model of Impact Measuring, to verify data reliability. The randomness joint test of Little was applied, according to the obtained value of Chi-square, not significant ($P > 0.05$), which confirms that data are missing completely at random, and makes possible the use of the attribution with regression analysis. Mahalanobis D^2 and D distance values of all the analyzed farms were shown. These last were compared to the percentiles of Chi-square test: $X^2_{(P < 0.05;20)} = 34.2$ distribution. The results showed the null hypothesis acceptance, so the farms were not considered as atypical cases. The used methods for the treatment of missing and atypical values allowed to obtain a complete data base of real values that will contribute to the application of the Statistical Model of Impact Measuring.

Key words: *Statistical Model of Impact Measuring, joint test of randomness, Mahalanobis distance.*

Missing values (absent or lack information) is a common problem in any research and it cannot be forgotten in data analysis, because it can have serious repercussion on the potency lost of the analysis, even on the appearance of unacceptable bias. The elimination of entities with this problem limits the external validity or representativeness of the study results, although it is something unavoidable in researches (Uriel and Aldas 2013).

Schafery and Graham (2002) showed that the reasons for data absence could have been diverse: faults in measuring instruments, persons that not attend to the interview or not answer to certain questions, or answer with the "do not know" option included on the questionnaire. For this and other reasons, missing data are ubiquitous in the research.

As Tabacknick and Fidell (1996) pointed out, the pattern of missing values is more important that its quantity. If their distribution is at random in data matrix, they cannot cause damage to the analysis. However, if they respond to a certain pattern, the analysis can be affected.

Little and Rubin (1989) presented a technique to prove if absent data constitutes or not a group of random numbers by means of the randomness joint test of Little, based on χ^2 (chi-square)distribution. Once it is proven that randomness in absent data exits, it can start any statistical analysis.

Basically, two fundamental procedures for missing values exit: the elimination of cases that contain them or the attribution of a rated value to the variable. (Uriel and Aldas 2013).

The elimination of all cases that have a missing value is the most used procedure, because it has mistake in most of the statistical programs. This causes that, if the researcher does not carry out a previous data studio, the program can be eliminating the cases with missing values, although these are in unused variables. The main limitation of this method is the lack of information that takes place at working with a lower sample.

In the attribution, the most usual is to replace the missing value by the variable mean that is calculated with the available cases .The attribution per regression is another alternative method, where the variable whose missing values wants to be considered, acts as dependent variable, while the remainder, as independent. The regression analysis can predict the variable absent values, from the relation with others of the data set (Perez 2004).

Atypical values are individuals that showed a value or values combination in the observed variables that differs from the remainder. These values can appear for diverse reasons, like mistakes in data coding, consequence of an extraordinary situation, or they can also due to unknown causes (Hawkins1980).

Certain atypical values can cause important distortion in the analysis results, that is why is necessary to find their appearance, to study the influence they has and, in case of being about influential observations, to analyze causes and to decide if they should be retained or excluded from the analysis. The atypical values detection can be carried out from an univariate or multivariate perspective. Uriel and Aldas(2013) recommended the statistical multidimensional measure Mahalanobis (D^2) for the treatment of the atypical multivariate values.

The objective of this research was to present the treatment of missing and atypical values in the collected information, from 90 dairy farms, in Pastaza province, Ecuador. The Little randomness test and Mahalanobis distance was used, before applying the Statistical Model of Impact Measuring (Torres *et al.*2008).

Materials and Methods

The missing and atypical values treatment was carried out from the data base of the dairy farms system in Pastaza province, with a matrix of 90 farms and 20 variables.

The joint test of Little randomness was applied, formal contrast based on χ^2 (Chi-Square test) distribution, to check if values are randomly distribute according to the null hypothesis: data are missing totally at random, if P value is significant($P<0.05$, $P<0.01$ or $P<0.001$). Otherwise , data are not randomly missing.

The attribution of rated values to the variables with the absentees by means of regression analysis was carried out, that consists on predicting the variable omitted values, from their relation with others of the data set.

This method calculates the multiple lineal regression estimates and increases the estimates with random components. For each predicted value, the procedure can add a residue of a complete case (farm), selected at random way, a random standard deviation or a random deviation of T-Student distribution.

The detection of atypical values from the multivariate point of view with the use of Mahalanobis (D^2) distance was carried out, statistical measure of an individual multidimensional distance, regarding to the centroidal or observations mean , according to Cuadras(2012), by the following expression:

$$D^2=(X_i-\bar{X})'\hat{S}^{-1}(X_i-\bar{X})$$

Where:

X_i: Column vector with values of all variables for i-th observation

$\bar{X}$: Column vector of sample means

$\hat{S}^{-1}$: Inverse of matrix variance – sample covariance

To determine if every of 90 farms is an atypical case, it is assumed that the square of Mahalanobis distance is distribute according $\chi^2_{(m\ g.l.)}$. The null hypothesis is that the farm is not atypical. To calculate the test significance, the critical value of χ^2 distributions, with 20 freedom degree (variables number) was used.

All methods were processed with the statistical software IBM SPSS (2013), IBM SPSS Statistics 22. Algorithms. Chicago: IBM SPSS Inc.

Results and Discussion

In table 1is showed the number of observations and statistical values of each variable. Variables X3, X9 and X16 showed a missing value from the original data base in farms 39, 2 and 57, respectively.

The joint test of Little randomness contributed the Chi- Square test value, that was not significant

($P > 0.05$).It was proven that data are missing, totally at random. Then, the attribution per regression was used.

The attributed values, means and standard deviation (SD) are showed in table 2. In accordance with the results previously shown, it can be concluded that means and SD values in variables with missing values are practically the same, main objective of the attribution method.

Table 1. Missing and statistical values

Variable	N	Mean	SD
X1	90		
X2	90	52.95	39.11
X3	89	39.12	**27.95**
X4	90	65.72	29.01
X5	90	16.87	9.73
X6	90	37.79	22.94
X7	90	5.17	15.87
X8	90	19.13	11.59
X9	89	12.09	**7.10**
X10	90	68.45	18.57
X11	90	10.27	5.52
X12	90	2.53	2.37
X13	90	2.94	2.58
X14	90	10.67	7.12
X15	90	25.83	19.18
X16	89	6.71	**2.79**
X17	90	2.50	2.84
X18	90	1.13	1.45
X19	90	63.39	44.16
X20	90	19.00	7.51

Lohr (1999) showed , that the attribution procedures per regression to produce a data base without missing data are very important. It should not abused of attribution methods, because does not increase the available information, but it is generated from the one that is having. It should be consider as a basic idea that the attribution does not replace or omit some previous phase, like collecting data and digitalization. It is necessary to try to obtain the original data from the different variables, and in case of not obtaining it, to turn to the data attribution.

The distance D^2 and D values for all analyzed farms are showed in table 3.These values are compared with the distribution Chi-Square test: $X^2_{(P<0.05;20)}= 34,2$ percentiles. The results showed that there was not significant difference ($P > 0.05$), that allowed to accept the null hypothesis and to state that farms are not atypical cases.

Hair *et al.* (1999) suggested using a preservative level, maybe 0.001as threshold value to be designating as atypical case. The value 0.05 to check that there were not atypical data was considered in this research.

If significant differences in determined farms exits, they should be studied and to interpret by the researchers, to decide if any is eliminated or not, and to inform the reasons for it.

The used methods for the treatment of missing values and the atypical farms analysis allowed to obtain

Table 2. Attributed values

Variable	No farm	Attributed value	Mean	S.D.
X3	39	39.52	39.12	27.79
X9	2	12.09	12.09	7.06
X16	57	6.73	6.71	2.77

Little test: Chi-square=72.26; P = 0.084

Table 3. Results of Mahalanobis contrast

Farm	D^2	D	Sign.	Farm	D2	D	Sign.	Farm	D^2	D	Sign.
1	11.82	3.44	NS	31	40.76	6.38	NS	61	18.74	4.33	NS
2	33.78	5.81	NS	32	11.08	3.33	NS	62	50.74	7.12	NS
3	9.34	3.06	NS	33	17.53	4.19	NS	63	14.88	3.86	NS
4	6.78	2.6	NS	34	13.22	3.64	NS	64	14.52	3.81	NS
5	17.41	4.17	NS	35	14.47	3.8	NS	65	9.82	3.13	NS
6	9.34	3.06	NS	36	10.44	3.23	NS	66	12.55	3.54	NS
7	10.26	3.2	NS	37	16.32	4.04	NS	67	67.32	8.2	NS
8	6.78	2.6	NS	38	18.21	4.27	NS	68	17.97	4.24	NS
9	14.47	3.8	NS	39	28.16	5.31	NS	69	35.85	5.99	NS
10	10.93	3.31	NS	40	14.02	3.74	NS	70	8.8	2.97	NS
11	8.68	2.95	NS	41	39.47	6.28	NS	71	60.25	7.76	NS
12	29.1	5.39	NS	42	12.51	3.54	NS	72	9.49	3.08	NS
13	17.06	4.13	NS	43	6.05	2.46	NS	73	51.37	7.17	NS
14	18.11	4.26	NS	44	14.39	3.79	NS	74	20.16	4.49	NS
15	12.37	3.52	NS	45	13.12	3.62	NS	75	13.6	3.69	NS
16	18.11	4.26	NS	46	21.65	4.65	NS	76	12.34	3.51	NS
17	23.1	4.81	NS	47	13.7	3.7	NS	77	14.9	3.86	NS
18	18.41	4.29	NS	48	34.68	5.89	NS	78	13.43	3.66	NS
19	9.45	3.07	NS	49	10.71	3.27	NS	79	3.7	1.92	NS
20	14.33	3.79	NS	50	32.13	5.67	NS	80	14.29	3.78	NS
21	18.4	4.29	NS	51	18.8	4.34	NS	81	64.2	8.01	NS
22	28.67	5.35	NS	52	32.8	5.73	NS	82	28.39	5.33	NS
23	17.76	4.21	NS	53	17.48	4.18	NS	83	41.01	6.4	NS
24	18.72	4.33	NS	54	40.78	6.39	NS	84	16.87	4.11	NS
25	14.36	3.79	NS	55	6.95	2.64	NS	85	7.19	2.68	NS
26	8.56	2.93	NS	56	12.65	3.56	NS	86	11.67	3.42	NS
27	16.07	4.01	NS	57	12.27	3.5	NS	87	17.16	4.14	NS
28	12.14	3.48	NS	58	44.72	6.69	NS	88	43.13	6.57	NS
29	11.08	3.33	NS	59	17.35	4.17	NS	89	13.28	3.64	NS
30	2.24	1.5	NS	60	17.35	4.17	NS	90	13.28	3.64	NS

a complete data base of real values that make possible to develop the application of the Statistical Model of Impact Measuring.

References

Cuadras, C. M. 2012. Nuevos Métodos de Análisis Multivariado. © C.M. Cuadras Editions Manacor 30 08023. Barcelona-Spain. p.20

Hair, J., Anderson, R. & Tatham, R. 1999. Multivariate Data Analysis (5taEd.). Ed. Englewood Cliffs Prentice Hall. Iberia, Madrid, España. p. 57:59

Hawkins, D. M. 1980. Identification of outliers. Ed. Chapman and Hall.Londres. p.13

Little, R. J. A. & Rubin, D. B. 1989.The Analysis of Social Science Data with Missing Values. Sociological Methods and Research 18: 292

Lohr, Sh. 1999. Muestro: Diseño y Análisis. Editorial Thomson.Nueva York. p. 31-42.

Pérez, C. 2004. Técnicas de Análisis Multivariante de Datos. Editorial Pearson Educación. Madrid, España. p. 25-33

Schafer, J.L. & Graham, J.W. 2002. Missing Data: Our View of the State of the Art. Psychological Methods 7:147

Tabacknick, B. G. & Fidell, L. S. 1996. Using multivariate statistics. 3° edición. Editorial Harper Collins.Nueva York. p. 26

Torres, V., Ramos N., Lizazo, D., Monteagudo, F. & Noda, A. 2008. Statistical model forme a suring the impact of innovation or technology transfer in agriculture. Cuban J. Agric. Sci. 42:13

Uriel, E. & Aldás, J. 2005. Análisis Multivariante Aplicado. Thomson Ed. Madrid, España. p. 13-28

Received: April 10, 2014