

Dedication

I dedicate my thesis work to my family and my students.

Acknowledgement

I would take this opportunity to thank my research supervisor Dr. Amin Ibrahim, family and friends (Dr. Mohammed Suleman Gibreel , Dr. Zakria Mohamed Salih) for their support and guidance without them this research would not have been possible.

Abstract

The missing data in household health survey was a problem for the researchers because it leads to incomplete analysis. The statistical tool of cluster analysis methodology was implemented in the collected data of Sudan's household health survey in 2006.

This research focuses specifically on the analysis of the collected data and the objective is to deal with the missing values in cluster analysis. Two-Step Cluster Analysis is applied in which each participant is classified into one of the identified pattern and the optimal number of classes is determined using SPSS Statistics/IBM. Any observation with missing data is excluded in the Cluster Analysis as in the multi-variable statistical techniques. Therefore, before performing the cluster analysis, missing values is imputed using multiple imputations (SPSS Statistics/IBM). The clustering result is displayed in tables. The descriptive statistics and cluster frequencies are produced for the final cluster model, while the information criterion table displayed results for a range of cluster solutions.

Furthermore, the objective is extended to include the reduction of biases arising from the fact that non-respondents may be different from those who participate and to bring sample data up to the dimensions of the target population totals.

المستخلص

لقد شهدت مشكلة البيانات المفقودة اهتماماً في السنوات الأخيرة ، ومع التطور السريع لأجهزة الحاسوب والبرمجيات في معالجة العمليات أصبح تطوير طرق تحليل البيانات المفقودة ممكن نظرياً وعلى الرغم من ذلك ما زال العديد منها بحاجة للتطوير ويعاني من مشاكل عديدة. من هنا تأتي أهمية تسليط الضوء على طرق معالجة القيم المفقودة لإيجاد أفضل الطرق التي تلائم البيانات من ناحية تقدير القيمة المفقودة، أخذين بعين الاعتبار نسب الفقدان و آلية الفقدان ونمط الفقدان. ان البيانات المفقودة في مسح صحة الأسرة تمثل مشكلة للباحثين لأنه يؤدي إلى تحليل غير كامل. تم تنفيذ أداة إحصائية منهجية التحليل العنقودي في البيانات التي تم جمعها من مسح صحة الأسرة في السودان في عام 2006.

ويركز هذا البحث على وجه التحديد على تحليل البيانات التي تم جمعها والهدف من ذلك هو التعامل مع القيم المفقودة في التحليل العنقودي. تم تطبيق التحليل العنقودي ذو البعدين في كل صنف من النماذج المعرفه وحدد العدد الأمثل من الطبقات باستخدام حزم البرامج الإحصائية spss. تم استبعاد أي ملاحظات مع البيانات المفقودة في التحليل العنقودي كما في التقنيات الإحصائية متعددة المتغيرات. لذلك، قبل تنفيذ التحليل العنقودي، نعوض القيم المفقودة باستخدام التعويض المتعدد باستخدام (IBM / SPSS).

تم عرض نتائج المجموعات في جداول ، ونتائج إحصاءات وصفية عنقوديه، كما أظهرت النتائج وجود عناقيد ذات جوده ما بين قوي وقوي جدا ولا يوجد عنقود ضعيف.

Table of contents

subject	page
Dedication	i
Acknowledgement	ii
Abstract	iii
Abstract in Arabic	iv
Table of contents	viii
List of tables	ix
List of figures	x
Chapter one The introduction	
1.1 Preface	1
1.2 The research problem	1
1.2.1 Statement of the Problem	1
1.3 Methodology	2
1.4 Sudan Household Health Survey (SHHS)	2
1.4.1 Data Sources	3
1.5 Objective of the study	3
1.6 Importance of the study	5
1.7 Questionnaires	5
1.7.1 Questionnaires Sample	6
1.8 Cluster Analysis	7
1.9 Scope and Limitations	8
1.10 Previous Studies	9
1.8 Organization of The Study	9

Chapter 2 Literature Review	
2.1 Criticism on Data Collection of Household Health Surveys	11
2.2 Suggestions for Analysing Survey Data	13
2.3 Missing Data Treatment	14
2.3.1 “Ad-hoc” Methods	15
2.3.2 Multiple Imputation	16
2.3.4 Conditional Gaussian	18
2.3.5 Chained Equations	18
2.3.6 Methods for Monotone Data sets	19
2.3.7 Issues with Imputation	19
2.3.8 Methods of Weighting	20
2.3.9 Bayesian Approaches	20
2.4 Cluster Analysis	21
Chapter 3 Methodology	
3.2 Sudan Household Health Survey (SHHS)	30
3.2.1 Sample Design	30
3.2.2 Sampling frame and units of analysis	31
3.2.3 Stratification	31
3.2.4 Size and Allocation of Samples	32
3.2.5 Sample selection procedures	33
3.3 Estimation and weighting procedures	36
3.4 Data analysis	36
3.4.1 Two-Step Cluster Analysis	37
3.4.2 Assumptions of Data in Two-Step Cluster Analysis	39

3.4.3 Two-Step Cluster Analysis Plots	40
3.4.4 Two-Step Cluster Analysis Output	41
Chapter 4. Results & Discussion	
4.1 Characteristics of woman respondents	42
4.1.1 Describing the Pattern of Missing Data	45
4.1.2 Using Multiple Imputations to Complete and Analyze a Dataset	50
4.1.3 Imputation Models	51
4.1.4 Custom Imputation Model	53
4.1.5 Nominal Regression	54
4.1.6 Two-step Cluster Analysis	58
4.2 Knowledge of meanings of HIV/AIDS of women	65
4.2.1 Multiple Imputations	69
4.2.2 Descriptive Statistics acknowledge HIV/AIDS	73
4.2.3 Checking FCS Convergence	78
4.2.4 Two-step Cluster Analysis	79
4.2.4.1 Model Summary and Cluster Quality	80
Chapter 5	
Conclusions	98
Recommendations	99
References	
Appendix	

LIST OF TABLES

List title	page
Table 1.1 Questionnaires Sample	6
Table 4.1: distribution Number of women response rates	43
Table 4.2 : Women's characteristics	44
Table 4.3 : Univariate Statistics Pattern of Missing Data	45
Table 4.4 : Separate Variance t Tests ^a Pattern of Missing Data	46
Table 4.5 : mstatus(Marital status) Pattern of Missing Data	47
Table 4.6 : melevel(education) Pattern of Missing Data	48
Table 4.7 : wlthind5(wealth) Pattern of Missing Data	48
Table 4.8 : EM Estimated Statistics	49
Table 4.9 : Imputation Specifications	51
Table 4.10 : Imputation Results	51
Table 4.11 : Imputation Models	52
Table 4.12 : WM9(age of woman) imputed values	52
Table 4.13 : logage	53
Table 4.14 Case Processing Summary	54
Table 4.15 Model Fitting Information	55
Table 4.16 Pseudo R-Square	55
Table 4.17 Likelihood Ratio Tests	55
Table 4.18 : Model Fitting Information	56
Table 4.19 Likelihood Ratio Tests	56
Table 4.20. Knowledge of HIV/AIDS Percentage of woman Year of birth (1951-1991)	57
Table 4.21. Case Processing Summary	65
Table 4.22. Variable Summary	68
Table 4.23. Imputation Specifications	70

Table 4.24. Imputation Results	73
Table 4.25. Imputation Models	73
Table 4.26. HA3_X (Can AIDS be avoided?)	74
Table 4.27. HA9A(AIDS from mother to child during pregnancy)	75
Table 4.28. HA9B(AIDS from mother to child at delivery)	76

LIST OF FIGURES

List tilte	page
Fig.4.1 over all summary of missing data	50
Fig.4.1 Model Summary of cluster	58
Fig. 4.2 Custer	59
Fig. 4.3 Model Summary of imputation 1	60
Fig. 4.6 Model Summary of imputaion 2	61
Fig. 4.4 Clusters of imputation 2	62
Fig. 4.5 Model Summary Imputation 3	62
Fig. 4.9 Clusters of imputation 3	63
Fig. 4.6 Model Summary Imputation 4	63
Fig. 4.11 Clusters of imputation 4	64
Fig. 4.7 Model Summary Imputation 5	64
Fig. 4.13 Clusters of imputation 5	65
Fig.4.14 chart distribution ever heard of HIV or AIDS and Fig.4.15 can AIDS be avoid?	66
Fig.4.16 chart Distronution of ever heard of HIV or AIDS and Fig.4.17 can AIDS be avoid?	67
Fig. 4.18 Summary missing values	69
Fig. 4.19 <i>missing value patterns</i> by variables analysis	71
Fig.4. 20 <i>missing value pattern</i>	72
Fig. 4.21 FCS Iteration number	78
Fig. 4.22 imputation original Model Summary and Fig. 4.23 imputation original	80

Fig. 4.24 imputation original data Custers	80
Fig. 4.25 imputation number 1 Model Summary and Fig. 4.26 imputation number 1 cluster size	81
Fig. 4.27 imputation number 1 clusters	81
Fig. 4.28 imputation number 2 Model Summary and Fig. 4.29 imputation number 2 cluster size	82
Fig. 4.30 imputation number 2 clusters	82
Fig. 4.31 imputation number 3 Model Summary and Fig. 4.32 imputation number 3 cluster size	83
Fig. 33 imputation number 3 clusters	83
Fig. 4.34 imputation number 4 Model Summary and Fig. 4.35 imputation number 4 cluster size	84
Fig.4.36 imputation number 4 clusters	84
Fig. 4.37 imputation number 5 Model Summary and Fig. 4.38 imputation number 5 cluster size	85
Fig. 4.39 imputation number 5 clusters	85