CATEGORICAL DATA ANALYSIS PROJECT

WOMEN'S CHOICE OF CONTRACEPTIVES

Table of Contents:

1. INTRODUCTION

Contraceptive choices have been provided to women in a way that respects and fulfills their human rights necessities enabling them to make informed choices for themselves. Women's choices, however, are often taken away from them or limited by direct or indirect social, economic and cultural factors. From a women's point of view, her choices are made at a particular time, in a particular societal and cultural context; choices are complex, multifactorial and subject to change. Decision-making for contraceptive methods usually requires making trade-offs among the advantages and disadvantages of different methods, and these vary according to individual circumstances, perceptions and interpretations. Factors to consider when choosing a particular contraceptive method include the characteristics of the potential user, the risk of disease, the adverse effects profile of different products, cost, availability and preferences. In this project, I am interested in finding out what type of contraceptive methods are frequently chosen by women of different demographics. This information is valuable to help discern what kind of women are using birth control as it could be used to create targeting marketing campaigns or help improve a products appeal to specific types of women allowing for an easier decision-making process.

2. DESCRIPTION OF DATA

The dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey with 1473 observations was taken from *Kaggle*. The survey samples were taken from married women who were not pregnant at the time of the interview. The response variable is the contraceptive method used by women which consist of 3 categories: **1**:*no use,* **2**: *long-term and* **3**: *short-term use*. The potential explanatory variables of interest to me are solely based on the women's socio-economic and demographic background which are age, education level, number of children, religion, employment status, standard of living index and media exposure. The age of women was measured in years where the average age for a married woman in this study was 32.54 years with the minimum and maximum age of 16 years and 49 years respectively. This data did not include women above 49 years because they are most likely in menopause and may possibly skew the data (*see histogram figure 2.1*). The education level for women were measured which was as follows: 1: low, 2: medium, 3: medium-high, 4: high. The women in this study were quite educated with an average education of 2.959. Women in this dataset had an average of 3.261 children in their family and looking at the boxplot (*see boxplot 2.3),* there are potential outliers; however, as there is no strong reason to remove them from the model, I will include them as they could be important in showing model inadequacy. Furthermore, in the dataset, there was a majority of 1253 women belonging to Islam group (1) and a minority of 220 belonging to other religion (non-Islam is 0). A majority of the women were unemployed (not employed-1, employed-0), had a good media-exposure (good media exposure-0, not a good exposure-1) with an average standard of living index as 3.134 (index level 1 to 4 which goes form low to high) as seen in *summary output 2.4*. Interestingly, long-term contraceptives proportion count increases as index level goes up (*see figure 2.2*). Information on media exposure in the dataset is limited as a

good media exposure and bad media exposure for women is quite vague. Furthermore, the data was fairly clean and did not require any transformations or imputation as it did not have any missing values. (*see figure 2.5*)

3. STATISTICAL ANALYSIS (*SHORT RECAP*)/ INFERENCES

The statistical analysis was conducted using SAS and R. A baseline-category logit model was fit with *short-term* contraceptive use ("3") as the reference category. There was no quasi-complete or complete separation in the model; hence, use of exact methods was not required. Model was selected through *purposeful model* selection. From *output 3.2*, looking at conditional significance, all the predictors were significantly associated with the response (holding other predictors constant), and also the model overall was significant with a small p-value (*see 3.3*). Furthermore, checking model fit using Hosmer-Lemeshow Test, the model with Age+Education+Children+Index+Age: Index+Age^2 proved to be adequate. (*see output 3.5*)

Table 3.1 **ODDS RATIO ESTIMATES (from final model)**

Effect	Contraceptive	Point Estimate	95% Wald C.I.
Education	1	0.715	(0.621, 0.824)
Education	2	1.734	(1.438, 2.090)
Children	1	0.693	(0.641, 0.749)
Children	2	0.967	(0.891, 1.049)

Age*	1	1.115	(1.091, 1.139)
Age*	2	1.070	(1.045, 1.097)
Index*	1	0.784	(0.685, 0.898)
Index*	2	1.155	(0.962, 1.386)

Note: * gives a rough estimates and confidence intervals as model includes interactions and higher order term.

The fitted model: *(see output 3.4)*

$\boldsymbol{ln}\left(\frac{\hat{\pi}_1}{\hat{\pi}_3}\right) = 1.9764 - 0.1842 * \text{Age} - 0.3352 * \text{Education} - 0.3668 * \text{Children} + 1.1835 * \text{Index} + 0.00689 * Age^2 - 0.0484 * (\text{Age} * \text{Index})$

$\boldsymbol{ln}\left(\frac{\hat{\pi}_2}{\hat{\pi}_3}\right) = -6.8231 - 0.1073 * \text{Age} - 0.5504 * \text{Education} - 0.0337 * \text{Children} + 0.9826 * \text{Index} + 0.0011 * Age^2 - 0.0294 * (\text{Age} * \text{Index})$

where, $\hat{\boldsymbol{\pi}}_1, \hat{\boldsymbol{\pi}}_2, \hat{\boldsymbol{\pi}}_3$ are estimated probabilities of No use, Long term use Short term (*reference*) contraceptive methods

From table 3.1, we can see that for each additional child, the estimated odds of women using no contraceptives as opposed to short terms use change by a factor of 0.693, and the estimated odds of women using long term contraceptives as opposed to short terms use change by a factor of 0.891 holding other variables fixed in the model. However, as education level increases there is no prominent choice of contraceptive use in women. Furthermore adjusting for other predictors in the model, for every year increase in age, the estimated odds of women using no contraceptives as opposed to short terms use increase by a factor of *roughly* 1.115, and the estimated odds of women using no contraceptives as opposed to long terms use change by a

factor of *roughly* 1.04. Moreover, women are also more likely to use long term contraceptive methods over none as their standard of living increases. Also, for each level increase in index, the estimated odds of women using long term contraceptives as opposed to short terms use change by a factor of *roughly* 1.155, holding other variables fixed.

5. DISCUSSION

In my statistical analysis, I have found that each socioeconomic and demographic variable has a unique influence on the contraceptive used. It appears by my analysis that as women grow older, they are more likely to use no contraception than either short- or long-term contraceptive methods. Interestingly, my model shows that women of a higher standard of living prefer to use long term birth control methods; also, the more children a woman has, the more likely she is to use short-term methods of contraception. Most of these finds match what is understood logically, that as women age, they are more likely to want children and therefore less likely to use any form of birth control. Additionally, women with children would prefer short term methods because they may want more children and the ability to have them whenever they choose. From this information companies may choose to target younger demographics when marketing either form of contraceptive. Utilizing this analysis, short term contraceptive companies may want to see how they can become more useful to women with higher standard of living index, or long-term contraceptive companies may want to see how they can appeal more to women with multiple children. There are several possible applications that could be used to improve the contraceptive industry and how it allows women to make their choices.

As far as limitations, it is important to consider this data is dated in 1987, nearly 34 years ago. As the contraception methods have become safer and more varied in the last several years the results could be very different than data that could be collected today. Geographical location also holds a strong weight over the validity of this data. Indonesia has a vastly different culture than the US as they are more conservative. This could have a negative effect on how the culture these women belong to see contraception as a whole. Further studies should be done taking surveys from different women internationally to see if such results can be duplicated. For further research, it would be important to consider analyzing residuals for multicategory logit models to see which observations are influential in the model. However, after some research, I found doing so in SAS is not *simple*. Additionally, for this data we used a baseline logit model where our response was treated nominally. For future work, I believe working with a cumulative logit model where the response variable is treated as ordinal will potentially have a power advantage over baseline logit model, also making interpretation much easier.

APPENDIX

Figure 2.1 Histogram of Age

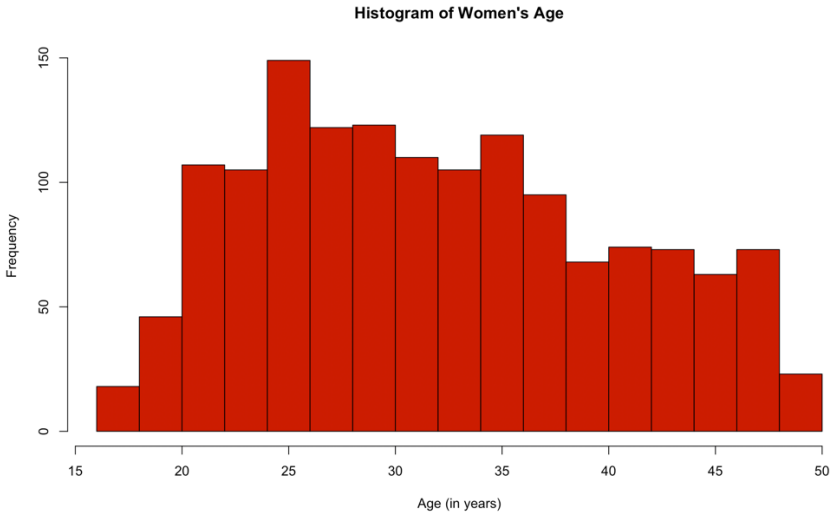


Figure 2.2 Bar plot of contraceptive_method for each index

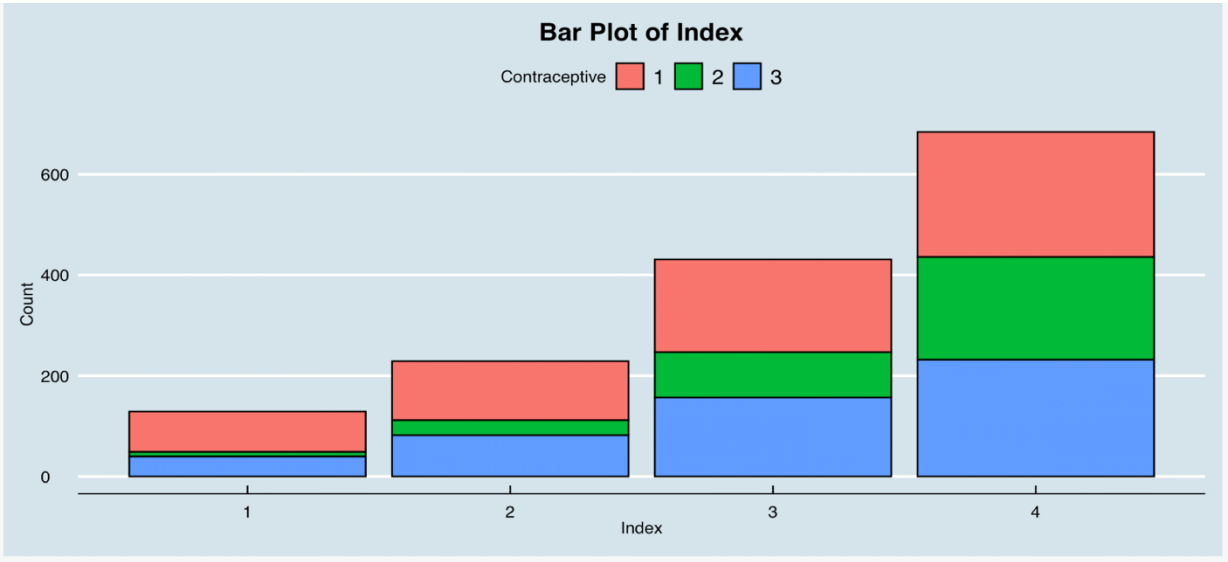
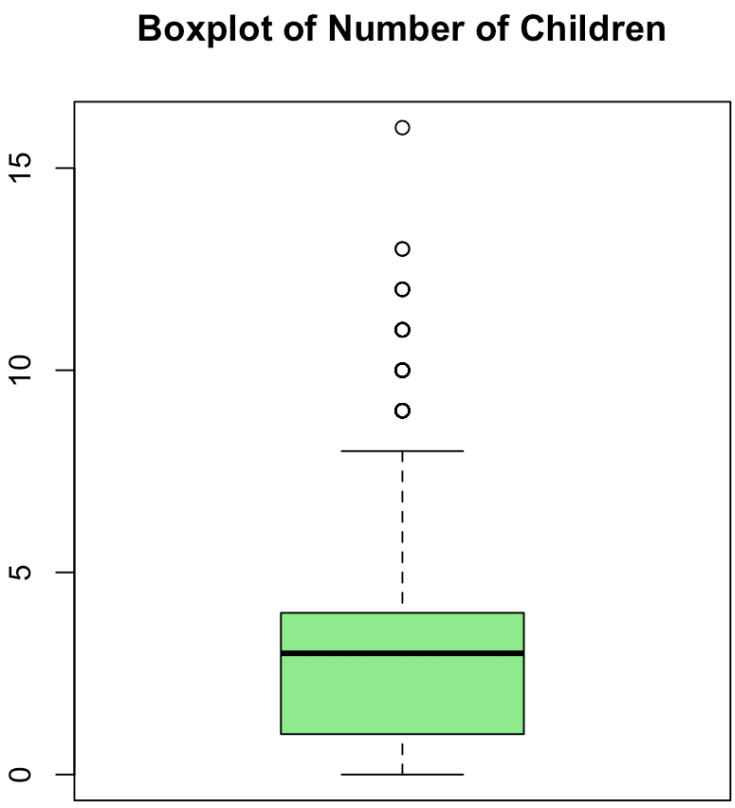


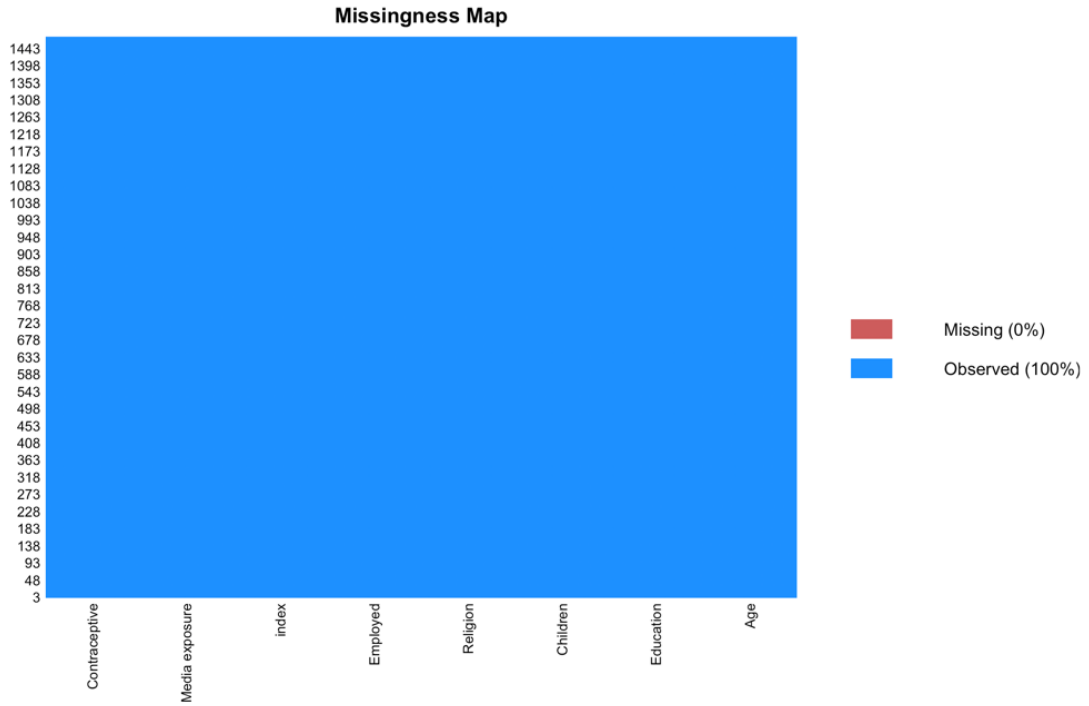
Figure 2.3 Boxplot of Children



Output 2.4 Summary

```
> summary(CMC)
      Age          Education       Children      Religion Employed     index       Media exposure
 Min.   :16.00   Min.   :1.000   Min.   : 0.000   0: 220   0: 369   Min.   :1.000   0:1364
 1st Qu.:26.00   1st Qu.:2.000   1st Qu.: 1.000   1:1253   1:1104   1st Qu.:3.000   1: 109
 Median :32.00   Median :3.000   Median : 3.000                     Median :3.000
 Mean   :32.54   Mean   :2.959   Mean   : 3.261                     Mean   :3.134
 3rd Qu.:39.00   3rd Qu.:4.000   3rd Qu.: 4.000                     3rd Qu.:4.000
 Max.   :49.00   Max.   :4.000   Max.   :16.000                     Max.   :4.000
 Contraceptive
 1:629
 2:333
 3:511
```

Figure 2.5 Missing value Map



Output 3.2 Wald Test shows Conditional Independence

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	2	14.2661	0.0008
Education	2	89.8483	<.0001
Children	2	98.7917	<.0001
index	2	16.0987	0.0003
Age*Age	2	46.2920	<.0001
Age*index	2	22.9287	<.0001

Output 3.3 Overall model significance

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	401.1146	12	<.0001
Score	349.4518	12	<.0001
Wald	279.3186	12	<.0001

Output 3.4

Analysis of Maximum Likelihood Estimates						
Parameter	Contraceptive	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	1.9764	1.2176	2.6349	0.1045
Intercept	2	1	-6.8231	1.7675	14.9014	0.0001
Age	1	1	-0.1842	0.0687	7.1850	0.0074
Age	2	1	0.1073	0.0887	1.4619	0.2266
Education	1	1	-0.3352	0.0724	21.4581	<.0001
Education	2	1	0.5504	0.0953	33.3266	<.0001
Children	1	1	-0.3668	0.0397	85.1762	<.0001
Children	2	1	-0.0337	0.0414	0.6628	0.4156
index	1	1	1.1835	0.2981	15.7654	<.0001
index	2	1	0.9826	0.4202	5.4671	0.0194
Age*Age	1	1	0.00689	0.00111	38.7115	<.0001
Age*Age	2	1	0.00110	0.00132	0.6996	0.4029
Age*index	1	1	-0.0484	0.0101	22.8313	<.0001
Age*index	2	1	-0.0294	0.0133	4.8666	0.0274

Output 3.5 Hosmer-Lemeshow test for Model Fit

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
22.6077	16	0.1246