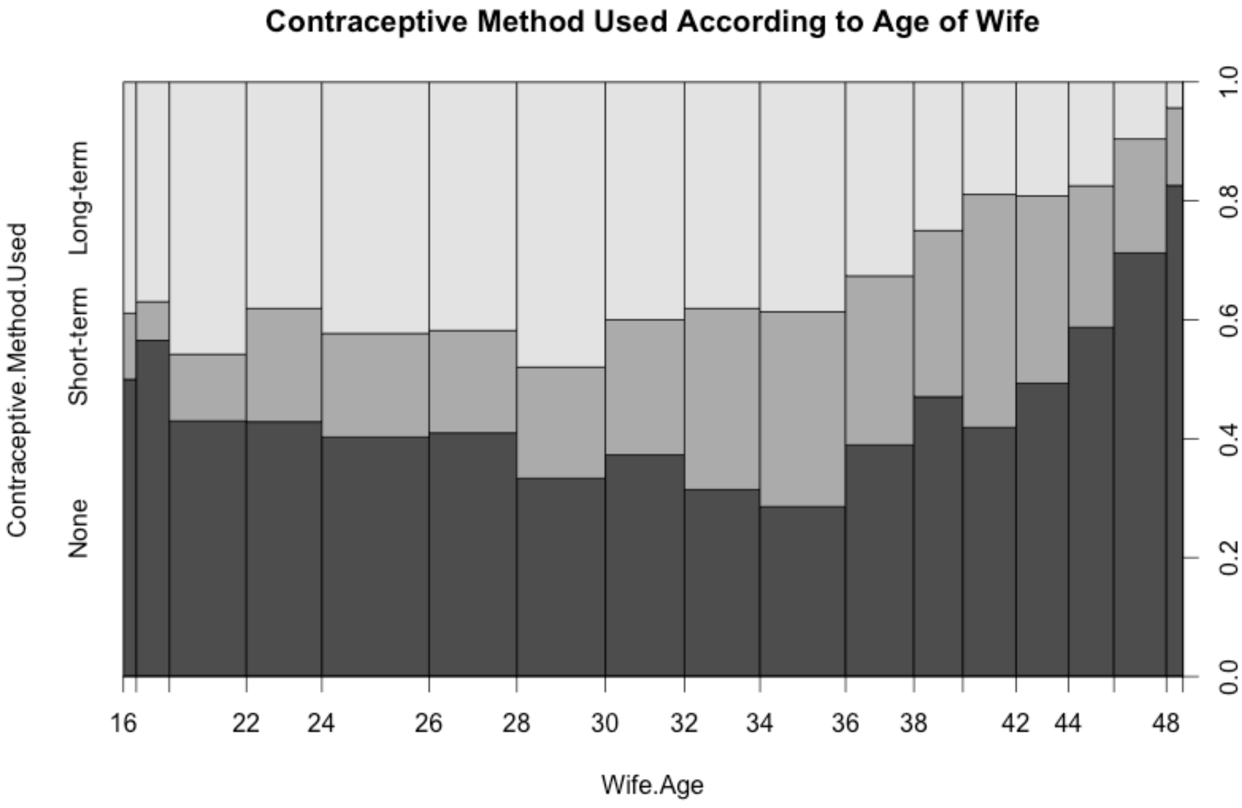
Exploratory Data Analysis on Contraceptive Method Choice in Indonesia

Shaheed Shihan
Spring 2016

Background

The dataset used in this report is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of the interview. The problem is to predict the current contraceptive method choices – no use, long-term methods or short-term methods of a woman based on her demographic and socio-economic characteristics. Predicting the contraceptive method choice of Indonesian women can help the government in making critical decisions with how, where and who to target to provide information on contraceptive choices for the female population. This is important because currently Indonesia is the 4th most populated country in the world with a population of approximately 234 million. In order to be a sustainable economy, the population growth needed to be sustainable as well and thus the government was interested in finding out what affected a woman’s choice of contraceptive use.

Exploratory Data Analysis

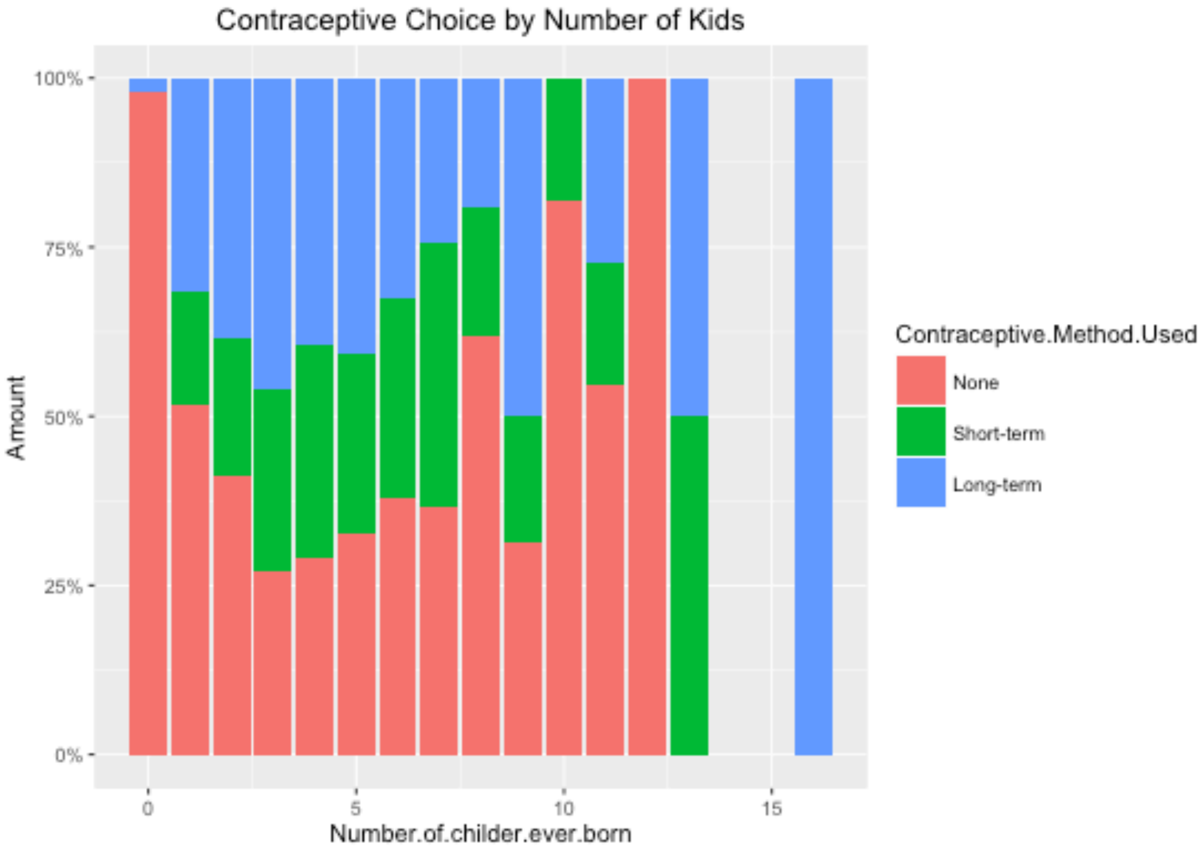


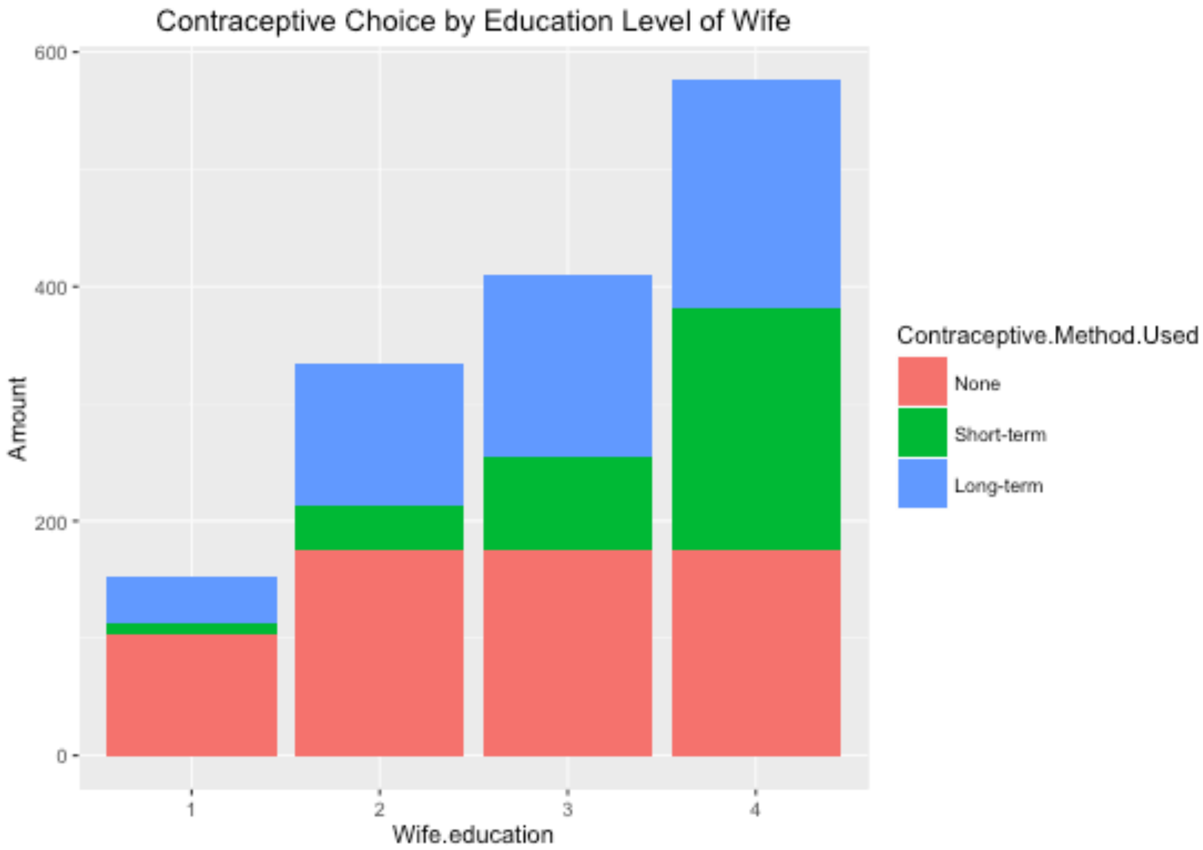
```
Contraceptive.Method.Used
 1       2         3
```

```
629      333      511
> prop.table(table(Contraceptive.Method.Used))
Contraceptive.Method.Used
      1                  2                3
0.4270197      0.2260692      0.3469111
>
```

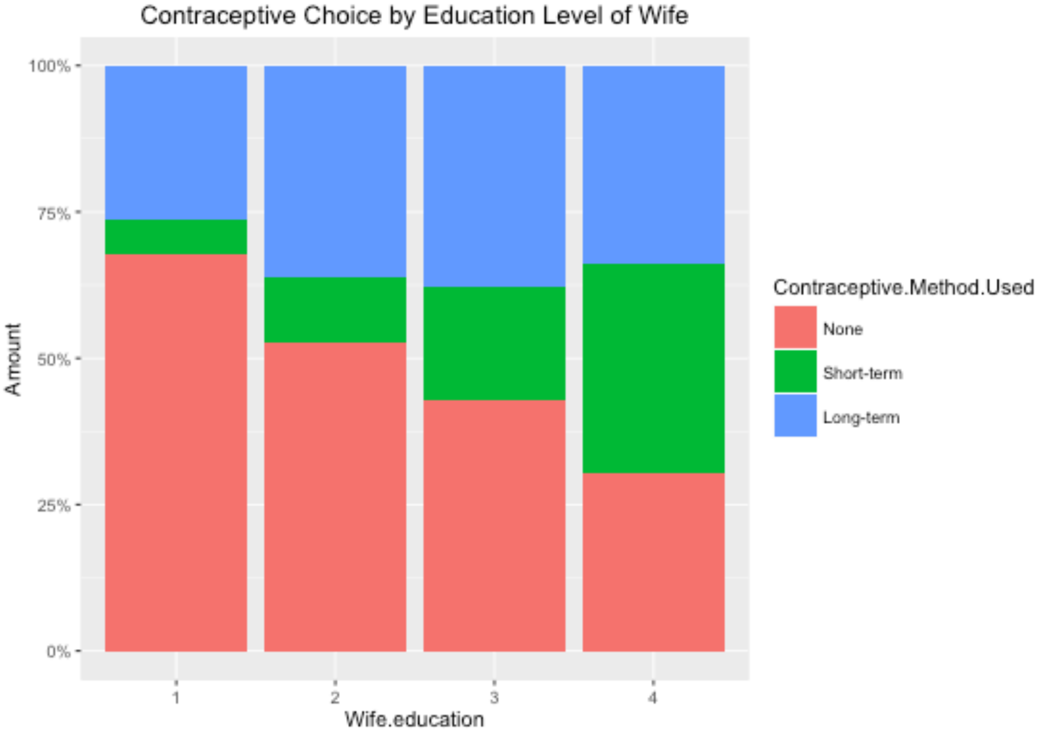
The largest proportion of people (42.7%) in the survey admitted to using no contraception at all.

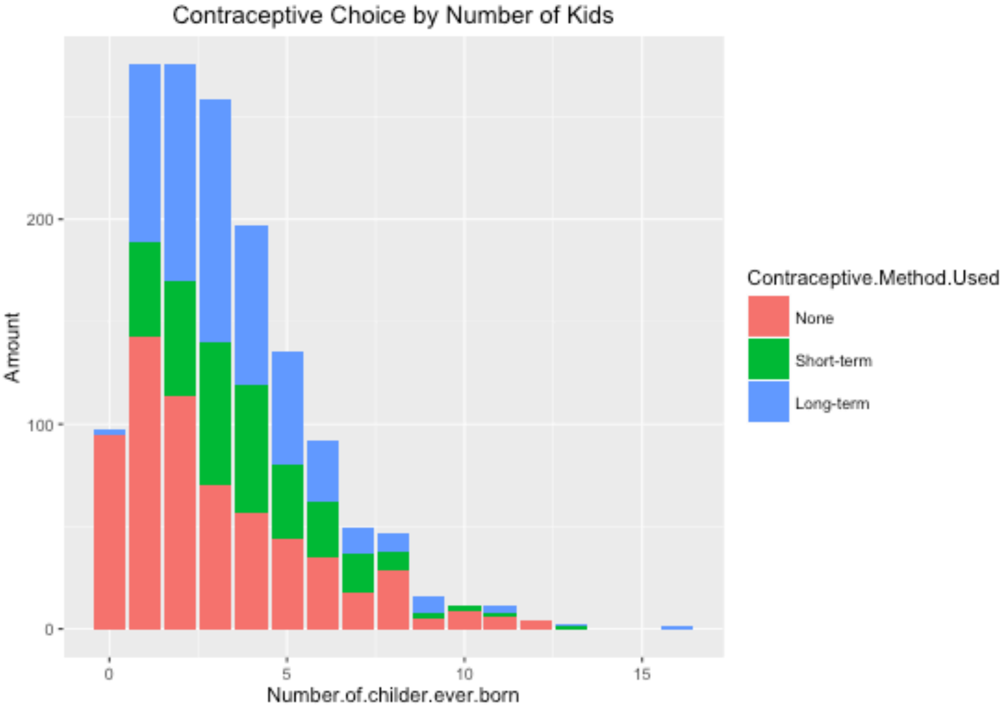
Next some stacked bar graphs were plotted to visualize the relationships between Contraceptive choice by number of kids and the education level of the wife.





As women got more educated, a larger number started using short or long term methods of contraception. The first stacked plot shows the relationship between the contraceptive method used according to the age of the wife. The highest contraceptive use seems to occur when the average age of the women is between 34-36. It is difficult to establish a relationship between contraceptive choice and the number of kids. There seems to be a decline with the lowest occurring at 4-5 kids and then on a gradual rise again.


Contraceptive Choice by Education Level of Wife
Amount
100%
75%
50%
25%
0%
1
2
3
4
Wife.education
Contraceptive.Method.Used
None
Short-term
Long-term


Contraceptive Choice by Number of Kids
Amount
200
100
0
0
5
10
15
Number.of.childer.ever.born
Contraceptive.Method.Used
None
Short-term
Long-term

Statistical Analysis:
Statistical Analysis was conducted on the data set to check the error rate of the various models. This was done in an effort to better understand which method was more suitable in this particular study.

The test error was calculated using three different approaches:

- Validation Set Approach: This approach counters the effect of overfitting. In addition to having a test set and a training set, we had a validation set. This set was used to compare the performances and the test set was used to obtain the performance characteristics such as accuracy, sensitivity, specificity, F-measure and so on.
- Leave One Out Cross Validation: In terms of computational power, this was the most intensive. It takes one observation out as the the validation set and uses the remaining observations to make up the training set. It then proceeds to repeat the process for the entire data set. The drawback of this method is that since only one observation is used the MSE is highly variable.
- 10-fold Cross Validation: This process involves randomly dividing the set of observations into k groups or fold of equal size. The first fold is treated as a validations et and the method is fit on the remaining k-1 folds. The entire process is then repeated k times.
- The caret package was used extensively in calculating the errors and results. The caret package is essentially a wrapper package that was made to streamline the process of classification and regression techniques.

	Test Error		
Method	**VSA**	**LOOCV**	**10-fold CV**
Multinomial Logistic Regression	67.27%	Didn't Converge	49.5%
KNN	51.59%	47.25%	48.80%
LDA	49.77%	49.08%	49.70%
QDA	52.95%	53.36%	53.84%
MclustDA	45.50%	49.22%	47.27% (VVE Model = ellipsoidal, equal orientation)
EDDA	50.92%	54.18%	53.16%
Evolutionary Tree	46.95%	Didn't converge	48.46%
Random Forest	50.45% with m=2	NA	NA
Regular Tree	Pruned : 55.6% Unpruned:51.36%	NA	NA
Bagging	53.64%	NA	NA
SVM's	Linear: 51.18% Radial : 44.77% Polynomial: 45%		

Interpretation:

- None of these methods gave us an acceptable error rate. Most of them were near 50% which means that we might as well have flipped a coin to get the result.
- The LOOCV approach didn't converge for the multinomial logistic regression and the evolutionary tree.
- Because of the way KNN, LDA, QDA and the Mclust model works, the only two predictor variables that were used are the continuous variables: wife's age and number of children.
- The two biggest issues that I assume caused such a high error rate:
    - The presence of 7 categorical variables as predictor variables. Most of the models suggested by the problem use continuous variables as predictors
    - A three-way classification problem. Binary classification was the main focus of our studies thus far and this proved to be difficult.

I dived into researching this data set a little more. In my research I came across another study: "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms" that used this data set in their study. According to their findings, "The most difficult to classify are Contraceptive Method Choice…with minimum error rates of 0.4%."

This corresponds to my findings as well.

Thus in my second attempt, I changed the classification to a binary system where I grouped short term and long term users of contraception into a single group of users. The error rates dropped dramatically.

	Test Error		
Method	**VSA**	**LOOCV**	**10-fold CV**
glm	28.86%	32.72%	32.45%
KNN	32.27%	31.23%	31.43%
LDA	32.27%	32.93%	33.4%
QDA	33.18%	35.71%	35.36%
MclustDA	31.85%	30.89%	29.09% (VVE Model = ellipsoidal, equal orientation)
EDDA	Didn't converge	Didn't converge	31.59%
Evolutionary Tree	29.77%	Didn't converge	28.17%

Regular Tree	31.82% (Both pruned and unpruned)	NA	NA
Random Forest	29.55% for m =2	NA	NA
Bagging	32.73%	NA	NA
SVM	Linear: 32.95% Radial: 27.95% Polynomial: 32.05%		32.38%

The error rates definitely improved after using a binary classification. The lowest error rate for the validation set approach was achieved by the glm model which makes sense because glm function makes use of all the predictor variables as compared to the rest. The evolutionary tree performed quite well for the 10-fold CV, achieving an error rate of 28.17%.

Evolutionary Tree algorithm:

The evtree package interested me because in my research I found my dataset built into this package. This was one of the data sets they analyzed. The evtree (Evolutionary learning of Globally Optimum Classification and Regression Trees) was introduced as an alternative to the CART algorithm which uses a recursive partitioning method that builds models in a forward stepwise search. The CART algorithm is efficient; however, it only produces locally optimal results since the splits are chosen to maximize homogeneity at the next step.

A pseudo code of the evtree algorithm is as follows:

```
• Initialize the Population
• Evaluate each individual
• While (termination condition not satisfied) do:
    • Select parents
    • Alter selected individuals through variation operators
        • Variation operators include splitting, pruning and crossovers.
    • Evaluate new solutions
    • Select survivors for the next generation
        • Uses a deterministic crowding approach
• Terminates if the best 5% of the trees stabilizes for 100 iterations but not before 1000
  iterations.
```

Conclusion

The findings of this study are still ongoing. The question that we were trying to address – predicting the choice of contraceptive use depended mostly on number of children and wife’s age. At the same time, these were the two numerical predictor variables so there really is a question of how effective the categorical variables are in predicting the response variable. I am curious to find out how better to use categorical predictor variables in terms of choosing models and modelling techniques.

References

"Indonesia | Data". *Data.worldbank.org*. N.p., 2016. Web. 8 Apr. 2016.

Grubinger, Thomas, Achim Zeileis, and Karl-Peter Pfeiffer. "Evtree : Evolutionary Learning Of Globally Optimal Classification And Regression Trees In R". *Journal of Statistical Software* 61.1 (2014): n. pag. Web.

Lim, Tjen-Sien, Wei-Yin Loh, and Yu-Shan Shih. "A Comparison Of Prediction Accuracy, Complexity, And Training Time Of Thirty-Three Old And New Classification Algorithms". *Machine Learning*40.3 (2000): 203-228. Web. 8 Apr. 2016.