Joelle Strom
STAT 410
Final Project
April 29, 2021

Inference and Predictions of the Contraceptive Method Choice Data Set

Introduction: The data set used in this report is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The survey collected demographic and socio-economic characteristics of married women who were not knowingly pregnant at the time of survey, as well as their choice of contraceptive method. It is currently well understood that contraceptive use to prevent unplanned pregnancies benefits women, their immediate families, and the greater society by reducing maternal deaths and child-rearing costs and improving women's education and capacity to enter and remain in the workforce. An analysis of the factors that impact contraceptive use would be helpful to medical and public health professionals by informing future marketing campaigns, research, and patient services.

Previous research incorporating this data set have employed it as a "real-world" test data set to evaluate various predictive methods, such as decision trees (Harris, 2001; Ray & Page, 2005) or association rules (Balcázar). The paper that most informed this report was the research study by Lim, Loh, and Shih (2000), where 33 different prediction methods were tested and compared, including various decision tree algorithms, neural nets, and several statistical models such as Polyclass and polychtomous logistic regression.

As much of the existing research focused on the utilization of this data set to demonstrate efficacy of prediction methods, this report will switch the focus back to the data set and attempt to perform inference on the explanatory variables. In addition, a comparison between the popular random forest model and polyclass will be performed. Polyclass was indicated as one of the top performing models in the previously mentioned paper, but this paper was published before the random forest model became widely used. As such, it was not mentioned as one of the decision tree models and it will be interesting to see its prediction ability compared to an already identified effective method for this data set.

Data and Methods: The data set contains an ordinal response variable (contraceptive method, 1 = none, 2 = short-term, 3 = long-term) and nine explanatory variables and/or confounders. These include 2 numerical variables (Age and Number of Children), 4 nominal-categorical variables (Husband Occupation, Religion, Employment, and Media Exposure), and 3 ordinal-categorical variables (Education, Husband Education, and Standard of Living, each with 4 levels, 1 = Low to 4 = High). There are 1473 observations in total and no missing data. Religion, Employment, and Media Exposure are binary variables: for Religion, 0 = Not Islam and 1 = Islam; for Employment, 0 = Not Employed and 1 = Employed; for Media Exposure, 0 = Good and 1 = Not Good. The Husband Occupation variable has 4 categories, but no further information is provided about what these categories mean, so this variable was left out of models designed for interpretable parameters. Ordinal variables are treated as numerical, according to the argument by Pasta (2009). Data obtained from the UCI Machine Learning Repository (Dua & Graff, 2019).

Previous studies have either compressed the response variable into a binary response (0=None, 1=Some form of contraception) or have treated the three categories as nominal via polychtomous logistic regression. As such, the starting point for this report was to attempt to fit a

Proportional Odds (PO) model to account for the ordinal nature of the response variable. The Purposeful Selection process presented by Agresti (2018) was used to find an appropriate model. However this model did not stand up to the proportional odds assumption, so this was followed by the purposeful selection process for the Baseline Category Logit (BCL) model and binary logistic regression using the simplified response variable for comparison. All significance tests for model comparison were based on Likelihood Ratio Tests.

The random forest and polyclass algorithms were implemented “out-of-the-box”, with no intentional selection of variables or tuning parameters, as a baseline comparison between the two methods. Both models were again implemented on the dataset with the binary response variable instead of the original response variable.

Results: As mentioned above, the PO model turned out to be a poor fit for the data, as the proportional odds assumption was not met ($\chi^2 = 1955$, p-value < 0.0001). The variables included in the final PO model were Age, Children, Education, Standard of Living, Media Exposure, and an interaction term for Age and Children. This tells us that these variables likely have a significant effect on the choice of contraceptive method, even if we do not retain this model. The non-parallel cumulative model also fit poorly, with warnings provided by R that too many predicted values were close to 0 or 1.

Thus, the BCL model was the next choice to try to obtain some information about significant effects. Although it ignores the ordinal nature of the response, it can still provide relevant information. The response variable was re-leveled so that the reference group was women with no contraceptive use. After purposeful selection, the variables remaining in the model were Age, Children, Education, Standard of Living, and the Age:Children interaction term. Coefficients are shown in the appendix but will be summarized here. The Age and Age:Children effects considered together show that the odds of choosing short-term or long-term contraceptives over none decrease with an increase in age, assuming all other things are equal (this effect is stronger for short-term vs. none than long-term vs. none). An increase in the number of children, the level of education, or standard of living is associated with an increase in odds of both short-term and long-term vs. none. This effect is stronger for long-term vs. none with standard of living and education variables but is stronger for short-term vs. none with the children variable. The binary model shows similar effects for these variables when considering *any* method use vs. none. In addition, this model shows that identification as Muslim or as having poor media exposure is associated with a decrease in the odds of choosing any contraceptive use over none.

A brief foray into predictive models shows that the polyclass model slightly outperforms the random forest model based on the confusion matrices produced by each. A reduction of the response variable to a binary case improved prediction accuracy of both models.

Discussion: Despite the difficulty in finding an appropriate model to fit these data, some clear effects emerge from logistic regression in terms of how contraceptive use relates to these demographic and socio-economic factors which could inform future research. In particular it would be interesting to conduct a subsequent survey which attempts to identify factors more directly impacting contraceptive use, such as cost, side effect concerns, stigma, or lack of perceived benefit. The response variable could also be modified so that the categories are specific classes of contraceptive, rather than broad “short-term” and “long-term”. The comparison of predictive models shows that random forest and polyclass perform similarly, and that reduction of the response variable in this case reduces predictive error.

References

Agresti, A. (2018). *An Introduction to Categorical Data Analysis, 3rd ed.* Wiley.
Balcázar, J. L. Rules with bounded negations and the coverage inference scheme. Dept. LSI, UPC.
Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science.
Harris, Earl. (2001). Information gain versus gain ratio: A study of split method biases. The MITRE Corporation.
Lim, T. S., Loh, W. Y. & Shih, Y. S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, 40, 203-228.
Pasta, D. J. (2009). Learning when to be discrete: Continuous vs. categorical predictors. *SAS Global Forum 2009*, Paper 248-2009.
Ray, S. & Page, D. (2005). Generalized skewing for functions with continuous and nominal attributes. *Proceedings of the 22nd International Conference on Machine Learning.*

Appendix

Example of Data:

```
##   Age Education HusbandEd Children Religion Employment HusbandOcc SOL MediaExp Method
## 1  24         2         3        3        1          1          2   3        0      1
## 2  45         1         3       10        1          1          3   4        0      1
## 3  43         2         3        7        1          1          3   4        0      1
## 4  42         3         2        9        1          1          3   3        0      1
## 5  36         3         3        8        1          1          3   2        0      1
## 6  19         4         4        0        1          1          3   3        0      1
```

Summary of Data:

```
##       Age        Education HusbandEd    Children       Religion Employment
##  Min.   :16.00   1:152     1: 44     Min.   : 0.000   0: 220   0: 369
##  1st Qu.:26.00   2:334     2:178     1st Qu.: 1.000   1:1253   1:1104
##  Median :32.00   3:410     3:352     Median : 3.000
##  Mean   :32.54   4:577     4:899     Mean   : 3.261
##  3rd Qu.:39.00                       3rd Qu.: 4.000
##  Max.   :49.00                       Max.   :16.000

##  HusbandOcc SOL     MediaExp Method
##  1:436      1:129   0:1364   1:629
##  2:425      2:229   1: 109   2:333
##  3:585      3:431            3:511
##  4: 27      4:684
```

Proportional Odds: Models fit in the process of purposeful selection (Method here is an ordered variable where 1 = none, 2 = short-term, and 3 = long-term); final model is Italicized

Formula	Deviance	Log-Likelihood	Df	Significance

Method ~ 1 (Null)	3143	-1571	2944	
				Compared to Null Model
Method ~ Age	3105	-1552	2943	***
Method ~ Children	3132	-1566	2943	***
Method ~ Education	3105	-1552	2943	***
Method ~ HusbandEd	3125	-1562	2943	***
Method ~ Religion	3142	-1571	2943	--
Method ~ Employment	3139	-1569	2943	*
Method ~ SOL	3128	-1564	2943	***
Method ~ MediaExp	3118	-1559	2943	***
Method ~ Age + Children + Education + HusbandEd + Employment + SOL + MediaExp	2948	-1474	2937	
				Compared to the Previous Nested Model
Method ~ Age + Children + Education + HusbandEd + Employment + SOL	2953	-1476	2938	*
Method ~ Age + Children + Education + HusbandEd + Employment	2966	-1483	2939	***
Method ~ Age + Children + Education + HusbandEd	2967	-1483	2940	--
Method ~ Age + Children + Education	2967	-1484	2941	--
Method ~ Age + Children	3035	-1518	2942	***
(Method ~ Age)	(3105)	(-1552)	(2943)	***
Method ~ Age + Children + Education + SOL + MediaExp	2949	-1475	2939	
Method ~ Age + Children + Education + SOL + MediaExp + Religion	2946	-1473	2938	--
				Compared to Main Effects Model
Method ~ Age + Children + Education + SOL + MediaExp + Age:Children	*2888*	*-1444*	*2938*	***

Method ~ Age + Children + Education + SOL + MediaExp + Age:Education	2931	-1466	2938	***
Method ~ Age + Children + Education + SOL + MediaExp + Children:Education	2937	-1469	2938	***

Summary of PO Model 10: This model contains all main effects that were significant when compared to the null model individually

```
## Call:
## vglm(formula = Method ~ Age + Children + Education + HusbandEd +
##     Employment + SOL + MediaExp, family = cumulative(parallel = TRUE),
##     data = dat)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -0.487457   0.333310  -1.462 0.143611
## (Intercept):2  0.540237   0.333399   1.620 0.105149
## Age            0.087190   0.008315  10.485  < 2e-16 ***
## Children      -0.278720   0.028842  -9.664  < 2e-16 ***
## Education     -0.304068   0.067338  -4.516 6.32e-06 ***
## HusbandEd     -0.044960   0.082295  -0.546 0.584842
## Employment    -0.115617   0.117977  -0.980 0.327091
## SOL           -0.198897   0.059097  -3.366 0.000764 ***
## MediaExp       0.506745   0.226944   2.233 0.025555 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 2947.887 on 2937 degrees of freedom
##
## Log-likelihood: -1473.943 on 2937 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##       Age  Children  Education  HusbandEd Employment       SOL  MediaExp
## 1.0911045 0.7567518 0.7378104 0.9560356 0.8908167 0.8196347 1.6598802
```

Summary of PO Model 16: Final main effects model after backward elimination

```
## Call:
## vglm(formula = Method ~ Age + Children + Education + SOL + MediaExp,
##     family = cumulative(parallel = TRUE), data = dat)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -0.682406   0.275090  -2.481 0.013114 *
## (Intercept):2  0.344514   0.274694   1.254 0.209778
## Age            0.087863   0.008284  10.606  < 2e-16 ***
```

```
## Children      -0.280964   0.028520  -9.852  < 2e-16 ***
## Education     -0.321412   0.057980  -5.543 2.97e-08 ***
## SOL           -0.201930   0.058106  -3.475 0.000511 ***
## MediaExp       0.522209   0.226165   2.309 0.020945 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 2949.235 on 2939 degrees of freedom
##
## Log-likelihood: -1474.617 on 2939 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##       Age  Children Education       SOL  MediaExp
## 1.0918388 0.7550552 0.7251244 0.8171523 1.6857476
```

Summary of PO Model 18: Final model, including interaction

```
## Call:
## vglm(formula = Method ~ Age + Children + Education + SOL + MediaExp +
##     Age:Children, family = cumulative(parallel = TRUE), data = dat)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  1.387784   0.385245   3.602 0.000315 ***
## (Intercept):2  2.453290   0.388979   6.307 2.84e-10 ***
## Age            0.027745   0.011096   2.500 0.012406 *
## Children      -1.126841   0.114339  -9.855  < 2e-16 ***
## Education     -0.340401   0.058766  -5.792 6.94e-09 ***
## SOL           -0.183753   0.059375  -3.095 0.001970 **
## MediaExp       0.553388   0.231147   2.394 0.016661 *
## Age:Children   0.022490   0.002937   7.657 1.90e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 2888.497 on 2938 degrees of freedom
##
## Log-likelihood: -1444.248 on 2938 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##          Age     Children    Education          SOL     MediaExp Age:Children
##    1.0281335    0.3240554    0.7114852    0.8321410    1.7391358    1.0227453
```

Summary of Non-parallel Cumulative Logit Model:

```
## Call:
## vglm(formula = Method ~ Age + Children + Education + SOL + MediaExp +
##     Age:Children, family = cumulative, data = dat)
##
## Coefficients:
##                  Estimate Std. Error  z value Pr(>|z|)
## (Intercept):1   2.397e+00  1.365e-06  1755608   <2e-16 ***
## (Intercept):2   6.949e-01  1.299e-06   535134   <2e-16 ***
## Age:1           1.202e-02  3.676e-08   327031   <2e-16 ***
## Age:2           3.101e-02  3.410e-08   909200   <2e-16 ***
## Children:1     -1.030e+00  5.969e-07 -1725443   <2e-16 ***
## Children:2     -8.955e-01  5.930e-07 -1509929   <2e-16 ***
## Education:1    -4.882e-01  4.660e-07 -1047831   <2e-16 ***
## Education:2    -4.693e-02  4.960e-07   -94617   <2e-16 ***
## SOL:1          -2.100e-01  1.639e-07 -1281079   <2e-16 ***
## SOL:2          -7.602e-02  1.626e-07  -467603   <2e-16 ***
## MediaExp:1      3.670e-01  3.605e-07  1018155   <2e-16 ***
## MediaExp:2      3.583e-01  3.378e-07  1060510   <2e-16 ***
## Age:Children:1  2.017e-02  2.203e-08   915478   <2e-16 ***
## Age:Children:2  1.887e-02  2.136e-08   883126   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 4843.169 on 2932 degrees of freedom
##
## Log-likelihood: NA on 2932 degrees of freedom
##
## Number of Fisher scoring iterations: 2
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2', 'Age:2', 'Children:1', 'Education:1', 'SOL:1', 'MediaExp:2',
'Age:Children:2'
```

Baseline Category Logit: Formula began with the final PO model and was modified as below (Method here is an unordered factored variable with the same levels as above; 1 = none is the reference group); final model is Italicized

Formula	**Deviance**	**Log-Likelihood**	**Df**	**Significance**
Method ~ Age + Education + Children + SOL + MediaExp + Age:Children	2739	-1369	2932	
Method ~ Age + Education + Children + SOL + Age:Children	*2742*	*-1371*	*2934*	-- (compared to above model)
Method ~ Age + Education + Children + SOL	2803	-1401	2936	*** (compared to above model)

Method ~ Age + Education + Children + Age:Children	2762	-1381	2936	*** (compared to model 2)
Method ~ Age + Children + Age:Children	2945	-1472	2938	*** (compared to above model)
				Compared to Model 2
Method ~ Age + Children + Education + SOL + Age:Children + HusbandEd	2742	-1371	2932	--
Method ~ Age + Children + Education + SOL + Age:Children + Religion	2738	-1369	2932	--
Method ~ Age + Children + Education + SOL + Age:Children + Employment	2741	-1370	2932	--

Summary of BCL Model 1: Formula is identical to the final PO model

```
## Call:
## vglm(formula = Methodfac ~ Age + Education + Children + SOL +
##     MediaExp + Age:Children, family = multinomial, data = dat)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  -2.043011   0.491726  -4.155 3.26e-05 ***
## (Intercept):2  -6.274504   0.625796 -10.026  < 2e-16 ***
## Age:1          -0.033091   0.014623  -2.263 0.023644 *
## Age:2           0.014396   0.015863   0.908 0.364129
## Education:1     0.371748   0.074249   5.007 5.53e-07 ***
## Education:2     0.910413   0.095892   9.494  < 2e-16 ***
## Children:1      1.388055   0.151499   9.162  < 2e-16 ***
## Children:2      1.157172   0.170843   6.773 1.26e-11 ***
## SOL:1           0.203234   0.072916   2.787 0.005316 **
## SOL:2           0.358870   0.095296   3.766 0.000166 ***
## MediaExp:1     -0.515094   0.280918  -1.834 0.066712 .
## MediaExp:2     -0.379432   0.387727  -0.979 0.327775
## Age:Children:1 -0.028312   0.003975  -7.122 1.06e-12 ***
## Age:Children:2 -0.021608   0.004231  -5.108 3.26e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])
##
## Residual deviance: 2738.708 on 2932 degrees of freedom
##
## Log-likelihood: -1369.354 on 2932 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
```

```
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2', 'Age:Children:1'
##
##
## Reference group is level  3  of the response
```

Summary of BCL Model 2: Final BCL model

```
## Call:
## vglm(formula = Methodfac ~ Age + Education + Children + SOL +
##     Age:Children, family = multinomial, data = dat)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1  -2.182800   0.486952  -4.483 7.37e-06 ***
## (Intercept):2  -6.383614   0.619143 -10.310  < 2e-16 ***
## Age:1          -0.034811   0.014587  -2.386   0.0170 *
## Age:2           0.013115   0.015835   0.828   0.4075
## Education:1     0.402593   0.072415   5.560 2.70e-08 ***
## Education:2     0.933752   0.094098   9.923  < 2e-16 ***
## Children:1      1.383802   0.151119   9.157  < 2e-16 ***
## Children:2      1.152731   0.170634   6.756 1.42e-11 ***
## SOL:1           0.227425   0.071649   3.174   0.0015 **
## SOL:2           0.377680   0.093743   4.029 5.60e-05 ***
## Age:Children:1 -0.028251   0.003966  -7.124 1.05e-12 ***
## Age:Children:2 -0.021520   0.004225  -5.093 3.52e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])
##
## Residual deviance: 2742.382 on 2934 degrees of freedom
##
## Log-likelihood: -1371.191 on 2934 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2', 'Age:Children:1'
##
##
## Reference group is level  3  of the response
```

Binomial Logistic: Models fit in the process of purposeful selection (Method here is a binary variable, where 0 = none and 1 = any method); final model is Italicized

Formula	Deviance	AIC	Df	Significance
Method ~ 1 (Null)	2011	2012	1472	
				Compared to Null Model
Method ~ Age	1998	2002	1471	***
Method ~ Children	1989	1993	1471	***
Method ~ Education	1922	1923	1471	***
Method ~ HusbandEd	1974	1978	1471	***
Method ~ Religion	2003	2007	1471	**

Method ~ Employment	2008	2012	1471	-- (<0.2)
Method ~ SOL	1973	1977	1471	***
Method ~ MediaExp	1980	1984	1471	***
				Compared to Next Nested Model
Method ~ Age + Education + SOL + HusbandEd + MediaExp + Children + Religion + Employment	1772	1790	1464	--
Method ~ Age + Education + SOL + HusbandEd + MediaExp + Children + Religion	1773	1789	1465	*
Method ~ Age + Education + SOL + HusbandEd + MediaExp + Children	1779	1793	1466	***
Method ~ Age + Education + SOL + HusbandEd + MediaExp	1894	1906	1467	--
Method ~ Age + Education + SOL + HusbandEd	1897	1907	1468	--
Method ~ Age + Education + SOL	1897	1905	1469	***
Method ~ Age + Education	1911	1917	1470	***
Method ~ Age + Education + SOL + MediaExp + Children + Religion	1773	1787	1466	
Method ~ Age + Education + SOL + MediaExp + Children + Religion + Age:Children	*1718*	*1734*	*1465*	*** (compared to above model)

Summary of Logistic Model 10: Contains all main effects that were significant when compared to null model individually

```
## Call:
## glm(formula = Bin ~ Age + Education + SOL + HusbandEd + MediaExp +
##     Children + Religion + Employment, family = binomial, data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1637  -1.0795   0.6747   0.9462   2.1115
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.193026   0.439001  -0.440 0.660158
## Age         -0.082314   0.009382  -8.773  < 2e-16 ***
## Education    0.525115   0.076995   6.820  9.1e-12 ***
## SOL          0.255535   0.066515   3.842 0.000122 ***
## HusbandEd    0.001260   0.091855   0.014 0.989052
## MediaExp    -0.478374   0.247976  -1.929 0.053717 .
## Children     0.340041   0.034410   9.882  < 2e-16 ***
## Religion    -0.400578   0.170992  -2.343 0.019146 *
## Employment   0.123195   0.133007   0.926 0.354329
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2010.5  on 1472  degrees of freedom
## Residual deviance: 1772.4  on 1464  degrees of freedom
## AIC: 1790.4
##
## Number of Fisher Scoring iterations: 4
```

Summary of Logistic Model 17: Full main effects model after backward elimination

```
## Call:
## glm(formula = Bin ~ Age + Education + SOL + MediaExp + Children +
##     Religion, family = binomial, data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1605  -1.0872   0.6794   0.9435   2.1284
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.074795   0.384512  -0.195  0.84577
## Age         -0.082992   0.009356  -8.871  < 2e-16 ***
## Education    0.523775   0.066603   7.864 3.72e-15 ***
## SOL          0.252755   0.065361   3.867  0.00011 ***
## MediaExp    -0.487846   0.247176  -1.974  0.04842 *
## Children     0.343435   0.034148  10.057  < 2e-16 ***
## Religion    -0.397899   0.170938  -2.328  0.01993 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2010.5  on 1472  degrees of freedom
## Residual deviance: 1773.3  on 1466  degrees of freedom
## AIC: 1787.3
##
## Number of Fisher Scoring iterations: 4
```

Summary of Logistic Model 18: Final model, including interaction

```
## Call:
## glm(formula = Bin ~ Age + Education + SOL + MediaExp + Children +
##     Religion + Age:Children, family = binomial, data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6322  -1.0267   0.5972   0.9260   2.0850
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.316968   0.496183  -4.670 3.02e-06 ***
## Age          -0.019848   0.012314  -1.612 0.106991
## Education     0.546834   0.068023   8.039 9.06e-16 ***
## SOL           0.239642   0.067644   3.543 0.000396 ***
## MediaExp     -0.502717   0.255611  -1.967 0.049214 *
## Children      1.262868   0.133432   9.465  < 2e-16 ***
## Religion     -0.359972   0.171397  -2.100 0.035709 *
## Age:Children -0.024425   0.003341  -7.311 2.66e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2010.5  on 1472  degrees of freedom
## Residual deviance: 1718.3  on 1465  degrees of freedom
## AIC: 1734.3
##
## Number of Fisher Scoring iterations: 4
```

Confusion Matrix from Polyclass: This model was fit with the 3-level response variable

```
##         predicted
## observed   1   2   3
##        1 422  41 166
##        2  78 133 122
##        3 137  80 294
```

Random Forest Output: This model was fit with the 3-level response variable

```
## Call:
##  randomForest(formula = factor(Method) ~ Age + Children + Education +
HusbandOcc + HusbandEd + SOL + MediaExp + Religion + Employment,      data =
dat, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 45.62%
## Confusion matrix:
##     1   2   3 class.error
## 1 415  56 158   0.3402226
```

```
## 2  93 132 108   0.6036036
## 3 161  96 254   0.5029354
```

Confusion Matrix from Polyclass: This model was fit with the binary response variable

```
##          predicted
## observed   0   1
##        0 340 289
##        1 114 730
```

Random Forest Output: This model was fit with the binary response variable

```
## Call:
##  randomForest(formula = factor(Bin) ~ Age + Children + Education + 
HusbandOcc + HusbandEd + SOL + MediaExp + Religion + Employment,      data = 
dat, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 30.55%
## Confusion matrix:
##     0   1 class.error
## 0 338 291   0.4626391
## 1 159 685   0.1883886
```