STAT 637 Mini Project 2

Sean Miner, Stephen Merrill & William White

February 11, 2017

1 Introduction and Exploratory Data Analysis

The goal of this study is to determine if there are any attitudes or actions within the LDS Church that influence a member's chance of getting and staying married. Analysis was done on a survey given to two geographical groups of church membership (Salt Lake City and San Francisco) from 1967-1969. This dataset contains responses to over 250 different questions, so the first step of the analysis was to determine which attitudes and actions were of interest. Some 15-20 variables, each with multiple factor levels, were selected, and the data was then subset accordingly. Additionally, the response variable, marriage, was clarified to be binary with a positive result assigned to those reporting as married or widowed, and a negative result given to those that were single or divorced. Figure 1 shows a preliminary view of a couple variables of interest.

2 Model Selection

The final model was selected using a two-step variable reduction process. First, we used a LASSO algorithm to reduce the number of variables under consideration by imposing a constraint that shrunk some coefficients to zero. Then, we performed a best-subset selection algorithm that evaluated all possible combinations of the remaining variables and ranked them according to their BIC. We considered a few of the highest-ranked models and were able to identify one that meets the goal of the study and also accurately models the data. Finally, we determined which link function on the binary response to use by comparing the resulting residuals and model deviances.

Logit	Probit	Cloglog	Loglog
764.74	765.87	769.41	764.54

Table 1: Deviances of Model under different link functions

The logit and log-log models have the lowest deviances, and we selected the logit link for interpretability. The model is given as follows:

$$p = Prob(Y = 1)$$

$$Y_i \overset{ind}{\sim} Bernoulli(p_i)$$

$$log(\tfrac{p_i}{1-p_i}) = x_i'\beta \implies p_i = \tfrac{exp(x_i'\beta)}{1+exp(x_i'\beta)}$$

- β_0: Intercept
- β_1: Effect of number of LDS friends
- β_2: Effect of age
- β_3: Effect of attitude towards a non-LDS marriage
- β_4: Effect of attitude towards not attending sacrament meeting

3 Final Model and Parameter Inference

The final model includes both continuous and categorical variables. Summary results are given in Table 2.

Interpreting the results can be done as follows: β_1 and β_2 are treated as continuous variables and can be interpreted as on average, for every one unit increase in x_i, the log-odds ratio increases by β, or being married is e^{β_j} more likely. β_3 and β_4 are treated as categorical variables and can be interpreted as if that situation is present, then on average, the log-odds ratio increases by β, or being married is e^{β_j} more likely.

The confidence intervals are shown in Table 4. The coefficients and confidence intervals were transformed so that they measure the expected effect on the odds increase/decrease of marriage. For example, we would

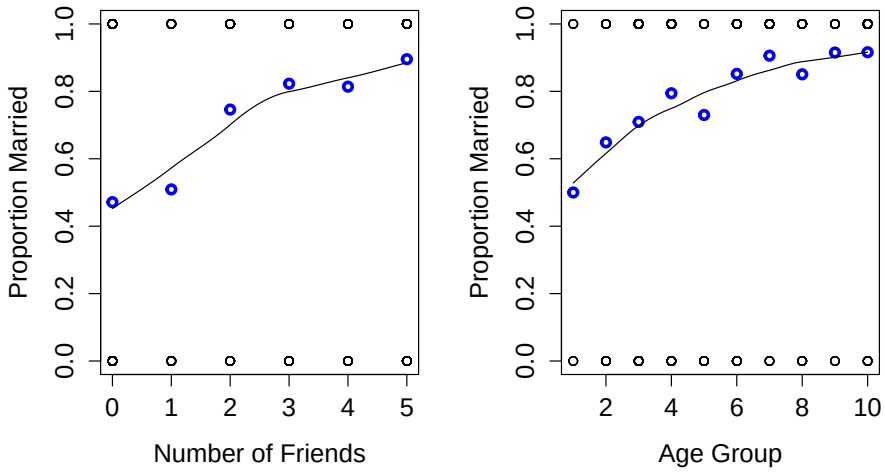

Figure 1: Proportion of married individuals at each level of friends and age. An upward trend can be seen with both variables.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.3675	0.3137	-4.36	0.0000
Number Friends LDS	0.4941	0.0600	8.24	0.0000
Age	0.2092	0.0373	5.61	0.0000
Marriage Not LDS [2]	-0.9751	0.3657	-2.67	0.0077
Marriage Not LDS [3]	-2.0157	0.5398	-3.73	0.0002
Marriage Not LDS [4]	-1.0804	0.3830	-2.82	0.0048
Marriage Not LDS [5]	-0.9809	0.3967	-2.47	0.0134
Not Attndng Scrmt [2]	0.9770	0.4026	2.43	0.0152
Not Attndng Scrmt [3]	2.0908	0.8583	2.44	0.0149
Not Attndng Scrmt [4]	1.0185	0.3993	2.55	0.0108
Not Attndng Scrmt [5]	0.7710	0.4224	1.83	0.0680

Table 2: Final Model

expect that someone who increased their number of Mormon friends by one among their top 5 closest friends would be 1.64 times more likely to be married. Further we are 95% confident that the true increase of the likelihood of marriage for one increased Mormon friend to be in between 1.46 and 1.84 times higher.

From Figure 2 we see some of the predicted probabilities for a variety of factors. For these predicted probabilities, we chose to fix the opinion of the importance of sacrament meeting attendance as "can't decide". As the age of someone increases, we see a general increase in the expected probability of marriage. The points right next to each other, show that there is a increase of the probability of marriage when one has more of their top 5 friends as Mormon. Figure 1 also shows the effect of the opinion on whether marrying someone who is not LDS matters. If someone thinks it scarcely matters, our model shows they are more likely to be married. If they are undecided, then our model shows they are less likely to be married.

In Figure 3, we show the effect of friends, age and one's opinion of the importance of sacrament meeting attendance on the likelihood of being married. For this figure, we fixed the opinion of how important it is to not be married to someone who is not LDS as "can't decide". The effects of age and friends are the same. We see two important opinions on the importance of sacrament meeting attendance are "Scarcely Matters" and "Can't Decide". If someone put that it scarcely matters, we would expect that the probability of marriage to decrease. If someone put that they couldn't decide on how important it was, we would expect the probability of them being married to increase.

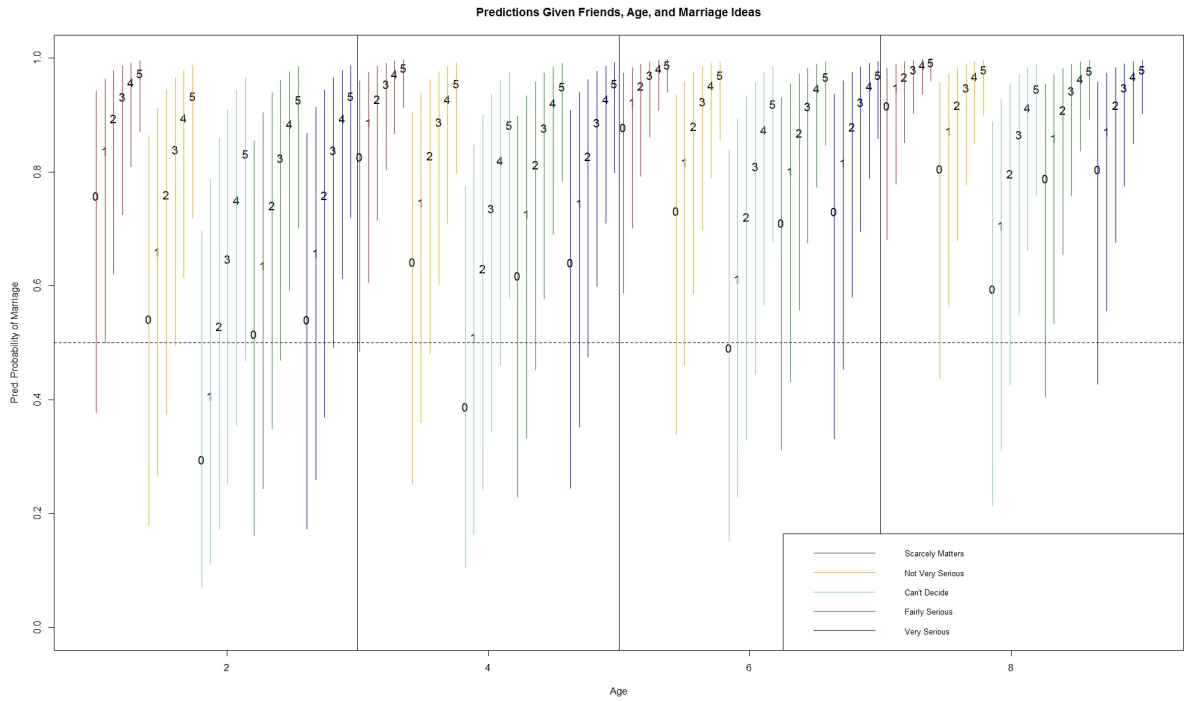

Figure 2: Predicted probability of marital status by age, number of friends and belief about interfaith marriage. Numbers on each line placed at true proportion of married members and represent number of friends.

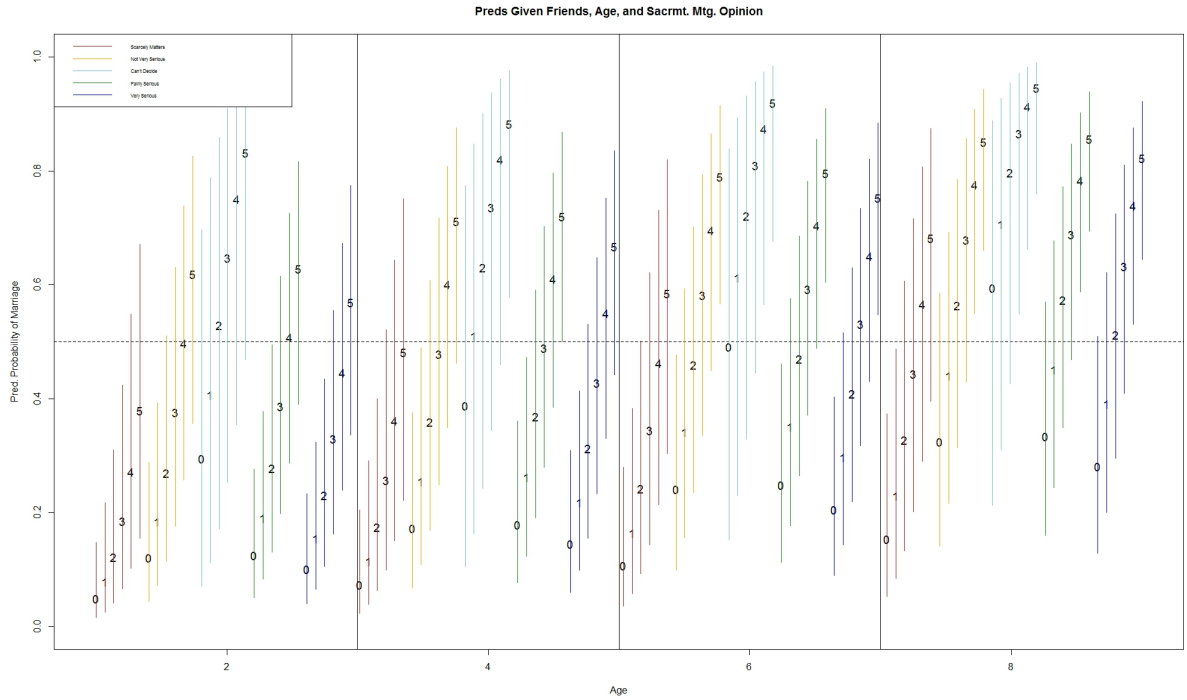

Figure 3: Predicted probability of marital status by age, number of friends and belief about sacrament meeting importance. Numbers on each line placed at true proportion of married members and represent number of friends.

1	Scarcely Matters at all
2	Not Very Serious
3	Can't Decide
4	Fairly Serious
5	Very Serious

Table 3: Code for Opinion Questions

	Lower	Coefficients	Upper
(Intercept)	0.14	0.25	0.47
FRIEND	1.46	1.64	1.84
AGE	1.15	1.23	1.33
Marriage Not LDS [2]	0.18	0.38	0.77
Marriage Not LDS [3]	0.05	0.13	0.38
Marriage Not LDS [4]	0.16	0.34	0.72
Marriage Not LDS [5]	0.17	0.37	0.82
Not Attndng Scrmt [2]	1.21	2.66	5.85
Not Attndng Scrmt [3]	1.50	8.09	43.51
Not Attndng Scrmt [4]	1.27	2.77	6.06
Not Attndng Scrmt [5]	0.94	2.16	4.95

Table 4: Confidence intervals on the effect on odds (transformed)

4 Model Evaluation

Table 5 shows a confusion matrix. From this table, we see that our model fit is fairly good. We have a sensitivity rate of 0.85 and a specificity rate of 0.671. We also looked at the ROC curve in Figure 5 to evaluate the true and false positive rate. From the ROC curve we can see that our model does fairly well with the true positive rate and the false positive rate.

We also fit a model, using the same variables, on a training dataset and used it to predict on a testing dataset. We found the misclassification rate to be 18.4%.

The QQ-Plot in Figure 4 shows the normality of our residuals. We'd expect our model to have some deviation from normality since we have a lot of possible combinations of factors with limited data, but even so, the deviation from normality isn't too bad.

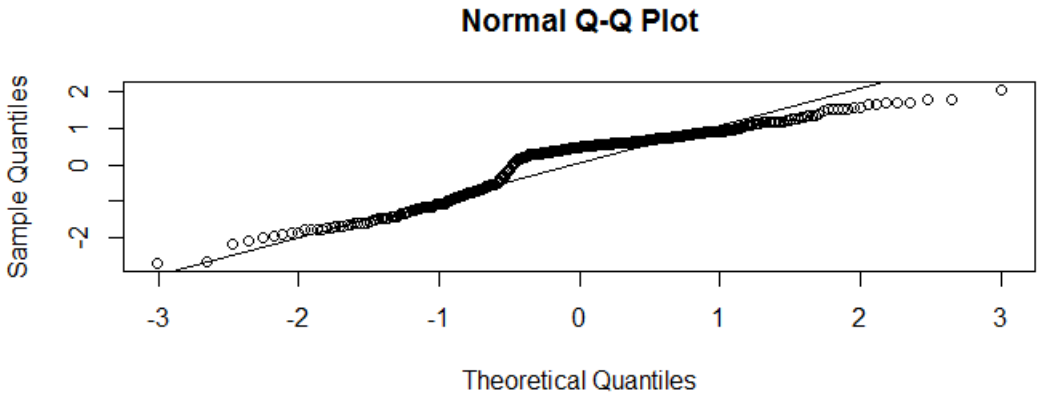

Figure 4: QQplot

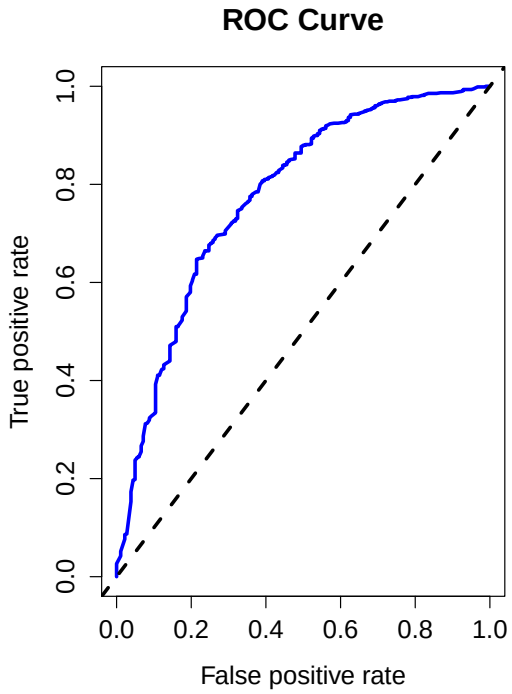

Figure 5: ROC Curve for the proposed model. The model does well maximizing the area under the curve.

	Pred. Success	Pred. Failure
Actual Success	731	129
Actual Failure	26	53

Table 5: Confusion Matrix

5 Conclusion

Based on our assessment we would recommend this model to predict whether members of the LDS Church are married or not. This model can also be used to infer relationships between probability of being married and a member's number of close LDS friends, age group, belief on interfaith marriage and belief on the importance of sacrament meeting attendance. We feel our model may have flaws but it is a fairly simple model that can be understood and interpreted with ease. A more complicated model could be built with more variables or more advanced statistical methods if prediction is deemed of greater worth than inference. In a future study, we would like to use a multinomial response to separate out divorce and widow rates. We would also like to use age as a truly continuous variable instead of using five year wide bins.