How Couples Meet and Stay Together

Catherine Weldon, Angie Pinilla, and Josh Robbins

Abstract

Relationship fulfillment is considered one of the most important contributors to happiness. Factors that determine the success of relationships, specifically romantic relationships, are usually looked at from a psychological and qualitative perspective. Through the large data set called How Couples Meet and Stay Together, we set out to discover the quantitative factors that make a successful relationship. The ability to create a model that quantitatively determines the factors that are important in a successful relationship could change the way we seek partners and maintain relationships. This model could serve as evidence for advice psychologists and couple's counselors give and could also help online dating services improve algorithms that bring people together.

The Data Set

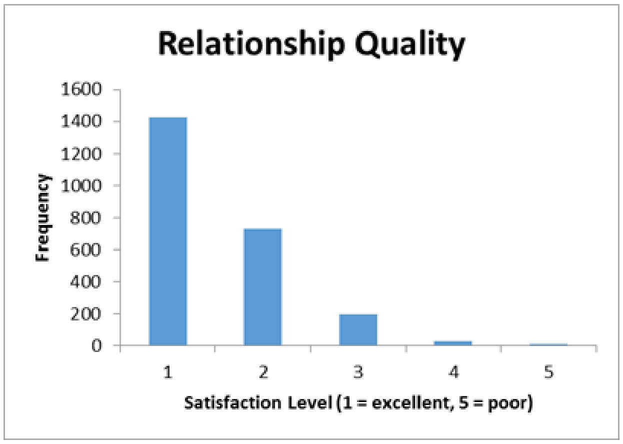
The data set, How Couples Meet and Stay Together, includes a large amount of factors that could determine the success of a relationship/marriage. The data comes from a survey that has 5 waves (over 6 years). The first survey has 35 questions -- many questions having multiple parts -- and focuses on demographic information such as income, race, age, sexual orientation, and various other questions that could provide insight into how successful a relationship will be.

For the first wave, out of the original 4,002 participants, only 2,996 had data for quality of relationship. For the second wave, only 2,007 of those participants were able to be contacted for the follow-up survey, the third wave had 1,553 respondents, the fourth wave had 1,216 respondents, and the fifth and final wave only had 891 respondents. We can tell this data is missing because each wave has a field for how many people completed the survey. The reasons for this missing data is either that they no longer wanted to participate in the study, or they were ineligible for follow-up waves because they weren't in their original relationship. There were also text responses in this survey that were missing from the data set so all the analysis will revolve around quantitative data. There doesn't appear to be any corrupted data as all the entries are in the range of possible responses as indicated on the survey. It is not clear if any of these values were every incorrectly inputted or modified. For these reasons, we will assume that the data has not been corrupted.

Objective Measurement

Originally, we wanted to make our objective measurement the perceived relationship quality of the couple. We wanted to focus on the first wave and find a relationship between demographic information and a field where participants rate their relationship quality on a scale of 1 to 5 (with 1 being excellent and 5 being poor). After seeing the distribution of this field, we determined that it was not a good success metric as nearly all couples rated their relationship as very good or excellent.

Figure I: The skewness and subjectiveness of Relationship Quality made it a poor choice for our response variable



We also thought this measurement was too subjective and could be easily influenced by mood or individual personality.

In order to gain more objective and informative results, the metric we ended up using to determine the success of each relationship is if the couple broke up within the six year study. Each wave has a field that indicates if the couple has broken up, and through this data, we can predict whether or not a couple will still be together in six years.

Cleaning The Data

Because there were over 500 fields in the initial dataset -- many of these being repetitive or irrelevant to our objective -- we omitted the majority of the fields to focus on 120 of the fields. The fields omitted were either other fields recorded in alternate ways or fields that had too small of a sample size (under 30) to provide reliable insight. Other fields omitted were unrelated to our objective. For example, whether or not they included text in their responses was not a field we found relevant to our objective.

Most of the fields were categorical (like race or religion), so our next step was to change these fields to one-hot encoding for each category. We also combined fields that had small samples that could fit in a broader category. For example, the data set included many different websites couples could meet, but because some of these fields had less than 10 respondents, we combined the fields under “met on any online site”. Although there were only 891 respondents in the final wave, this was, for the most part, not because they hadn’t responded to the follow-ups but because they had broken up in a previous waves and no longer qualified for the study. The break-up data was combined for each year in two fields, one indicating if they broke-up or not, and one indicating what year of the study they broke-up (if applicable).

There was also a large amount of missing data. If we did not have data for if the couple stayed together or not, the participant was omitted from the analysis. For other missing data, if the majority of participants refused to answer the question, the field was omitted, and if the participant had refused to answer most of the survey, the participant was omitted. For participants that only refused to answer a few of the questions, we predicted what their answer would have been based on responses to related fields or the general distribution of responses for that field. Overall, there weren’t many missing values to predict so this missing data likely doesn’t have a large effect on the end results.

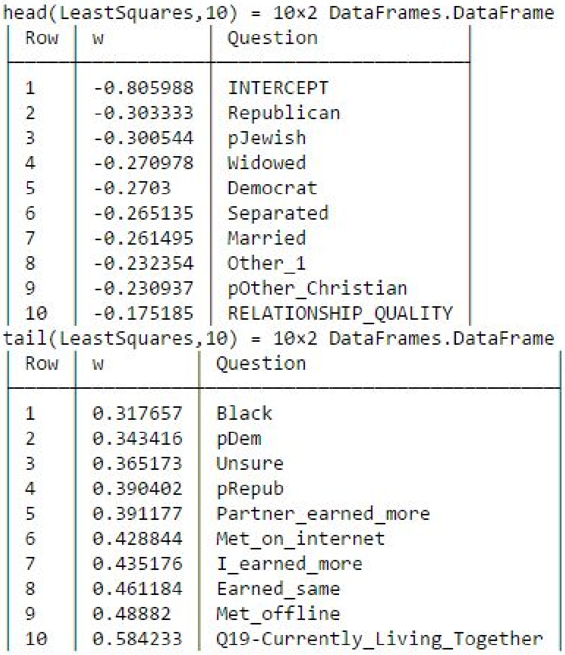
Once the data was sufficiently cleaned, we separated the dataset into a train set (80%) and a test set (20%) so that we could assess the accuracy of our models.

Modeling

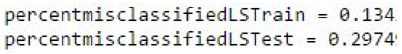
The first models attempted were classification models based on our initial objective (relationship quality). To perform this classification, we first used the following from scikit-learn's modules: a 'Multiclass learning model using One-vs-Rest' (train error-rate = 0.55, test error-rate = 0.56), and then implemented a 'RandomForestClassifier' (train error-rate = 0.01, test error-rate = 0.12) which proved more accurate. These classification models were performed while columns were still in a categorical form and included 158 features. Although the error rates for the Random Forest Classifier were low, it was determined that this resulted from the large skew in perceived relationship quality (most people rated their relationship as excellent). Because of this skew, the objective function used for the rest of the models was whether or not the couple had broken up. These models were also discarded not only because the train and test error rates were very different, but because the data was still categorical and didn't include all the fields that were determined as potentially relevant. From analysis of these models, we decided to further clean the data so ensure future models would be more accurate.

Next, we fit a least squares model to the clean data using Julia's backslash operator to develop a baseline for our future models.

Figure II: Head and tail of Julia output for least squares coefficients, misclassification rates



```
head(LeastSquares,10) = 10×2 DataFrames.DataFrame
│ Row │ w         │ Question             │
├─────┼───────────┼──────────────────────┤
│ 1   │ -0.805988 │ INTERCEPT            │
│ 2   │ -0.303333 │ Republican           │
│ 3   │ -0.300544 │ pJewish              │
│ 4   │ -0.270978 │ Widowed              │
│ 5   │ -0.2703   │ Democrat             │
│ 6   │ -0.265135 │ Separated            │
│ 7   │ -0.261495 │ Married              │
│ 8   │ -0.232354 │ Other_1              │
│ 9   │ -0.230937 │ pOther_Christian     │
│ 10  │ -0.175185 │ RELATIONSHIP_QUALITY │
tail(LeastSquares,10) = 10×2 DataFrames.DataFrame
│ Row │ w        │ Question                     │
├─────┼──────────┼──────────────────────────────┤
│ 1   │ 0.317657 │ Black                        │
│ 2   │ 0.343416 │ pDem                         │
│ 3   │ 0.365173 │ Unsure                       │
│ 4   │ 0.390402 │ pRepub                       │
│ 5   │ 0.391177 │ Partner_earned_more          │
│ 6   │ 0.428844 │ Met_on_internet              │
│ 7   │ 0.435176 │ I_earned_more                │
│ 8   │ 0.461184 │ Earned_same                  │
│ 9   │ 0.48882  │ Met_offline                  │
│ 10  │ 0.584233 │ Q19-Currently_Living_Together │
```

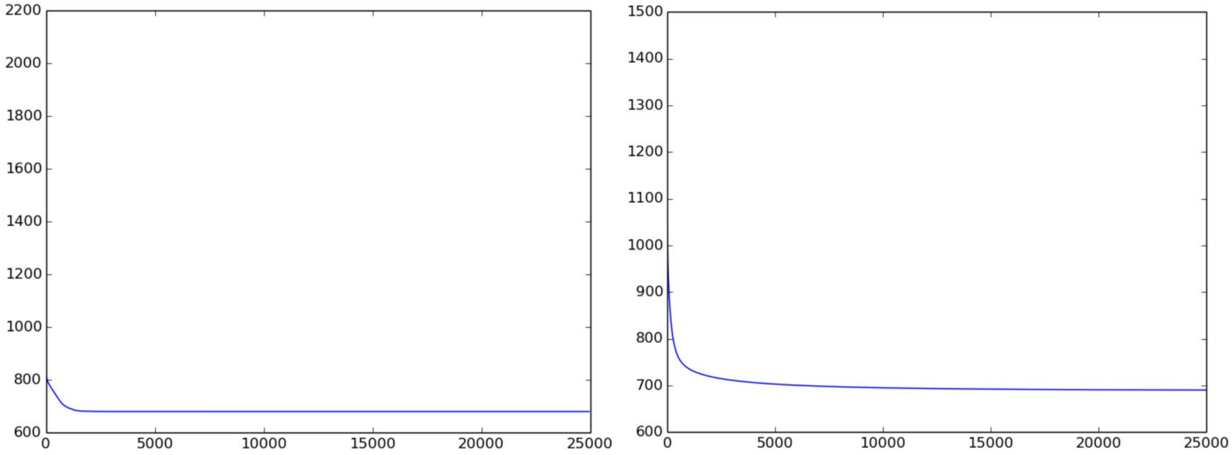


```
percentmisclassifiedLSTrain = 0.134
percentmisclassifiedLSTest = 0.2974
```

This least squares solution performed fairly well, giving us an error rate of 13% on the training set and 30% on the test set. This discrepancy clearly shows that the least squares model is overfitting on the training set.

Because our clean data set had so many fields, most of them being binary, we next tried low rank models to find a sparser solution. First we used the proximal gradient method with both a hinge and a logistic loss function. Initially, the proximal gradient method was having difficulty converging even when using the proxgrad_linesearch function or the Lipschitz constant for our step size. We determined the lack of converging for the proximal gradient method was due to large differences in scaling between the binary and numeric variables; fields such as distance apart ranged from 0 to 9330 miles apart. After scaling down the variables with large values, the proximal gradient method seemed to converge (Figure III).

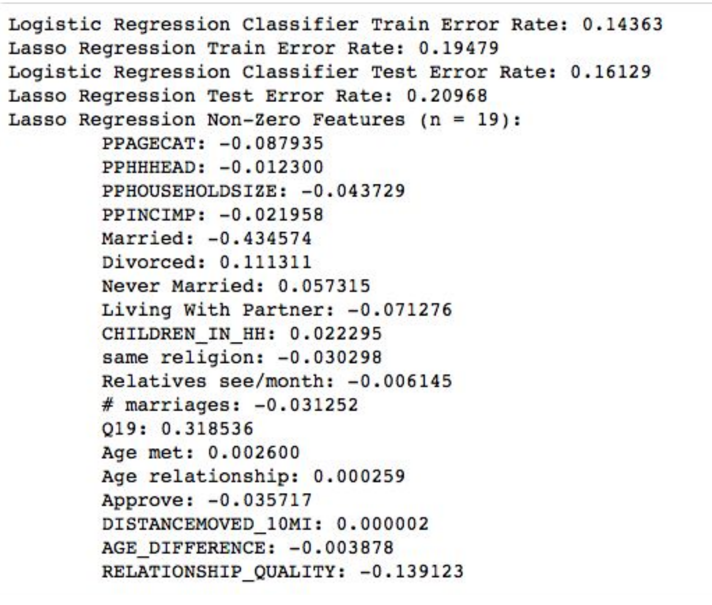
Figure III: Convergence history for the proximal gradient method with hinge loss and logistic loss, respectively



Although the train error rate was relatively good for the proximal gradient methods with both loss functions, the test error rate was not, indicating that the model overfits the data, at least partly a result of our fairly small dataset. Because of this overfitting, these models cannot accurately predict the success of a relationship.

The solutions to these models were still not sparse enough to produce reliable and interpretable results so we turned to the Lasso method in the Scikit Learn Package as a method for variable selection; in the case of lasso, the l_1 penalty has the effect of reducing the magnitudes of some of the coefficients to be exactly zero when the tuning parameter, lambda, is sufficiently large. Hence, the lasso, with 5-fold cross validation on the training set and prediction on the test set's response variable 'Breakup Binary', yielded a sparse model with a subset (p = 19) of predictor variables having non-zero coefficients as listed in Figure IV. For comparison of model accuracy, a logistic regression model, which uses the general method of maximum likelihood rather than least squares to estimate the unknown linear regression coefficients and the l_2 norm as the penalty term, was fit with 5-fold cross-validation on our training data and also used to predict on our test data. Although the train and test error rates for the logistic regression (0.14 and 0.16, respectively) were lower than those for the lasso regression (0.19 and 0.21), the lasso regression produces a model only containing 19 features, which gives us more information about the factors that influence relationship success.

Figure IV: Python output for logistic regression and lasso

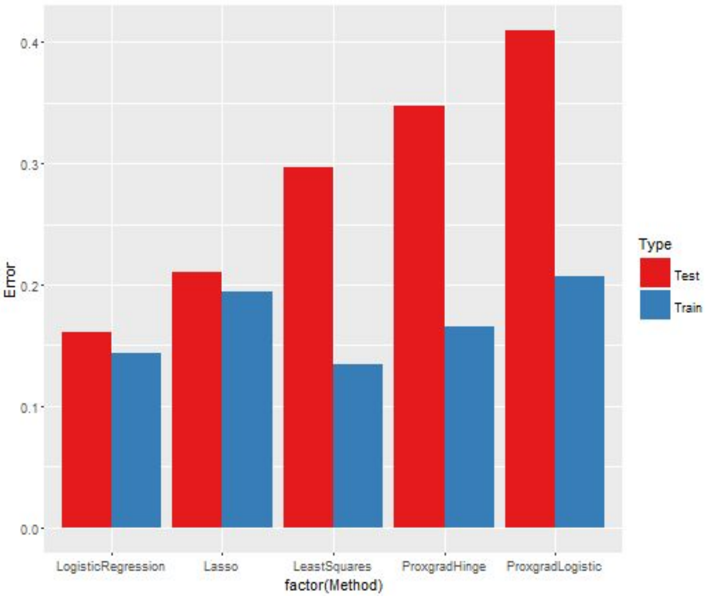


```
Logistic Regression Classifier Train Error Rate: 0.14363
Lasso Regression Train Error Rate: 0.19479
Logistic Regression Classifier Test Error Rate: 0.16129
Lasso Regression Test Error Rate: 0.20968
Lasso Regression Non-Zero Features (n = 19):
        PPAGECAT: -0.087935
        PPHHHEAD: -0.012300
        PPHOUSEHOLDSIZE: -0.043729
        PPINCIMP: -0.021958
        Married: -0.434574
        Divorced: 0.111311
        Never Married: 0.057315
        Living With Partner: -0.071276
        CHILDREN_IN_HH: 0.022295
        same religion: -0.030298
        Relatives see/month: -0.006145
        # marriages: -0.031252
        Q19: 0.318536
        Age met: 0.002600
        Age relationship: 0.000259
        Approve: -0.035717
        DISTANCEMOVED_10MI: 0.000002
        AGE_DIFFERENCE: -0.003878
        RELATIONSHIP_QUALITY: -0.139123
```

Conclusion

Overall, because the lasso regression gave the most sparse solution, it is likely the most useful model. Logistic regression performed marginally better (4%) on predicting relationship success. However, the logistic regression uses 114 features, significantly more than the 19 features required for our lasso model. When looking for an economical solution in practice, having 6 times as many features for an additional 4% of accuracy is unlikely to be an ideal model. The lasso regression also had the smallest difference in error between the test and train datasets, indicating that it overfits the least as it represents the test dataset nearly as well as the train data set (Figure V).

Figure V: Train and test error rates for each method



Our model, given these 19 features, can predict to nearly 80% accuracy whether or not a couple will still be in their relationship after 6 years. Because of this accuracy, we would recommend using this data, whether that be for modifying data algorithms, or for personal use. A larger sample would make us more confident in our results and allow us to quantify the uncertainty surrounding our predictors variables, but we still do believe that there is much value gained from this study.

From the lasso regression, one can see the factors that are most important in a successful relationship. Variables such as living together or being married greatly added to the success of a couple, but these seem obvious as they would indicate a higher level of commitment. Less obvious variables that are seen as contributing to relationship success are having approval from parents, seeing more of their relatives per month, having more education, a higher income, and sharing the same religion. We also interestingly see that having a larger age gap between partners, but meeting earlier on in life, contributes to staying together.

It is important to note that there may be some confounding variables that create these results. For example, more traditional marriages that disapprove of divorce more may meet earlier on in life and have a larger age gap. It would be interesting to see if these variables hold up in relationship happiness as well. We were also surprised to see that the model didn't indicate variables we would have thought of as important contributing to relationship longevity. For example, the lasso regression did not indicate that sharing the same race was important. Other variables that the model didn't indicate as significant were how the couple met or their sexual orientation. Takeaways society can have from that information are that despite the negative views on homosexual relationships some people hold, they are just as successful at maintaining a relationship, and that people are also finding lasting love on the internet.