Isabelle Rennenberg Final Report SAS Model Building

**Introduction**

The purpose for this study is to take the data from a speed dating study and create the model(s) that best predict what factors will cause someone to say that they like the person they are dating. Equation(s) will be created in order to predict a dater’s opinion(like)based on the factors: attractiveness, sincerity, intelligence, fun, ambitiousness, and shared interests with their partner. It will also be determined if race and age are factors in determining the daters opinion. This study is important in creating a streamlined and highly usable equation for a dating app which will aid in the building of a new site that will optimally match couples.

**Descriptive Statistics**

**A screenshot of a computer

Description automatically generated with low confidence**For the data given, the experimental unit is one couple, each row (observation) being a couple’s date. Each value is the ranking of one individual. The data was collected from 276 heterosexual couples with 22 different variables for evaluation. There are two character variables: race male and race female. The rest are numerical, and either are represented as 1 or 0 or on a scale of 1 to 10. Age is on a scale of 18-55. All variables are divided into two, one which is the female score of the male, and vise versa. For this dataset: 1=Yes 0=No.

A screenshot of a computer

Description automatically generated with medium confidenceThere were missing variables present in the dataset. Refer to Table 1. Missing values should be looked at in order to determine possible reasons for why the person declined to enter information about the date.

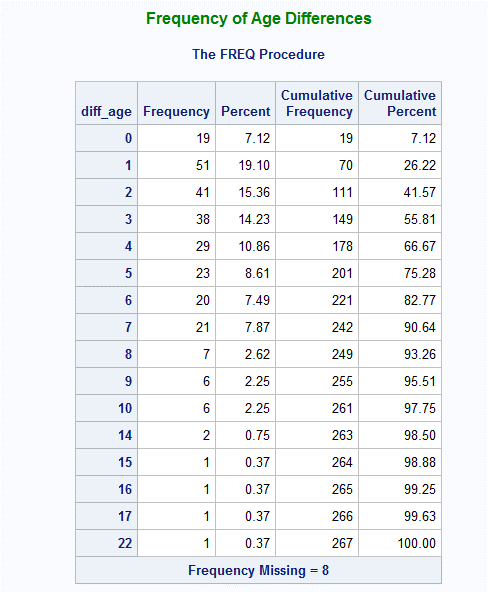
Table . Missing Values in Dataset

For race of the couple, Table 2 was created. In this figure, 20 couples were both Asian, 0 couples were both Black, 99 couples were both Caucasian, and 1 couple was both Latino. It is important to note that there were 6 missing race values, so those numbers could be higher or lower. 149 couples were with a different race, or race was missing or in the other category. I did not find that being of the same race mattered for the outcome of the date. Refer to table 3.

Table 2. Race Frequency

***Table 3. Correlation Between being on a Date with the Same Race and Like***

|  |  |
| --- | --- |
| Here, the correlation coefficient is 0.04. This is incredibly low. A high correlation coefficient is -1 or 1, which would indicate a strong relationship. Since this is much lower than 1, there is little to no linear relationship. | Here, the correlation coefficient is 0.04. This is incredibly low. A high correlation coefficient is -1 or 1, which would indicate a strong relationship. Since this is much lower than 1, there is little to no linear relationship. |

A screenshot of a computer

Description automatically generated with low confidence

For the comparison of ages, Table 4 was created. 19% of couples were within 1 year of each other. From this, if 2 years is defined as close in age: 111 couples are close in age, and 156 are not close in age. In addition, Table 5 showed that the average age for males and females was 26. The minimum age for females was 19, maximum 55, and minimum males 18 and maximum 42. I found that being close in age was not significant enough for males and females to say that there was a relationship. Refer to Table 6.

Table 4. Frequency of Age Differences

Table 5. Age Statistics

***Table 6. Correlation Between being close in Age and Like***

|  |  |
| --- | --- |
| A picture containing text, screenshot, font, number  Description automatically generatedHere, the correlation coefficient is 0.097. This is incredibly low. A high correlation coefficient is -1 or 1, which would indicate a strong relationship. Since this is much lower than 1, there is little to no linear relationship. | A picture containing text, screenshot, font, number  Description automatically generated Here, the correlation coefficient is 0.055. This is incredibly low. A high correlation coefficient is -1 or 1, which would indicate a strong relationship. Since this is much lower than 1, there is little to no linear relationship. |

**Selection of the Models and Type of Analysis Employed**

For this study, I created two models, one for males and one for females. Prior to beginning my model building, I split the dataset into a 180 training sample and 96 holdout sample. This was done in order to later understand the reliability of the model that was created.

To begin, I created polynomial terms for all of the independent variables: attractiveness, sincerity, intelligence, fun, ambitiousness, and shared interests with their partner. I did so in case there was a quadratic or cubic relationship between LikeM/LikeF and any of the independent variables. The sample size for this study was 276, so it would be comfortable to have 10\*k number of variables of interest. In order to add polynomial terms in the model, I employed both forward selection and backwards elimination. For forward selection, I ran regression diagnostics on the model, and if the variable was significant, I added the quadratic term. If the quadratic term was significant, I added the cubic. For backwards elimination, I started with all 18 terms (18\*10 is less than 276, so this is still optimal for the study) and removed any nonsignificant terms until all left were significant. The base terms were added in order of positive correlation, with the highest correlated variables first, and lowest last. Significance, in this case, means that the independent variable interacts/predicts Like. For reference, at a 95% confidence of significance, the p value should be less than 0.05. If the value is less than 0.05, then the variable of interest is significant to the model and aids in predicting Like. I employed this logic for all polynomial evaluations. I started with the male model.

**Table 7. Male Forward Selection**

|  |  |  |
| --- | --- | --- |
| Attractiveness, Fun, Shared Interests and Sincerity were all significant in accordance to being 95% confident. | The only polynomial term that was added and remained significant was FunM2 and FunM3 | **Model Decision:** Polynomial terms to consider adding to the model include FunM2 and FunM3 **Suggest having Attractive, Fun, FunM2, Shared Interests, and Sincerity. Consider Fun cubic.** |

**Table 8. Male Backwards Elimination**

|  |  |
| --- | --- |
|  | **Model Decision:** AttractiveM2, AttractiveM3, SharedInterestsM2, SharedInterestsM3, Intelligent(all), SincereM2, SincereM3, Ambitious(all) were removed. |

Including Polynomial Terms runs the risk of collinearity, which occurs when independent variables are related to one another. I ran collinearity analysis on FunM2 and FunM3 and found that they were highly related to one another. I attempted centering the base (subtracting the mean) variable but had no effect. Just adding FunM2 resulted in no large violations as before. I would therefore include FunM2 as the only polynomial term for Male Like. Refer to Table 9 for collinearity diagnostics.

**Table 9. Collinearity for Fun (Males)**

|  |  |
| --- | --- |
| The inclusion of both quadratic and cubic terms for fun resulted in a condition index of almost 90. The condition index should be below 30. In addition, Variance inflation was higher than the critical value of 10. | Here, the condition index has fallen below 30. Variance inflation remains higher than 10, but not as much as before. I feel that the significance of FunM2 as well as the lower condition index warrants its inclusion in the model. |

I now ran the following model building techniques in SAS: All possible models, forward selection, backwards elimination, and stepwise forward selection. Refer to table 10.

**Table 10. SAS Model Building**

|  |  |  |
| --- | --- | --- |
| The significance level for forward selection if 0.50, so variables such as ambitious and attractive M3 were included. | In this approach, SAS did all the same as Stepwise, but it kept Shared Interest quadratic instead of the base term. This is not ever suggested. | Model includes **Attractive, Fun, Shared Interests, Sincere, and Fun Quadratic.** |
| All Possible Models showed that 5 variables was the optimal model (as shown by the third row, Cp, -3.288 being less than p, 5). The variables included in this model are the same of interest as above: **Attractive Fun Fun2 Shared Interests and Sincere.** | | |
| **Model Decision: The variables to be included for LikeM are: Attractive, Fun, Fun^2, Shared Interests, and Sincere.** | | |

Thus, our regression equation to predict whether or not a male likes a female depends most on attractiveness, fun (and fun^2), Shared Interests, and Sincerity. This is a 5 variable model as all possible models and stepwise forward selection approach predicted. This model matches my prediction from the forwards approach. Now, I will perform the same steps with the female model.

**Table 11. Female Forward Selection**

|  |  |  |
| --- | --- | --- |
| From this, I would add polynomial terms to Fun, Attractive, Shared Interests, Sincere and Intelligent. | The only significant polynomial term is Shared Interests^2. | **Model Decision:** Only adding SharedInterestsF2 is significant to the model**. Suggest having Fun, Attractive, Shared Interests, Shared InterestsF2, Sincere and Intelligent.** |

**Table 12. Female Backwards Elimination**

|  |  |
| --- | --- |
|  | **Model Decision:** All polynomial terms were removed as well as all of ambitious. |

My forwards model leads me towards adding shared interests^2. Including polynomial terms will run the risk of collinearity (as above). I will run collinearity diagnostics to understand if adding SharedInterestsF2 adds this risk.

**Table 12. Collinearity Evaluation**

|  |  |
| --- | --- |
|  | Here, the condition index is much lower than 30 (2.6), so I am not much concerned with collinearity issues. None of my variance inflation values are higher than 10. Looks great! |

I now ran model building techniques in SAS: All possible models, forward selection, backwards elimination, and stepwise forward selection. Refer to table 13.

**Table 13. SAS Model Building**

|  |  |  |
| --- | --- | --- |
| A picture containing text, screenshot, number, font  Description automatically generatedThe significance level for forward selection if 0.50, so variables such as SincereF and Shared Interests cubic were included. | A picture containing text, screenshot, font, number  Description automatically generatedIn this backwards approach, SAS had FunF2 (it is not suggested to have a polynomial term without a quadratic), as well as all of attractive variables. | A picture containing text, screenshot, font, number  Description automatically generatedModel includes **Fun, Shared Interests, Attractive, Intelligent and Shared Interests quadratic.** |
| A picture containing text, screenshot, font, line  Description automatically generatedAll Possible Models showed that 5 variables was the optimal model (as shown by the third row, Cp, -1.11 being less than p, 5). The variables included in this model are the same of interest as above with the stepwise; **FunF AttractiveF2 SharedInterestF2 and IntelligentF. Note- no sincerity is included in either of those models (including backwards).** | | |
| **Model Decision: The variables to be included for LikeF are: Fun Attractive, Shared Interests, Shared Interests^2 and Intelligent.** | | |

Sincerity was not included in any of the later models except for forward, so I looked at its significance in the Type III SS. I found that it was not significant above and beyond other variables and agreed with the SAS approach (over 0.05). Thus, our regression equation to predict whether or not a female likes a male depends most on fun, attractiveness, Shared Interests, Shared Interests^2 and Intelligence. This is a 5 variable model as all possible models approach and stepwise forward selection predicted. This model closely matched my prediction from the forwards approach.

**Diagnostics**

To determine the reliability of these models, cross-validation was employed. This compares the R^2 of our training model to our holdout sample. When subtracting the two, you do not want the value to be more than 0.10. Essentially, this is making sure our model can handle future predictions and do so accurately.

For the male model: The R^2 for our training model was 0.67. Our R^2 for the holdout sample was (0.77)^2=0.59. 0.67-0.59=0.07, which is less than 0.10. This is a reliable model! Refer to Table 14 for output used to do this calculation.

**Table 14. Reliable Model Calculation**

|  |  |
| --- | --- |
| **A screenshot of a computer  Description automatically generated with low confidence** | **A screenshot of a computer  Description automatically generated with medium confidence0.77^2=0.59** |

For the female model: The R^2 for the holdout sample was (0.77)^2=0.59. 0.64-0.59=0.05, which is less than 0.10. This is a reliable model! Refer to Table 15 for output used to do this calculation.

**Table 15. Reliable Model Calculation**

|  |  |
| --- | --- |
| **A screenshot of a computer  Description automatically generated with medium confidence** | **A screenshot of a computer  Description automatically generated with medium confidence0.77^2=0.59** |

A screen shot of a graph

Description automatically generated with medium confidenceA screen shot of a graph

Description automatically generated with medium confidenceUsing the chosen models, jackknife residuals were evaluated. This is a way to assess the stability of the regression model and “goodness” of a fit. Jackknife values over 2 are of concern to a model. In addition, plotting jackknife residuals and predicted values allows one to visually inspect this. Obvious patterns such as funneling, or sinusoidal curves would indicate a problem with the data. The following scatterplots of the regression models show no pronounced patterns. There were parallel diagonal lines observed, but this is likely due to gap in the data as it was retrieved on a scale of 1-10, or a discrete way of measuring the variables. It is not indicative of a pattern that would raise any alarms. Refer to the following scatterplots:

Figure 2. Jackknife Residuals for Males

Figure 1. Jackknife Residuals for Females

The data was normally distributed, but there were indications of the possibility of outliers. For females and males, there were jackknife values over 2. I ran more diagnostics on these observations, noted that the values did not seem to be made in error, and found that none of the observations had violated Cook’s distance, which should never be over 1. None of the observations were above 1. However, some were high for leverage, which is 2(1+5)/276 for both of the models (0.043). Because Cook’s distance was not violated, I would not suggest removing any terms.

**Summary of Findings**

Two models, one for males and one for females, were chosen based on the variables given for evaluation.

Based on polynomial forward selection and backwards elimination, the model that would best predict like for males included Attractiveness, Fun, Fun^2, potentially Fun^3, Shared Interests and Sincerity. It was found that including both the quadratic and cubic terms for Fun resulted in large issues with collinearity. The all models approach and the stepwise forward selection model approach did not include FunM3. In addition, when looking at the regression model, it did not prove that the cubic term for fun added anything above and beyond the quadratic term for fun. At a 95% confidence, it could be proven that FunM3 was not needed for the model. That paired with the collinearity issues removed it from the model. Adding FunM2 did prove a higher than desired variance inflation, but the condition index remained below 30 (which is desirable). Thus, the equation for the prediction of male like is the following:

**(Ŷ(LikeM)) = β0 + β1 \*Attractiveness (1-10) + β2 \*Fun (1-10) + β3 \*Fun^2 (1-10) +β4 \*Shared Interests (1-10) + β5 \*Sincerity (1-10)**

**Ŷ=-1.09+0.52 β1+0.41 β2-0.020 β3+0.14 β4+0.21 β5**

Based on polynomial forward selection and backwards elimination, the model of interest for females included Fun, Attractiveness, Shared Interests, Shared Interests^2, Intelligence and potentially Sincerity. When looking at the significance of sincerity above and beyond other variables, it was not deemed to be significant when 95% confident. It was shown that the maximal optimal model had 5 variables, so sincerity was removed. All possible models and stepwise forward selection models both had the same variables. It was found that including both the quadratic term for shared interests was significant and did not show issues with collinearity. Thus, the equation for the prediction of male like is the following:

**(Ŷ(LikeF)) = β0 + β1 \*Fun (1-10) + β2 \*Attractiveness (1-10) + β3 \*Shared Interests (1-10) +β4 \*Shared Interests^2 (1-10) + β5 \*Intelligence (1-10)**

**Ŷ=-0.44+0.26 β1+0.28 β2+0.44 β3-0.023 β4+0.22 β5**

My evaluation of reliability of the models compared the R^2 of a training sample to a holdout sample, which essentially tested how well the model can predict values. For both, the value was below 0.10, which means they are reliable models. The model for males predicts Like at an R^2 of 0.64. The model for females predicts Like at an R^2 of 0.63. For these models, ~60% of variation in Like depends on the selected variables.

Finally, it was determined that 19% of the entire population were close in age, but that being close in age did not influence the decision of couples when liking the other. Exactly 111 dates occurred where the couples were close in age. It was determined that matches were most commonly of the same race for Caucasian heterosexual dates. There were 20 dates where both male and females were Asian, and only 1 date where both the male and female were Latino. There were no couples during the entirety of the study that were both Black. It was found that it did not matter if the date was of the same race for the outcome of the date.