A Statistical Report of a Short Speed Couple data

Introduction

This study utilized data from 276 heterosexual couples who were randomly paired with one another for a short speed data. The main purpose of this project is to create a model that can predict a dater's opinion about the person they are dating based on the dater's evaluation of some of the individual's qualities assessed in the study. Effects of the differences in the race and age of partners were also studied. These results will be used to optimally match couples correctly on a dating site.

Descriptive Statistics

This dataset included 552 respondents who were randomly paired with one another. Out of this number of respondents, only 546 had valid data on the response variable, Like. Table 1 reports the valid total number and percentages for each variable and for each group under the variables (Race and Age), and also the means and standard deviations for the numeric variables. The number of respondents who were close in age was 222 (314 not close in age) while the number that were of the same race was 242 (298 not of the same race). The average Like rating was 6.53 with a standard deviation of 1.77. Correlation coefficients for some of the variables were computed. Significant moderately strong positive linear relationships were found between Like and Attractive with r = .684 (p<.0001), and Like and Fun (r = .651; p<.0001).

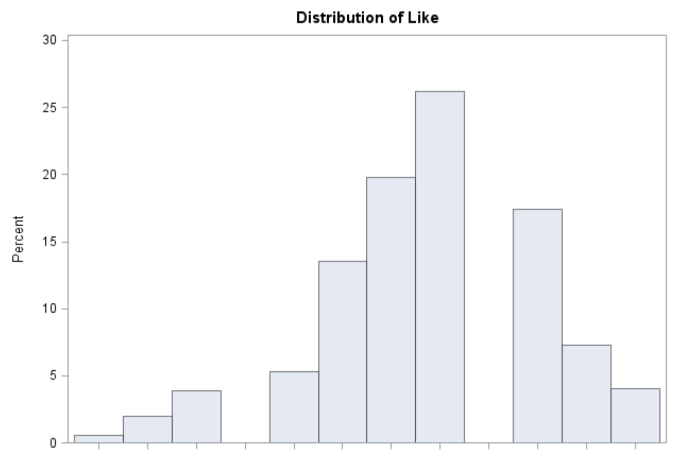

Figure 1. Histogram Showing the Distribution of Like

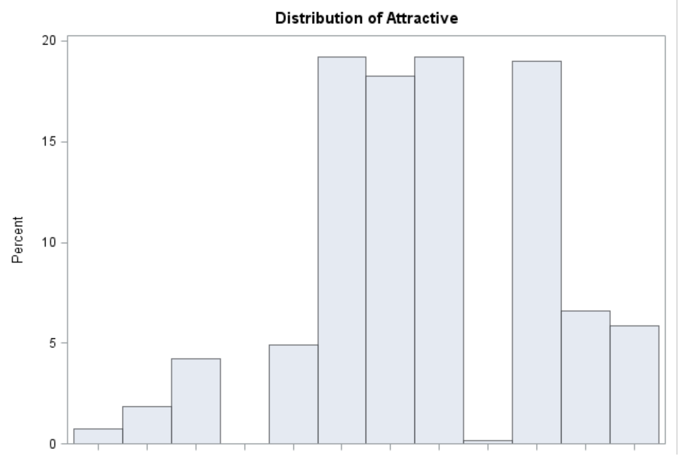

Figure 2. Histogram Showing the Distribution of Attractive

Figures 1 and 2 show the distribution of Like and Attractiveness respectively. Like appears skewed to the left and most of the values are between 5and 9. Attractive appears bimodal with most of its values between 5 and 9 too.

Variables	N	Percent or Mean (SD)
Sex		
Female	276	50
Age		
Close in age	222	41.42
Not close in age	314	58.58
Race		
Same Race	242	44.81
Not Same Race	298	55.19
Like	546	6.525 (1.772)
Attractive	547	6.480 (1.873)
Sincere	544	7.820 (1.598)
Intelligent	541	7.774 (1.407)
Fun	540	6.713 (1.885)
Ambitious	525	7.103 (1.781)
Shared Interests	495	5.529 (2.177)

Table 1. Showing the Descriptive Statistics of the Variables

Model Selection and Analysis

To choose the maximum model, the linear effects of Like, Attractive, Sincere, Intelligent, Fun, Ambitious, Shared Interests, Sex, Race and Age were included in the model. After testing the significance of the linear terms, squared terms for Attractive, Sincere, Fun and Shared Interests were also included. Interactions terms involving Attractiveness were also added as previous studies have suggested that such interactions were important. Hence a total of 21 variables were included in the maximum model as predictors of LIKE: ATT; ATT2 = ATT^2 ; SIN; SIN2 = SIN^2 ; INT; FUN; FUN2 = FUN^2 ; AMB; SHA; SHA2 = SHA^2 ; SEX (0 for Male and 1 for Female); RACE (1 for same and 0 for not same); AGE (1for close age and 0 for not close age); ATT x SIN; ATT x INT; ATT x FUN; ATT x AMB; ATT x SHA; ATT x RACE; ATT x AGE and ATT x SEX. All linear variables except SEX, RACE, AGE and LIKE were centered to remove collinearity. Interactions and quadratic terms were calculated from the centered variables. The dataset was divided into two and each person was randomly assigned to either dataset A (training dataset) or dataset B (holdout dataset). Data set A (containing 276 individuals), which is 50% of the entire dataset will be used to build the best model and then dataset B will be used to compute cross-validation correlation and shrinkage statistics.

Selection Criteria and Strategy

The main criteria used to select the best model was Mallows C_p . Here, p (the number of predictors in a model) was allowed to increase up to a point where C_p was less than p. Then all models with that p number of variables were evaluated using R^2 and MSE. The all-possible models strategy was used. SAS was used to generate all possible models ($2^k – 1$; where k = 21) that can be used in the prediction of Like. At the end of the selection, the best model had 10 variables. The variables were: ATT; SIN; INT; FUN; AMB; SHA ; ATT x FUN; ATT x SHA; FUN2; and SHA2.

Best Model

$$Y_{LIKE} = 6.665 + 0.383X_{ATT} + 0.176X_{SIN} + 0.133X_{INT} + 0.101X_{FUN} + 0.00369X_{AMB} + 0.192X_{SHA} - 0.0722X_{ATT} X_{FUN} + 0.0573X_{ATT} X_{SHA} + 0.0304X_{FUN} X_{FUN} - 0.0296X_{SHA} X_{SHA}$$

Based on all 10 predictors, R-squared was found to be 0.6311. This means that this model explained 63.11% of the variation in the LIKE variable. C_p was 8.9710 and MSE was 1.20054. Race (p=.6370) and Age (p=.3553) were not useful in the prediction of Like above and beyond other variables in the sample.

Variable	Parameter	Standard	T value	P value
Intercept	6.66545	0.09633	69.19	<.0001
Attractive	.38266	0.05001	7.65	<.0001
Sincere	.17638	0.06057	2.91	.0039
Intelligent	.13292	0.07021	1.89	.0596
Fun	.10136	0.05639	1.80	.0736
Ambitious	.00369	0.04854	.08	.9395
Shared Interests	.19244	0.03930	4.90	<.0001
AttFun	-.07220	0.02727	-2.65	.0087
AttSha	.05729	0.02060	2.78	.0059
Fun2	.03040	0.02056	1.48	.1405
Sha2	-.02961	0.01252	-2.37	.0188

Table 2. Showing Regression Parameter Estimates and P-Values

Other models considered included a 12-predictor model (ATT; SIN; INT; FUN; AMB; SHA; ATT x SIN; ATT x FUN; ATT x SEX; ATT x SHA; FUN2 and SHA2) which had an R-square of .6410, C_p = 8.3024 and MSE = 1.17870 and a 17-predictor model (ATT, SIN, INT, FUN, AMB, SHA, SEX, RACE, AGE, ATT x FUN; ATT x SIN; ATT x SHA, ATT x RACE, ATT x SEX, and ATT2, FUN2 and SHA2) which had an R-square of .6355, C_p = 14.8543 and MSE = 1.19752. At the end, the 10-predictor model was chosen because each subsequent predictor addition to the model produced a very little increase in R-square.

The overall regression model was found to be statistically significant with F(10, 228) = 39.00, p <.0001. All selected predictors irrespective of their significance at α=.05 were included in the model. Attractive (ATT, p<.0001), Sincere (SIN, p=.0039), Shared Interests (SHA, p<.0001), and interactions such as ATTFUN (p=.0087), ATTSHA (p=.0059) and SHA2 (p=.0188) were found to be significant (α = .05) and useful in predicting a dater's opinion (LIKE) while Intelligent (INT, p=.0596), FUN (p=.0736), Ambitious (AMB, p=.9395) and FUN2 (p=.1405) were found not significantly useful in the prediction of LIKE. Also, there was no evidence for statistically significant differences in the average predicted outcomes of the LIKE variable between males and females at the .05 significance level (p=.3237).

From the chosen model, the average Like score of someone whose qualities were equal to the average qualities in the sample was found to be 6.665 (p<.0001). For every one-unit increase in Attractiveness, we expect the average Like score to increase by .383 (p<.0001) when the Shared Interests and Fun are 0. Also, for every one-unit increase in the difference between the Attractive score and average Attractiveness, we expect the effect of Shared Interests to increase by 0.0573 (p=.0059) and the effect of Fun to decrease by .0722 (p=0087). For every one-unit increase in the Sincere rating, we expect the average Like score to increase by .176 (p=.0039). Fun was not significantly useful (p=.0736) in predicting Like when Attractiveness was 0 but was useful for other values of Attractiveness (p=.0087).

Assessing Reliability of the Model

Cross-validation analysis was carried out by using the regression equation estimated from data set A to predict Like values for data set B. The linear variables in data set B were centered before analysis was done. The R-square between these predicted values and the observed LIKE values in data set B was 0.6198, and this is the cross-validation correlation. This means that 61.98% of the variation seen in the Like variable is explained by this model. Since R-squared for data set A was 0.6311, shrinkage is 0.6311 – 0.6198 = 0.0113 which is quite small (less than 0.10) and indicates excellent reliability of prediction.

Model Diagnostics

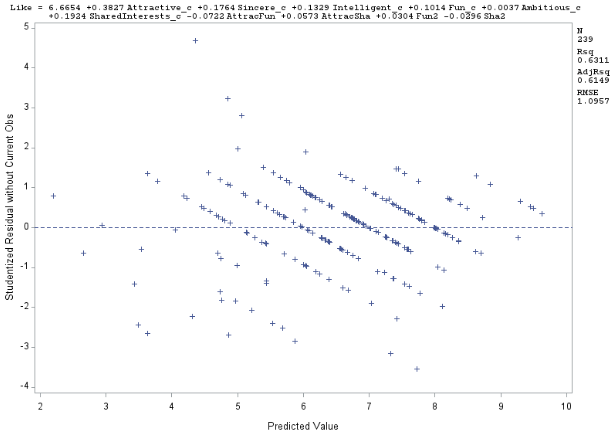

Figure 3. A scatterplot of jackknife residuals versus the predicted values for the model

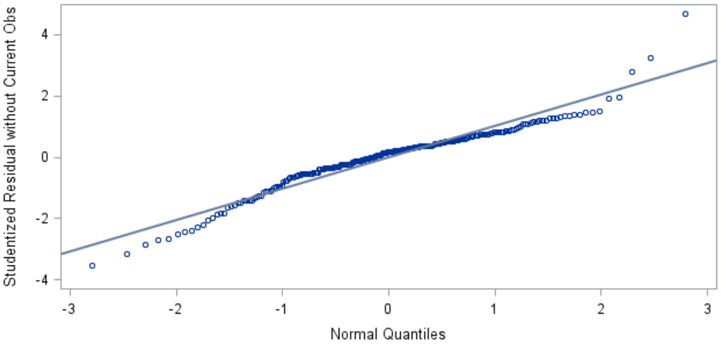

Figure 4. Normal Probability Plot of the Jackknife Residuals

Based on figure 3, the plot appears to be randomly scattered with no obvious funneling or systematic pattern. Thus, assumptions of linearity, homoscedasticity and independence are not violated. Figure 4 shows the normal probability plot of the jackknife residuals which appear relatively linear. This supports the normality assumption. The skewness statistic value of -0.24 and kurtosis statistic value of 2.74 also do not suggest a gross violation of the normality assumption. The histogram (figure 5) of the distribution of the jackknife residuals appears to be normally distributed. From the boxplot, it can be seen that the mean is slightly below the median. Also, there are a few empty circles outside the box indicating values that are more than 1.5 times the interquartile range.

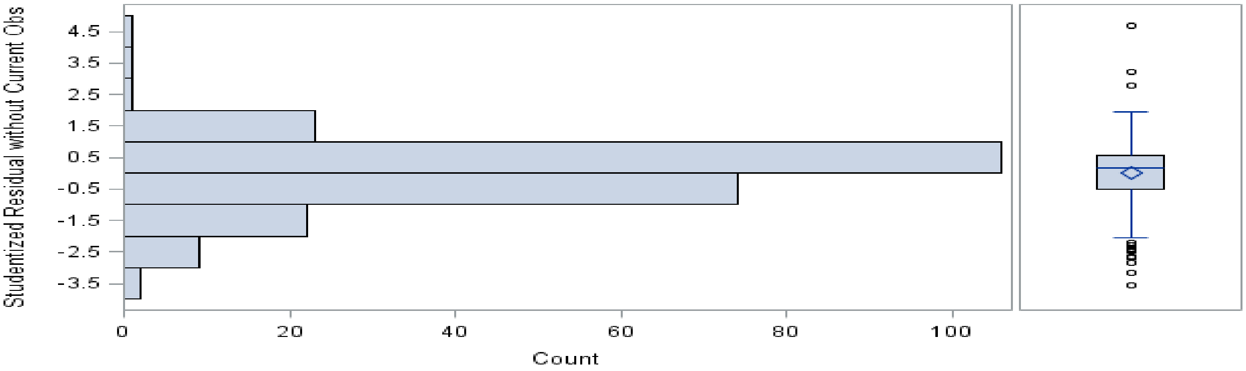

Figure 5. Distribution of Jackknife Residuals with mean=-0.00014; Sd=1.02348; and median=0.16961

Formal outlier diagnostics revealed 36 outliers in the training dataset. These values should be scrutinized further to ensure that they are not data entry errors. The criteria used to assess outliers include cook's distance greater than 1, jackknife residuals greater then 2 (alpha approximately equal to .05) and leverage greater than .092 (calculated using the formula [2(k+1)]/n where k is the number of predictor variables and n is the number of observations in the dataset).

Collinearity diagnostics were carried out and all Variance Inflation Factors were less than 10. The highest Variance Inflation Factor obtained was 2.83. Also, the condition number was found to be 3.84. Based on these, collinearity does not appear to be an issue among the variables.

Summary of Findings

This data analysis aimed to create a model that can be used to predict a dater's opinion about the person they are dating based on the dater's evaluation of some of the individual's qualities assessed in the study. In total, 121 couples (44.81%) were of the same race and 111 couples (41.42%) were close in age as defined by being within 2 years of one another.

Race and Age were not useful in the prediction of Like above and beyond other variables. Useful moderately strong linear relationships were found among some of the variables (Like & Attractive; Like & Fun). The estimated ratings when compared in closeness to the actual ratings were found to be 0.7944 in the first dataset and 0.7873 in the second dataset. Thus, we can conclude that the predicted scores are quite close to the actual values. The average Like score of someone whose qualities were equal to the average qualities in the sample was found to be 6.665. There was no evidence for statistically significant differences in the average predicted outcomes between males and females.

The interaction terms included in the model produced the effect a predictor has on the dater's opinion depending on the values of another predictor. The effect of attractiveness on Like depends on different values of shared interests and Fun; and the effects of Shared Interests and Fun are dependent on Attractiveness. The effect of shared interests on a dater's opinion was found to be high for its lower values and low for its higher values when attractiveness was 0.

Further scrutiny of extreme observations or data values which were found to be far outside the average value is recommended to make sure that these are not data entry errors. Also, for this study it was not stated whether this data was collected at an online or in-person event. Since this is for an online dating site, it is recommended that data should be collected from an online event since ratings from an in-person event may give erroneous results.