Report On Speed dating

Name: Cholleti, Sai Vignesh Reddy

Date: 04/18/2022

Course: STA 512 Take Home Exam

Table of Contents

[Introduction 1](#_Toc101217216)

[2. Descriptive Statistics: 2](#_Toc101217217)

[3. Selection of the Models and Type of Analysis Employed 7](#_Toc101217218)

[4. Diagnostics 11](#_Toc101217219)

[4.1 Collinearity Test 11](#_Toc101217220)

[4.2 Normality 12](#_Toc101217221)

[4.3 Homogeneity and Independence 16](#_Toc101217222)

[4.4 Outliers 16](#_Toc101217223)

[4.5 Effects of same race and closeness in age 18](#_Toc101217224)

[4.6 Reliability 18](#_Toc101217225)

[5. Summary of Findings 19](#_Toc101217226)

# 1. Introduction

Online dating is a growing industry with recent quarterly profits well in excess of millions. Entrepreneur goal is to break into this industry with his/her new site PIZZAZ.COM that uses the power of statistics to optimally match couples. To help with this he/she recently hosted a speed dating event where 276 couples were randomly paired up with one another for a short speed date. The results are recorded in the accompanying dataset.

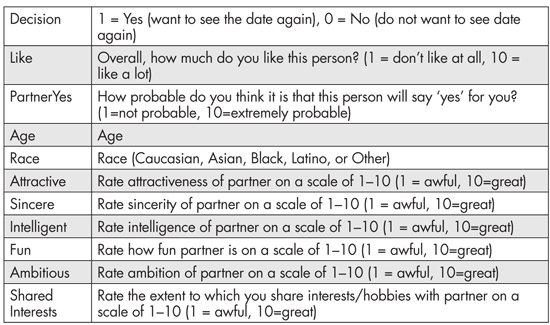
The ultimate goals are:

* Create an equation that can predict a dater’s opinion of the person they are dating (As indicated by the Like variable) based on the attractiveness, sincerity, intelligence, fun, ambitiousness, and shared interests of their partner. In addition I want you to see whether or not it matters if the partners are the same race. As part of your descriptive statistics please let me know how many couples had the same race and how many did not. Finally, I want you to also see whether or not it matters if the partners are close in age (As defined by being within 2 years of one another -- ie 26 and 28). As part of your descriptive statistics please let me know how many couples are "close" in age and how many are not.
* It would help me if you could give me a good idea of how close your estimated like ratings were to the actual like ratings. I could then take this information to potential investors when I decide to take my company public. If it works really well perhaps I can give you some shares of the stock!
* Lets keep this last point between the two of us. Im not entirely sure about the integrity of the sampled speed daters. Im nervous that some of them might have paid their partner to inflate their score. If you find any evidence of situations where like ratings are suspiciously high, I hope that you will please let me know which couple they are from so that I can look into them further.

# 2. Descriptive Statistics:

Note that variables like Decisionm refers to the Male's decision regarding whether he wants to see the female again. Likem refers to how much the male liked the female. Attractivem refers to the male's rating of the female's attractiveness.

However, agem and racem refers to the male's age and race. The description of the dataset is shown below (table 2.1).



*Table 2.1*

The variables are rated on a scale of 0 to 10.

Instead of considering the interested variables for two different analyses (male and female), the mean of male and female ratings is taken into consideration and created new variables to make the further analysis/model building simple.

For ex: if LikeM = 6, LikeF = 8 then Like = 7 (mean of 6 and 8)

The below table (table 2.2) shows the basic measurement statistics of interested variables Like, Attractive, Intelligent, Ambitious, Fun, SharedInterests and Sincere.

| **Variable** | **N** | **N Miss** | **Mean** | **Median** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- |
| Like Ambitious Attractive Intelligent Fun SharedInterests Sincere | 274 273 274 274 274 268 274 | 1 2 1 1 1 7 1 | 6.5392336 7.1410256 6.4771898 7.7636861 6.7153285 5.5410448 7.8175182 | 6.5000000 7.0000000 6.5000000 8.0000000 7.0000000 5.5000000 8.0000000 | 3.0000000 3.5000000 3.0000000 3.0000000 1.0000000 1.0000000 4.5000000 | 10.0000000 10.0000000 10.0000000 10.0000000 10.0000000 10.0000000 10.0000000 |

*table 2.2*

The below table (table 2.3) shows the descriptive statistics for the interested variables such as mean, median, minimum, maximum and count etc.

The below table shows the frequency of the couples with same race and couples with different race. The couple with same race has a value 1 for variable same\_race and a value of 0 otherwise.

| **same\_race** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 148 | 55.02 | 148 | 55.02 |
| **1** | 121 | 44.98 | 269 | 100.00 |
| **Frequency Missing = 6** | | | | |

*table 2.3*

The below table (table 2.4) shows the frequency of the couples who are close in age (defined by being within 2 years of one another). The couple who are close in age has a value of 1 for variable age\_close and a value of 0 otherwise.

| **age\_close** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| --- | --- | --- | --- | --- |
| **0** | 157 | 58.80 | 157 | 58.80 |
| **1** | 110 | 41.20 | 267 | 100.00 |
| **Frequency Missing = 8** | | | | |

*table 2.4*

The below table (table 2.5) shows the 2\*2 contingency table between the race of male and females. This table gives the count of couples and count of male/females by each race.

| **Table of RaceM by RaceF** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **RaceM** | **RaceF** | | | | | |
| **Frequency Percent Row Pct Col Pct** | **Asian** | **Black** | **Caucasian** | **Latino** | **Other** | **Total** |
| **Asian** | 20 7.43 31.75 28.99 | 4 1.49 6.35 28.57 | 28 10.41 44.44 19.05 | 8 2.97 12.70 34.78 | 3 1.12 4.76 18.75 | 63 23.42 |
| **Black** | 5 1.86 55.56 7.25 | 0 0.00 0.00 0.00 | 2 0.74 22.22 1.36 | 1 0.37 11.11 4.35 | 1 0.37 11.11 6.25 | 9 3.35 |
| **Caucasian** | 31 11.52 19.50 44.93 | 9 3.35 5.66 64.29 | 99 36.80 62.26 67.35 | 11 4.09 6.92 47.83 | 9 3.35 5.66 56.25 | 159 59.11 |
| **Latino** | 4 1.49 23.53 5.80 | 1 0.37 5.88 7.14 | 9 3.35 52.94 6.12 | 1 0.37 5.88 4.35 | 2 0.74 11.76 12.50 | 17 6.32 |
| **Other** | 9 3.35 42.86 13.04 | 0 0.00 0.00 0.00 | 9 3.35 42.86 6.12 | 2 0.74 9.52 8.70 | 1 0.37 4.76 6.25 | 21 7.81 |
| **Total** | 69 25.65 | 14 5.20 | 147 54.65 | 23 8.55 | 16 5.95 | 269 100.00 |
| **Frequency Missing = 6** | | | | | | |

*table 2.5*

The below table (table 2.6) shows the 2\*2 contingency table between the age groups of male and females. This table gives the count of couples and count of males/females in each age group.

| **Table of age\_grp\_M by age\_grp\_F** | | | | |
| --- | --- | --- | --- | --- |
| **age\_grp\_M** | **age\_grp\_F** | | | |
| **Frequency Percent Row Pct Col Pct** | **25-30** | **<=24** | **>30** | **Total** |
| **25-30** | 84 31.46 52.83 60.00 | 60 22.47 37.74 59.41 | 15 5.62 9.43 57.69 | 159 59.55 |
| **<=24** | 39 14.61 48.15 27.86 | 34 12.73 41.98 33.66 | 8 3.00 9.88 30.77 | 81 30.34 |
| **>30** | 17 6.37 62.96 12.14 | 7 2.62 25.93 6.93 | 3 1.12 11.11 11.54 | 27 10.11 |
| **Total** | 140 52.43 | 101 37.83 | 26 9.74 | 267 100.00 |
| **Frequency Missing = 8** | | | | |

*table 2.6*

The below table (table 2.7) shows the 2\*2 contingency table between the Decision of males and females. This table gives the count of couples who decided to see the date again or not by gender.

| **Table of DecisionM by DecisionF** | | | |
| --- | --- | --- | --- |
| **DecisionM** | **DecisionF** | | |
| **Frequency Percent Row Pct Col Pct** | **0** | **1** | **Total** |
| **0** | 65 23.64 50.39 43.92 | 64 23.27 49.61 50.39 | 129 46.91 |
| **1** | 83 30.18 56.85 56.08 | 63 22.91 43.15 49.61 | 146 53.09 |
| **Total** | 148 53.82 | 127 46.18 | 275 100.00 |

*table 2.7*

The below box plots (figure 2.1) shows the distribution of Like variable in each category of couples with same race and close in age.



*figure 2.1*

The below table (table 2.8) shows the correlation between the interested variables for model. These correlation values help to know how strongly the variables are related to each other.

| **Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Like** | **Ambitious** | **Attractive** | **Intelligent** | **Fun** | **SharedInterests** | **Sincere** |
| **Like** | 1.00000  274 | 0.44085 <.0001 273 | 0.73220 <.0001 274 | 0.57137 <.0001 274 | 0.67899 <.0001 274 | 0.55582 <.0001 268 | 0.57263 <.0001 274 |
| **Ambitious** | 0.44085 <.0001 273 | 1.00000  273 | 0.41550 <.0001 273 | 0.56847 <.0001 273 | 0.48494 <.0001 273 | 0.29753 <.0001 267 | 0.38195 <.0001 273 |
| **Attractive** | 0.73220 <.0001 274 | 0.41550 <.0001 273 | 1.00000  274 | 0.48209 <.0001 274 | 0.62480 <.0001 274 | 0.47017 <.0001 268 | 0.48842 <.0001 274 |
| **Intelligent** | 0.57137 <.0001 274 | 0.56847 <.0001 273 | 0.48209 <.0001 274 | 1.00000  274 | 0.54152 <.0001 274 | 0.29638 <.0001 268 | 0.56896 <.0001 274 |
| **Fun** | 0.67899 <.0001 274 | 0.48494 <.0001 273 | 0.62480 <.0001 274 | 0.54152 <.0001 274 | 1.00000  274 | 0.54897 <.0001 268 | 0.49521 <.0001 274 |
| **SharedInterests** | 0.55582 <.0001 268 | 0.29753 <.0001 267 | 0.47017 <.0001 268 | 0.29638 <.0001 268 | 0.54897 <.0001 268 | 1.00000  268 | 0.30420 <.0001 268 |
| **Sincere** | 0.57263 <.0001 274 | 0.38195 <.0001 273 | 0.48842 <.0001 274 | 0.56896 <.0001 274 | 0.49521 <.0001 274 | 0.30420 <.0001 268 | 1.00000  274 |

*table 2.8*

From the table (table 2.8), it seems like the variable Like is highly correlated with the variables Attractive and Fun among all other variables. This means the partner likes to choose other partner who has high ratings of attractive and fun.

# 3. Selection of the Models and Type of Analysis Employed

Firstly, two chunkwise tests are performed on the maximum model. One chunkwise test for testing interaction terms whether these terms add anything to the prediction of like ratings, above and beyond what is there already or not.

The other chunkwise test is for testing polynomial terms whether these poly terms add anything to the prediction of like ratings, above and beyond what is there already or not.

| **Test test\_poly Results for Dependent Variable Like** | | | | |
| --- | --- | --- | --- | --- |
| **Source** | **DF** | **Mean Square** | **F Value** | **Pr > F** |
| **Numerator** | 12 | 1.13516 | 1.94 | 0.0305 |
| **Denominator** | 234 | 0.58441 |  |  |

*Table 3.1*

| **Test test\_int Results for Dependent Variable Like** | | | | |
| --- | --- | --- | --- | --- |
| **Source** | **DF** | **Mean Square** | **F Value** | **Pr > F** |
| **Numerator** | 15 | 0.89912 | 1.54 | 0.0927 |
| **Denominator** | 234 | 0.58441 |  |  |

*Table 3.2*

From the above tables (table 3.1 and 3.2), it is clear that the interaction terms do not add anything to the prediction of like ratings whereas the polynomial terms add something significant amount to the prediction of like ratings.

So, the maximum model now consists of all the possible polynomial terms with degree no more than 3 along with linear base terms.

There are four strategies for selecting the variables in the best regression model. They are

1) All Possible model: This model generates all the possible models and displays every possible model.

| **Number in Model** | **R-Square** | **C(p)** | **MSE** | **Variables in Model** |
| --- | --- | --- | --- | --- |
| **6** | 0.6726 | 14.4291 | 0.62055 | Ambitious Attractive Intelligent Fun SharedInterests Sincere |
| **7** | 0.6807 | 9.8507 | 0.60767 | intg\_2 |
| **7** | 0.6794 | 10.9140 | 0.61013 | intg\_3 |
| **7** | 0.6791 | 11.1482 | 0.61068 | amb\_2 |
| **7** | 0.6786 | 11.5189 | 0.61154 | amb\_3 |
| **7** | 0.6742 | 15.1516 | 0.61997 | att\_2 |
| **7** | 0.6742 | 15.1633 | 0.61999 | att\_3 |
| **7** | 0.6741 | 15.2086 | 0.62010 | sharedintr\_2 |
| **7** | 0.6740 | 15.2770 | 0.62026 | sharedintr\_3 |
| **7** | 0.6731 | 16.0947 | 0.62216 | fun\_2 |
| **7** | 0.6730 | 16.1021 | 0.62217 | sin\_2 |
| **7** | 0.6730 | 16.1412 | 0.62226 | sin\_3 |
| **7** | 0.6728 | 16.2850 | 0.62260 | fun\_3 |
| **8** | 0.6858 | 7.6313 | 0.60018 | intg\_2 intg\_3 |
| **8** | 0.6842 | 8.9820 | 0.60333 | amb\_2 intg\_2 |

*Table 3.3*

The above table (table 3.3) gives the summary of all possible method. We should consider the variables where the mallows C\_p is less than or equal to number in model. From the table, the highlighted row satisfies the condition which includes the variables intg\_2 and intg\_3 in our model.

2) Forward selection: Selecting all the predictors one by one which are mostly correlated with the dependent variable until the most predictive variable not yet in the model is not significant at the predetermined significance level.

| **Summary of Forward Selection** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Variable Entered** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | intg\_2 | 7 | 0.0080 | 0.6807 | 9.8507 | 6.53 | 0.0112 |
| **2** | intg\_3 | 8 | 0.0051 | 0.6858 | 7.6313 | 4.24 | 0.0404 |
| **3** | amb\_2 | 9 | 0.0045 | 0.6903 | 5.9637 | 3.73 | 0.0547 |

*Table 3.4*

The above table (table 3.4) gives the summary of Forward selection process. This strategy includes the variables intg\_2, intg\_3 and amb\_2 along with all 6 linear terms in our model.

3) Backwards selection: Removing all the predictors from the maximum model one by one which are least significant with dependent variable until the least predictive variable is still significant at the prescribed significance level

| **Summary of Backward Elimination** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Variable Removed** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | att\_2 | 17 | 0.0000 | 0.6963 | 17.0102 | 0.01 | 0.9195 |
| **2** | sin\_3 | 16 | 0.0001 | 0.6963 | 15.0613 | 0.05 | 0.8210 |
| **3** | sharedintr\_3 | 15 | 0.0002 | 0.6960 | 13.2464 | 0.19 | 0.6662 |
| **4** | att\_3 | 14 | 0.0005 | 0.6956 | 11.6266 | 0.38 | 0.5358 |
| **5** | sin\_2 | 13 | 0.0005 | 0.6951 | 10.0485 | 0.43 | 0.5138 |
| **6** | fun\_3 | 12 | 0.0006 | 0.6945 | 8.5205 | 0.48 | 0.4893 |
| **7** | amb\_3 | 11 | 0.0005 | 0.6940 | 6.9433 | 0.43 | 0.5124 |
| **8** | sharedintr\_2 | 10 | 0.0020 | 0.6920 | 6.5848 | 1.67 | 0.1968 |
| **9** | fun\_2 | 9 | 0.0017 | 0.6903 | 5.9637 | 1.40 | 0.2373 |

*Table 3.5*

The above table (Table 3.5) gives the summary of Backward Elimination process. It is evident that those 9 variables are not significant by resulting in other variables intg\_2, intg\_3 and amb\_2 are significant for our model.

4) Stepwise Forward Selection: Same as forward selection except a variable selected into the model can later be removed from the model if the significance level falls below a prespecified level.

| **Summary of Stepwise Selection** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Step** | **Variable Entered** | **Variable Removed** | **Number Vars In** | **Partial R-Square** | **Model R-Square** | **C(p)** | **F Value** | **Pr > F** |
| **1** | intg\_2 |  | 7 | 0.0080 | 0.6807 | 9.8507 | 6.53 | 0.0112 |
| **2** | intg\_3 |  | 8 | 0.0051 | 0.6858 | 7.6313 | 4.24 | 0.0404 |
| **3** | amb\_2 |  | 9 | 0.0045 | 0.6903 | 5.9637 | 3.73 | 0.0547 |

*Table 3.6*

The above table (table 3.6) gives the summary of Stepwise selection process. From the table, the optimal model has three variables intg\_2, intg\_3 and amb\_2 along with all 6 linear variables.

By considering all four strategies, the below table compares the various aspects like R-square, MSE (Mean Square Error), Mallows C\_p for each model obtained.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| All possible Model | R-square = 0.6858 | Mallows C\_p = 7.6313 | MSE = 0.60018 | Variables: Ambitious Attractive Intelligent Fun SharedInterests Sincere intg\_2 intg\_3 |
| Forward selection | R-square = 0.6903 | Mallows C\_p = 5.9637 | MSE = 0.60562 | Variables: Ambitious Attractive Intelligent Fun SharedInterests Sincere amb\_2 intg\_2 intg\_3 |
| Backwards selection | R-square = 0.6903 | Mallows C\_p = 5.9637 | MSE = 0.60562 | Variables: Ambitious Attractive Intelligent Fun SharedInterests Sincere amb\_2 intg\_2 intg\_3 |
| Stepwise Forward Selection | R-square = 0.6903 | Mallows C\_p = 5.9637 | MSE = 0.60562 | Variables: Ambitious Attractive Intelligent Fun SharedInterests Sincere amb\_2 intg\_2 intg\_3 |

*Table 3.7*

From the table (table 3.7), almost all the strategies selected have same model and same statistical values.

So, without any doubt the model with 9 variables is selected as optimal model. Then this optimal model is further diagnosed to test whether all the assumptions are met or not.

# 4. Diagnostics

## 4.1 Collinearity Test

The collinearity exists when there is strong linear relationship between the independent variables and this collinearity causes the instability that makes regression coefficients estimates unreliable.

The above optimal model is tested for collinearity. We can say that collinearity exists in the predicted model by analysing the Variation Inflation Factors (VIF) and condition number. If either of the conditions VIF greater than 10 or Condition number greater than 30, the collinearity problem exists in the model. The VIF’s and condition number for our current model can be found in below tables (table 4.1.3 and 4.1.4).

Table

Description automatically generated

*Table 4.1.3 and 4.1.4*

From the tables (table 4.1.3 and 4.1.4), it is clear that collinearity problem exists in our model. This is because of the polynomial terms included in the model. As the polynomial terms are constructed based on linear terms, the variability in the like ratings which is explained by linear terms can also explained by these polynomials. So, this causes the collinearity when polynomial or interaction terms are included in the model.

This collinearity can be avoided by either one of those highly correlated variables or by centering the base variable. Here, the variables Intelligent and its polynomial terms intg\_2 and intg\_3 shows collinearity. Also, the other variables resulted in high VIF values are ambitious and its polynomial term amb\_2.

Now, after centering the base variables the final model looks like

Where = Average Attractive ratings of male and female in a couple

= Average Fun ratings of male and female in a couple

= Average SharedInterests ratings of male and female in a couple

= Average Sincere ratings of male and female in a couple

= centered variable of Average Intelligent ratings of male and female in a couple

= cubic term of centered variable

R-square for this model = 0.6859

Where is centered variable of Intelligent and is cubic term of centered variable . R-square for this model = 0.6859

Therefore, centering these variables and again the checking for collinearity in the model

Therefore, after centering the base variable Intelligent, the collinearity test is performed. The below tables (tables 4.1.1 and 4.1.2) show the VIF’s values and condition number for the model after centering.

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | 1 | 0.51861 | 0.43300 | 1.20 | 0.2321 | 0 |
| **Attractive** | 1 | 0.37108 | 0.04781 | 7.76 | <.0001 | 1.88687 |
| **Intelligent\_c** | 1 | 0.27720 | 0.06998 | 3.96 | <.0001 | 2.45452 |
| **Fun** | 1 | 0.19714 | 0.04780 | 4.12 | <.0001 | 2.18087 |
| **SharedInterests** | 1 | 0.13192 | 0.03202 | 4.12 | <.0001 | 1.48987 |
| **Sincere** | 1 | 0.19622 | 0.05350 | 3.67 | 0.0003 | 1.66987 |
| **intg\_3\_c** | 1 | -0.02553 | 0.00770 | -3.32 | 0.0010 | 1.70983 |

*Table 4.1.1*

| **Collinearity Diagnostics (intercept adjusted)** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Number** | **Eigenvalue** | **Condition Index** | **Proportion of Variation** | | | | | |
| **Attractive** | **Intelligent\_c** | **Fun** | **SharedInterests** | **Sincere** | **intg\_3\_c** |
| **1** | 3.22896 | 1.00000 | 0.03027 | 0.02565 | 0.02929 | 0.02453 | 0.03197 | 0.02096 |
| **2** | 1.06447 | 1.74166 | 0.04240 | 0.05552 | 0.03123 | 0.18620 | 0.00849 | 0.21256 |
| **3** | 0.60742 | 2.30561 | 0.06344 | 0.00142 | 0.00127 | 0.32441 | 0.40907 | 0.20699 |
| **4** | 0.45800 | 2.65521 | 0.40803 | 0.00479 | 0.05782 | 0.33016 | 0.46714 | 0.00215 |
| **5** | 0.35531 | 3.01460 | 0.45201 | 0.02280 | 0.74573 | 0.10103 | 0.00501 | 0.07042 |
| **6** | 0.28583 | 3.36105 | 0.00385 | 0.88982 | 0.13466 | 0.03367 | 0.07832 | 0.48691 |

*Table 4.1.2*

Now, all the VIF values are less than 10 and the condition number is less than 30 which means the above model does not show any symptoms for collinearity.

## 4.2 Normality



*Figure 4.2.1*

From the above figure (figure 4.2.1),

* The box-plot from the above figure seems symmetrically distributed except few outliers to bottom of plot one outlier being at extreme level.
* The distribution in Histogram looks normally distributed with few potential outliers to lower values of residual.
* The Normal probability plot shows few points away from the slope line on either ends of the line.

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.984663 | **Pr < W** | 0.0056 |
| **Kolmogorov-Smirnov** | **D** | 0.055382 | **Pr > D** | 0.0435 |
| **Cramer-von Mises** | **W-Sq** | 0.124684 | **Pr > W-Sq** | 0.0525 |
| **Anderson-Darling** | **A-Sq** | 0.821348 | **Pr > A-Sq** | 0.0352 |

*Table 4.2.1*

From the above table (table 4.2.1), the p-value for Kolmogorov-Smirnov test is > 0.05 which means the normality assumption is violated. As there is an extreme outlier observation with jackknife residual close to -4 which might result in violating the normality assumption. Then removing this observation from the data may result in normality. After removing this extreme outlier from the data, the test for normality table looks like below (figure 4.2.2).



*figure 4.2.2*

| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.989649 | **Pr < W** | 0.0533 |
| **Kolmogorov-Smirnov** | **D** | 0.044305 | **Pr > D** | >0.1500 |
| **Cramer-von Mises** | **W-Sq** | 0.095382 | **Pr > W-Sq** | 0.1325 |
| **Anderson-Darling** | **A-Sq** | 0.632662 | **Pr > A-Sq** | 0.0986 |

*Table 4.2.2*

After removing the outlier, now the assumption for normality is satisfied as the p-value > 0.05 from the table(table 4.2.2). But, as the outlier is removed from the data the prediction model gets affected due to the change in variation for the model. So, the re-visting the final model parameters and again checking all the diagnostics performed until now. The below tables (table 4.2.3 and 4.2.4) Shows no sign of collinearity but the parameters estimates are changed.

| **Parameter Estimates** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | 1 | 0.75163 | 0.42535 | 1.77 | 0.0784 | 0 |
| **Attractive** | 1 | 0.37456 | 0.04653 | 8.05 | <.0001 | 1.88386 |
| **Intelligent\_c** | 1 | 0.31237 | 0.06866 | 4.55 | <.0001 | 2.49150 |
| **Fun** | 1 | 0.16530 | 0.04720 | 3.50 | 0.0005 | 2.20516 |
| **SharedInterests** | 1 | 0.13722 | 0.03118 | 4.40 | <.0001 | 1.49074 |
| **Sincere** | 1 | 0.18839 | 0.05209 | 3.62 | 0.0004 | 1.66918 |
| **intg\_3\_c** | 1 | -0.02639 | 0.00750 | -3.52 | 0.0005 | 1.71116 |

*Table 4.2.3*

| **Collinearity Diagnostics (intercept adjusted)** | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Number** | **Eigenvalue** | **Condition Index** | **Proportion of Variation** | | | | | |
| **Attractive** | **Intelligent\_c** | **Fun** | **SharedInterests** | **Sincere** | **intg\_3\_c** |
| **1** | 3.23721 | 1.00000 | 0.03010 | 0.02539 | 0.02900 | 0.02432 | 0.03173 | 0.02091 |
| **2** | 1.05973 | 1.74779 | 0.04302 | 0.05343 | 0.02999 | 0.19024 | 0.00902 | 0.21313 |
| **3** | 0.60782 | 2.30780 | 0.06513 | 0.00134 | 0.00161 | 0.31672 | 0.40995 | 0.20991 |
| **4** | 0.45907 | 2.65549 | 0.40528 | 0.00504 | 0.05773 | 0.33216 | 0.46538 | 0.00175 |
| **5** | 0.35543 | 3.01792 | 0.45533 | 0.03500 | 0.70700 | 0.09571 | 0.00578 | 0.09183 |
| **6** | 0.28074 | 3.39575 | 0.00114 | 0.87981 | 0.17467 | 0.04085 | 0.07813 | 0.46248 |

*Table 4.2.4*

Now, the final model equation becomes

Where is centered variable of Intelligent and is cubic term of centered variable . R-square for this model = 0.6938

## 4.3 Homogeneity and Independence



*Figure 4.3.1*

From the above plot in the figure (figure 4.3.1)

* There is no funnelling pattern present in the above plot which means the Homogeneity test (one of the four assumption tests) is satisfied. There is no increase in the variance of the points as the predict value of like rating increases which strongly supports the Homogeneity assumption
* Also, there is not any clear-cut pattern in the above plot which means the Independence test is satisfied. Moreover, it is clear from the data that all the observations are independent to each other as those values correspond to each different couple.

Therefore, there are no apparent violations for the regression assumption

## 4.4 Outliers

Examined the outlier diagnostics, including Cook’s distance, leverage statistics, and

jackknife residuals and identified potential outliers. The below dataset (table 4.4.1) is the observations violating atleast one of three residual tests.

| **Obs** | **violated\_j** | **violated\_c** | **violated\_h** | **Like** | **preds** | **cooks** | **leverage** | **jackknife** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 0 | 0 | 1 | 7.5 | 7.69302 | 0.00060 | 0.05599 | -0.26459 |
| **2** | 0 | 0 | 1 | 4.0 | 4.40840 | 0.00407 | 0.08096 | -0.56765 |
| **3** | 0 | 0 | 1 | 4.0 | 4.52254 | 0.00500 | 0.06314 | -0.71964 |
| **4** | 0 | 0 | 1 | 4.5 | 5.19231 | 0.00698 | 0.05153 | -0.94828 |
| **5** | 1 | 0 | 0 | 4.5 | 6.39626 | 0.01362 | 0.01446 | -2.57591 |
| **6** | 1 | 0 | 0 | 6.0 | 8.08426 | 0.01754 | 0.01539 | -2.84022 |
| **7** | 1 | 0 | 0 | 9.0 | 7.46032 | 0.02492 | 0.03824 | 2.10852 |
| **8** | 0 | 0 | 1 | 4.0 | 4.46131 | 0.00652 | 0.09795 | -0.64734 |
| **9** | 0 | 0 | 1 | 6.0 | 5.72370 | 0.00138 | 0.06248 | 0.38012 |
| **10** | 0 | 0 | 1 | 6.5 | 5.79262 | 0.01164 | 0.07779 | 0.98275 |
| **11** | 1 | 0 | 1 | 7.5 | 5.67396 | 0.04975 | 0.05266 | 2.52889 |
| **12** | 1 | 0 | 0 | 8.5 | 7.02187 | 0.01748 | 0.02963 | 2.01374 |
| **13** | 1 | 0 | 0 | 4.5 | 6.36053 | 0.04104 | 0.04272 | -2.56412 |
| **14** | 0 | 0 | 1 | 8.5 | 7.46370 | 0.02338 | 0.07349 | 1.43940 |
| **15** | 1 | 0 | 1 | 7.5 | 5.87673 | 0.05604 | 0.07202 | 2.26604 |
| **16** | 1 | 0 | 0 | 5.5 | 7.18701 | 0.02868 | 0.03677 | -2.31245 |
| **17** | 1 | 0 | 0 | 3.5 | 5.40118 | 0.02819 | 0.02892 | -2.60243 |
| **18** | 0 | 0 | 1 | 3.5 | 4.92482 | 0.02947 | 0.05137 | -1.96247 |
| **19** | 0 | 0 | 1 | 3.0 | 4.00687 | 0.02595 | 0.08439 | -1.40657 |
| **20** | 1 | 0 | 0 | 5.0 | 6.98423 | 0.01069 | 0.01045 | -2.69307 |
| **21** | 0 | 0 | 1 | 6.5 | 5.97502 | 0.00454 | 0.05755 | 0.72086 |
| **22** | 0 | 0 | 1 | 6.5 | 6.07662 | 0.81909 | 0.79028 | 1.23476 |
| **23** | 0 | 0 | 1 | 6.0 | 5.63108 | 0.00318 | 0.07810 | 0.51192 |
| **24** | 0 | 0 | 1 | 7.5 | 7.15528 | 0.00245 | 0.07019 | 0.47628 |
| **25** | 1 | 0 | 0 | 6.5 | 8.41657 | 0.02020 | 0.02074 | -2.61278 |
| **26** | 0 | 0 | 1 | 7.0 | 5.90894 | 0.01872 | 0.05521 | 1.50125 |
| **27** | 1 | 0 | 0 | 3.5 | 5.70006 | 0.01384 | 0.01100 | -2.99645 |
| **28** | 0 | 0 | 1 | 8.0 | 7.50673 | 0.00688 | 0.09176 | 0.68989 |
| **29** | 0 | 0 | 1 | 6.5 | 6.37486 | 0.00028 | 0.06095 | 0.17198 |
| **30** | 1 | 0 | 0 | 7.0 | 5.35197 | 0.01483 | 0.02060 | 2.23887 |

*Table 4.4.1*

From the above dataset (Table 4.4.1),

* There are multiple observations (13 observations) violating the Jackknife test (Jackknife residuals with an absolute value more than 2).
* There are no observations violating the Cook’s distance test (Cook’s distance more than 1).
* There are multiple observations (19 observations) violating the leverage test (leverage more than 0.055).

These observations are violating the test for outliers. Therefore, these observations seems to be suspicious with low like ratings regardless all other decent ratings and need to be checked again.

## 4.5 Effects of same race and closeness in age

Effect of same race:

| **Test test\_race Results for Dependent Variable Like** | | | | |
| --- | --- | --- | --- | --- |
| **Source** | **DF** | **Mean Square** | **F Value** | **Pr > F** |
| **Numerator** | 1 | 0.05978 | 0.11 | 0.7454 |
| **Denominator** | 254 | 0.56582 |  |  |

*Table 4.5.1*

From the above table (table 4.5.1), we can conclude that the effect on same race is not significant on the predicted model.

As the p-value is way greater than 0.05, we fail to reject the null hypothesis and conclude that the same race does not have any significant difference on average like ratings.

Effect of closeness in age:

| **Test test\_age Results for Dependent Variable Like** | | | | |
| --- | --- | --- | --- | --- |
| **Source** | **DF** | **Mean Square** | **F Value** | **Pr > F** |
| **Numerator** | 1 | 1.26409 | 2.24 | 0.1357 |
| **Denominator** | 252 | 0.56426 |  |  |

*Table 4.5.2*

From the above table (table 4.5.2), we can conclude that the effect of closeness in age is not significant on the prediction model.

As the p-value is greater than 0.05, we fail to reject the null hypothesis and conclude that the closeness in age does not have any significant difference on average like ratings.

## 4.6 Reliability

Reliability is a measurement of how well our model handles future predictions. Here, the reliability for our model is evaluated by performing a split sample analysis. This analysis splits the data into two groups training and holdout groups. Then implementing the cross validation method that uses shrinkage measurement to assess reliability. The shrinkage is defined as the difference between the R-squared values of model built on training data and predicted values obtained on holdout data.

| **Simple Statistics** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** |
| **Like** | 92 | 6.60054 | 1.32654 | 607.25000 | 3.50000 | 10.00000 |
| **Yhat** | 91 | 6.57564 | 1.12645 | 598.38340 | 3.94519 | 9.81687 |

*Table 4.6.1*

| **Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations** | | |
| --- | --- | --- |
|  | **Like** | **Yhat** |
| **Like** | 1.00000  92 | 0.86384 <.0001 91 |
| **Yhat** | 0.86384 <.0001 91 | 1.00000  91 |

*Table 4.6.2*

From the table (table 4.6.2), the correlation coefficient between the variables Like and Yhat, r = 0.86384.

Therefore, R-square from holdout data = 0.7422

Now, Shrinkage = 0.7422 – 0.6938 = 0.0484.

Since this value is less than 0.10, our model is reliable.

# 5. Summary of Findings

* An Equation is created that can predict a dater’s opinion of the person they are dating based on the attractiveness, sincerity, intelligence, fun, ambitiousness and sharedinterests of their partner. The final equation created for prediction is:

* From the equation, it is evident that ambitiousness does not have significant amount in predicting the like ratings of partner. Also, from the descriptive statistics stated previously, It is evident that the people likes to choose other partner who is more attractive and fun.
* In addition, it does not matter whether the partners are same race or not. The interaction term is not significant which means the average like ratings does not significantly differ between same race couples and different race couples. As part of descriptive statistics, there are 121 couples with same race and 148 couples with different race.
* Also, checked whether it matters or not if partners are close in age. However, it can be concluded that the closeness in age between the partners does not have any impact. The interaction term for this model is not significant which means that the average like ratings does not significantly differ whether the partners are close in age or not. As part of descriptive statistics, there are 110 partners close in age whereas 157 partners are not close in age.
* From the table (table 4.6.2), the correlation between the variables Yhat and Like describes how estimated like ratings are related to actual like ratings. For our model the correlation coefficient is 0.8638 which means there is a strong linear relationship between the estimated and actual like ratings. Also, the R-square for the predicted model = 0.6938 which means that roughly 70% of the data fits the regression model employed. The shrinkage measurement from reliability test of our model describes how estimated like ratings are related to actual like ratings. For our model, the shrinkage value is 0.04 which is less than 0.10 and this shrinkage value states the difference between the coefficients of determination on training and holdout data. This means our model predicts the values which has very less difference from actual values.
* From the outlier section in diagnostics, it is clear that there are few like ratings which seems to be suspicious and need to recheck the data once. The like ratings for these observations are extremely low irrespective all decent ratings in attractiveness, sincerity, fun and intelligent. The number of observations and their respective data are mentioned on outlier section of Diagnostics. (refer section 4.4)