Ching Yi Lam

Introduction

The online dating industry is growing with recent quarterly profits well more than millions. The purpose of this report is to analyze the data for speed dating. We will use the data that we collected from the speed dating event where 276 heterosexual couples were randomly paired up with one another for a short speed date to optimally match couples.

In the first part of the report, we will have the descriptive statistics for this data. Then, we will see the selection of the models and the type of analysis employed. Afterward, it will be the diagnostics of the study. Last but not least, it will be the summary of the findings.

Descriptive Statistics

This dataset has 276 observations and 18 variables. We put M or F after each variable to show that variable refers to male's or female's opinions. The Like variable indicates how much the observer likes the partner. Age is the age of the observers. Race variable contains 5 different options: Caucasian, Asian, Black, Latino, or Other. Attractive is the rating of attractiveness of partner. Sincere is the rating of sincerity of partner. Intelligent is the rating of intelligence of partner. Fun is the rating of how fun partner is. Ambitious is the rating of ambition of partner. Shared Interests is the rating of the extent to which they share interests or hobbies. The variables of Attractive, Sincere, Intelligent, Fun, Ambitious, and Shared Interest are on a scale of 1 to 10 with 1 equal to awful and 10 equal to great.

Here is the output of the total number of observations that do not contain missing data, mean, standard deviation, minimum, and maximum of the variables:

TABLE 1: Descriptive Statistics Output for the variables

Descriptive Statistics Output for the variables

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
LikeM	274	6.6824818	1.7793853	1.0000000	10.0000000
LikeF	272	6.3658088	1.7551737	1.0000000	10.0000000
AgeM	273	26.6043956	3.5132495	18.0000000	42.0000000
AgeF	271	26.1881919	3.9802299	19.0000000	55.0000000
AttractiveM	273	6.6868132	1.8073304	1.0000000	10.0000000
AttractiveF	274	6.2737226	1.9176936	1.0000000	10.0000000
SincereM	271	7.8560886	1.4973896	1.0000000	10.0000000
SincereF	273	7.7838828	1.6936832	1.0000000	10.0000000
IntelligentM	268	7.6212687	1.3750162	4.0000000	10.0000000
IntelligentF	273	7.9230769	1.4237799	2.0000000	10.0000000
FunM	270	6.8629630	1.7921033	0	10.0000000
FunF	270	6.5629630	1.9647731	1.0000000	10.0000000
AmbitiousM	259	6.7683398	1.7299240	2.0000000	10.0000000
AmbitiousF	266	7.4285714	1.7731146	1.0000000	10.0000000
SharedInterestsM	249	5.5883534	2.0781663	0	10.0000000
SharedInterestsF	246	5.4695122	2.2758908	0	10.0000000

The data included some data entry errors. Our rating scale is between 1 to 10 and in the dataset, the variables of variables Fun for male, Shared Interests for male, and Shared Interests for female have a rating of 0 which is not part of the rating scale that we provided.

The table below shows the number of people in each race for male and female:

TABLE 2: Number of People in Each Race for Male and Female

NUMBER OF PEOPLE IN EACH RACE FOR MALE & FEMALE

The FREQ Procedure

RaceM	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asian	65	23.72	65	23.72
Black	10	3.65	75	27.37
Caucasian	161	58.76	236	86.13
Latino	17	6.20	253	92.34
Other	21	7.66	274	100.00
Frequency Missing = 2				

RaceF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asian	70	25.74	70	25.74
Black	15	5.51	85	31.25
Caucasian	148	54.41	233	85.66
Latino	23	8.46	256	94.12
Other	16	5.88	272	100.00
Frequency Missing = 4				

121 partners are the same race and 155 partners are not the same race. 111 partners are close in age, and we are defined as within 2 years of one another. There are 165 partners that their age differences are more than 2 years.

The data separated the variables for male and female so we will use 2 models to show how the result will be.

Models and Type of Analysis

We will build 2 models by separating male and female ratings using variables Age different, Race same, Attractive, Sincere, Intelligent, Fun, Ambitious, Shared Interests to predict how much the male liked the female or how much the female liked the male. The first model is the output of male's rating in the speed dating event. The second model is the output of female's rating in the

speed dating event. I used Stepwise Forward selection strategy to find the best model for both output of male and female. We use alpha = 0.05 as the significance level.

TABLE 3: The Output of Male

Stepwise Selection: Step 4

Variable FunM Entered: R-Square = 0.6449 and C(p) = 6.2737

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	469.44523	117.36131	101.68	<.0001
Error	224	258.53948	1.15419		
Corrected Total	228	727.98472			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.21501	0.39640	0.33958	0.29	0.5881
AttractiveM	0.53065	0.05160	122.05491	105.75	<.0001
SincereM	0.19627	0.05407	15.21113	13.18	0.0004
FunM	0.14226	0.05393	8.03229	6.96	0.0089
SharedInterestsM	0.13828	0.04021	13.64883	11.83	0.0007

Based on the stepwise forward selection strategy, we found out that the variables of Attractive, Sincere, Fun, and Shared Interests are significant in affecting the rating of how much the male liked the female with the significance level of alpha = 0.05.

The equation for predicting like variable of male will be: $\hat{Y}_M = -0.22 + 0.53X_A + 0.20X_s + 0.14X_F + 0.14X_{Share}$.

TABLE 4: The Output for Female

Stepwise Selection: Step 4

Variable IntelligentF Entered: R-Square = 0.6176 and C(p) = 4.4888

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	438.41534	109.60383	90.86	<.0001
Error	225	271.42814	1.20635		
Corrected Total	229	709.84348			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.17071	0.42754	0.19233	0.16	0.6901
AttractiveF	0.28402	0.04517	47.69181	39.53	<.0001
IntelligentF	0.21621	0.05396	19.37074	16.06	<.0001
FunF	0.23601	0.04931	27.64083	22.91	<.0001
SharedInterestsF	0.21685	0.04000	35.46348	29.40	<.0001

Bounds on condition number: 1.8203, 24.051

Based on the stepwise forward selection strategy, we found out that the variables of Attractive, Intelligent, Fun, and Shared Interests are significant in affecting the rating of how much the female liked the male with the significance level of alpha = 0.05.

The equation for predicting like variable of female will be: $\hat{Y}_F = 0.17 + 0.28X_A + 0.22X_I + 0.24X_F + 0.22X_S$.

For male and female models, forward selection, backwards elimination, and stepwise forward selection came out with the same optimal variables. If the same variables appeared when I used different methods, I think it is more accurate for our best model. For the best model of both

male and female models, they do not include age and race variables, so it does not matter if the partners are the same race or close in age by 2 years.

Diagnostics

FIGURE 1: Jackknife Residual Plot for Male

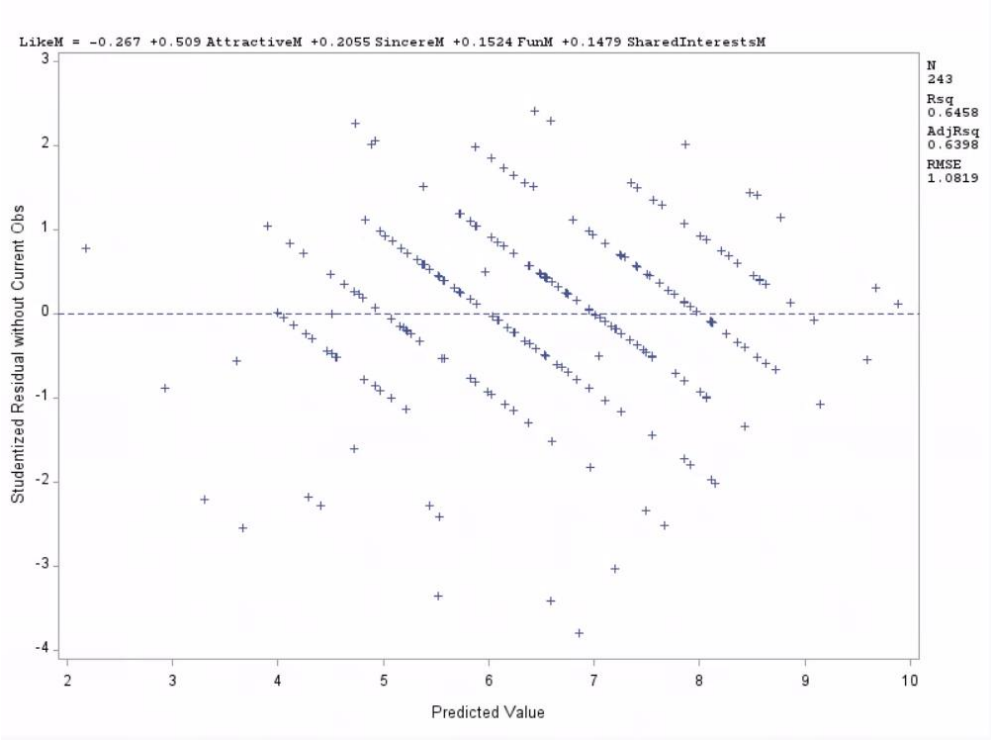
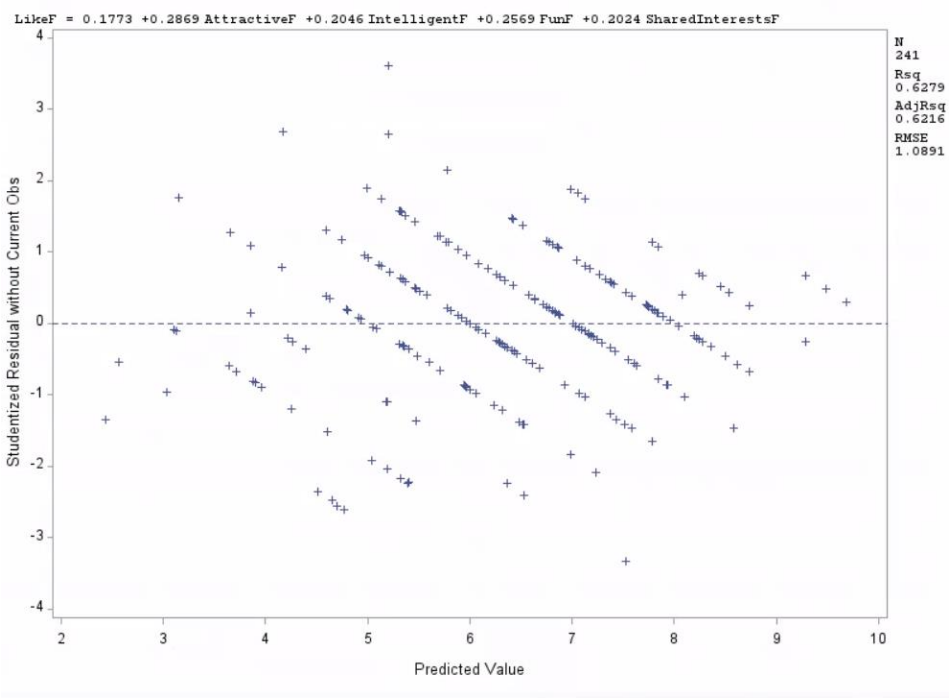
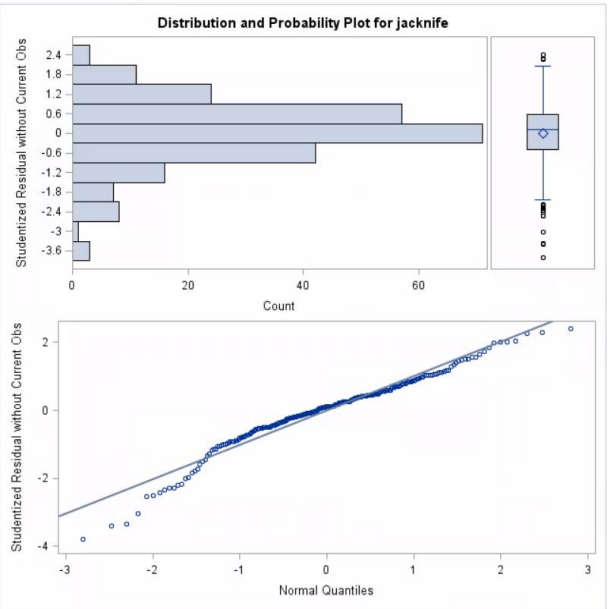


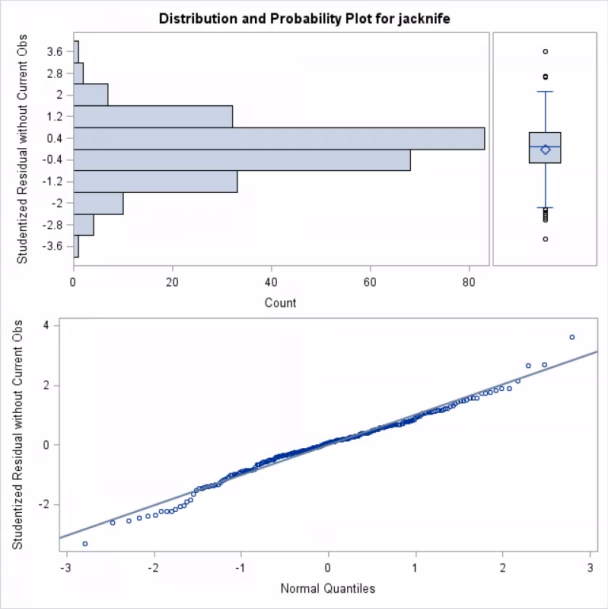
FIGURE 2: Jackknife Residual Plot for Female



When we look at the jackknife residual plot for male and female, we can see that the data satisfies all the regression assumptions. These are ideal scatterplots of residuals versus the predicted values should look like a random scatter about the line with mean = 0 and they also do not show any systematic trends.

FIGURE 3: Distribution and Probability Plot for Jackknife of Male (Left) and Female (Right)





When we look at the histogram for male and female, we can see that the data follows a normal distribution. The normal probability plot for male and female shows a linear pattern which indicates that the normality assumption is not violated. It has 19 outliers for male and 16 outliers for female based on jackknife residuals with an absolute value of more than 2.

TABLE 5: Collinearity Diagnostics Table for Male

Collinearity Diagnostics (intercept adjusted)						
			Proportion of Variation			
Number	Eigenvalue	Condition Index	AttractiveM	SincereM	FunM	SharedInterestsM
1	2.37654	1.00000	0.06787	0.06355	0.06736	0.06751
2	0.68383	1.86422	0.05571	0.71587	0.00063509	0.29186
3	0.56102	2.05818	0.34311	0.11271	0.08787	0.63791
4	0.37861	2.50541	0.53332	0.10787	0.84413	0.00272

TABLE 6: Collinearity Diagnostics Table for Female

Collinearity Diagnostics (intercept adjusted)						
			Proportion of Variation			
Number	Eigenvalue	Condition Index	AttractiveF	IntelligentF	FunF	SharedInterestsF
1	2.25328	1.00000	0.07811	0.05156	0.07813	0.07782
2	0.82652	1.65113	0.06712	0.84837	0.00087848	0.07507
3	0.52304	2.07558	0.62885	0.00017167	0.00000855	0.61506
4	0.39716	2.38190	0.22592	0.09990	0.92098	0.23206

TABLE 7: Variance Inflation Factor Table for Male

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.26701	0.39086	-0.68	0.4952	0
AttractiveM	1	0.50903	0.05040	10.10	<.0001	1.68410
SincereM	1	0.20551	0.05324	3.86	0.0001	1.34293
FunM	1	0.15245	0.05289	2.88	0.0043	1.88888
SharedInterestsM	1	0.14793	0.03945	3.75	0.0002	1.39095

TABLE 8: Variance Inflation Factor Table for Female

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.17728	0.41482	0.43	0.6695	0
AttractiveF	1	0.28693	0.04420	6.49	<.0001	1.53815
IntelligentF	1	0.20456	0.05263	3.89	0.0001	1.16668
FunF	1	0.25687	0.04772	5.38	<.0001	1.84312
SharedInterestsF	1	0.20236	0.03854	5.25	<.0001	1.53549

Ching Yi Lam

When we checked for the collinearity for male and female models, we looked at the condition index column. We generally only consider the largest number which is located at the bottom of the column, and we can see that the condition number of the male's model is 2.51 and the female's model is 2.38 which is less than 30. We also checked that none of the variables in both models have a variance inflation factor of more than 10, so collinearity does not appear to be an issue.

Summary of Findings

We found out that for male to match with the female, the important variables are attractiveness, sincerity, fun, and shared interests. For female to match with the male, the important variables are attractiveness, intelligence, fun, and shared interests. They share 3 variables to match with the partner.

We can try to put an observation to our male's and female's equation to see if the predicted value will be close to how the actual rating is. To understand the models, we can look at a random observation. For example, observation 4 has the following results:

For male: $\hat{Y}_M = -0.22 + 0.53X_A + 0.20X_s + 0.14X_F + 0.14X_{Share}$

$= -0.22 + 0.53(9) + 0.20(9) + 0.14(9) + 0.14(9)$

$= 8.87$

Observation 4 $Y_M = 9$

For female: $\hat{Y}_F = 0.17 + 0.28X_A + 0.22X_I + 0.24X_F + 0.22X_S$

$= 0.17 + 0.28(7) + 0.22(7) + 0.24(6) + 0.22(7)$

= 6.65

Observation 4 $Y_F = 7$

Our calculation above shows that the predicted value is close to our actual value which is the rating of how much male liked the female is 9 and how much female liked the male is 7. Using the variables of attractiveness, sincerity, fun, and shared interests can predict how much male will like the female. Using the variables of attractiveness, intelligence, fun, and shared interests can predict how much female will like the male.

By performing hypothesis tests, checking residual plots, and diagnosing collinearity, we found 2 models listed above. We are confident these two models can match couples optimally.