A strategy for speed Dating

YuTing Tian

ABSTRACT

Online dating is a growing industry with recent quarterly profits well in excess of millions. The goal of speed dating is to break into this industry by using the power of statistics to optimally match couples.

In order to attain this target, we will use a speed dating dataset to generate models to test feasibility of using statistics to match couples. This dataset is an open source on R package. The participants were students at Columbia's graduate and professional schools, recruited by mass email, posted fliers, and fliers handed out by research assistants. Each participant attended one speed dating session, in which they met with a participant of the opposite sex for four minutes. Order and session assignments were randomly determined. After each four minute "speed dating" encounter , participants filled out a form rating their date on scale of 1 to 10 on various attributes.

In this paper I use R software to download the dataset package, then use SAS® to analyze the data.

INTRODUCTION

In this paper, the author acknowledges that some of the figures are small; t is suggested that a reader might download the PDF and print the paper .

This paper is divided into the following four main sections:

1) Brief introduction for the speed dating data.

2) Some Basic Hypothesis

2.1) Impact of race2.2) Impact of age range

2.3) the correlation between like ratings and other independent variables

3) Building a model

3.1) Strategy to build the model

3.2) Collinearity test for the candidate model

3.3) Regression diagnostics for the final model

3.4) reliability evaluation for the final model

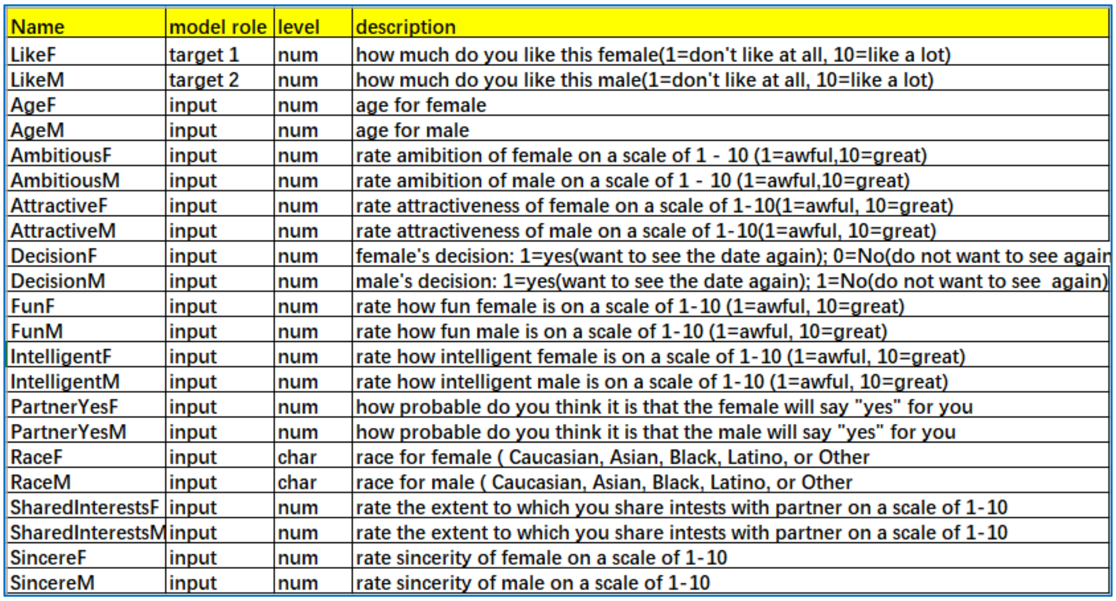
4) Conclusion

1) BRIEF INTRODUCTION FOR OUR SPEED DATING DATA

First the original dataset was cleaned, for example, converting the “NA” in the character variable, to missing value in the numerical variable (see Appendix), all the polynomial and interaction variables were added, for example Ambitious2=Ambitious**2, AA3=Attractive*Ambitious.

Then I split the data into two parts: training dataset 1/3 and test dataset 2/3

The chart in Figure 1 presents all the variables in the dataset. The target variables (likef likem) (numeric), indicating how much do you like this person



Name	model role	level	description
LikeF	target 1	num	how much do you like this female(1=don't like at all, 10=like a lot)
LikeM	target 2	num	how much do you like this male(1=don't like at all, 10=like a lot)
AgeF	input	num	age for female
AgeM	input	num	age for male
AmbitiousF	input	num	rate amibition of female on a scale of 1 - 10 (1=awful,10=great)
AmbitiousM	input	num	rate amibition of male on a scale of 1 - 10 (1=awful,10=great)
AttractiveF	input	num	rate attractiveness of female on a scale of 1-10(1=awful, 10=great)
AttractiveM	input	num	rate attractiveness of male on a scale of 1-10(1=awful, 10=great)
DecisionF	input	num	female's decision: 1=yes(want to see the date again); 0=No(do not want to see agair
DecisionM	input	num	male's decision: 1=yes(want to see the date again); 1=No(do not want to see again)
FunF	input	num	rate how fun female is on a scale of 1-10 (1=awful, 10=great)
FunM	input	num	rate how fun male is on a scale of 1-10 (1=awful, 10=great)
IntelligentF	input	num	rate how intelligent female is on a scale of 1-10 (1=awful, 10=great)
IntelligentM	input	num	rate how intelligent male is on a scale of 1-10 (1=awful, 10=great)
PartnerYesF	input	num	how probable do you think it is that the female will say "yes" for you
PartnerYesM	input	num	how probable do you think it is that the male will say "yes" for you
RaceF	input	char	race for female (Caucasian, Asian, Black, Latino, or Other
RaceM	input	char	race for male (Caucasian, Asian, Black, Latino, or Other
SharedInterestsF	input	num	rate the extent to which you share intests with partner on a scale of 1-10
SharedInterestsM	input	num	rate the extent to which you share intests with partner on a scale of 1-10
SincereF	input	num	rate sincerity of female on a scale of 1-10
SincereM	input	num	rate sincerity of male on a scale of 1-10

Figure1

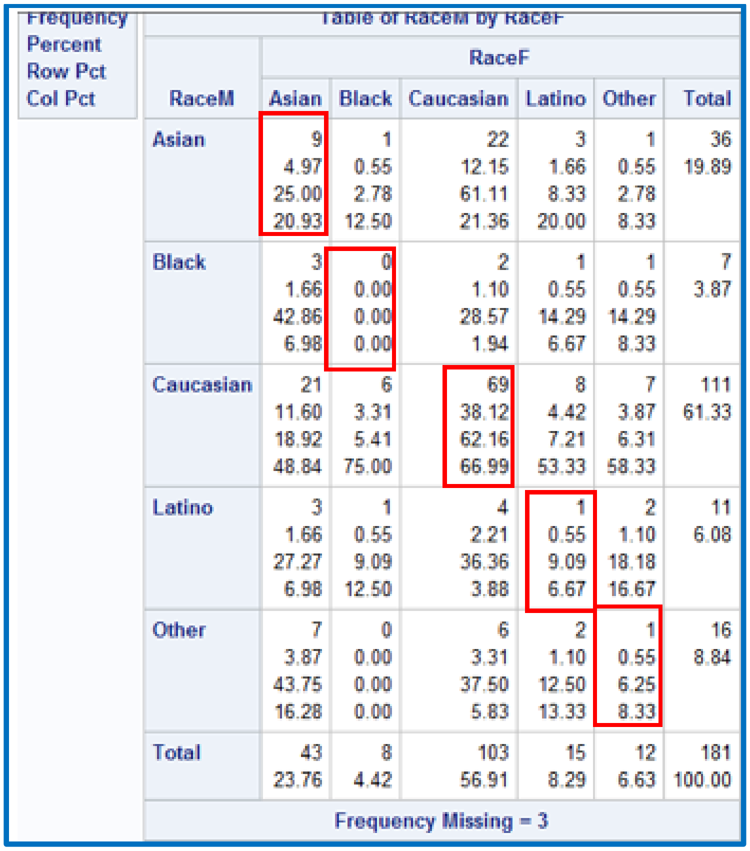
For each couple, 10 pair of input variables were recorded, for example ***AmbitiousF, AmbitiousM***, with each variable having a scale from 1 to 10. we will want to use each variable in the modeling to determine as a match

Each dater’s scale is from 0 to 10 based on the opinion of the person, as indicated by the like variable. based on the attractiveness, sincerity, Intelligence, Fun, ambitious and shared interest of their partner

2) SOME BASIC HYPOTHESIS

2.1) IMPACT OF RACE

The first test to be performed is the effect of same race on the ***Like*** variable



Frequency
Percent
Row Pct
Col Pct

Table of RaceM by RaceF						
	RaceF					
RaceM	Asian	Black	Caucasian	Latino	Other	Total
Asian	9 4.97 25.00 20.93	1 0.55 2.78 12.50	22 12.15 61.11 21.36	3 1.66 8.33 20.00	1 0.55 2.78 8.33	36 19.89
Black	3 1.66 42.86 6.98	0 0.00 0.00 0.00	2 1.10 28.57 1.94	1 0.55 14.29 6.67	1 0.55 14.29 8.33	7 3.87
Caucasian	21 11.60 18.92 48.84	6 3.31 5.41 75.00	69 38.12 62.16 66.99	8 4.42 7.21 53.33	7 3.87 6.31 58.33	111 61.33
Latino	3 1.66 27.27 6.98	1 0.55 9.09 12.50	4 2.21 36.36 3.88	1 0.55 9.09 6.67	2 1.10 18.18 16.67	11 6.08
Other	7 3.87 43.75 16.28	0 0.00 0.00 0.00	6 3.31 37.50 5.83	2 1.10 12.50 13.33	1 0.55 6.25 8.33	16 8.84
Total	43 23.76	8 4.42	103 56.91	15 8.29	12 6.63	181 100.00

Frequency Missing = 3

Figure 2

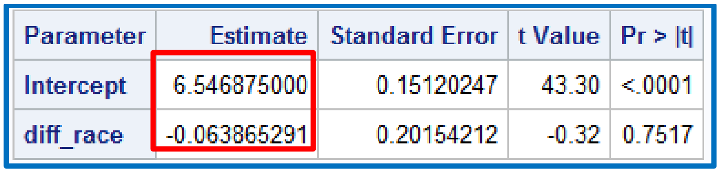
From the cross-tabulation above, we see there are 80 same-race couples (Figure 2).

Our model then looks like:

Model: like=β0+β1*diff_race;

where the β0 is the intercept and β1 is the first regression coefficient; measures how difference with the predicted value of ***Like*** will to be while the partners are the same race. I create the dummy variable called diff_race, among partners, when female's race different with male's race, the diff_race equals to 1; otherwise, the diff_race equals to 0.

Then, the Null hypothesis (H0) means there is no significant relationship between a partner of different race and like variable; otherwise, the alternative hypothesis (Ha) means there is significant relationship between a partner of different race and like variable existed in this case.



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	6.546875000	0.15120247	43.30	<.0001
diff_race	-0.063865291	0.20154212	-0.32	0.7517

Figure 3

I run the regression model, from the figure 3, we see the intercept is 6.5469, it means the expected value of ***"like"*** is 6.5469 when a partner's race is the same.

Then, we see the coefficient of diff_race is -0.0639, it means on average, the expected value of ***"like"*** will be 0.0639 less when a partner's race is the difference

We see the P value is 0.7517, greater than 0.05; it means there is no significantly difference on average "like" value whether race is the same or not. Then we accept the null hypothesis, because there is no significantly different.

2.2) THE IMPACT OF AGE RANGE

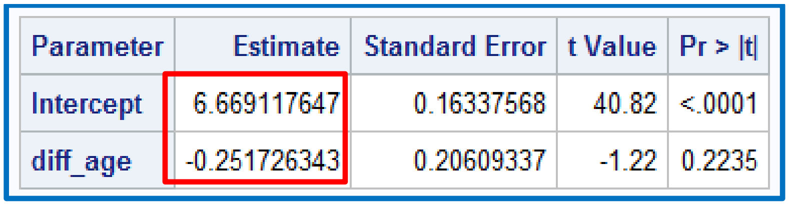
The second test to be performed the effect of age range on the ***Like*** variable. The same age range is defined by being within 2 years of one another.

The model then looks like:

Model: like=β0+β1*diff_age;

where the β0 is the intercept when the independent variables are equal to 0. β1 is the first regression coefficient, measures the difference with the predicted value of ***"like"*** when the partners are the same age range or not. binary dummy variable, diff_age, was created when the partner's the same age range within 2 years, the value 1; otherwise 0;

Then, the Null hypothesis (H0) means there is no significant relationship between a partner of different age range and ***like*** variable; otherwise, the alternative hypothesis (Ha) means there is significant relationship between a partner of different age range and ***like*** variable



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	6.669117647	0.16337568	40.82	<.0001
diff_age	-0.251726343	0.20609337	-1.22	0.2235

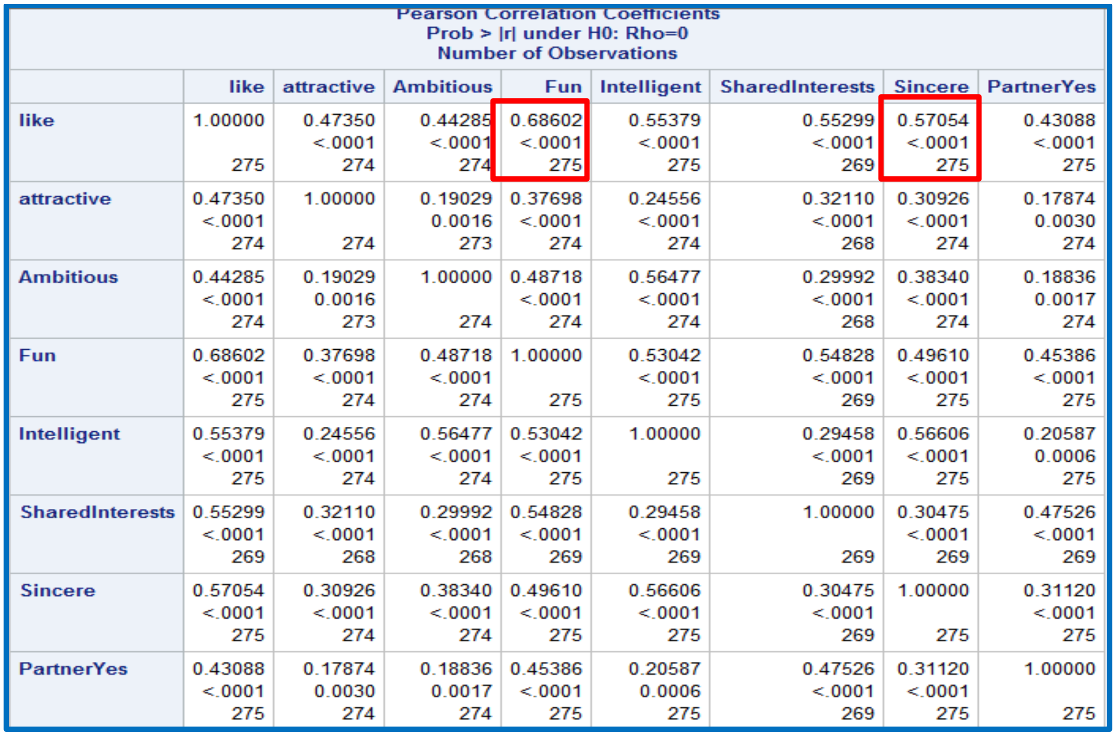
Figure 4

The regression model, shown figure 4 . we see the intercept is 6.6691, it means the expected value of **"like"** is 6.6691 when a partner's age range is the same.

Then, we see the coefficient of diff_race is -0.2517, it means on average, the expected value of *"like"* will be 0.2517 lower when a partner's age range is different. We see the P value is 0.2235, greater than 0.05; it means there is no significantly different on average "like" value whether the partner has same age range or not. Therefore, we will accept the H0

2.3) THE CORRELATION BETWEEN LIKE RATINGS AND OTHER INDEPENDENT VARIABLES

I want to see the whole correlation between independent variables and ***like*** without taking account into the factor gender, Therefore I get the mean ratings for selected predictor variables on each couple. Then create the Pearson Correlation Coefficient table shown on the figure 5;



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	like	attractive	Ambitious	Fun	Intelligent	SharedInterests	Sincere	PartnerYes
like	1.00000 275	0.47350 <.0001 274	0.44285 <.0001 274	0.68602 <.0001 275	0.55379 <.0001 275	0.55299 <.0001 269	0.57054 <.0001 275	0.43088 <.0001 275
attractive	0.47350 <.0001 274	1.00000 274	0.19029 0.0016 273	0.37698 <.0001 274	0.24556 <.0001 274	0.32110 <.0001 268	0.30926 <.0001 274	0.17874 0.0030 274
Ambitious	0.44285 <.0001 274	0.19029 0.0016 273	1.00000 274	0.48718 <.0001 274	0.56477 <.0001 274	0.29992 <.0001 268	0.38340 <.0001 274	0.18836 0.0017 274
Fun	0.68602 <.0001 275	0.37698 <.0001 274	0.48718 <.0001 274	1.00000 275	0.53042 <.0001 275	0.54828 <.0001 269	0.49610 <.0001 275	0.45386 <.0001 275
Intelligent	0.55379 <.0001 275	0.24556 <.0001 274	0.56477 <.0001 274	0.53042 <.0001 275	1.00000 275	0.29458 <.0001 269	0.56606 <.0001 275	0.20587 0.0006 275
SharedInterests	0.55299 <.0001 269	0.32110 <.0001 268	0.29992 <.0001 268	0.54828 <.0001 269	0.29458 <.0001 269	1.00000 269	0.30475 <.0001 269	0.47526 <.0001 269
Sincere	0.57054 <.0001 275	0.30926 <.0001 274	0.38340 <.0001 274	0.49610 <.0001 275	0.56606 <.0001 275	0.30475 <.0001 269	1.00000 275	0.31120 <.0001 275
PartnerYes	0.43088 <.0001 275	0.17874 0.0030 274	0.18836 0.0017 274	0.45386 <.0001 275	0.20587 0.0006 275	0.47526 <.0001 269	0.31120 <.0001 275	1.00000 275

Figure 5

There is a strong correlation coefficients between ***like*** and ***fun***, about 0.6861; and strong correlation coefficient between ***like*** and ***sincere***, around 0.5705 We see people are more likely to choose a partner who is ***fun, sincere*** and ***intelligent***.

.

All following procedures, I use training dataset;

In order to analyze the relationship between partner's ***like*** and ***age*** across over five different **races**, separated by gender. , so I use sgpanel procedure to give us a visual temptation.

The sgpanel procedure produces nearly the same types of graphics but instead of displaying only one plot per image, sgpanel can display several plots in a single image. For example , ***like by age by race***, see the figure 6 and figure 7

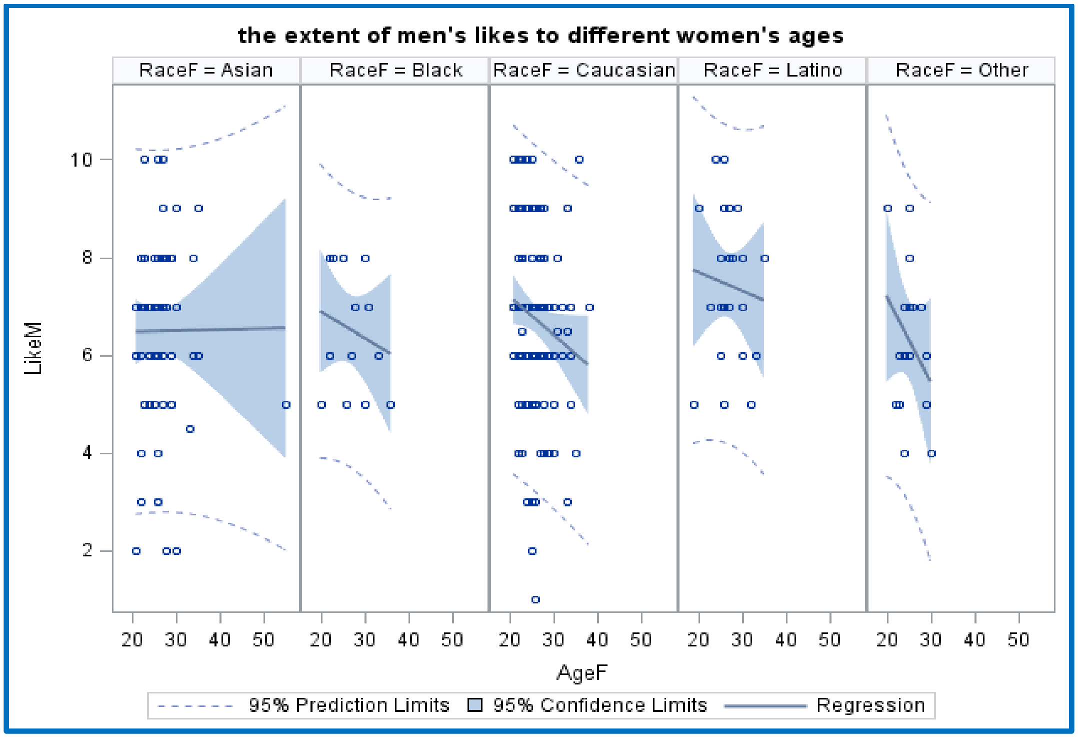


Figure 6

Figure 6 shows how the relationship looks like between females' ***ages*** and male's ***like*** across over different female races.

We see on average, the extent of men's like almost the same as women's ages growing, when the women's race are Asian. But the shadow of 95% confidence interval is wider as women's ages go up. It means the distribution of men's like are spread above and beyond the predicted value. As ages growing, the variance becomes greater

In addition, we see the extent of man's ***like*** drops down as woman's ***age*** goes up over across the other races.

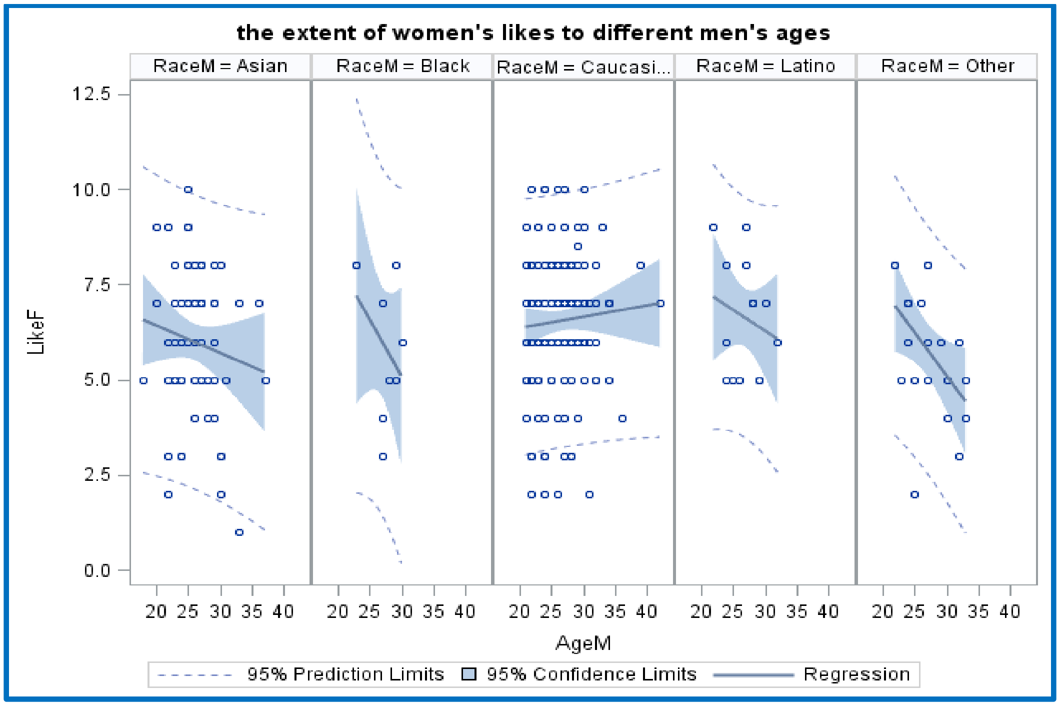


Figure 7

Figure 7 is similar as Figure 6, showing man's opinion to woman.
We see on average, the extent of men's ***like*** are increased gradually as women's ***age*** growing, when the women's race is Caucasian.

Thereafter, we see the extent of men's like is declined as women's ages rise up over across the other races.

3) BUILDING A MODEL

3.1) STRATEGY TO BUILD THE MODEL

When we face with a predicted modeling problem, we use statistical model selections searching for the best model. Methods include long familiar in the Reg procedure (all possible subsets regression, forward selection, backwards elimination, and stepwise forward selection). I will use these four different strategies to create different models for two reasons.

The first reason is to try and create enough examples so that the effect of different parameters could be deduced from different methods.

The second reason is that different strategy has its own procedure:

For the all possible subsets regression procedure, identify all possible regression models derived from all possible combinations of independent variable. With many independent variables, the number of combinations would be unwieldy.

For the forward selection method, it starts with no variable in the model. For each of independent variable, the forward method calculates F statistics that reflects the variable's contribution to the model. If the p value of F statistics has a significance level greater than the SLENTRY= value, the forward selection will be stopped otherwise the forward method adds this variable into the model.

Therefore, variables are added one at a time and remain in the model if they produce a significant F statistics. I set up the SLENTRY=0.05 in this method, because I want the variables whose significant values less than 0.05 can be added in the model. In the majority of analysis, an alpha is 0.05 is used as a cutoff value of significance.

For the backward selection method, it starts with all predicted variables. Then the variables are removed in the model one by one until all the variables remaining in the model produce F statistics significant at the SLSTAY= value. At each step with the variable shows the smallest contribution to the model is removed I set up the SLSTAY=0.05 in this method, because I want the variables whose significant values less than 0.05 can be kept in the model.

For the stepwise forward selection, it similar as forward selection, except a variable selected into the model can be removed later from the model if the significant level falls below a set up value. It combines options with both SLENTRY=value and SLSTAY=value. I set up the SLENTRY=0.5, SLSTAY=0.05 in this model, because I want the variables whose significant values less than 0.05 can be added, among those variables, the significant values less than 0.05 can be kept in the model.

The goal is to get the “best” model after I run the different selection processes are run. . The results from each model will be compared with the each model with some common criteria to determine which model is the most reasonable and useful in this case.

The following are some common criteria that we use to rank a model:

1) Multiple R^2

2) MSE(P)

3) Mallows $C_p = \frac{MSE(p)}{MSE(K)}[n-p]-n+2p$

The maximum R^2 is a technique that tries to find the best set of variables to include in the model, but it is not guaranteed to find the best model with highest R^2

The MSE(P) is the estimated error variance of the p variable model.

The Mallows C_p statistic is often used as a stopping rule for various forms of regression C_p has expectation nearly equal to P. We need to find the point where C_p is just less than or equal to P.

Method	Number in Model (P)	R square	C(p)	MSE(P)	Variables in Model
all possible subsets regression	4	0.6358	2.5502	0.6888	attractive SharedInterestes2 Sincere2 FI4
forward selection	3	0.6196	8.0737	0.7152	AS7 FI4 FS5
backward elimination	10	0.6612	2.7245	0.6639	Ambitious Fun Intelligent Sincere Ambious2 Fun2 Intelligent2 AI2 AS5 FS4
stepwise forward selection	3	0.6196	8.0737	0.7152	AS7 FI4 FS5

Figure 8

The Figure 8 shown above, after I run the regression models with four methods, I combine the outputs from SAS into the excel sheet, then compare the results across over model selection methods with those common criteria.

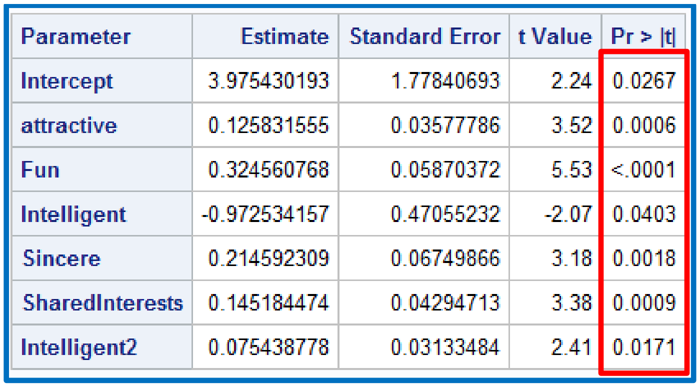
We notice some variables in the Figure 8 are not the original variables in the raw data, for example, AI2=Attractive*Intelligent Fun2=Fun**2, these are the polynomial and interaction variables that were created (see Appendix for code), then give them brief names for each created variable. All codes shown on the end of this paper.

First, look at Mallows C_p criterion, we will remove forward selection and stepwise forward selection out based on the standard with $C_p \leqslant P$; then followed by $R^{2,}$, the R^2 value of backward elimination model and all possible selection model are higher than other 2 models, and last thing I will consider MSE(P), comparing remaining two models between all possible subsets regression and backward elimination, we see the backward elimination model is better because of lower MSE(P).

Therefore, the candidate model is the one created from the backward elimination method.

In the model, we need to add base term manually if the automated model building program includes higher order polynomial term or interaction term. Based on this rule, the candidate model needed to be fixed with adding base term, for example we see there is an interaction variable called FS4, it is interaction between *Fun and SharedInterests*, but there is no base term *SharedInterests* in the candidate model, so I will add *SharedInterests* to fix the model.

After I added all base terms needed in the model, then run the updated regression model, then we see some variables' p values are greater than significant level 0.05, then remove one variable at a time until all remaining variables produce p values are less than significance level, 0.05.



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	3.975430193	1.77840693	2.24	0.0267
attractive	0.125831555	0.03577786	3.52	0.0006
Fun	0.324560768	0.05870372	5.53	<.0001
Intelligent	-0.972534157	0.47055232	-2.07	0.0403
Sincere	0.214592309	0.06749866	3.18	0.0018
SharedInterests	0.145184474	0.04294713	3.38	0.0009
Intelligent2	0.075438778	0.03133484	2.41	0.0171

Figure 9

After I adjusted the candidate model following the process prescribed above.

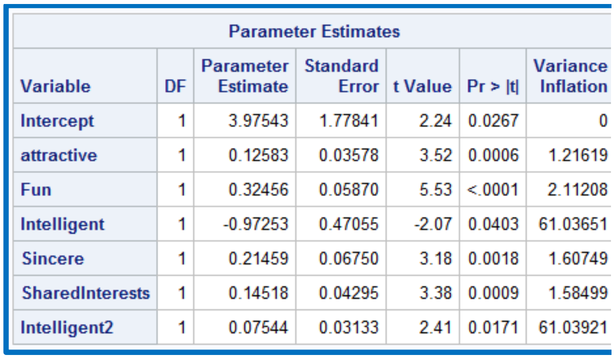
The adjusted candidate model is shown on the below:

Like~3.975+0.1258*attractive+0.3247*fun-0.9725*intelligent+0.2146*sincere+0.1452*SharedInterests

+0.0754*Intelligent2

3.2) COLLINEARITY TEST FOR THE CANDIDATE MODEL

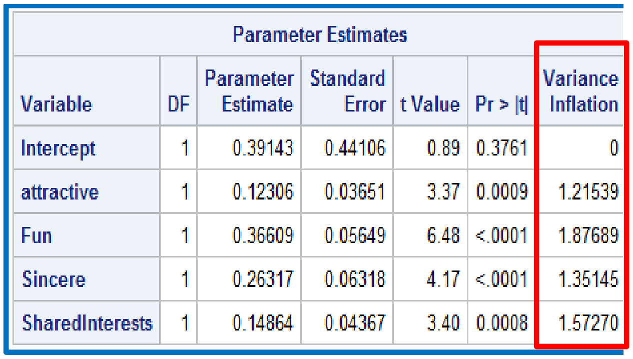
Collinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this step, I will focus on collinearity since this is often a major problem in polynomial regression. We always use two criteria to test collinearity issue, condition number and variance inflation factor (VIF), it is and index that measures how much the variance of estimated regression coefficient is increased because of collinearity. By default, although the condition number is less than 30 but the Variance Inflation Factors that exceeds 10, it means the collinearity exists in the model.



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	3.97543	1.77841	2.24	0.0267	0
attractive	1	0.12583	0.03578	3.52	0.0006	1.21619
Fun	1	0.32456	0.05870	5.53	<.0001	2.11208
Intelligent	1	-0.97253	0.47055	-2.07	0.0403	61.03651
Sincere	1	0.21459	0.06750	3.18	0.0018	1.60749
SharedInterests	1	0.14518	0.04295	3.38	0.0009	1.58499
Intelligent2	1	0.07544	0.03133	2.41	0.0171	61.03921

Figure 10

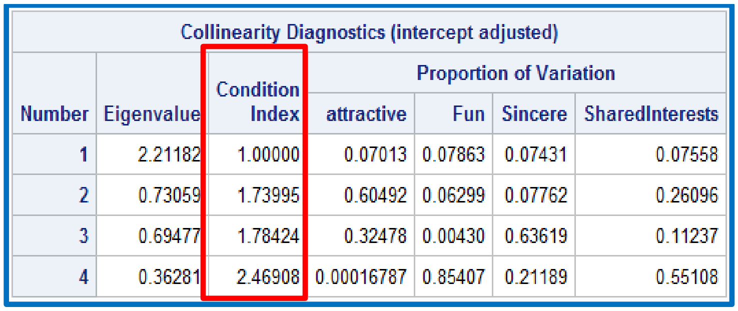


Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.39143	0.44106	0.89	0.3761	0
attractive	1	0.12306	0.03651	3.37	0.0009	1.21539
Fun	1	0.36609	0.05649	6.48	<.0001	1.87689
Sincere	1	0.26317	0.06318	4.17	<.0001	1.35145
SharedInterests	1	0.14864	0.04367	3.40	0.0008	1.57270

Figure 11

In the original candidate model (Figure10), we see there are two variables that have high variance inflation factors, intelligent and intelligent2; there is serious collinearity in the model because of the polynomial. Because of this collinearity, I use different methods trying to fix the model, one way is to centralize the independent variable, intelligent. Another way is to delete the base term and quadratic term on intelligent; this is the method used. After removing the term, the variance inflation values are all lower than 10 (Figure 11).



Collinearity Diagnostics (intercept adjusted)

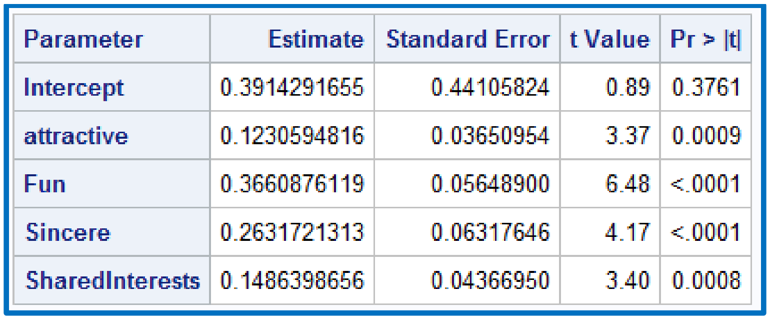
			Proportion of Variation			
Number	Eigenvalue	Condition Index	attractive	Fun	Sincere	SharedInterests
1	2.21182	1.00000	0.07013	0.07863	0.07431	0.07558
2	0.73059	1.73995	0.60492	0.06299	0.07762	0.26096
3	0.69477	1.78424	0.32478	0.00430	0.63619	0.11237
4	0.36281	2.46908	0.00016787	0.85407	0.21189	0.55108

Figure 12

Then from the Figure 12, we see the conditional number is less than 30, just 2.8126. Therefore, the collinearity issue has been solved based on two important criteria, condition number and variance inflation factor.

Now the final model shown on the below:

Like~3.9143+0.1231*attractive+0.3661*fun+0.2632*sincere+0.1486*SharedInterests



Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.3914291655	0.44105824	0.89	0.3761
attractive	0.1230594816	0.03650954	3.37	0.0009
Fun	0.3660876119	0.05648900	6.48	<.0001
Sincere	0.2631721313	0.06317646	4.17	<.0001
SharedInterests	0.1486398656	0.04366950	3.40	0.0008

Figure 13

3.3) REGRESSION DIAGNOSTICS FOR THE FINAL MODEL

Regression diagnostics are statistical techniques designed to detect conditions which can lead to inaccurate or invalid regression results.

Studentized or jackknife residual means we test the relationship between the residual and predicted value after removing the ***i***th observation.

It's an efficient and regular method using Jackknife residual to identify outliers in the model.

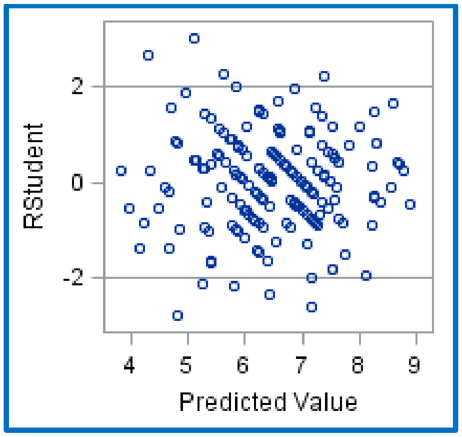
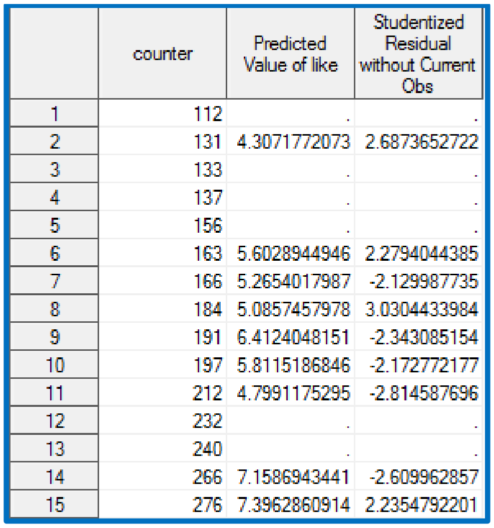


Figure 14

From the plot of the studentized or jackknife residual versus predicted values (Figure 14), there are some outliers with residuals above or lower -2; it means those observations are further away from our predicted line.



	counter	Predicted Value of like	Studentized Residual without Current Obs
1	112	.	.
2	131	4.3071772073	2.6873652722
3	133	.	.
4	137	.	.
5	156	.	.
6	163	5.6028944946	2.2794044385
7	166	5.2654017987	-2.129987735
8	184	5.0857457978	3.0304433984
9	191	6.4124048151	-2.343085154
10	197	5.8115186846	-2.172772177
11	212	4.7991175295	-2.814587696
12	232	.	.
13	240	.	.
14	266	7.1586943441	-2.609962857
15	276	7.3962860914	2.2354792201

Figure 15

Then I go back to check those outliers in the training dataset. There are 15 outliers shown on the Figure 15 asked on Jackknife residual criterion.

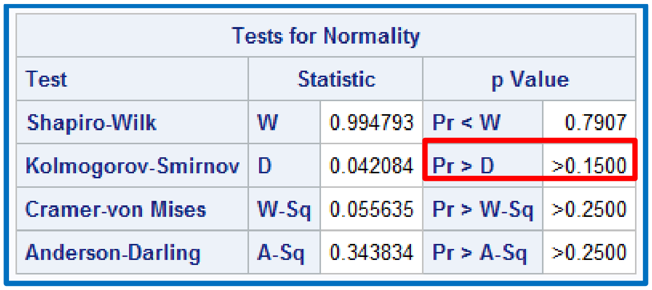
For the linear regression model, we have four test assumptions: contain Normality , Homogeneity , linearity and Independent .

First, I will test the normality assumption:

The normality hypothesis is:

H0:it follows the normal distribution;

Ha: otherwise



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.994793	Pr < W	0.7907
Kolmogorov-Smirnov	D	0.042084	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.055635	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.343834	Pr > A-Sq	>0.2500

Figure 16

We see the Kolmogorov-Smirnov indicator, the p value is >0.15;

So I will fail to reject the H0. It means the final model meets the normality assumption.

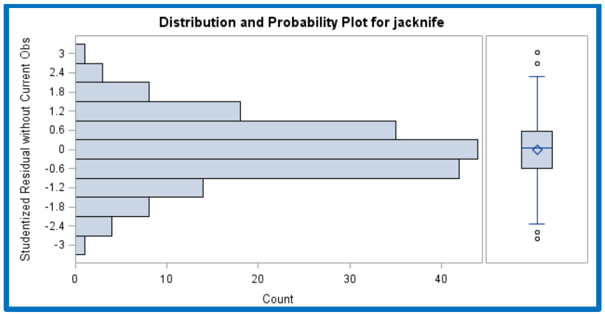


Figure 17

From the distribution and probability plot, the pattern on the left follows the normality shape;

Then I will test the homogeneity assumption

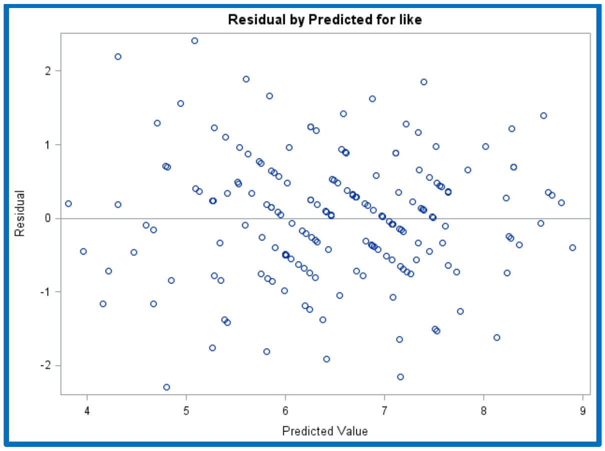
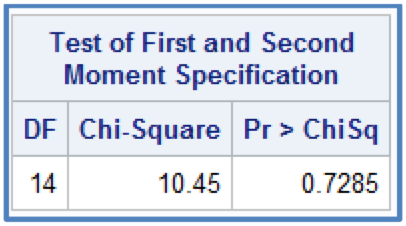


Figure 18

From the plot of residual and predicted value, Figure 18, I cannot see there is an obvious funnel shape exists as the predicted value increased, and the spots spread above and beyond the line with residual equals to 0. Therefore, it satisfies the homogeneity assumption from this plot.

Further, the test will be performed to check whether the homogeneity was violated or not. The SEPC option performs a model specification test in the REG model statement. The null hypothesis for the test maintains that the errors are homoscedastic and independent of the predictor variables. For details, see theorem 2 and assumptions 1-7 of White (1980), shown on the reference (Large Sample Properties of OLS, asymptotic confidence intervals and hypothesis test).



Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
14	10.45	0.7285

Figure 19

From the Figure 19, we see the Prob>ChiSq of 0.7285>0.05, it means we fail to reject the null hypothesis, the model dose not violate the homogeneity assumption.

The next step that I will test the Linearity assumption (Errors have mean of zero)

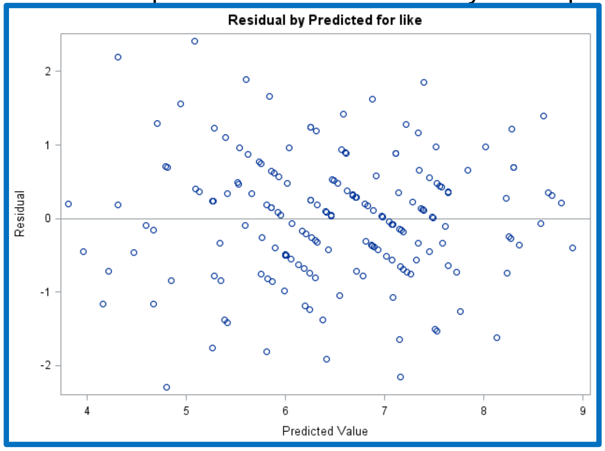


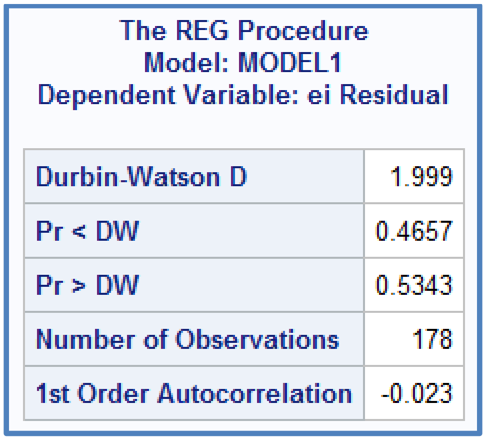
Figure 20

For the linearity, we observe the spots on the Figure 20, if a repeated pattern of y value falling above and below the Linear y=β0+β1*X1+β2*X2+…+βj*Xj; then its residual pattern will be a Repeated pattern falling above and below y=0

In this case, it does not like curvature pattern, so it does not violate linearity;

The final test is Independence assumption:

Durbin Watson statistics is a test for autocorrelation in the residual from a statistical regression analysis. The DW statistics will always have a value between 0 to 4. A value of 2 means there is no autocorrelation between observations in the model.



The REG Procedure
Model: MODEL1
Dependent Variable: ei Residual

Durbin-Watson D	1.999
Pr < DW	0.4657
Pr > DW	0.5343
Number of Observations	178
1st Order Autocorrelation	-0.023

Figure 21

On the Figure 21, we see the Durbin-Watson is 2, and Pr<DW, 0.4657 means there is no positive autocorrelation exists. In addition, we see the Pr>DW, 0.5343 means there is no negative autocorrelation exists. Therefore, it dose not violate the independent assumption.

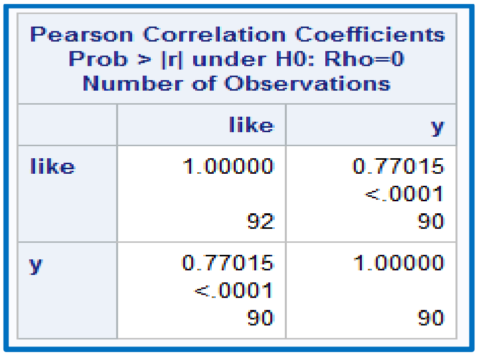
3.4) RELIABILITY EVALUATION FOR THE FINAL MODEL

Reliability is a measurement of how well the final model handles the future predictions.

I have already split the raw data into train and test dataset after cleaning all variables. I will use cross-validation method that uses a shrinkage measurement to assess reliability.

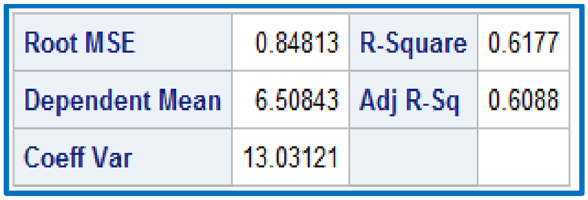
The regression coefficients from the final training model is to predict values for the test model. Then compute R^2 in the test model; Finally, compare the R^2 between the train model and test model to see how the shrinkage changes.

If the shrinkage changes less than 0.1 is seen as reliable for the final model.



Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	like	y
like	1.00000 92	0.77015 <.0001 90
y	0.77015 <.0001 90	1.00000 90



Root MSE	0.84813	R-Square	0.6177
Dependent Mean	6.50843	Adj R-Sq	0.6088
Coeff Var	13.03121		

Figure 22

The left table of Figure 22 shows the correlation coefficient between predicted dependent variable in the train model and predicted value for the test model, the R is 0.7702; then the compute R square is 0.60; The right table shows the R square of the final training model, it is 0.62;

Therefore, we see the shrinkage changes is 0.02, which is less than 0.1; the final model is reliable.

4) CONCLUSION

The paper has two goals. The first goal is to suggest a complete process, a series of steps, that could be followed to procedure how to build model with different strategy, how to revise the collinearity issue for the model, and how to run the diagnosis test. To figure out the reliable and useful model.

The second goal is to show how speed dating breaks into the industry with its own power of statistics to match couples efficiently.

5) REFRENCES

Introducing the GLMSELECT PROCEDURE for Model Selection

Robert A. Cohen, SAS Institute Inc. Cary, NC

https://support.sas.com/resources/papers/proceedings/proceedings/sugi31/207-31.pdf

SAS Code to Select the Best Multiple Linear Regression Model for Multivariate Data Using Information Criteria

Dennis J. Beal, Science Applications International Corporation, Oak Ridge,
http://www.biostat.umn.edu/~wguan/class/PUBH7402/notes/lecture8_SAS.pdf

Model-Selection Methods with SAS/STAT® Procedures https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_reg_sect030.htm

Assumptions of Linear Regression with Complete Dissertation™ By Statistics Solutions

https://www.statisticssolutions.com/assumptions-of-linear-regression/

Large Sample Properties of OLS, asymptotic confidence intervals and hypothesis testing

http://faculty.arts.ubc.ca/vmarmer/econ527/527_08.pdf

6) ACKNOWLEDGMENTS

Thanks to the people at SAS Tech Support.

7) CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

YuTing Tian, tianyangzi2017@gmail.com

8) APPENDIX

```
/*download R package*/
install.packages("Lock5withR")
library(Lock5withR)
View(Lock5withR::SpeedDating)
write.csv (SpeedDating,file="E:/hope/BK/SpeedDating.csv")
```

```
/*SAS code*/
libname in01 "\\Client\E$\hope\BK";

proc contents data=in01.Speed_Dating;
run;quit;

/*clean the data*/
data dating;
set in01.Speed_Dating;
  if funf ="NA" then funf =.;
  if PartnerYesM="NA" then PartnerYesM=.;
  if AttractiveM ="NA" then AttractiveM =.;
  if SharedInterestsM ="NA" then SharedInterestsM =.;
  if SharedInterestsF ="NA" then SharedInterestsF =.;
  AttractiveM_num=input(AttractiveM,Best32.);
  FunF_num=input(FunF,Best32.);
  partnerYesM_num=input(PartnerYesM,Best32.);
  sharedInterestsF_num=input(SharedInterestsF,Best32.);
  sharedInterestsM_num=input(SharedInterestsM,Best32.);
  drop AttractiveM FunF PartnerYesM SharedInterestsF SharedInterestsM;
  rename FunF_num=FunF partnerYesM_num=partnerYesM
  sharedInterestsF_num=sharedInterestsF
  sharedInterestsM_num=sharedInterestsM;
run;quit;
```

```
/*calculate mean values combing with gender*/
data dating2;
  set dating;
  like=mean(likem,likef);
  attractive=mean(AttractiveF,AttractiveM);
  Fun=mean(FunF,funm);
  Intelligent=mean(IntelligentF,IntelligentM);
  SharedInterests=mean(SharedInterestsM,SharedInterestsF);
  Sincere=mean(SincereM,SincereF);
  Ambitious=mean(AmbitiousM,AmbitiousF);
  PartnerYes=mean(PartnerYesM,PartnerYesF);
run;quit;
```

```
/*get the Pearson Correlation table*/
proc corr data=dating2;
  var like Attractive Ambitious Fun Intelligent SharedInterests Sincere
  PartnerYes;
run;quit;


/*add polynomial and interaction variables in the dataset*/

data dating3;
  set dating2;

  if racem^=racef then diff_race=1;
if racem = racef then diff_race=0;
  if 0=<agem-agef<=2 or 0=<agef-agem<=2 then diff_age=0;
  else diff_age=1;
  diff_age2=diff_age**2;
  Attractive2=Attractive**2;
  Ambitious2=Ambitious**2;
  Fun2=Fun**2;
  Intelligent2=Intelligent**2;
  SharedInterests2=SharedInterests**2;
  Sincere2=Sincere**2;
  *PartnerYes2=PartnerYes**2;

  AA1=diff_age*Attractive;
  AA2=diff_age*Ambitious;
  AF1=diff_age*Fun;
  AI1=diff_age*Intelligent;
  AS1=diff_age*SharedInterests;
  AS2=diff_age*Sincere;
  *AP1=diff_age*PartnerYes;


  AA3=Attractive*Ambitious; AF2=Attractive*Fun;
  AI2=Attractive*Intelligent;
  AS7=Attractive*SharedInterests;AS3=Attractive*Sincere;
  *AP2=Attractive*PartnerYes;

  AI3=Ambitious*Intelligent; AS5=Ambitious*SharedInterests;
  AS6=Ambitious*Sincere;
  *AP3=Ambitious*PartnerYes;

  FI4=Fun*Intelligent; FS4=Fun*SharedInterests;
  FS5=Fun*Sincere;*FP6=Fun*PartnerYes;
  IS5=Intelligent*SharedInterests;
  IS6=Intelligent*Sincere;*IP6=Intelligent*PartnerYes;
  SS7=SharedInterests*Sincere;*SP7=SharedInterests*PartnerYes;

run;quit;

proc glm data=training;
  model like=diff_race;
run;

proc freq data=training;
  table racem racef racem*racef;
```

```
run;quit;

proc glm data=training;
  model like=diff_age/solution;
run;quit;

proc sgpanel data=dating;
  panelby racef/colums=5;
  reg x=agef y=likem/cli clm;
  title "the extent of men's likes to different women's ages";
run;

proc sgpanel data=dating;
  panelby racem/colums=5;
  reg x=ageM y=likeF/cli clm;
  title "the extent of women's likes to different men's ages";
run;

/*split the dataset into 2 parts: traing-184 and holdout-92*/
data split;
   set dating3;
   ran_num=ranuni(2019);
   run;

proc sort data=split;
  by ran_num;
run;

data test training;
   set split;
   counter=_N_;
   if counter <=92 then output test;
   else output training;
run;

/*all possible model*/
proc reg data=training;
   model like= diff_age Ambitious Fun Attractive Intelligent SharedInterests
  Sincere diff_age2 Attractive2 Ambitious2 Fun2 Intelligent2 SharedInterests2
  Sincere2 AA1 AA2 AF1 AI1 AS1 AS2  AA3 AF2 AI2 AS7 AS3   AI3 AS5 AS6  FI4 FS4
  FS5 IS5 IS6  SS7 /selection=rsquare cp mse;
run;quit;
/*forward method*/
proc reg data=training;
   model like=diff_age Ambitious Fun Attractive Intelligent SharedInterests
  Sincere diff_age2 Attractive2 Ambitious2 Fun2 Intelligent2 SharedInterests2
  Sincere2 AA1 AA2 AF1 AI1 AS1 AS2  AA3 AF2 AI2 AS7 AS3   AI3 AS5 AS6  FI4 FS4
  FS5 IS5 IS6  SS7 /selection=forward slentry=0.05;
run;quit;


/*backward method*/
proc reg data=training;
   model like=diff_age Ambitious Fun Attractive Intelligent SharedInterests
  Sincere diff_age2 Attractive2 Ambitious2 Fun2 Intelligent2 SharedInterests2
  Sincere2 AA1 AA2 AF1 AI1 AS1 AS2  AA3 AF2 AI2 AS7 AS3   AI3 AS5 AS6  FI4 FS4
  FS5 IS5 IS6  SS7 /selection=backward slentry=0.05;
```

```
run;quit;


/*stepwise method*/
proc reg data=training;
  model like=diff_age Ambitious Fun Attractive Intelligent SharedInterests
  Sincere diff_age2 Attractive2 Ambitious2 Fun2 Intelligent2 SharedInterests2
  Sincere2 AA1 AA2 AF1 AI1 AS1 AS2  AA3 AF2 AI2 AS7 AS3   AI3 AS5 AS6  FI4 FS4
  FS5 IS5 IS6  SS7 /selection=stepwise slentry=0.5 slstay=0.05;
run;quit;


/*colinearity test for the candidate model*/
proc reg data=training plots=all;
  model like=Attractive fun intelligent Sincere SharedInterests intelligent2
  /collinoint vif;
run;quit;


/*colinearity test after adjusting the candidate model*/

proc reg data=training plots=all;

  model like=Attractive fun Sincere SharedInterests /collinoint vif;
run;quit;


/*regression final model*/
proc glm data=training plots=all;
  model like=Attractive fun  Sincere SharedInterests/solution;
run;quit;




/*diagnostic*/
ods graphics on;
ods listing close;
proc reg data=training plots=all;
  model like= Attractive fun  Sincere SharedInterests;
  output out=train_out p=pred residual=ei
  stdr=s_ei rstudent=jacknife student=student;
run;

/*find outliers*/
data outlier;
  set train_out(keep=counter pred jacknife);
  if jacknife>2 or jacknife<-2 then output;
run;quit;


/*normality test*/
proc univariate data=train_out
  normal plots;
  var jacknife;
run;

/*homogeneity test*/

proc reg data=training plots=all;
```

```
   model like=Attractive fun  Sincere SharedInterests/SPEC;
run;quit;


/*test independent*/
data train_out4;
  set train_out4 ;
  obsnr = _n_;
run;

proc gplot data=train_out4;
   plot  ei *obsnr=1/frame;
run;quit;
proc reg data=train_out4;
   model ei= obsnr /dw dwprob;
quit;
/*test reliability*/
data test2;
  set test;
  y=0.3914+0.1231*attractive+0.3661*fun+0.2632*sincere+0.1486*sharedinterests;
run;

proc corr data=test2;
  var like y;
run;
```