# 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 PIZZAZ.com is to break into this industry that using the power of statistics to optimally match couples.

In order to attain this target, we will use a fictional dataset speeding dataset to generate models to test feasibility of using statistics to match couples.

## INTRODUCTION

In this paper, the author acknowledge that some of the figures are rather small. It is suggested that a reader might print the paper and download the PDF.

This paper is divided into the following _4_ main sections:

1) Brief introduction for our speed dating data

2) Some Basic Hypothesis

  2.1) Impact of race

  2.2) 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 for male (dependent variable)

  3.2) Regression diagnostics for the candidate model for male (dependent variable)

  3.3) Collinearity test for the candidate model for male (dependent variable)

  3.4) Logistic regression model for male (dependent variable)

  3.5) **All the process** repeated **above for female**

4) Conclusion

## 1) BRIEF INTRODUCTION FOR OUR SPEED DATING DATA

The chart in Figure 1 illustrates The target (Y) is (a numerical), indicating how much do you like this person

For each male dater,
9 input (X) variables were recorded;

A dater's opinion of the person, as indicated by the like variable. based on the attractiveness, sincerity, Intelligence, Fun, ambitious And shared interest of their partner

| 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, 11=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,11=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, 11=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 1-11 |

Figure 1

## 2) SOME BASIC HYPOTHESIS

### 2.1) IMPACT OF RACE

Table of RaceM by RaceF

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

Frequency Missing = 3

In total, there are 80 couples have the same race

**Figure 2**

The first test I will perform the effect of same race on the Like variable. From a cross-tabulation we see there are 80 same-race couples (Figure 2). A dummy binary variable was created with the value 1 for same race, 0 for different; our model then looks like:

H0: there is no significant relationship between race and like variable;

Model: like=β0+β1*race;

so, we see for man: the extent to like this person will be less than 0.015 point if they are the same race, comparing to the diff race;

but the P value is 0.9525; so there is no significantly diff on the extent of like whether this lady is the same race or not;

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 6.647058824 | 0.17730216 | 37.49 | <.0001 |
| race | -0.015808824 | 0.26742668 | -0.06 | 0.9529 |

for woman: the extent to like this person with same race will be greater than 0.1475 point, comparing to the different race; but the P value is 0.5479;

so there is no significantly diff on the extent of like whether this man is the same race or not

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 6.315000000 | 0.17501445 | 36.08 | <.0001 |
| race | 0.147500000 | 0.26252167 | 0.56 | 0.5749 |

## 2.2) THE IMPACT OF AGE RANGE

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 6.838235294 | 0.21634889 | 31.61 | <.0001 |
| age | -0.316305470 | 0.27336188 | -1.16 | 0.2488 |

whether partner are the same age range (the same age range is defined by being within 2 years of one another)

H0: there is no significant relationship between age range and the extent of like

model : like=β0+β1*age； if they are the same age range, age will be 0, if they are not the same age range, age=1

therefore , we see for man: the extent to like this person without the same age range will be less than 0.316,comparing to the same age range;

but the P value is 0.2488; so there is no significantly diff on the extent of like whether this lady is the same age range or not

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 6.477611940 | 0.21380665 | 30.30 | <.0001 |
| age | -0.154603091 | 0.26984739 | -0.57 | 0.5674 |

for woman: the extent to like this person will be greater than 0.1546 point if they are not the same age range, comparing to the same age range;

but the P value is 0.5674; so there is no significantly diff on the extent of like whether this man is the same age range or not

| age | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
|---|---|---|---|---|
| 0 | 69 | 37.50 | 69 | 37.50 |
| 1 | 115 | 62.50 | 184 | 100.00 |

I create a new variable-age: when 0=<agem-agef<=2 or 0=<agef-agem<=2. then age is 0; otherwise;

So , we see there are 69 couples are in the same age range in our data

## 2.3) THE CORRELATION BETWEEN LIKE RATINGS AND OTHER INDEPENDENT VARIABLES

First, from the big visual:

We see: For men, they prefer to like younger woman with humor , and women who like to share interest and hobbies with partner, and they are more likely to predict this woman will say yes to them.

| | LikeM | AgeF | AttractiveF | AmbitiousF | FunF | IntelligentF | SharedInterestsF | SincereF | PartnerYesF |
|---|---|---|---|---|---|---|---|---|---|
| LikeM | 1.00000 | -0.13466 | 0.02966 | 0.04203 | 0.24317 | 0.08799 | 0.22337 | 0.11661 | 0.18701 |
| | | 0.0707 | 0.6918 | 0.5819 | 0.0011 | 0.2388 | 0.0043 | 0.1180 | 0.0119 |
| | 182 | 181 | 181 | 174 | 178 | 181 | 162 | 181 | 180 |

For woman, we see: they prefer to like men who are intelligent , more ambitious and more sincere, and they are more likely to predict this man will say yes to them

| | LikeF | AgeM | AttractiveM | AmbitiousM | FunM | IntelligentM | SharedInterestsM | SincereM | PartnerYesM |
|---|---|---|---|---|---|---|---|---|---|
| LikeF | 1.00000 | -0.06556 | 0.10564 | 0.16631 | 0.11490 | 0.17726 | 0.13749 | 0.22824 | 0.15141 |
| | | 0.3860 | 0.1605 | 0.0312 | 0.1289 | 0.0189 | 0.0792 | 0.0022 | 0.0442 |
| | 180 | 177 | 178 | 168 | 176 | 175 | 164 | 178 | 177 |

# 3) BUILDING A MODEL

## 3.1) STRATEGY TO BUILD THE MODEL FOR MALE (DEPENDENT VARIABLE)

First , I will focus on analyzing the man's like extent and build the model looks like :

likeM~β0+β1*X1.f+β2*X2.f+…+βj*Xi.f

where Xi are a series of variables.

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

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

For the Mallows: the $C_p$ statistic is often used as a stopping rule for various forms of regression

$C_p$ has expectation nearly equal to *P*

we find the point where Cp is less than or equal to p.

*If* $C_p < 0$ in extreme cases. It is suggested that one should choose a subset that has $C_p$ approaching *P*,

All possible models

We will consider all the conditions when we build the maximum model, single independent variable , all second order variables: then we use SAS to select the better model with all possible selection;

| Number in Model | R-Square | C(p) | MSE | Variables in Model |
|---|---|---|---|---|
| 2 | 0.1174 | 1.9649 | 2.57490 | AgeF AF1 |
| 2 | 0.1174 | 1.9730 | 2.57504 | PartnerYesF2 FS5 |
| 2 | 0.1170 | 2.0290 | 2.57601 | FunF2 SP7 |

So , we select 2 variables: partneryes$F^2$ and FS5 kept in our model based on the all possible selection method; Among the FS5 means interaction between funf and sinceref;

Then I will fix the model, because automated model building programs often will include higher order polynomial terms or interactions without including base terms;

Therefore, I add the base term in my model,

Y~β0+β1*funf+β2*partneryesF+β3*sinceref+β3*funf*sinceref+β4* partneryes$F^2$ , then ask SAS to run the regression.

After adjusting the model , the only variable I want to keep in the model is funf, which can significantly predict the dependent variable.

Y~5.2766 +0.2088*funf

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 5.276567072 | 0.43175691 | 12.22 | <.0001 |
| FunF | 0.208778685 | 0.06277594 | 3.33 | 0.0011 |

We see the coefficient estimate of funf (slope) is 0.2088, which means one unit increase in funf, the extent of like will increased by 0.2088 for man on average. We see the R square is 0.059, which means we just have 5.9% variability of the extent of likem can be explained by these X variables for our model funf;

| R-Square | Coeff Var | Root MSE | LikeM Mean |
|---|---|---|---|
| 0.059129 | 25.50191 | 1.695590 | 6.648876 |

Stepwise forward selection

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 11.41758409 | 2.18083826 | 5.24 | <.0001 |
| AgeF | -0.08019185 | 0.03464985 | -2.31 | 0.0219 |
| FunF | -0.55229984 | 0.30532461 | -1.81 | 0.0723 |
| IntelligentF | -0.62071382 | 0.23867101 | -2.60 | 0.0101 |
| FunF*IntelligentF | 0.09884746 | 0.03696605 | 2.67 | 0.0082 |
| racef1 | 0.17965034 | 0.53594246 | 0.34 | 0.7379 |
| racef2 | -0.22096039 | 0.74328824 | -0.30 | 0.7666 |
| racef3 | 0.75742700 | 0.49556538 | 1.53 | 0.1283 |
| racef4 | 1.52509537 | 0.64363444 | 2.37 | 0.0190 |

From the estimate coefficient table, we see the extent of likem for racef4 (latino) higher 1.53 than others on average, the extent of likem for Caucasian is higher 0.76 than other, the extent of likem for black is lower 0.22 than other, the extent of likem for Asian is higher 0.18 than other on average ;

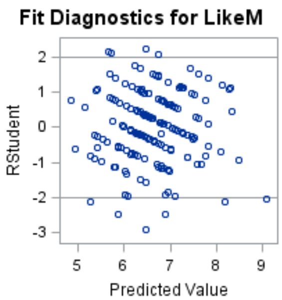
| R-Square | Coeff Var | Root MSE | LikeM Mean |
|---|---|---|---|
| 0.175900 | 24.40903 | 1.617792 | 6.627841 |

In this model, We see the R square is 17.60%, which means we just have 17.60% variability of the extent of likem can be explained by these X variables for our model agef, funf , intelligentf and racef ;

So , comparing r square between these two models, I will select the latter one would be better , because the variability of dependent variable can by explained more by our independent variables

## 3.2) REGRESSION DIAGNOSTICS FOR THE CANDIDATE MODEL FOR MALE (DEPENDENT VARIABLE)

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




From the plot of the studentized or jackknife residual versus predicted values, there are some outliers with residuals above or lower than -2; It means those observations are further away from our predicted line;

I suggest go back to check those observations

Studentized or jackknife residual means we test the relationship between the residual and predicted value after we removing the ith observation out.

Then I want to check whether outliers exists with cook's distance; leverage statistics and jackknife residual

Based on the rules, we know the cutoff value of cook's distance is 1, the leverage critical value is h>2(K+1)/n=2*(8+1)/176=0.102;

The jackknife residual cutoff value is 2; So , we see these two observations are outliers

| Obs | LikeM | jacknife | cooks | leverage |
|---|---|---|---|---|
| 128 | 6 | -2.05654 | 0.062465 | 0.11933 |
| 148 | 5 | -2.12224 | 0.065579 | 0.11801 |

Test assumptions

1) Normality.2) Homogeneity.3) Linearity (Errors have mean of zero) .4) Independence

| Tests for Normality | | | | |
|---|---|---|---|---|
| Test | Statistic | | p Value | |
| Shapiro-Wilk | W | 0.989221 | Pr < W | 0.2026 |
| Kolmogorov-Smirnov | D | 0.043174 | Pr > D | >0.1500 |
| Cramer-von Mises | W-Sq | 0.051852 | Pr > W-Sq | >0.2500 |
| Anderson-Darling | A-Sq | 0.431194 | Pr > A-Sq | >0.2500 |

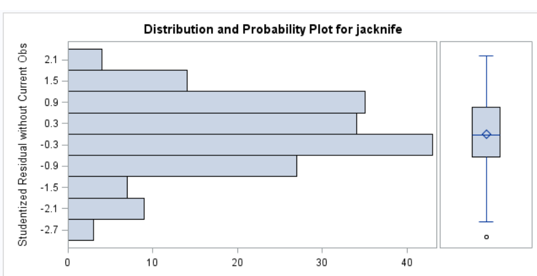
Our hypothesis:

H0- it follows the normal distribution;

Ha- it violate the normal distribution

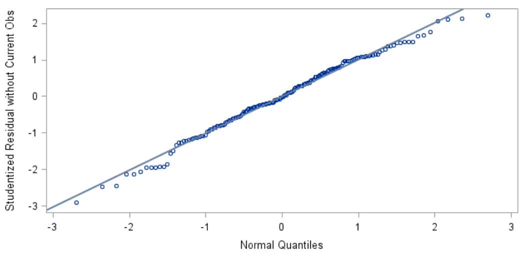
We see the Kolmogorov-Smirnov indicator, the p value is >0.15;

So I will fail to reject the H0. It means it meet the normality assumption for data



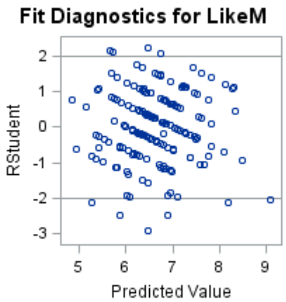

From the distribution and probability plot, the pattern follow the normality shape;

But one spot has too small and negative residual




From the Q-Q plot, we see there are several spots spread a little bit further from the linear line

2) Homogeneity




From the pattern, I cannot see there is an obvious funnel shape exists as the predicted value increase, and those spots do not spread out too much from each other;

Therefore , It seems satisfy the homogeneity assumption from this plot.

3) Linearity (Errors have mean of zero)

For the linearity , we observe if these plots are curvature in the plot;

It means 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 I think it does not violate linearity;

4) Independence

From the plot, it seems no obvious evidence to show there is correlation

Between each observations. So I think it does not violate the independence

## 3.3) COLLINEARITY TEST FOR THE CANDIDATE MODEL FOR MALE (DEPENDENT VARIABLE)

The next test is to test for collinearity. In statistics, 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. I will focus on collinearity since this is often a major problem in polynomial regression. Since all our variables are functions of other variables.

Two criterions to test collinearity: condition number and variance inflation factor

(VIF is and index that measures how much the variance of estimated regression coefficient is increased because of collinearity).

In this case, we see the condition number is less than 30 but we have very high Variance Inflation Factors that easily exceed 10.

Because of VIF exceeds 10, it means the collinearity exists. Therefore, I will fix the collinearity with center the X variables

| Parameter Estimates | | | | | | |
|---|---|---|---|---|---|---|
| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Variance Inflation |
| Intercept | 1 | 8.01840 | 0.98738 | 8.12 | <.0001 | 0 |
| AgeF | 1 | -0.08019 | 0.03465 | -2.31 | 0.0219 | 1.03844 |
| FunF | 1 | 0.22958 | 0.06554 | 3.50 | 0.0006 | 1.19258 |
| IntelligentF | 1 | 0.02970 | 0.09840 | 0.30 | 0.7631 | 1.29373 |
| FI4_fix | 1 | 0.09885 | 0.03697 | 2.67 | 0.0082 | 1.09686 |
| racef1 | 1 | 0.17965 | 0.53594 | 0.34 | 0.7379 | 3.50940 |
| racef2 | 1 | -0.22096 | 0.74329 | -0.30 | 0.7666 | 1.61197 |
| racef3 | 1 | 0.75743 | 0.49557 | 1.53 | 0.1283 | 4.05189 |
| racef4 | 1 | 1.52510 | 0.64363 | 2.37 | 0.0190 | 2.03969 |

| Collinearity Diagnostics (intercept adjusted) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Proportion of Variation | | | | | | | |
| Number | Eigenvalue | Condition Index | AgeF | FunF | IntelligentF | FI4_fix | racef1 | racef2 | racef3 | racef4 |
| 1 | 1.86636 | 1.00000 | 0.02042 | 0.00447 | 0.05108 | 0.04201 | 0.04726 | 0.00089628 | 0.05144 | 0.01047 |
| 2 | 1.45538 | 1.13243 | 0.01392 | 0.21821 | 0.14891 | 0.00327 | 0.00000263 | 0.08812 | 0.01543 | 0.00569 |
| 3 | 1.17683 | 1.25933 | 0.09286 | 0.01070 | 0.00095845 | 0.04611 | 0.06085 | 0.00056772 | 0.00056476 | 0.23140 |
| 4 | 1.04367 | 1.33726 | 0.05879 | 0.00007693 | 0.05522 | 0.28639 | 0.01769 | 0.23971 | 0.00206 | 0.02677 |
| 5 | 0.95318 | 1.39930 | 0.49350 | 0.04679 | 0.00169 | 0.16228 | 0.00199 | 0.03277 | 0.02100 | 0.07635 |
| 6 | 0.85111 | 1.48083 | 0.29968 | 0.21846 | 0.00310 | 0.23491 | 0.00165 | 0.18336 | 0.00021399 | 0.01928 |
| 7 | 0.53885 | 1.86108 | 0.01447 | 0.49956 | 0.73825 | 0.22380 | 0.00843 | 0.00043655 | 0.00254 | 0.00188 |
| 8 | 0.11463 | 4.03508 | 0.00636 | 0.00173 | 0.00079017 | 0.00121 | 0.86213 | 0.45413 | 0.90674 | 0.62817 |

Now, we see all VIF are less than 10, and condition number Is less than 30;

After fixing the collinearity , this is our final model shown on the below:

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 8.018397576 | 0.98738013 | 8.12 | <.0001 |
| AgeF | -0.080191849 | 0.03464985 | -2.31 | 0.0219 |
| FunF | 0.229583562 | 0.06553631 | 3.50 | 0.0006 |
| IntelligentF | 0.029702453 | 0.09839666 | 0.30 | 0.7631 |
| FI4_fix | 0.098847459 | 0.03696605 | 2.67 | 0.0082 |
| racef1 | 0.179650345 | 0.53594246 | 0.34 | 0.7379 |
| racef2 | -0.220960388 | 0.74328824 | -0.30 | 0.7666 |
| racef3 | 0.757427000 | 0.49556538 | 1.53 | 0.1283 |
| racef4 | 1.525095365 | 0.64363444 | 2.37 | 0.0190 |

From the estimate coefficient table, we see the extent of likem for racef4 (latino) higher 1.53 than others on average, the extent of likem for Caucasian is higher 0.76 than other, the extent of likem for black is lower 0.22 than other, the extent of likem for Asian is higher 0.18 than other on average ;

Each unit increase in the age of female, the extent of like will be decreased by 0.08 on average for man;

Each unit increase in the funf of female, the extent of like will be increased by 0.23 on average for man;

Each unit increase in the intelligent of female, the extent of like will be increased by 0.03 on average for man;

There is interaction between funf and intelliengencef, which can add a significant prediction in our model

## 3.4) LOGISTIC REGRESSION MODEL FOR MALE (DEPENDENT VARIABLE)

Finally , I will use logistic regression model to test the relationship between man's decision and man's like extent; from the output table above, we know Wald $X^2$ is 35.53. is substantially greater than df=1, so we may reject the H0; P-value <0.0001<0.05(alpha value), so I will reject the H0.

Conclusion : We are convincing evidence for significant effect due to the extent of like for male.

Probability modeled is DecisionM='1'.

| Analysis of Maximum Likelihood Estimates | | | | | |
|---|---|---|---|---|---|
| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
| Intercept | 1 | -5.1091 | 0.8921 | 32.8017 | <.0001 |
| LikeM | 1 | 0.7815 | 0.1311 | 35.5340 | <.0001 |

pi =P(Yi=1) =P(the probability of decisionM=yes for female i in measure of the extent of like)

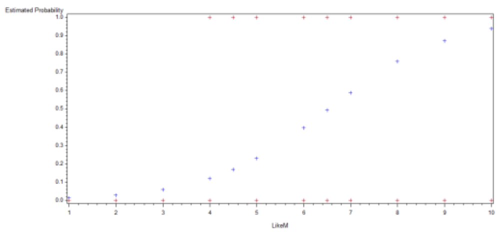
E(ln(pi/(1-pi))) =β0 + β1 Xi (β0: population intercept; β1: population slope)

| Odds Ratio Estimates | | | |
| --- | --- | --- | --- |
| Effect | Point Estimate | 95% Wald Confidence Limits | |
| LikeM | 2.185 | 1.690 | 2.825 |

We see the odds ratio is 2.185; means one unit change in odds favor of "decision=yes" for each one unit increase in the value of LikeM

we can be 95% confident that the OR associated with a one unit increase in likeM falls between 1.690 and 2.825; This is a set of plausible values for the population odds ratio associated a one-unit increase likeM, comparing to the relative odds.

Note: that 1 is not in the interval, it means that is not a plausible for the odds in favor of decision =yes associated with a one-unit increase to be equal.



We see the estimated probability of decisonM increased as the likeM goes up; after the likeM point above 8, the trend of increase becomes slower and a bit flatter from this curve pattern.

## 3.5) ALL THE PROCESS REPEATED ABOVE FOR FEMALE

I build model with these two methods: Stepwise forward selection and All possible models:

then , I compare r square between these two models, I will select the latter one would be better-all possible models , because the variability of dependent variable can by explained more by our independent variables.

| Tests for Normality | | | | |
| --- | --- | --- | --- | --- |
| Test | Statistic | | p Value | |
| Shapiro-Wilk | W | 0.974697 | Pr < W | 0.0044 |
| Kolmogorov-Smirnov | D | 0.097693 | Pr > D | <0.0100 |
| Cramer-von Mises | W-Sq | 0.238718 | Pr > W-Sq | <0.0050 |
| Anderson-Darling | A-Sq | 1.469066 | Pr > A-Sq | <0.0050 |

Our hypothesis:

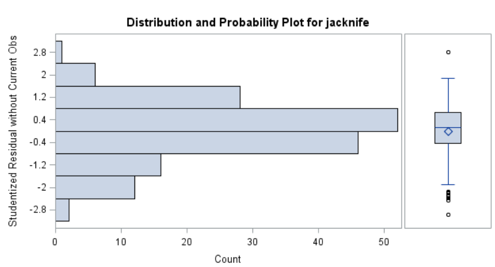
H0- it follows the normal distribution;

Ha- it violate the normal distribution

We see the Kolmogorov-Smirnov indicator, the p value is <0.01;

So I will reject the H0. It means it violate the normality assumption for data

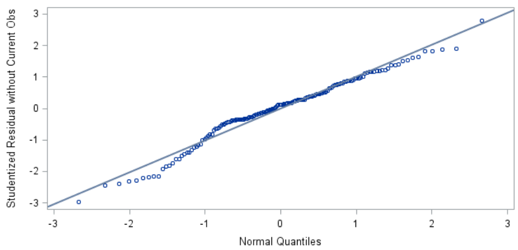
We can try to fix the normality with square transform



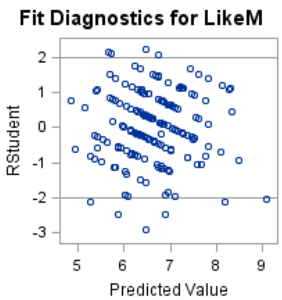

From the distribution and probability plot, the pattern follow the normality shape;

But some spots has too small and negative residual

One spot is too high and positive




From the Q-Q plot, we see there are several spots spread a little bit further from the linear line




From the plot of the studentized or jackknife residual versus predicted values, there are some outliers with residuals above or lower than -2; It means those observations are further away from our predicted line;

I suggest go back to check those observations

Based on the rules, we know the cutoff value of cook's distance is 1, the leverage critical value is h>2(K+1)/n=2*(6+1)/163=0.086;

The jackknife residual cutoff value is 2; we see there are almost 45 observations have problem either violate three of them (cooks, leverage, jackknife) one observations are outliers

we see there are almost 45 observations have problem either violate three of them (cooks, leverage, jackknife) one observations are outliers

and the same conclusion with male test with other three assumptions (homogeneity, independence and linearity)

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 2.199067308 | 0.97492089 | 2.26 | 0.0255 |
| AgeM | 0.222762931 | 0.10142487 | 2.20 | 0.0295 |
| ageM2 | -0.022751719 | 0.00816272 | -2.79 | 0.0060 |
| AttractiveM | 0.110381837 | 0.14313205 | 0.77 | 0.4418 |
| AA1_fix | -0.051144626 | 0.02102493 | -2.43 | 0.0161 |
| AmbitiousM | 0.192417223 | 0.09558670 | 2.01 | 0.0458 |
| SincereM | 0.322095848 | 0.09445177 | 3.41 | 0.0008 |

Each unit increase in the age of male, the extent of like will be increased by 0.22 on average for female;

Each unit increase in the attractive of male, the extent of like will be increased by 0.11 on average for female;

Each unit increase in the ambitious of male, the extent of like will be increased by 0.19 on average for female;

Each unit increase in the sincere of male, the extent of like will be increased by 0.322 on average for female;

There is interaction between age and attractive, which can add a significant prediction in our model

## 4) CONCLUSION

In this paper, we know the PIZZAZ.com get people together using statistics. The author hopes that this paper, by showing the process of how to build models and test our models in statistics.