**SPEED DATING MODEL TO PREDICT THE DATERS OPINION**

1. **Introduction**

Online dating is a growing industry with recent quarterly profits well more than millions. The goal of PIZZAZ.com is to break into this industry that uses the power of statistics to optimally match couples. To achieve the target, a speed dating event was conducted where 276 couples were randomly paired up with one another for a short speed date and the results were recorded.

1.1 Problem statement:

The Goal of the model is to predict the daters opinion of the person they are dating based on the other factors- attractiveness, sincerity, intelligence, fun, ambitiousness, and shared interests of their partner. The model will be used by PIZZAZ.com to show people who closely match with each and would predict the likelihood of dater liking the other person. This will help business to maximize matches and improve customer experience.

The other things to look at will be:

* Accuracy of Predicted Like rating to the Actual Ratings
* Situations where like ratings are suspiciously high.

1. **Descriptive Statistics**

The data was captured separately for males and females and recorded in different variables. The below table (Figure1) illustrates the information obtained from the survey along with its description.

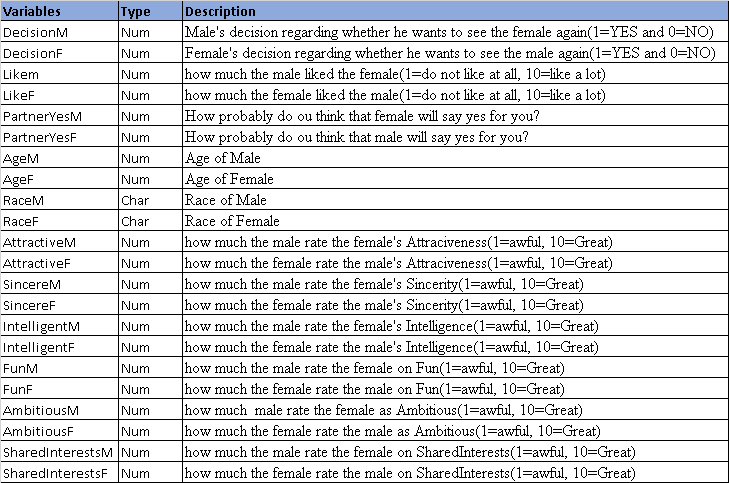


Figure1: Variables Description

There are two subgroups described in the study-Male and Female. The data is recorded for both across the similar characteristics and their corresponding responses. Looking at the type of variables its noticed that only Race variable is categorical (Nominal-­ do not have a specific order to them) in nature. However, rest of the variables listed in the table above are Numeric. Out of numeric variables, Age is continuous Numeric variable and others have Interval scale.

*Dependent Variable –* Like variable will be used as dependent variable that says how much the dater liked the other person.

*Predictor Variables –*Attractive, sincere, intelligent, fun, ambitious, shared interests.

**2.1 Variable Distributions:**

First step of building a model is to understand the data for which we look at descriptive like variable distributions, mean, median and measure of variability like variance, standard deviation etc.

Figure2 shows the descriptive statistics of the variables in the model:

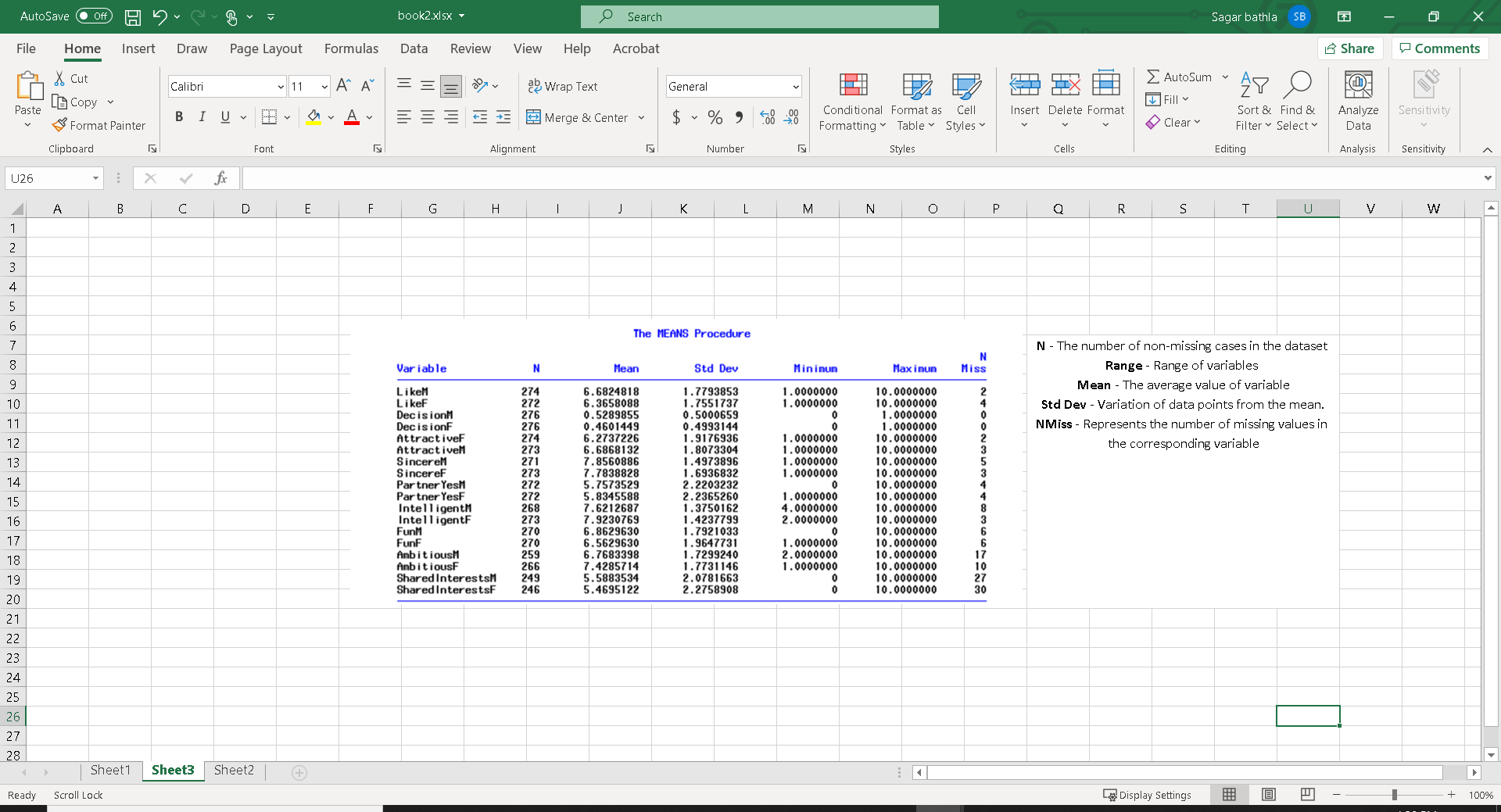
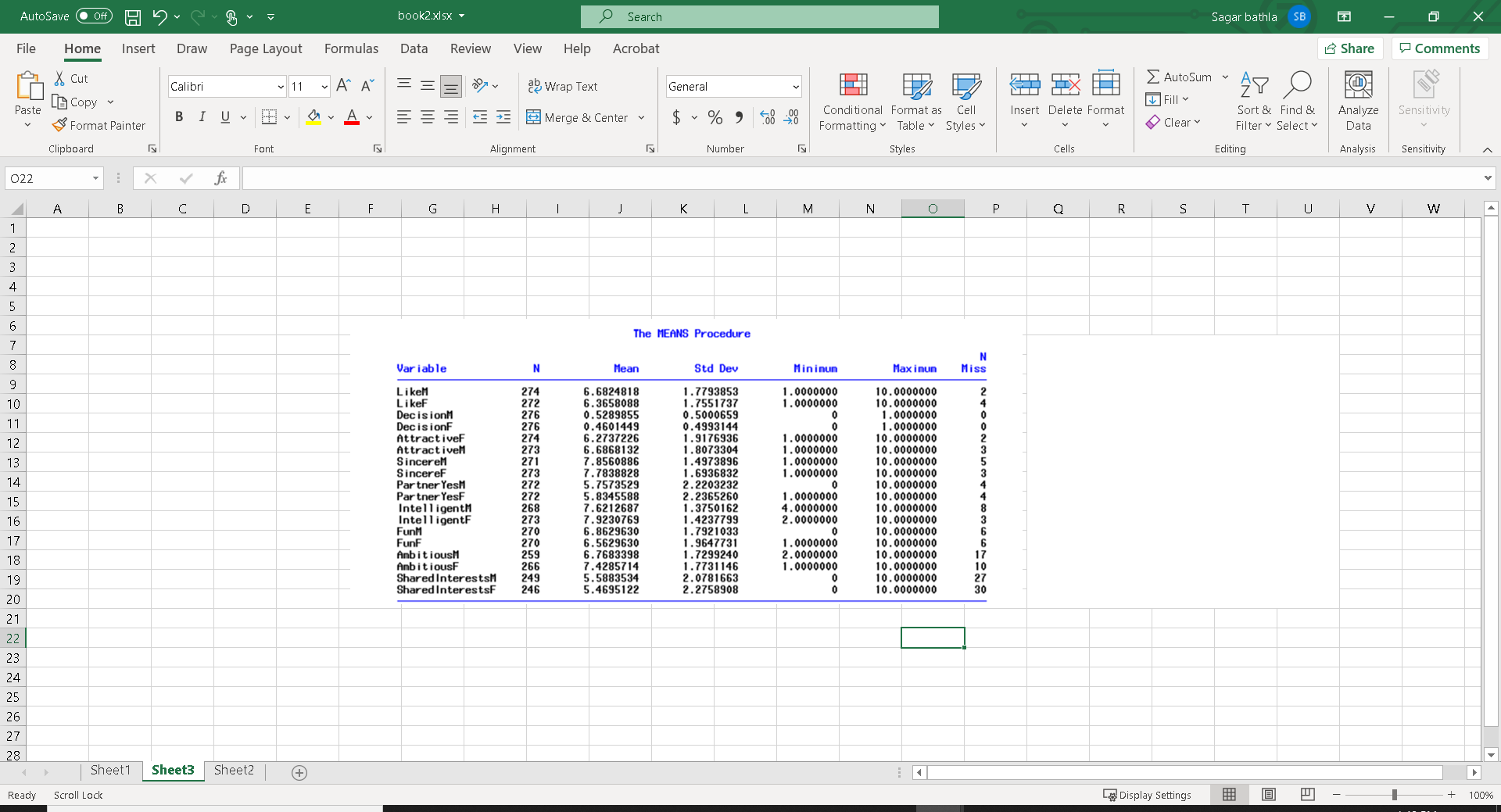
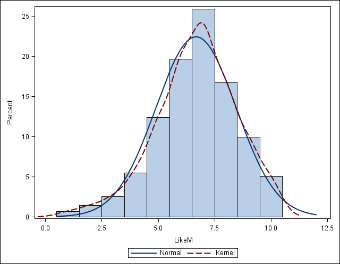
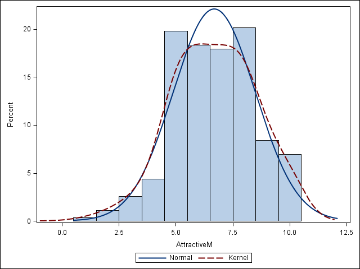
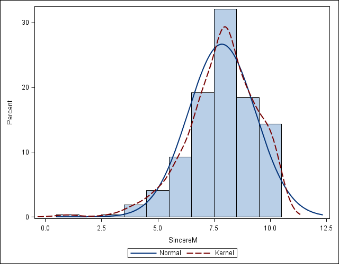
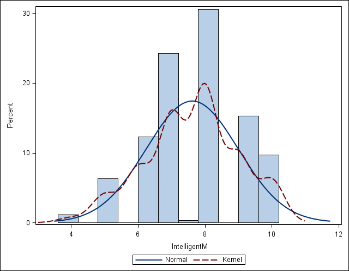


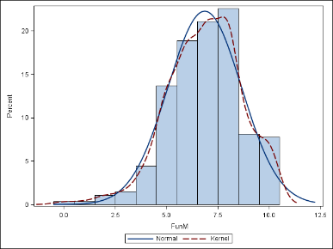
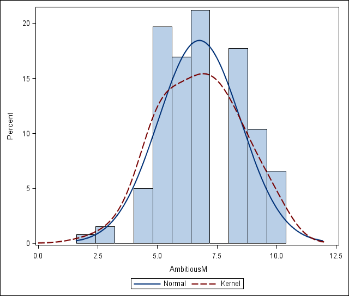
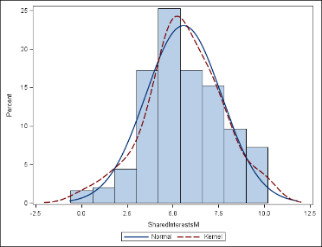
Figure2: Descriptive Statistics of Numeric Variables

By looking at the below histograms, we see that some of the variables are skewed and very few daters rated their partner with very low ratings on different dimensions. ­­We can also notice that some of the daters rated in decimal digits (example in sharedintrestsM rating 6.5 was given by one person) due to which graph looks distorted. Also, in rare instances, we can see 0 ratings as well.

**LikeM AttractiveM Sincere IntelligentM**

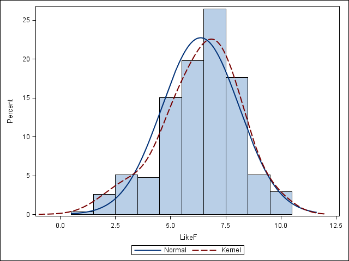
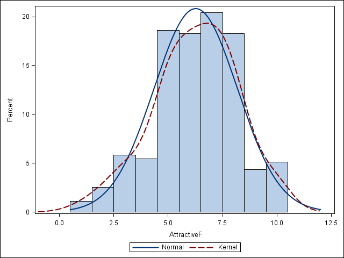
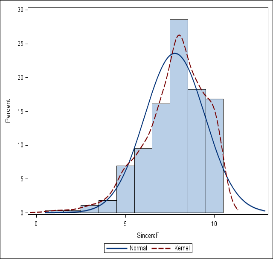
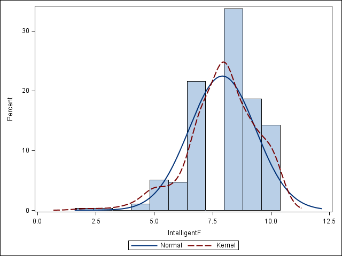
   

**FunM AmbitiousM SharedInterestM**

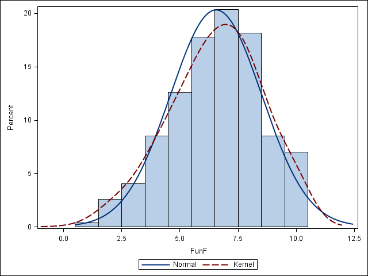
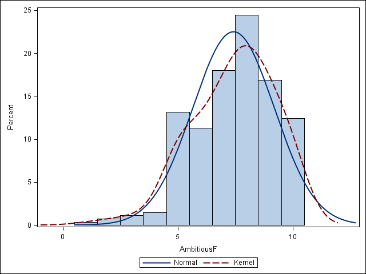
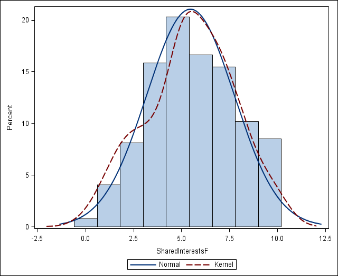
  

For Females, the variabls distributions as plotted by histogram looks like:

**LikeF AttractiveF SincereF IntelligentF**

**FunF AmbitiousF sharedInterestM**

*  *

Dependent Variable’s (Response) Box Plot: The left-most line of the boxplot is Q1, where 25% of data is less than this value, the median is represented by the line across the center of the box and the right-most line of the box is Q3, where 25% of data is greater than this value. The smallest and largest observations (excluding outliers) are illustrated by the whiskers (from left to right, respectively). Any data points outside of the range of the whiskers are denoted as outliers. Here, there are two outliers depicted in this box and whisker plot as open circles (Figure 3).

**LikeM LikeF**

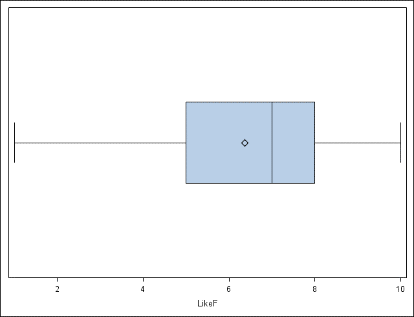
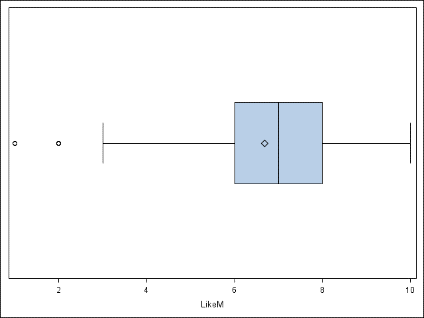


Figure3 Figure4

Figure5 shows the descriptive statistics of Categorical Variable. It represents that most daters were Caucasian followed by Asian.

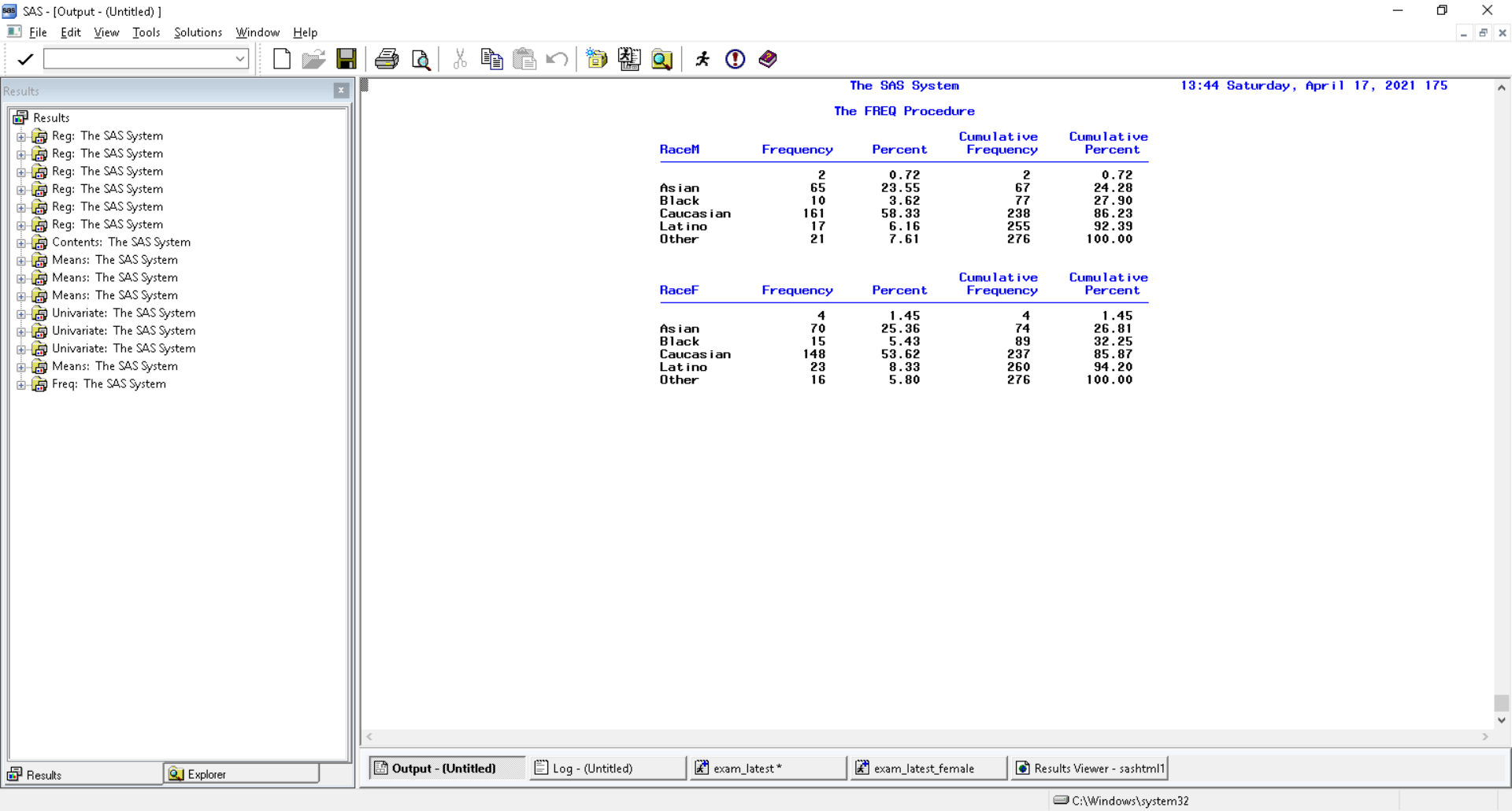


Figure5: Descriptive Statistics of Categorical Variables

**2.2 Variable Correlation:**

Correlation shows how strongly are variables correlated to each other and corresponding direction of a linear relationship. It takes value from -1 to 1 where sign gives us direction. Strength is observed from the value of correlation coefficient. 0 tells us that there is no linear relation whereas 1 means the strong linear relationship.

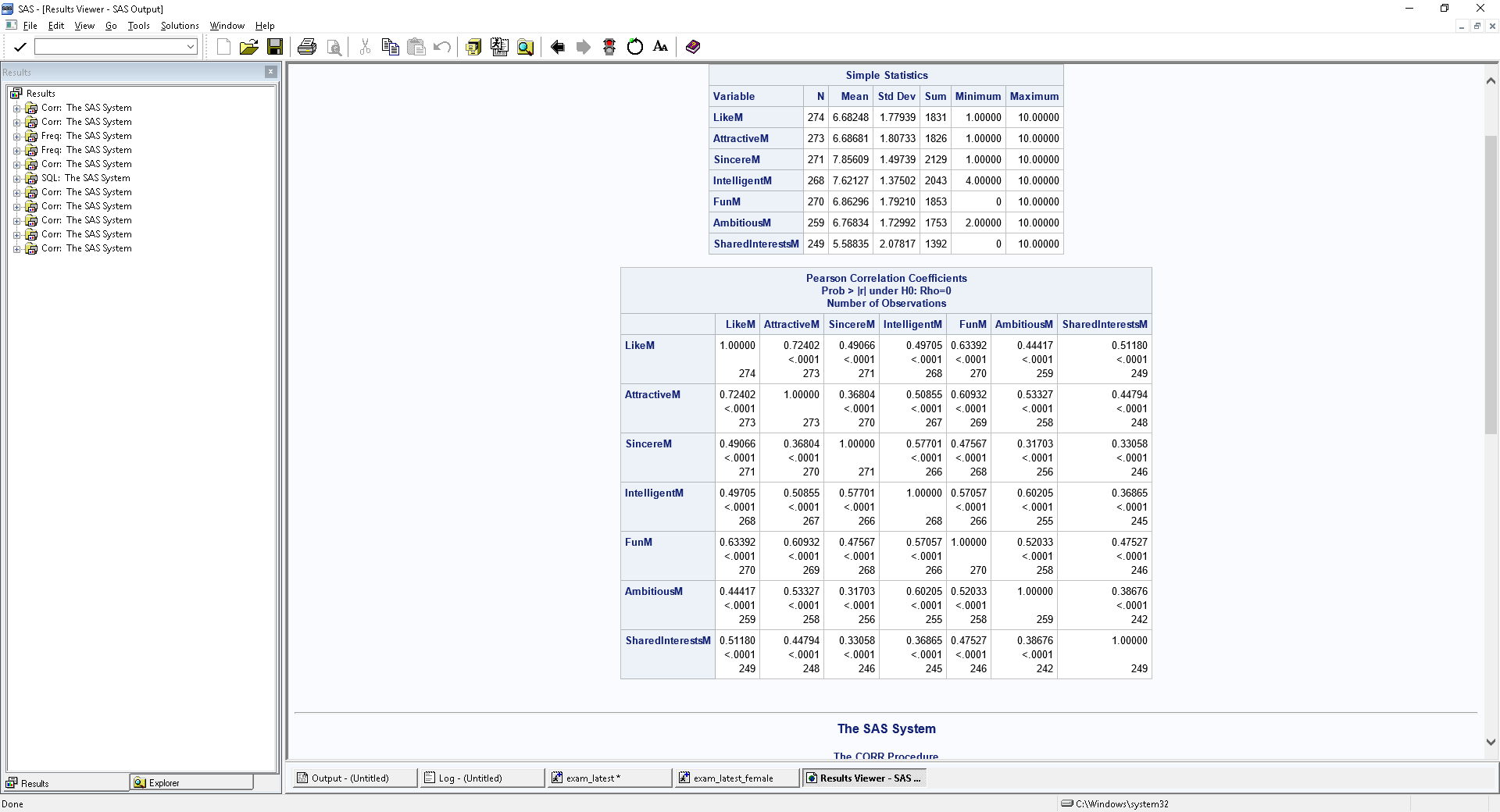
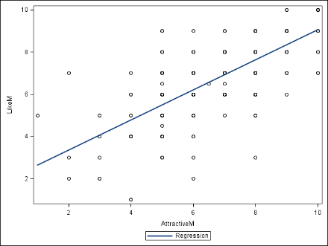
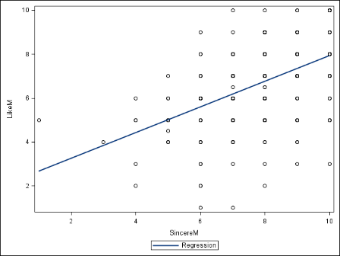
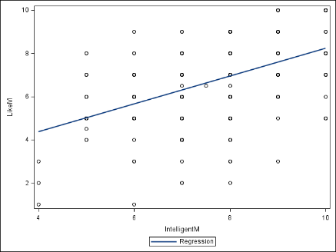


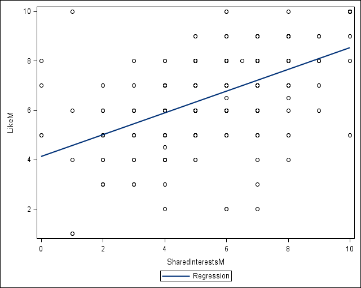
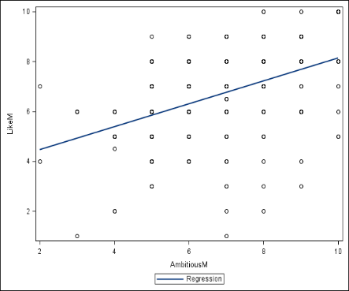
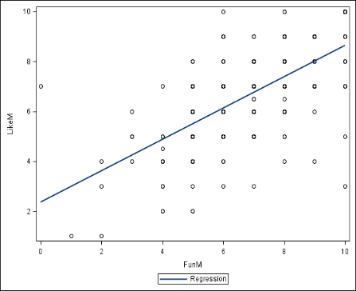
Figure6: Correlation Matrix for Male

*Scatterplots:*

**AttractiveM SincereM IntelligentM**

**FunM AmbitiousM SharedInterstM**



The correlation score and scatter plot diagrams show that like is strongly related with Attractive and Fun. Importance of the attributes ranked high to low: Attractive Fun, Shared Interest, Intelligent, Sincere, and Ambitious. Also looking at the matrix it seems that Intelligent and Ambitious, Fun and Attractive variables are somewhat intercorrelated.

**For Female:**

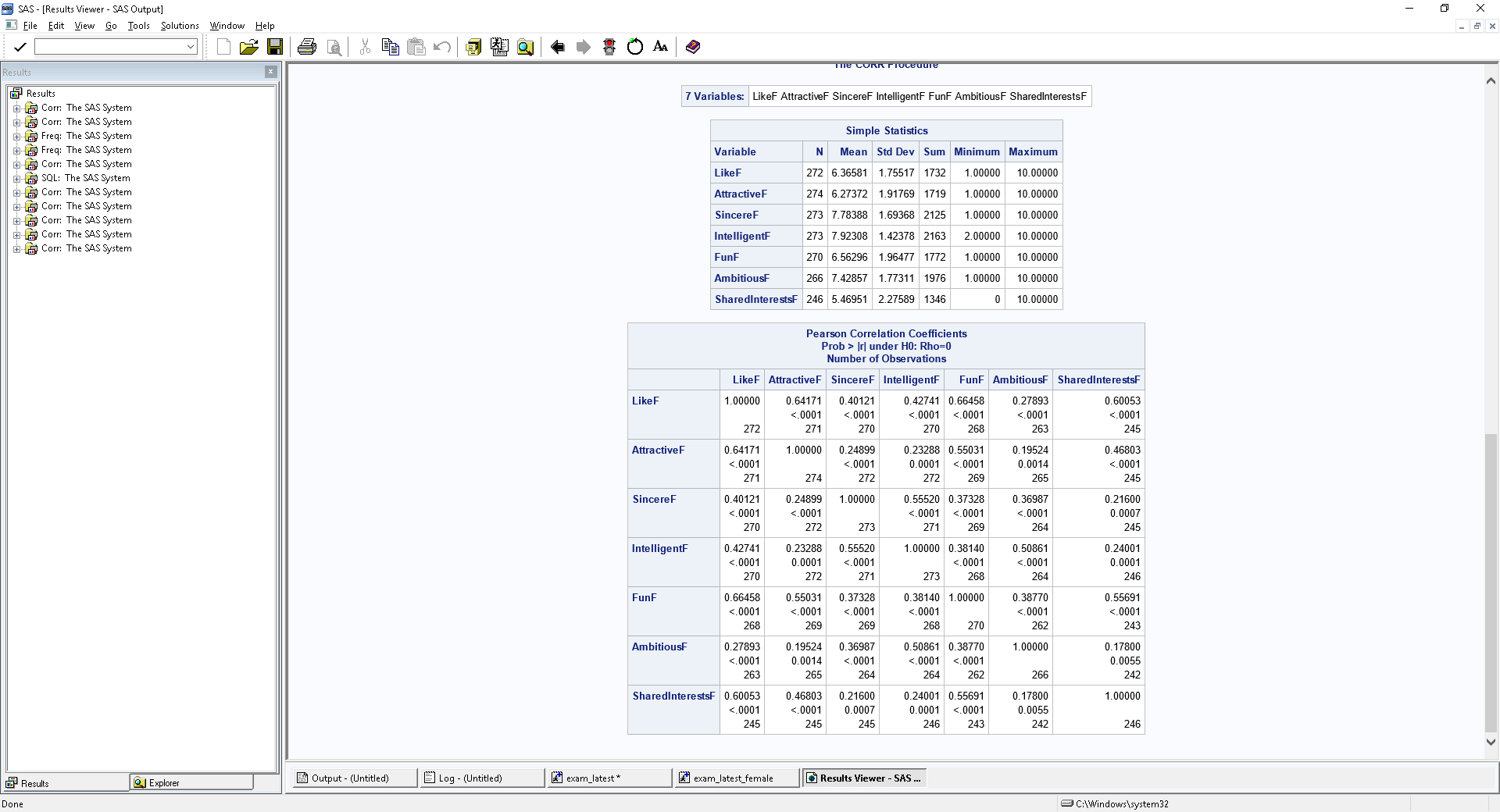
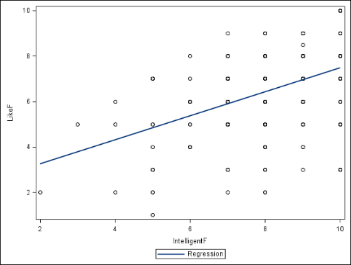
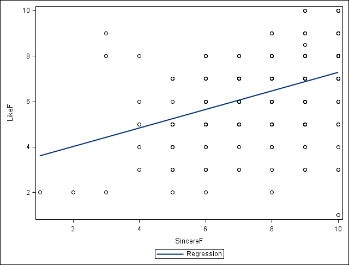
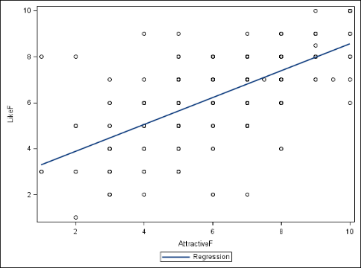
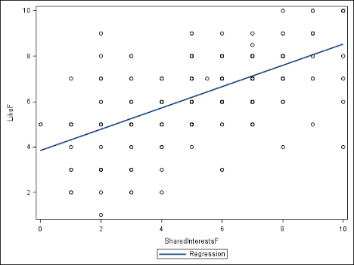
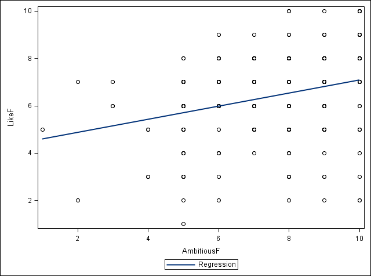
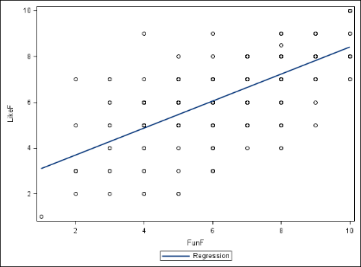


Figure7: Correlation Matrix for Female

*Scatterplots:*

**AttractiveF SincereF IntelligentF** 

**FunF AmbitiousF SharedInterestF** 

The correlation score and scatter plot diagrams show that like for females is strongly related with Attractive and Fun and shared Interest. Importance of the attributes ranked high to low: Fun, Attractive, Shared Interest, Intelligent, Sincere, and Ambitious. No variable seems to be intercorrelated with other.

We can see that there is a great difference between what male and female participants are looking for.

* For male participants, the attractiveness of the female is given a lot more weight, and then comes fun but that is not rated as high as attractive.
* For females, the points are more evenly distributed across Fun attractive and shared Interests, with fun ranked slightly higher compared to others.

Data cleaning:

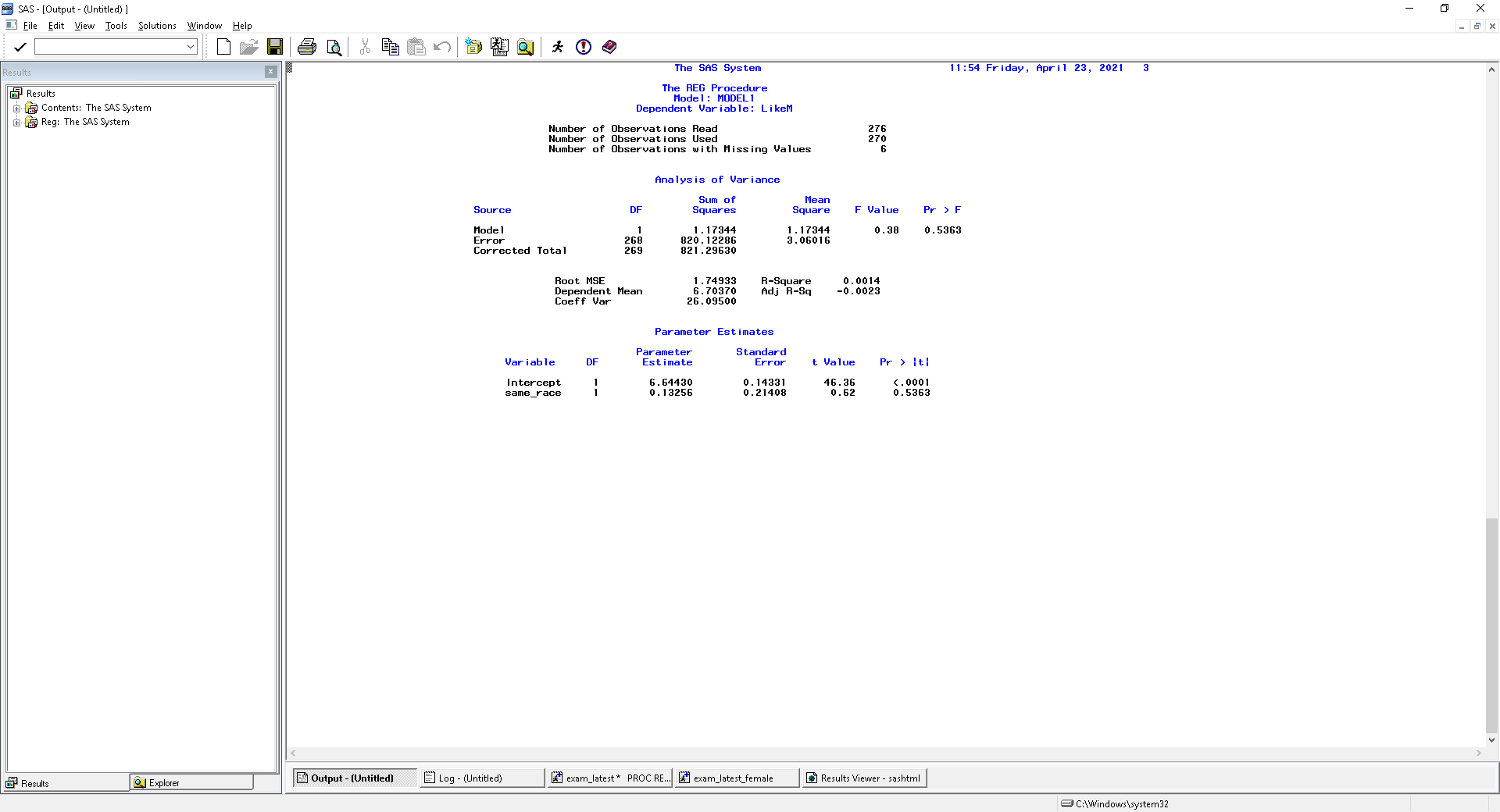
Next important step is cleaning the dataset, I looked at the missing values present in the dataset. Few outliers were observed in the box plot during diagnostics. The observed outliers are closely examined because they can bias our results. Thus, to avoid that I have replaced the missing values with the median of that variable.

**2.3 Impact of Same Race:**

Test is performed to check the impact of Race on the Like variable.

I have created the dummy variable called same\_race with 1 if both of them are from same race and 0 if they are not.

For Male:



Model equation is Y=β0+β1\*same\_race i.e., Y=6.64+0.132\*X

where the β0 is the intercept and β1 is the first regression coefficient.

Then, the Null hypothesis (H0): β1=0 means there is no significant relationship between a partner of same race and like variable

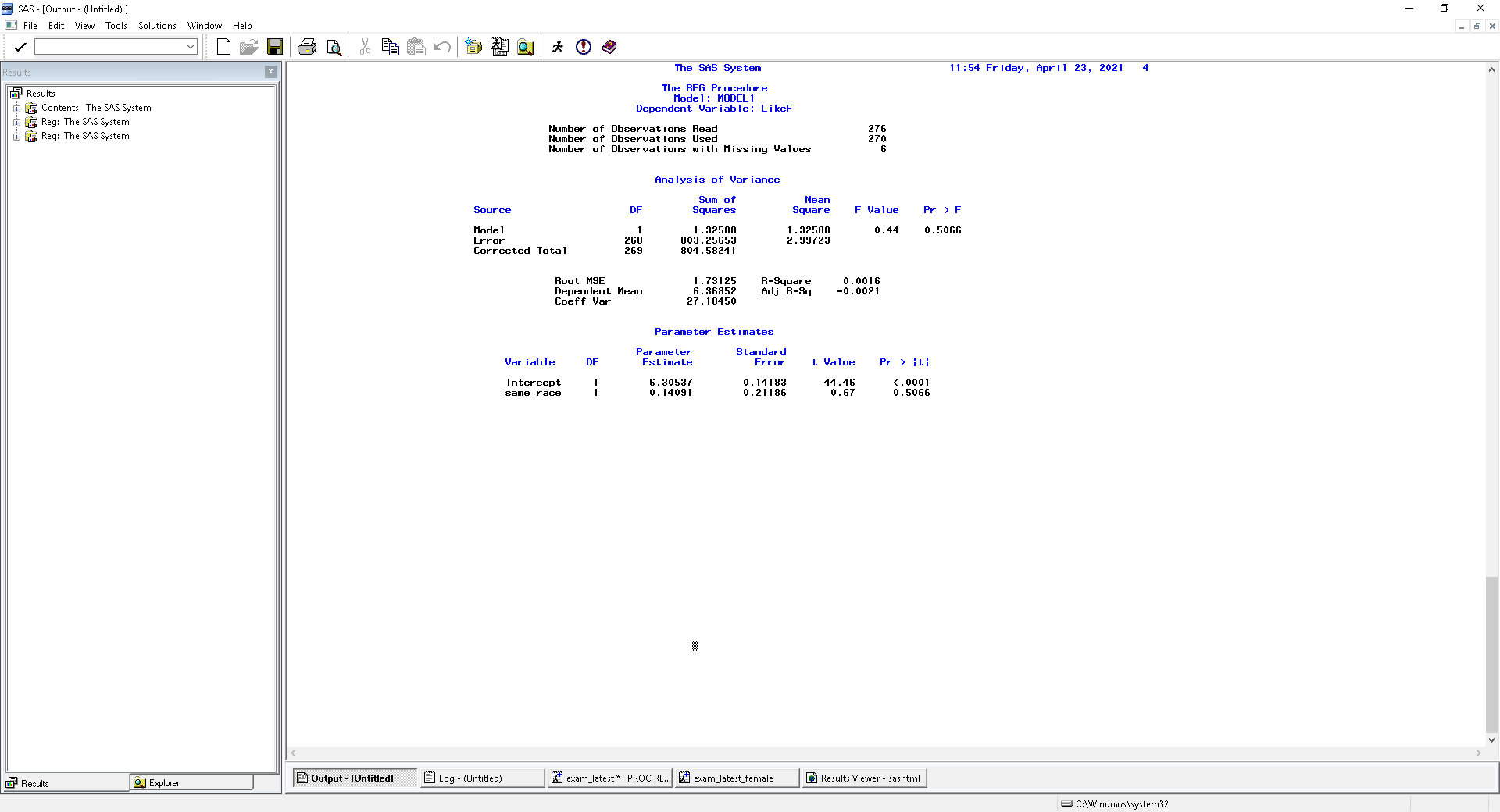
The alternative hypothesis (Ha): β1! =0 means there is significant relationship between a partner of same race and like variable.

If male and female have different race then then X=0 which means expected value of like is Y=6.64+0=6.64

If male and female have same race then X=1 which means expected value of like is Y=6.64+0.13=6.77

The test statistics is 0.62 and we see the P value is 0.5363, greater than 0.05. Since p-value is greater than the significance level (0.05) we do not reject the null hypothesis which concludes that there is no significant relationship between a partner of same race and like variable. Therefore, it does not matter if the partner belongs to same race or not.

For Female:



Model equation is Y=β0+β1\*same\_race i.e., Y=6.30+0.14\*X

Similarly for females, (H0): β1=0 and (Ha): β1! =0

The test statistics is 0.67 and we see the P value is 0.5066 > 0.05. Hence, we do not reject the null hypothesis which concludes that there is no significant relationship between a partner of same race and like variable. Therefore, it does not matter if the partner belongs to same race or not.

Thus, for both male and female, Race has no impact on the Like variable.

Figure6 shows that in our sample dataset 121 couples have same race.

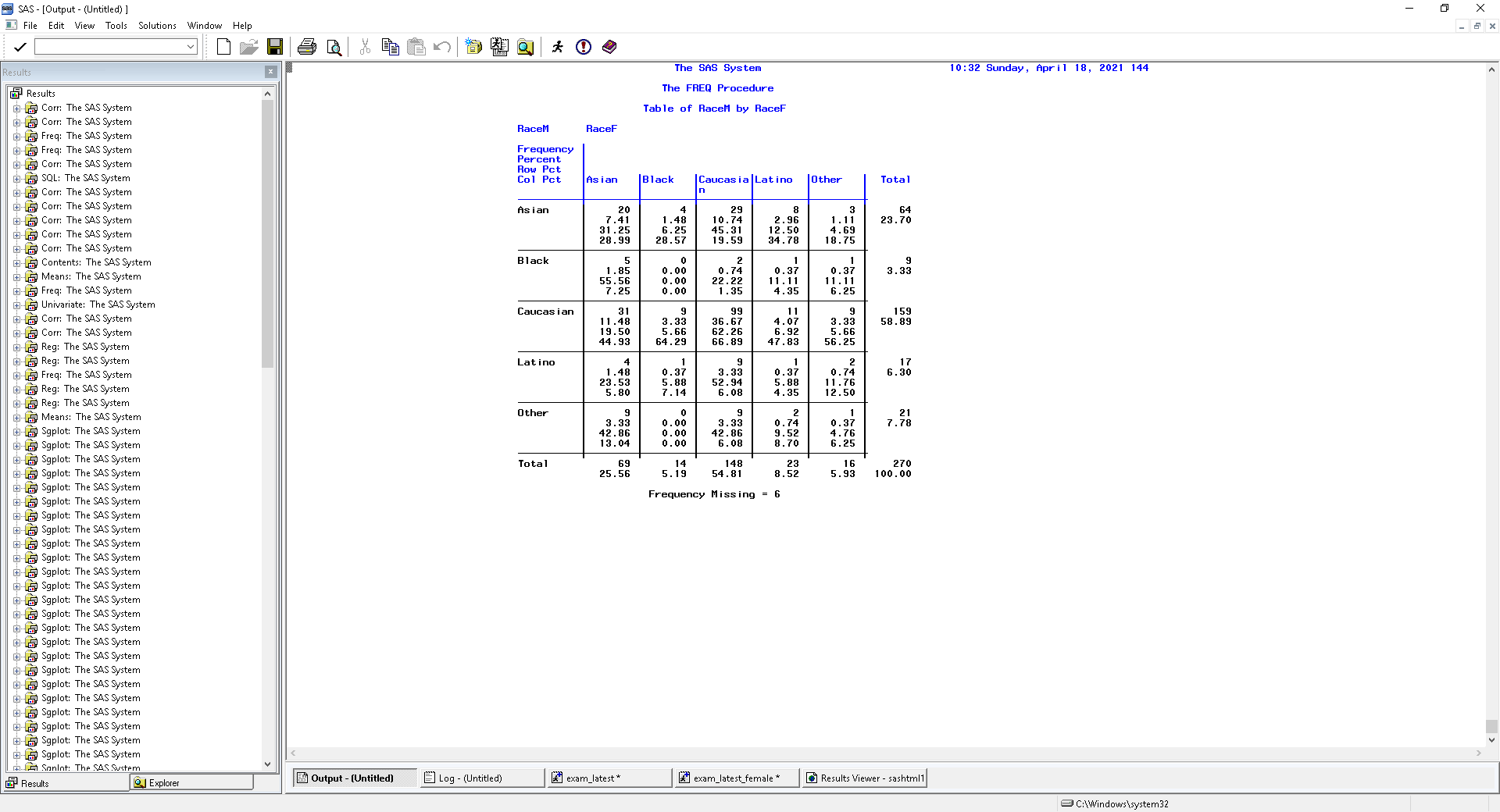
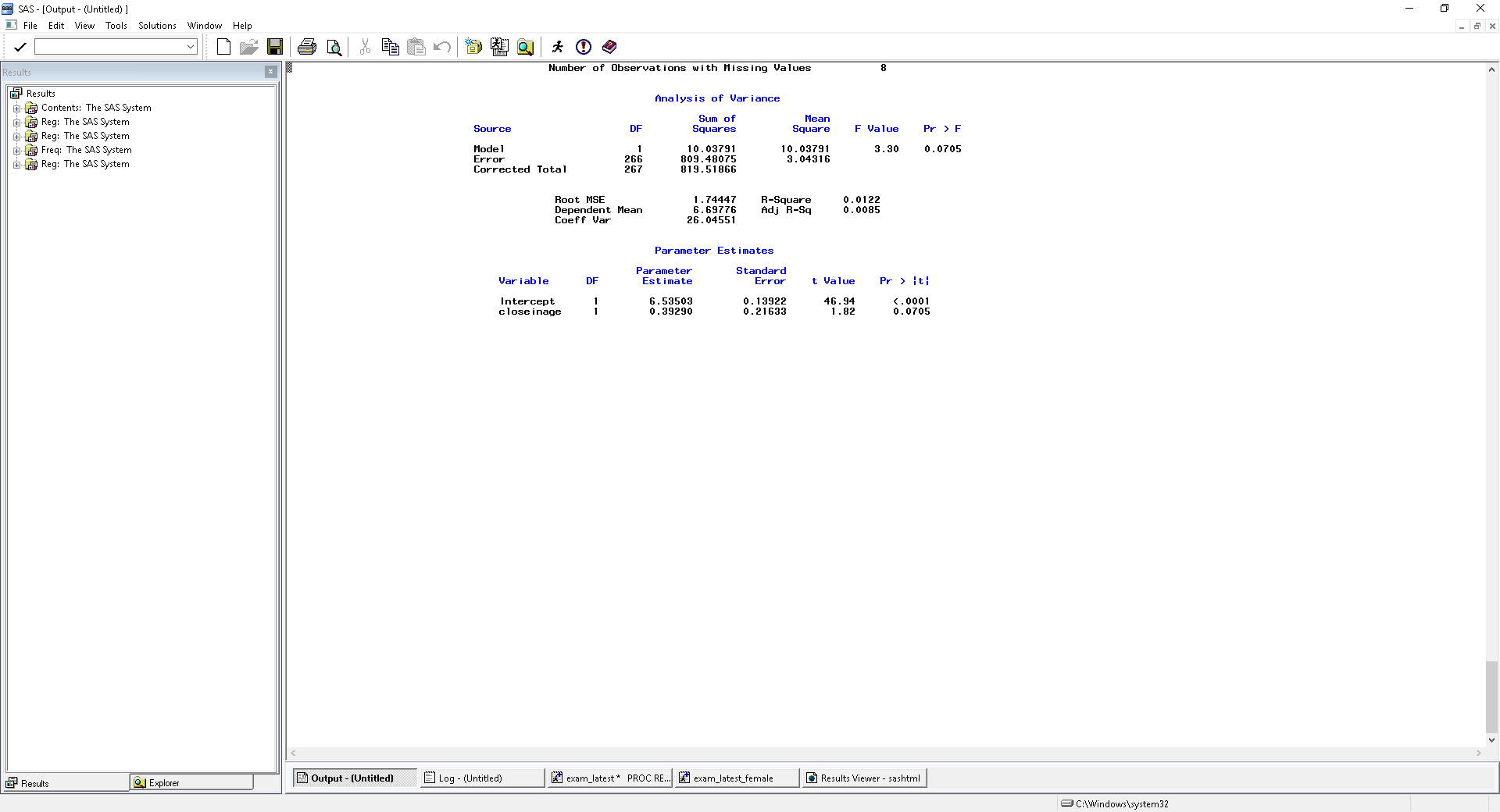


Figure8: Race distribution of couples

**2.4 Impact of close age**

This test is performed to check the effect of age range on the Like variable. The same age range is defined by being within 2 years of one another. I created the dummy variable called close in age. It is set to 1 if the age of male and female is within the 2 years otherwise it is set to 0.

**For Male:**



Model equation is Y=β0+β1\*closeinage i.e., Y=6.53+0.39\*X

where the β0 is the intercept and β1 is the first regression coefficient which measures the difference in predicted value of “like” when the partners are the same age and when they are not.

The Null hypothesis (H0): β1=0 no significant relationship between like and partner of same age

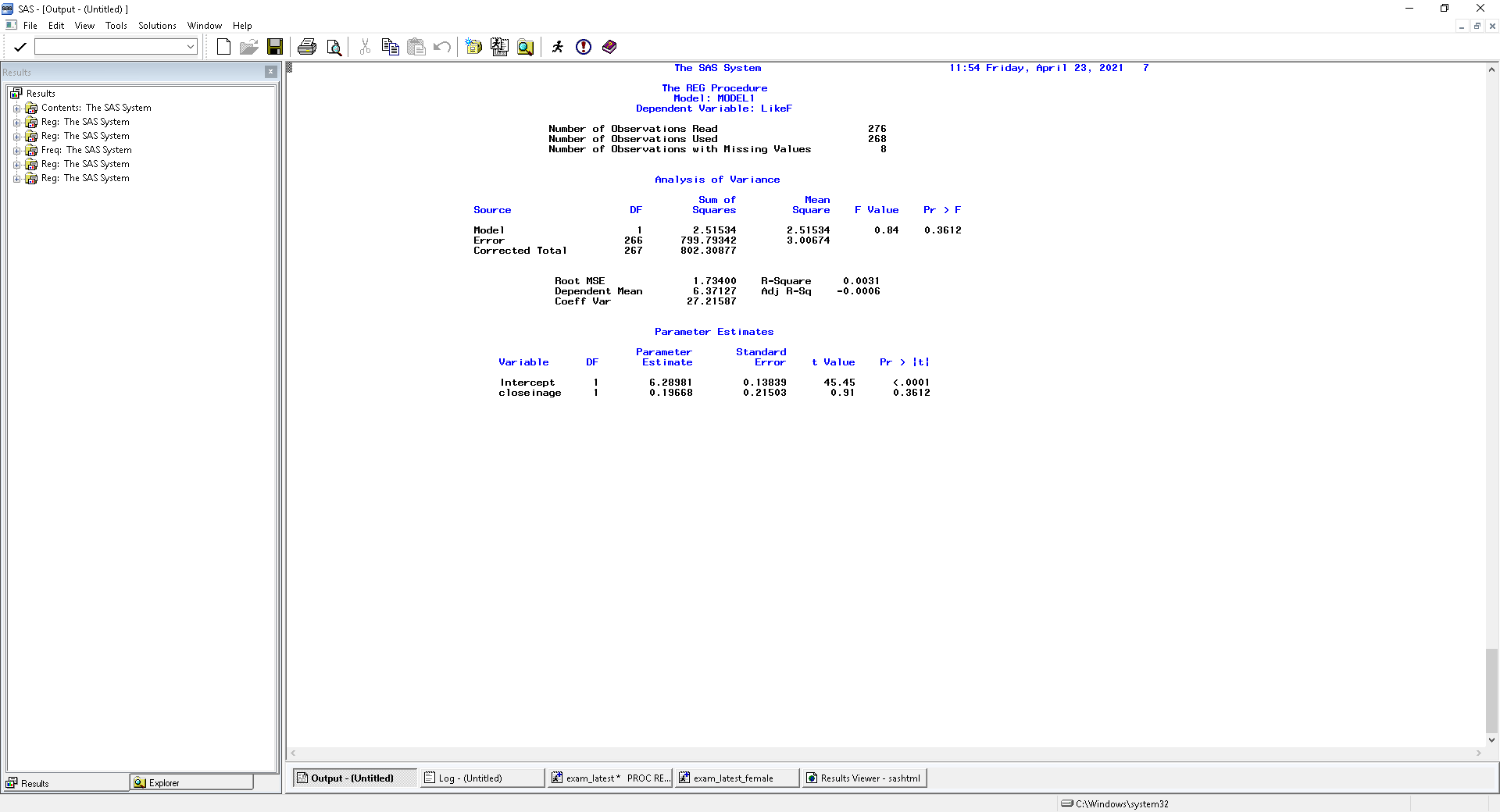
The alternative hypothesis (Ha): β1! =0 significant relationship between like and partner of same age

If male and female have different age then then X=0 which means expected value of like is Y=6.53+0=6.53

If male and female have same age then X=1 which means expected value of like is Y=6.53+0.39=6.92

The test statistics is 1.82 and we see the P value is 0.07 > 0.05. Hence, we do not reject the null hypothesis which concludes that there is no significant relationship between a partner of same age and like variable. Therefore, it does not matter if the partner belongs to same age or not.

**For Female:**



Model equation is Y=β0+β1\*closeinage i.e. Y=6.289+0.19\*X

(H0): β1=0 no significant relationship between like and partner of same age

(Ha): β1! =0 means significant relationship between like and partner of same age.

The test statistics is 0.91 and we see the P value is 0.36 > 0.05. Hence, we do not reject the null hypothesis which concludes that there is no significant relationship between a partner of same age and like variable. Therefore, it does not matter if the partner belongs to same age or not.

Into our dataset, there are 111 couples within the same age group and 157 couples belong to different age group.

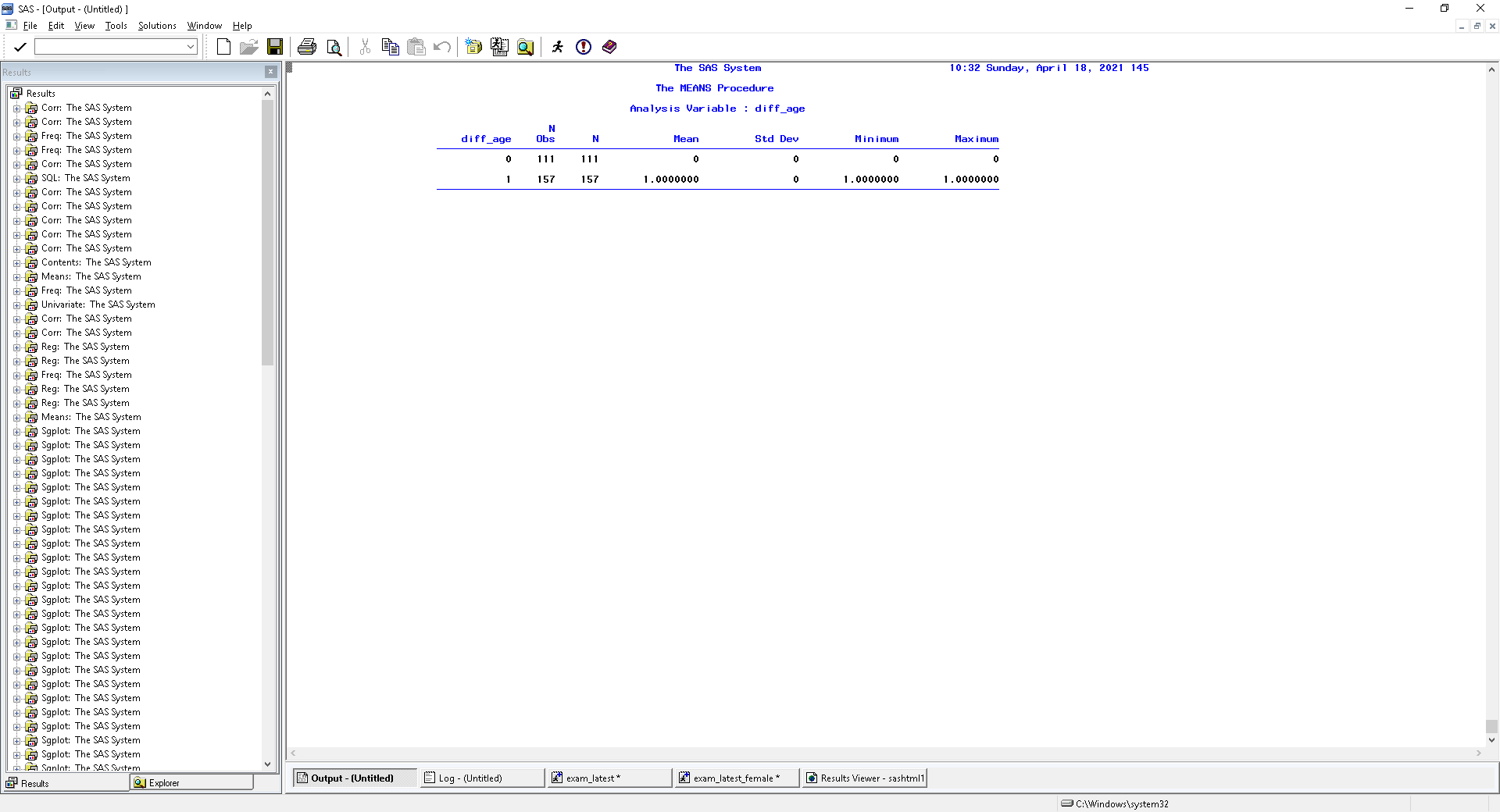


Figure9: Age Group distribution

1. **Model Development**

I created all possible polynomial terms as well as Interaction terms to add into the model. Let us start with the Male model first.

Interaction Effects in Equations: Interaction occurs among independent variables (X’s). We say there is interaction between two variables X1 and X2 when the relationship between X1 and Y is different at different levels of X2. It is represented as the product of two or more independent variables.

For Example, in below equation X1X2 is the interaction:

Y = b0 + b1X1 + b2X2 + b3X1X2. It is called a two-way interaction because it is the interaction between two independent variables. In our model we consider interaction up to two ways for simplicity.

Polynomial: For single independent variable X, we can generate the kth degree polynomial.

We use polynomial regression when the relationship between a predictor and response variable is nonlinear. Also, we consider polynomial up to cubic terms for simplicity.

Y = β0 + β1\*X + β2\*X^2 + … + βk\*X^k + ε

For our model, I have added all the possible pairwise interactions and third order polynomial for each variable. So, my dataset has new variables listed below(Figure 10) in addition to Attractive Sincere Fun Intelligence Ambitious and Shared Interests.

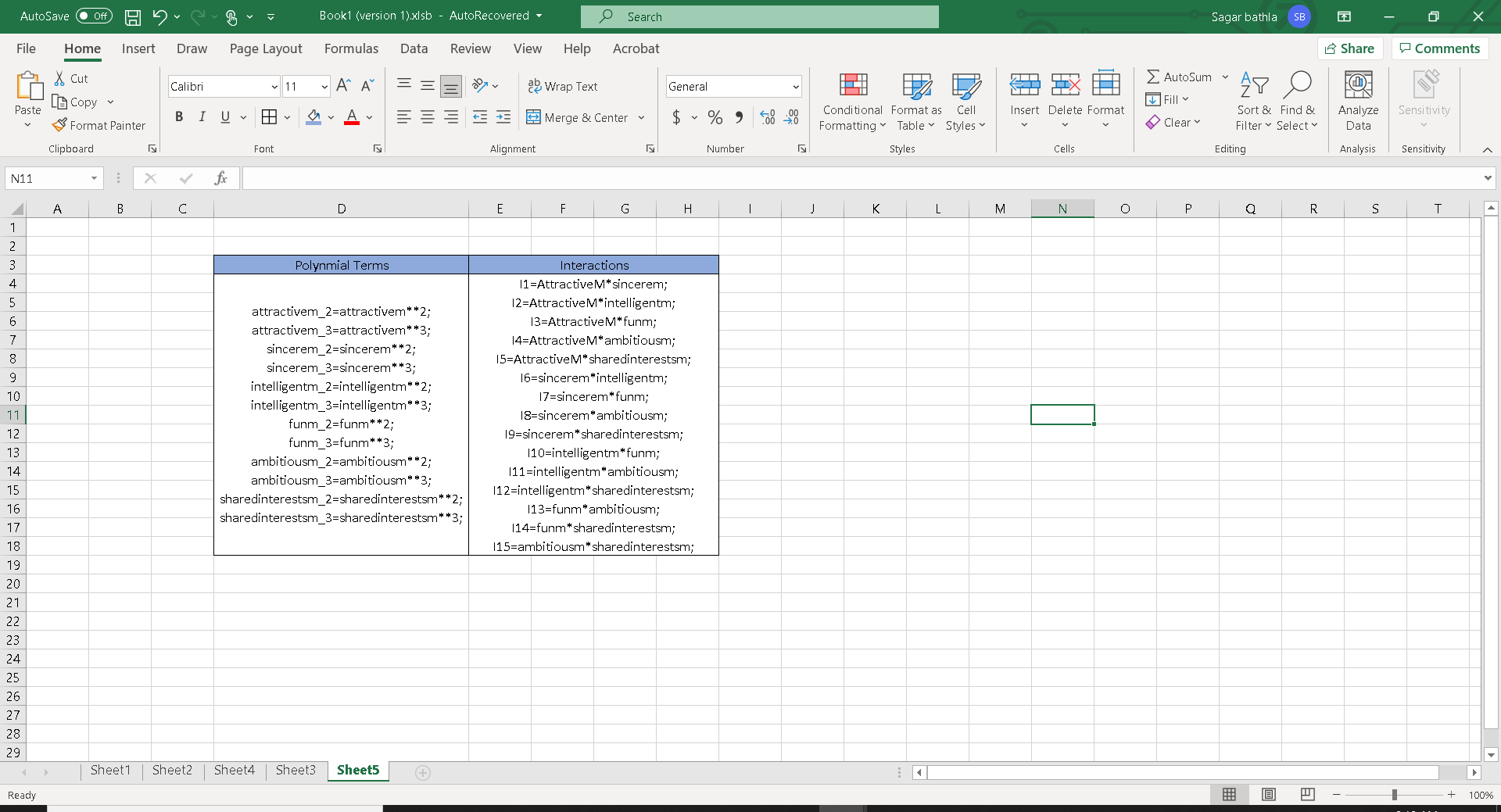


Figure10: Polynomial and Interactions

Next step is to divide the data into two parts- Training and Holdout. I used Training data to build the model and will be testing the final model on the holdout group. Training group has 2/3 of the data and holdout has 1/3 of the data.

**3.1 Variable Selection**

In multiple regression a common goal is to determine which independent variables contribute significantly to explaining the variability in the dependent variable. A goal in determining the best model is to minimize the mean square error, which would intern maximize the multiple correlation value R-square.

There are four methods of building the model and I will be using all four different strategies to create different models because every strategy has its own procedure.

1. *ALL POSSIBLE MODEL*: This method will fit (2^k-1) model for k independent variables for example, if k =10, then (2^10 -1) =1023.

SAS will list all the possible combination of variables and we select the most appropriate one using Mallow C(p), Multiple R-Square and MSE.

1. Mallows Cp is the measurement of mean square predictor error (p<=k).

We let p getting larger and larger until we find a point where cp<=p. This helps us to determine the correct number of variables for our model. Once we know the number of variables(p) for our model, then we use R^2 and MSE to select the variables for our model.

1. R square is the percent of variability in Y explained by X. R-Square ranges from zero to one. The larger the R-Square, the better the model.
2. MSE (Mean square error): This is estimated error variance of p variable model. Smaller the better.

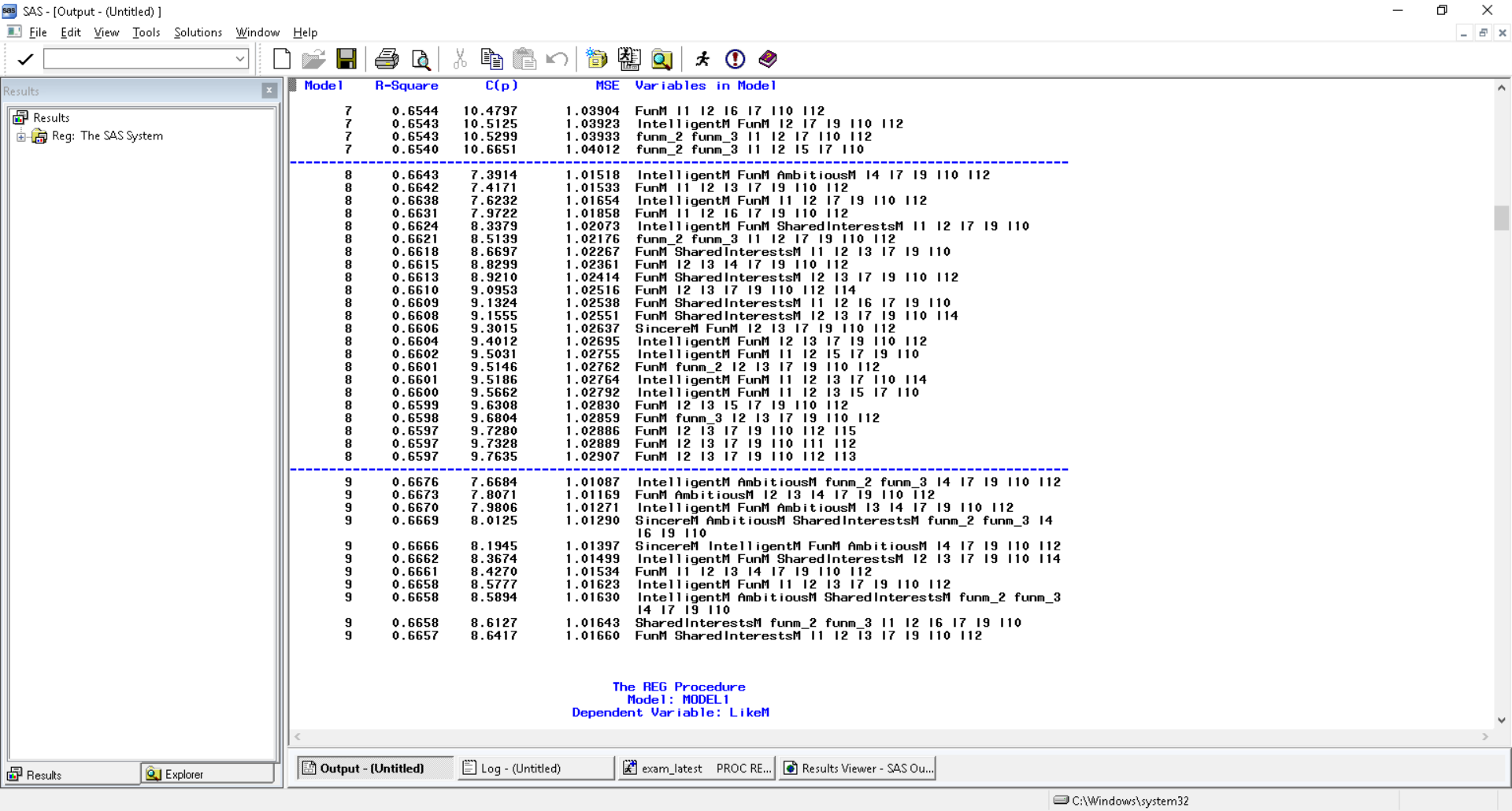


Figure 11 shows the point where cp<=p. Number of variables are 8. Out of all the combinations of 8 variables, select one with max R square and Minimum MSE.

Figure11: Variable selection from All possible Approach

1. *FORWARD SELECTION:*

It is constructive approach, i.e., it starts with no variable in the model. It will find the highly correlated variable among all the variables added to the model. Then out of the remaining variables, it will determine the partial F statistics for the variables not yet in the model, and then add to the model the variable with the largest partial F-value (if it is statistically significant). At any step, if the largest partial F is not significant, no more variables are included in the model, and the process is terminated.

It uses SLENTRY= option to mention the significance level entry (it uses default as .50). If the p value 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. 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 most of the analysis, an alpha is 0.05 is used as a cutoff value of significance.

1. *BACKWARD SELECTION:*

It is destructive approach; 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.

The default significance level is .10 but I will be using .05 which means the variables whose significant values are less than 0.05 can be kept in the model.

1. *STEPWISE APPROACH:*

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=0.50, SLSTAY=0.05 in this model, because I want the variables whose significant values less than 0.50 can be added, among those variables, the significant values less than 0.05 can be kept in the model. The summary of the variables obtained from different processes for male model is shown in figure12.

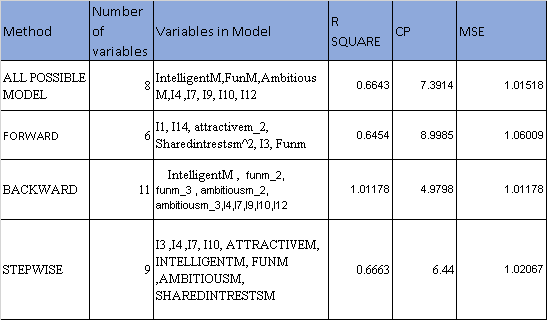


Figure12: Variable Selection Outcome from the above four approaches

Out of all the four methods, I eliminated the Forward approach because cp>p. From the remaining three I looked at the least Mean Squared Error (MSE) in remaining approaches which are close.

So, I looked at all the approaches one by one. As we know in case of polynomial and interaction, if higher order terms are present in the model, then base terms should also be there. For example, in backward approach we see funm^2 but there is no base term funm similarly I7 is interaction of attractivem\*sincerem and sincerem variable is not selected. I tried to fix the model by including the base terms.

Thus, final variables after fixing the base terms are:

*In stepwise approach*: I3, I4, I7, I10, ATTRACTIVEM, INTELLIGENTM, FUNM, AMBITIOUSM, SHAREDINTRESTSM

*In backward approach*: INTELLIGENTM, FUNM, FUNM\_2 FUNM\_3 AMBITIOUSM AMBITIOUSM\_2 AMBITIOUSM\_3, ATTRACTIVEM SINCEREM SHAREDINTRESTSM I4, I7, I9, I10, II2

*All Possible Model*: ATTRACTIVEM, INTELLIGENTM, FUNM, AMBITIOUSM, SHAREDINTRESTSM, SINCEREM, I4, I7, I9, I10, I12

After adding all base terms in the model, I ran the updated regression model, then we see some variables’ p values are greater than significance level 0.05, I started removing one variable at a time until all remaining variables produce p values are less than significance level, 0.05.

Therefore, final variables from stepwise are: ATTRACTIVEM, INTELLIGENTM, FUNM, SHAREDINTRESTSM, SINCEREM,I10.

I got the same variables after elimination of nonsignificant variable from backward approach and All Possible Model.

Thus, I came to my conclusion that the variable selected are correct because all the three approaches are giving same result. Below Figure13 shows that all the variables in the model are significant.

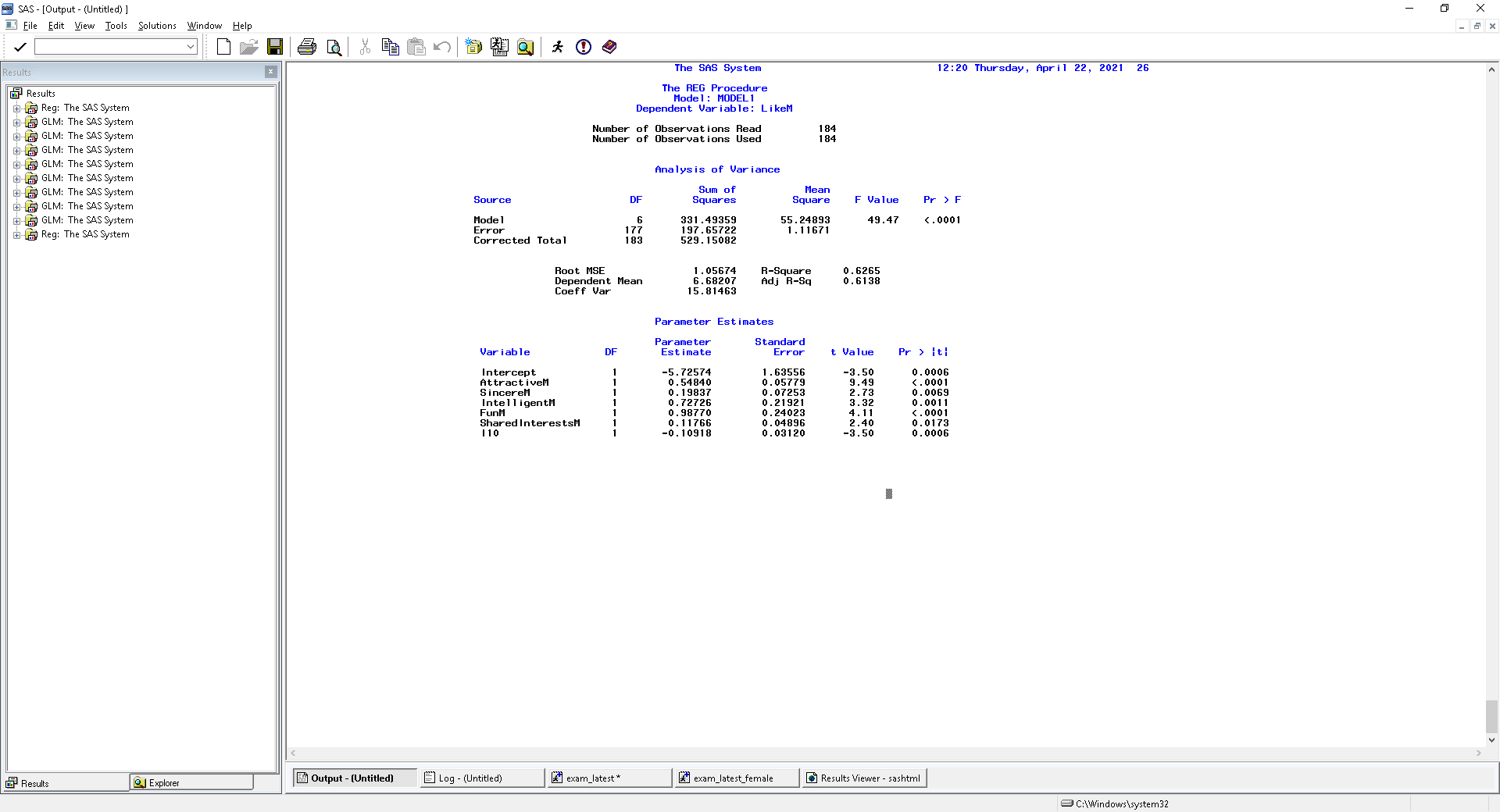


Figure13: Regression Output

Model Equation for male:

Likem = -5.72 + 0.548\*attractivem +.727\*Intelligentm +0.98770\*funm +0.1176\*SharedInterestsm +0.198\*sincerem -0.109\*I10

Interpretation:

With one unit increase in the attractive rating, we expect likem to increase by .548 if all other variables are kept constant.

With one unit increase in intelligent rating we expect likem to increase by .72 if all other variables are kept constant

With one unit increase in fun rating we expect likem to increase by .98770 if all other variables are kept constant.

With one unit increase in sharedinterests rating we expect likem to increase by .117 if all other variables are kept constant.

With one unit increase in sincere rating we expect likem to increase by .198 if all other variables are kept constant

I10 is interaction between intelligentm and funm. We can say that effect of intelligence on like variable depend negatively on the fun variable. It implies that the higher the fun, the lesser (more negative) the effect of intelligence on like is and vice versa.

1. **Regression diagnostics for the final model**

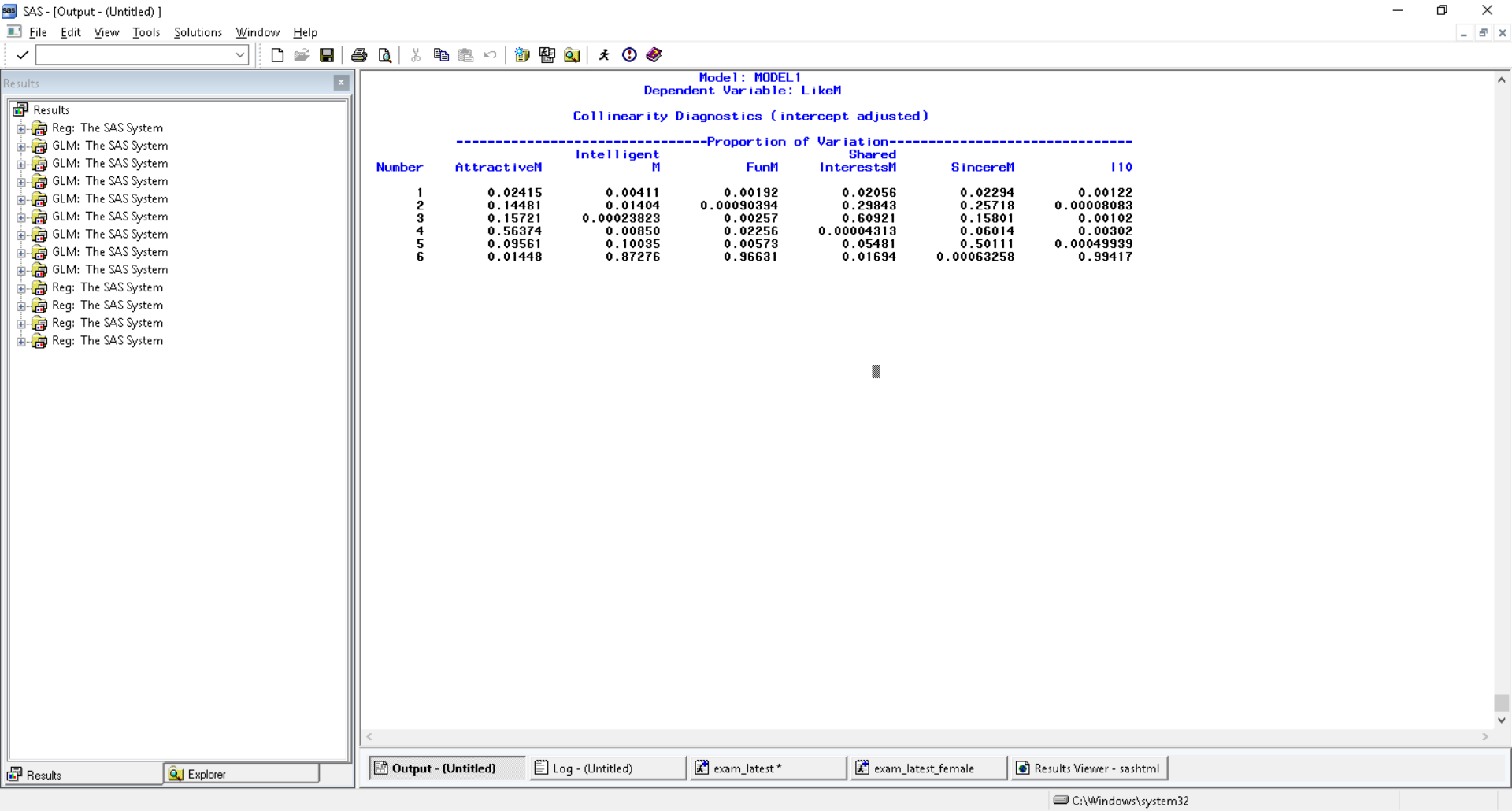
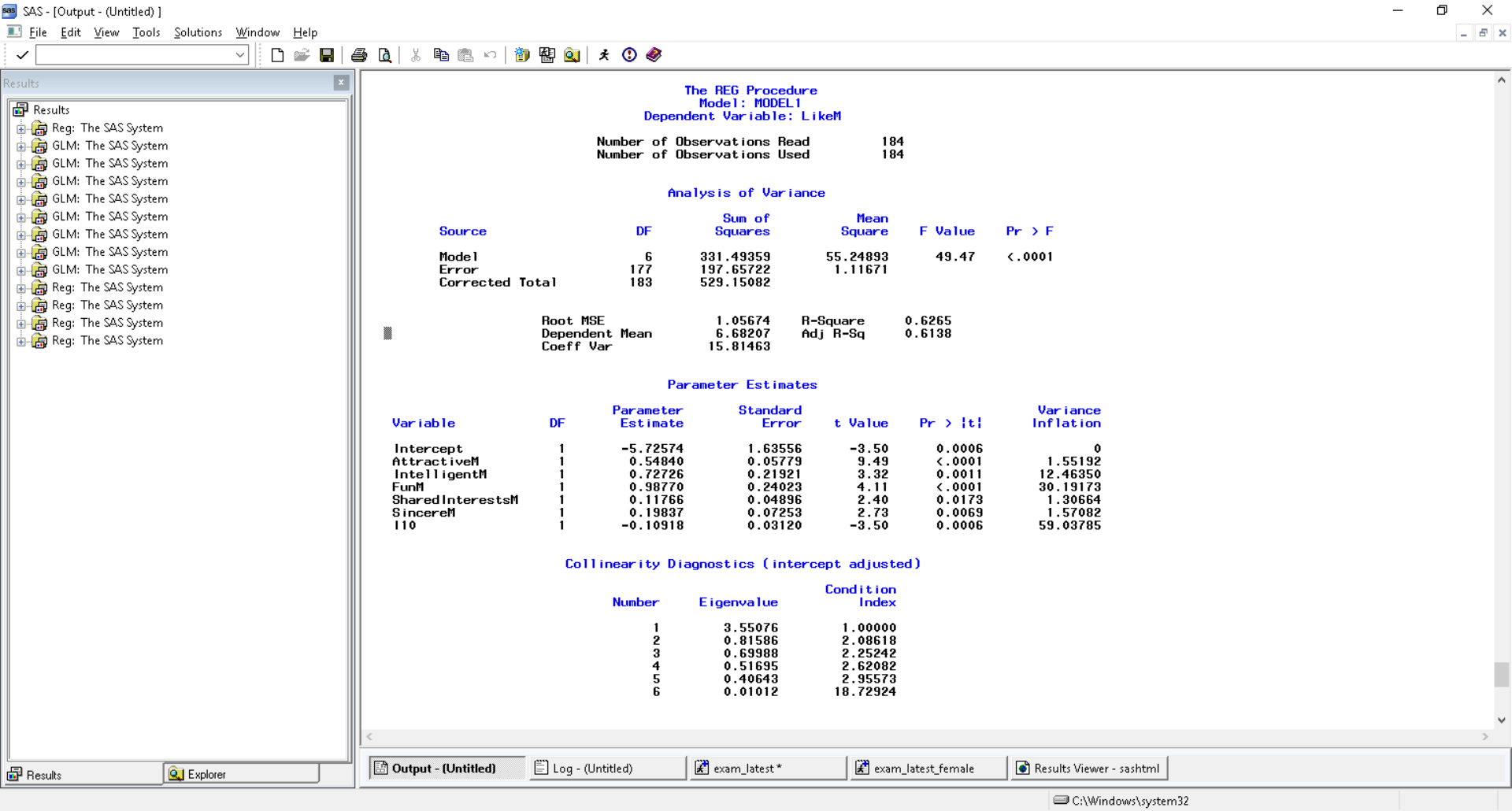
Regression diagnostics are statistical techniques designed to detect conditions which can lead to inaccurate or invalid regression results. We have already checked the data prior to start of process. Let us move on to next step for detecting collinearity.

**4.1 Collinearity test**

Collinearity exists when there is strong linear relationship between independent variables. It is often a major problem in polynomial regression and interactions. We always use two criteria to test collinearity issue, Eigen Analysis, and variance inflation factor (VIF).

Variance Inflation Factor: The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model. The VIF for k predictor variables and the jth regression coefficient is proportional to 1/(1-Rj2). Variance inflation factors range from 1 upwards. In general, a VIF above 10 or R-square above 0.90 indicates high correlation and is cause for concern.

Eigen Analysis: Condition indexes and condition numbers are calculated from eigenvalues and are found through the formulas and . The condition number is the largest condition index and condition indices are non-decreasing. Condition numbers greater than 30 suggest potential collinearity and variables which are responsible for collinearity is determined by the variable > .50 under Proportion Index. Figure 14 shows that multicollinearity is present in my model.

 Figure14-Detection of Multicollinearity

We see that Intelligent, Fun and I10(intelligent\*fun) are highly correlated.

To reduce high VIFs produced by interaction and higher-order terms, one way is to standardize (center) the continuous predictor variables. This method removes the multicollinearity produced by interaction and higher-order terms and it has the added benefit of not changing the interpretation of the coefficients. In my model I centered the Intelligent and Fun variable and after transformation my variables have VIF<10 and condition index <30. The results obtained after transformation is shown in Figure15.

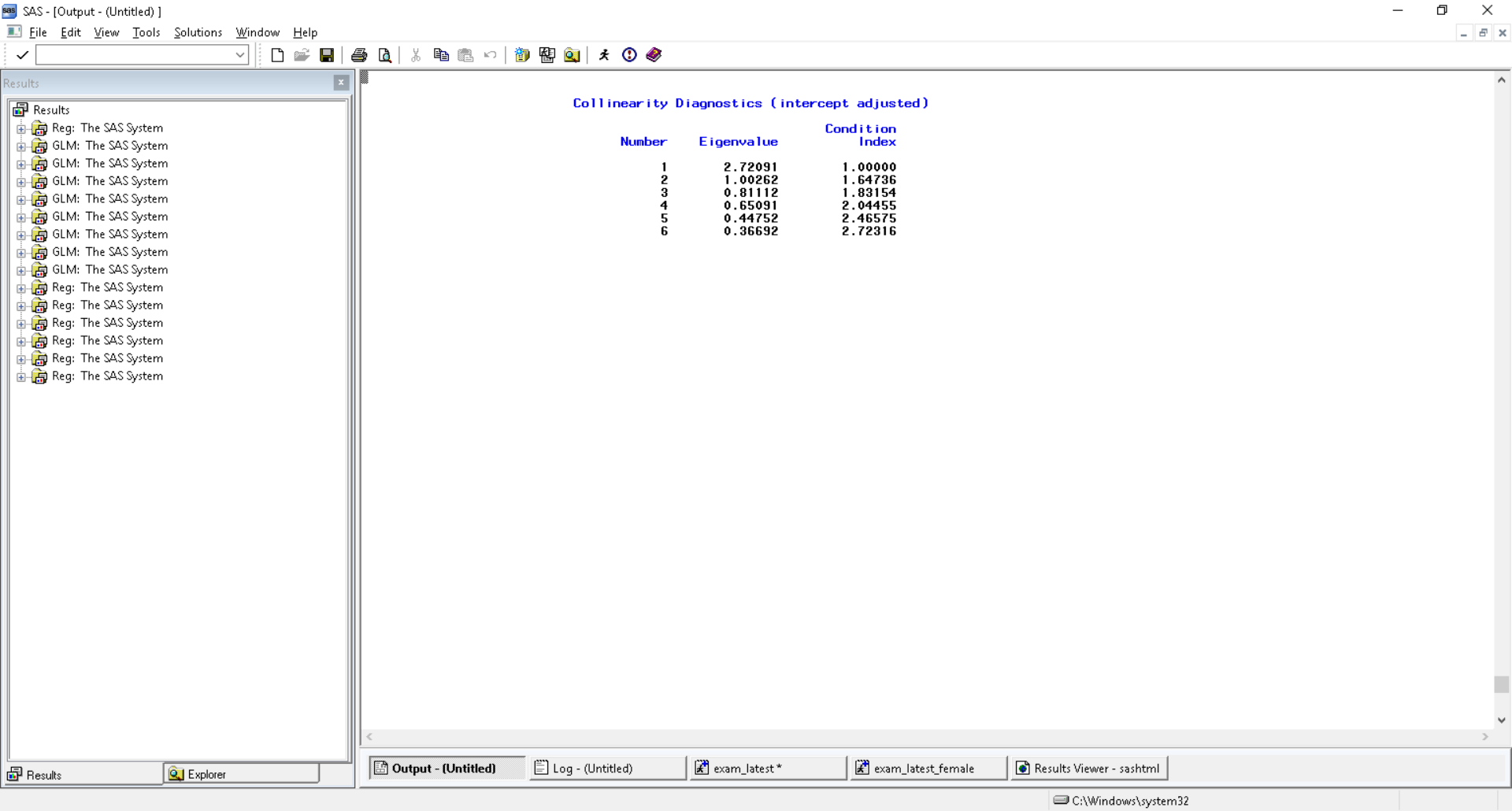
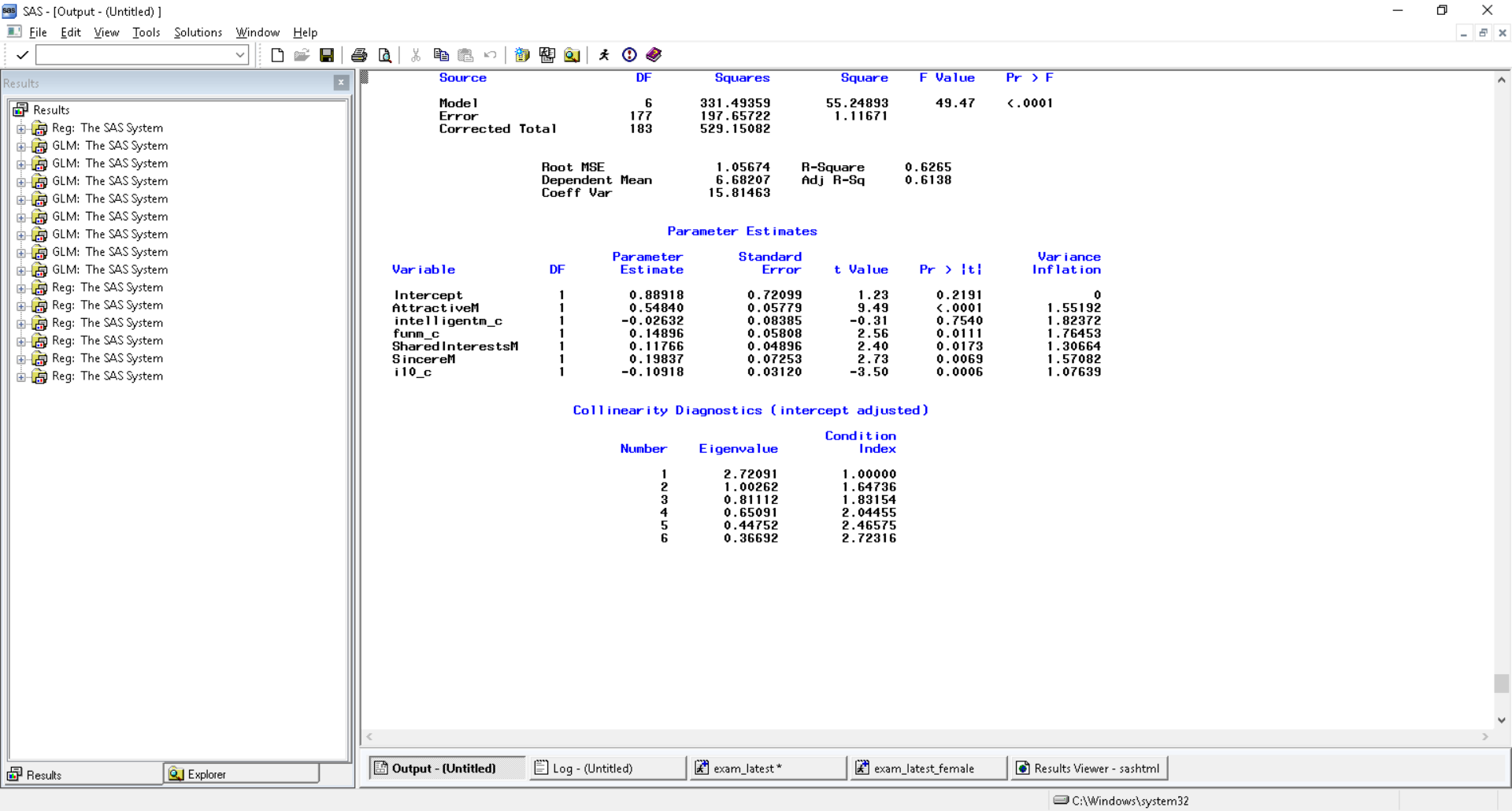


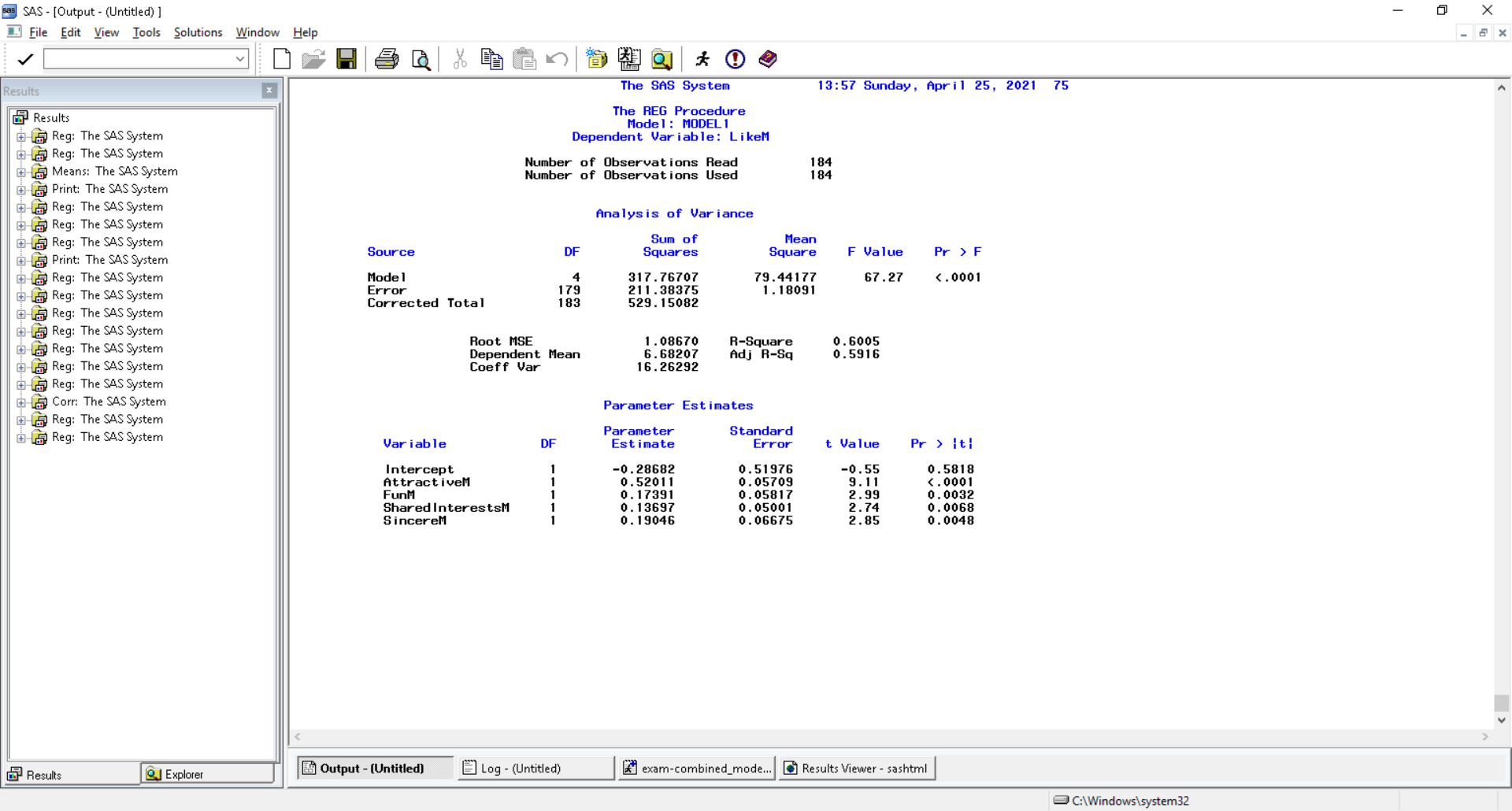
Figure15-Resolving Multicollinearity

Therefore, the collinearity issue has been solved but after transforming, one of my variable intelligentm\_c became insignificant, so I went another way of resolving multicollinearity by removing the intelligentm and interaction associated with it (I10) from the model because interaction cannot be present in model without base term.

Now the final model is:

Y= -0.28682+ 0.52011\*attractivem+ 0.17391\*funm+ 0.19046\*sincerem+ 0.13697\*sharedinterestsm

**Interpretation**: With one unit increase in attractive value, we would expect Y to increase by 0.52011 when all other variables are held constant. With one unit increase in fun we expect funm to increase by 0.17 when all other are held constant. Same goes for other variables (sincere and sharedinterests) in the model.



**4.2 Checking the assumptions:**

Below are the assumptions of linear regression model.

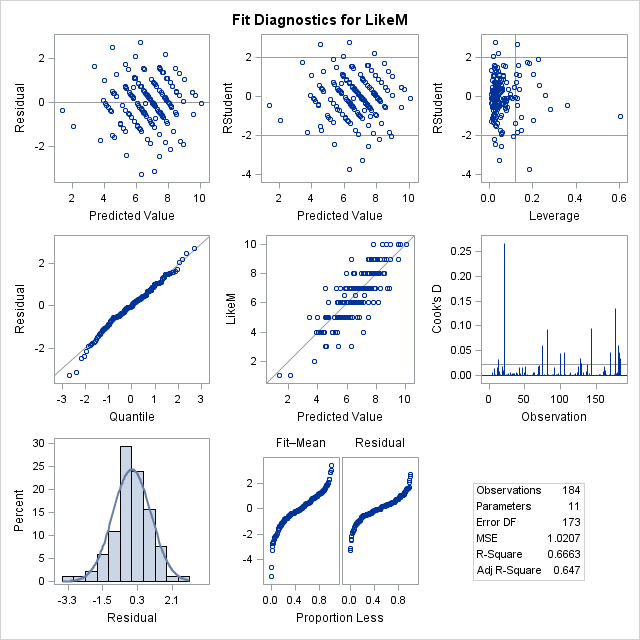
1. Independent: The Variable Y are independent of one another.  
2. Linearity: The mean value of Y is straight line function of X. There exists a linear relation between independent variable Y and dependent variable X.  
3. Homoscedasticity: Homo means same, and Homoscedasticity means scattered thus the variable Y has same variance for any value of X.  
4.Normal Distribution: The variable Y follows the Normal distribution.

Linearity, homoscedasticity, and independence:

In General, Residual = (observed y – model predicted y) and standardized residual=Residual divide by SD.

Jackknife residual (which is the most efficient way) means we test the relationship between the residual and predicted value after removing the ith observation.

A plot of jackknife residuals versus predicted values indicates whether or not assumptions of homoscedasticity, independence, and linearity are violated. If the model does not meet the linear model assumption, we would expect to see residuals that are very large (big positive value or big negative value). To assess the assumption of linearity we want to ensure that the residuals are not too far away from 0 (standardized values less than -2 or greater than 2 is problematic). To assess if the homoscedasticity assumption is met, we look to make sure that there is no pattern in the residuals and that they are equally spread around the y = 0 line.



Based on this plot, the scatter appears to be quite random with no evident funneling or sinuous pattern. Thus, assumptions of homoscedasticity, independence, and linearity are not notably violated.

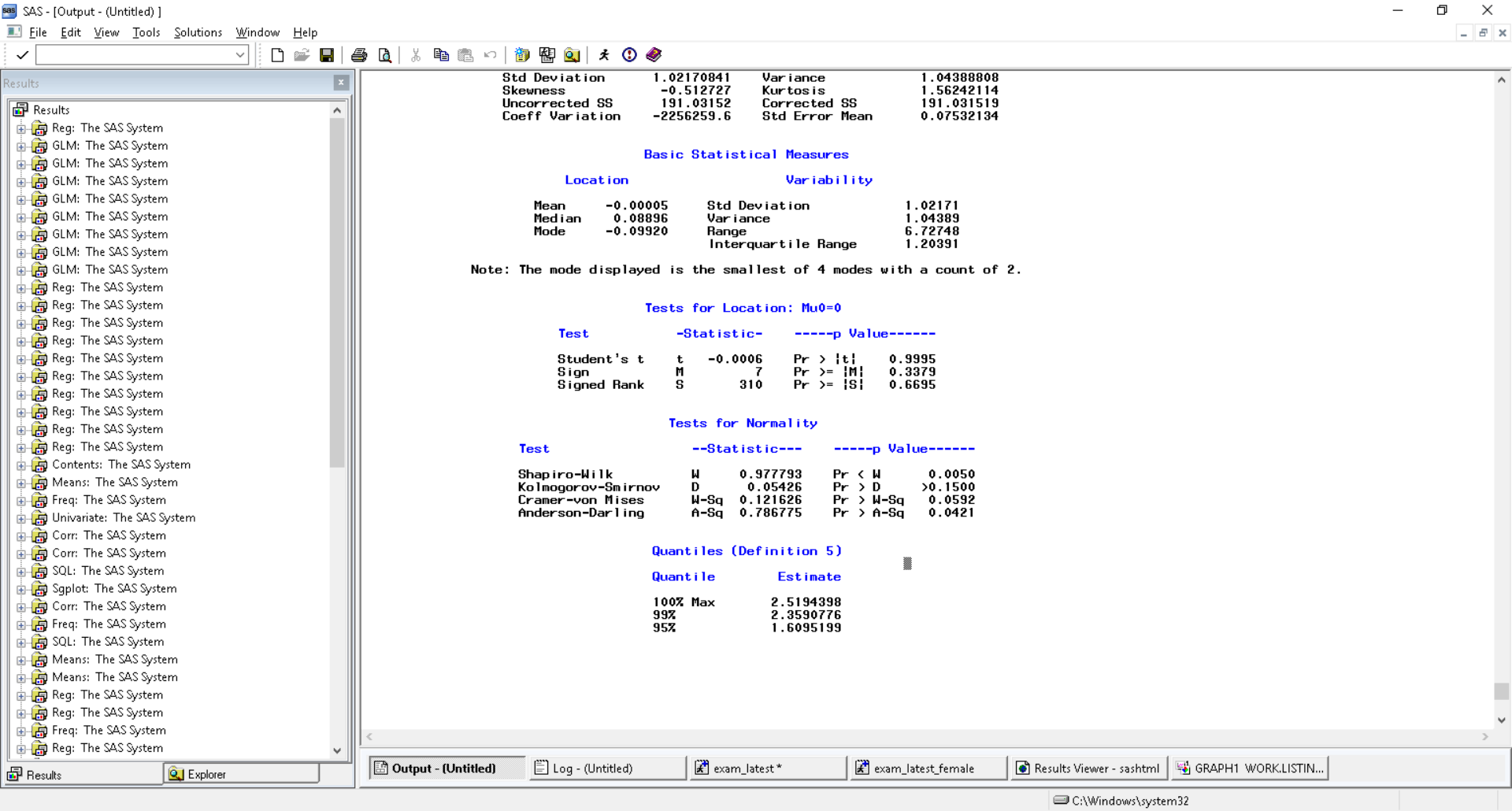
Normality assumption:

Here I used different methods to check the normality assumptions:

1. *Several graphs, plots, and tables can be created via PROC UNIVARIATE*. An investigation into basic descriptive statistics of Jackknife residuals yields a 25% Q1 value of -0.58, a median of 0.0889, a 75% value Q3 of 0.622, a mean of -.00005, and a standard deviation of 1.02. Skewness is -0.05 and it measure the degree and direction of asymmetry. Looking at these numbers, it looks like normal because Mean is close to 0, variance is close to 1 and skewness is also close to zero.
2. *Goodness of fit test (Hypothesis Testing):*

H0: It follows the normal distribution

Ha: It do not follow Normal distribution



We will consider the result from Kolmogorov-Smirnov because our sample data is more than 50 otherwise, we would have considered Shapiro-Wilk test. Since the p value is greater than the significance level (.05) it means do not reject the null hypothesis i.e., the data follows normal distribution.

1. *Normal Quantile Plot (QQ Plot):*

This tells us that if the data is normal then it should be close to straight line. The distribution looks normal as it is close to the diagonal line so we may assume that normality holds here (Figure16).

1. *Histogram:*

To confirm the normality of residuals I checked the Histogram, it looks little left skewed but that is very minor. Overall, it looks normally distributed(Figure17).

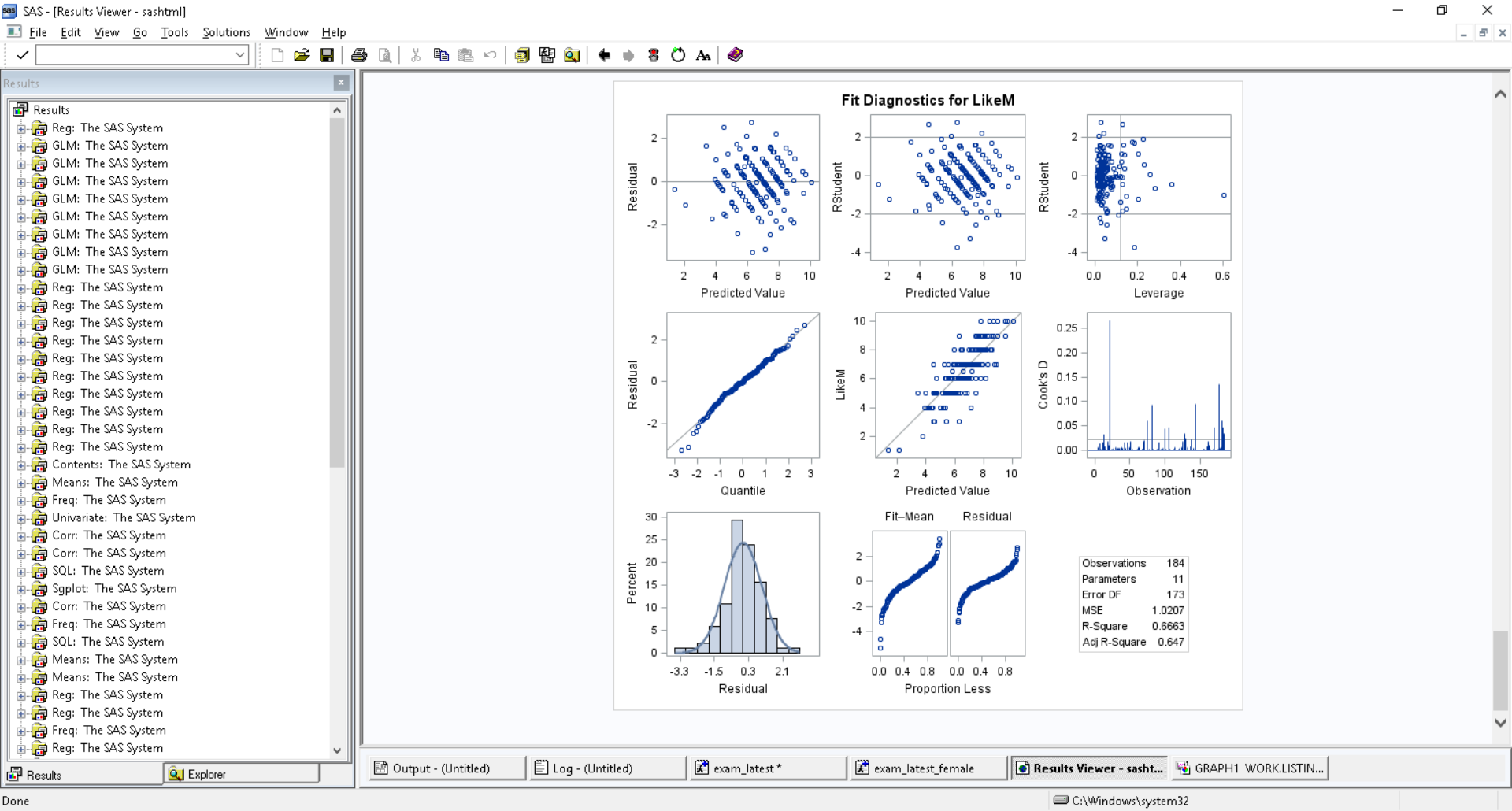
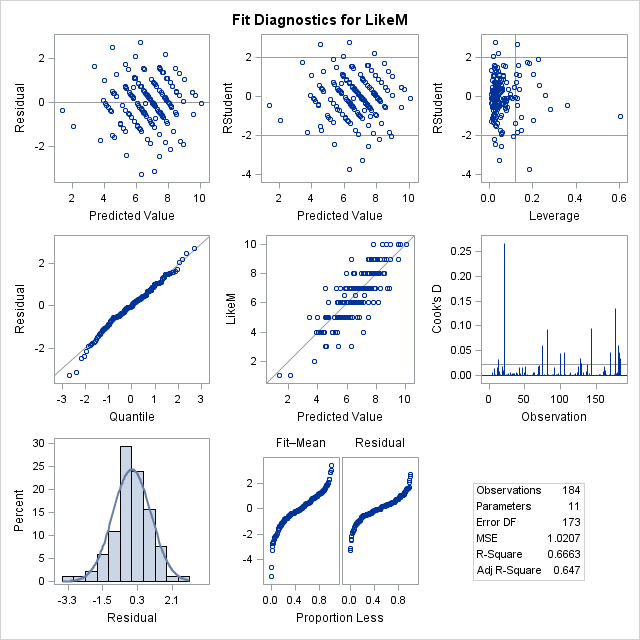
 

Figure16-QQ Plot Figure17-Histograms

Therefore, all the four assumptions of linear regression model are satisfied.

**4.3 Outlier Detection:**

Outliers are observations that are extreme in the sense that they are noticeably different than the other data points. Large residual/Outlier means that observed /actual value is far from predicted. We test the outlier using three things (Jackknife residual, Cooks Distance, and leverage). Using these three methods explained in Figure18, 44 outliers are observed which are shown in Figure19.

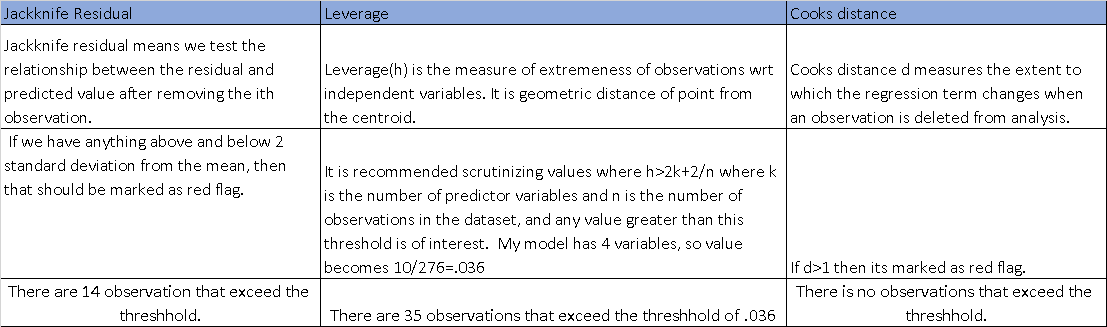


Figure18-Oulier detection methods

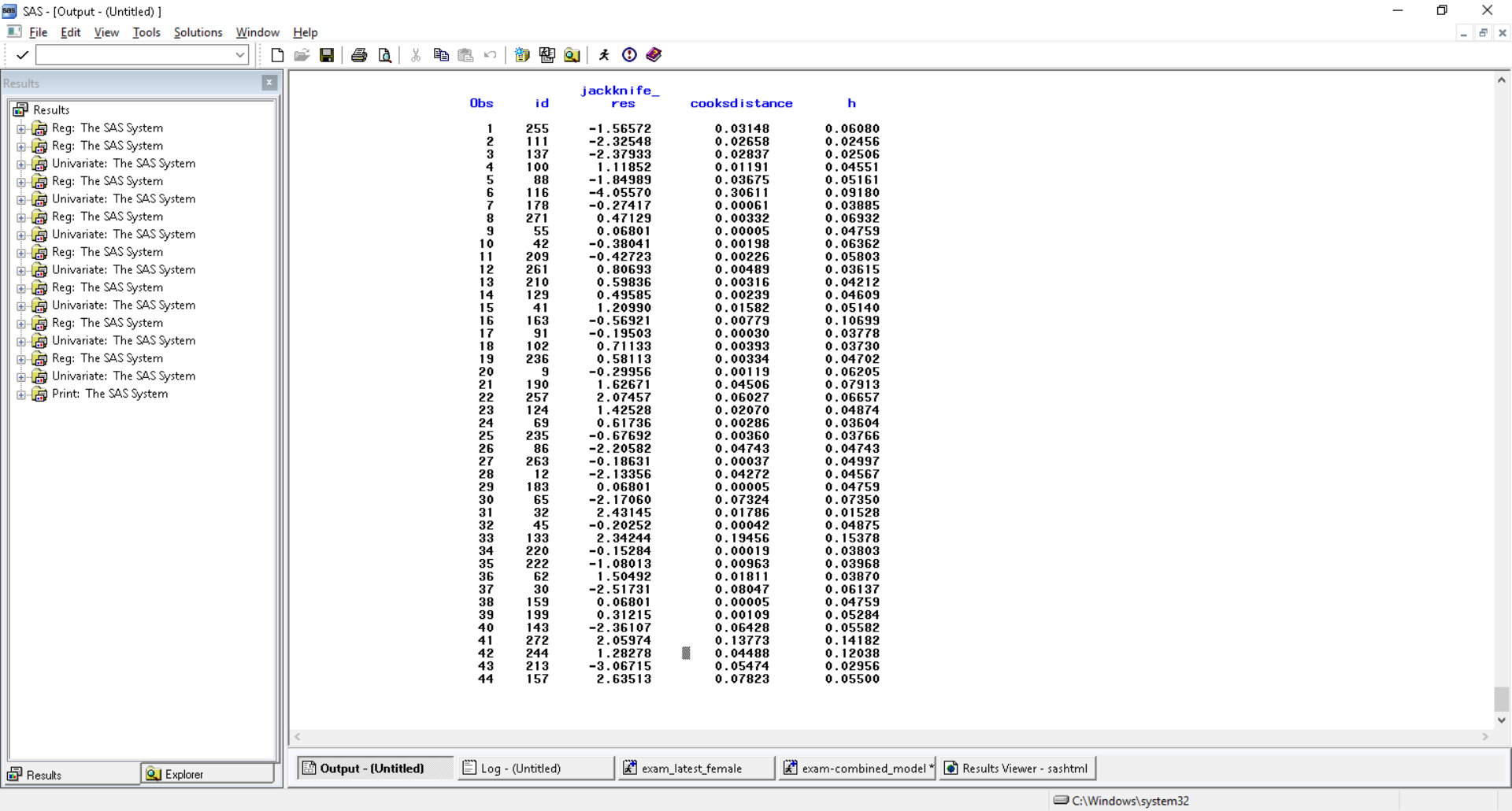
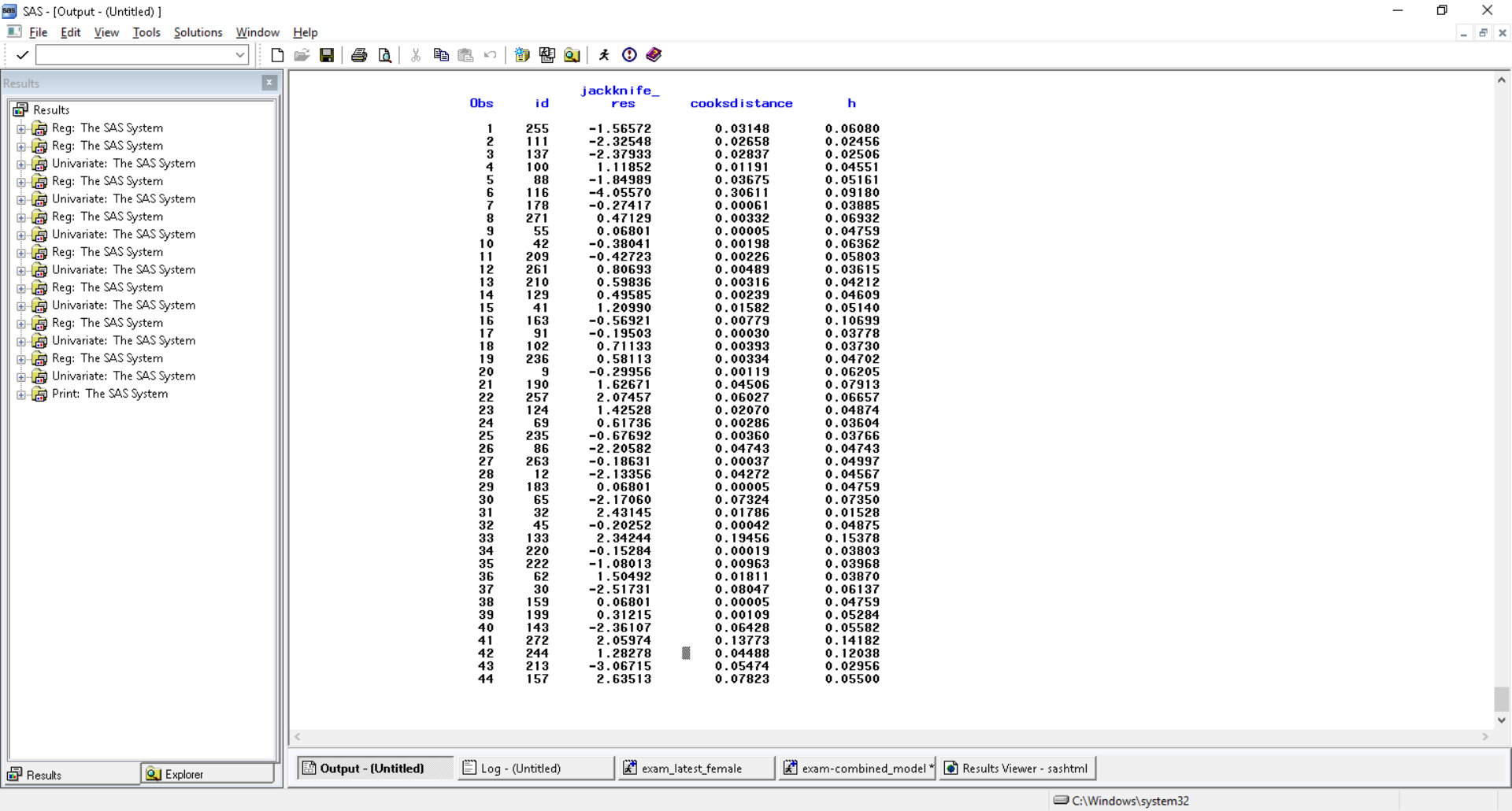
 

Figure19- Observations containing outliers identified via a Cook’s distance of greater than one, leverage values greater than 10/276, and Jackknife residuals with absolute values greater than 2. No observations were noted for suspicious di, but nine observations had |Jackknife residuals| > 2 and h>036. These observations should be further explored.

High residual can mean two things:

1. Data entry error: In general, outliers identified in this analysis should be more closely examined for plausibility, which is commonly determined prior to the start of the study. After consulting with the researcher, any data entry errors should be corrected. In my model I am assuming that there is no data entry error.
2. It can mean other thing as well. Example if any one of them have paid their partner to inflate the rating, then in that case residual will be positive and will be high. Observation **157** in original dataset seems to be inflated because it has **residual 2.6** and **h=0.0550** and it has been rated high as compared to predicted.
   1. **ACCURACY OF MODEL:**

I have checked the closeness of the actual and predicted values using methods expplained in Figure20.

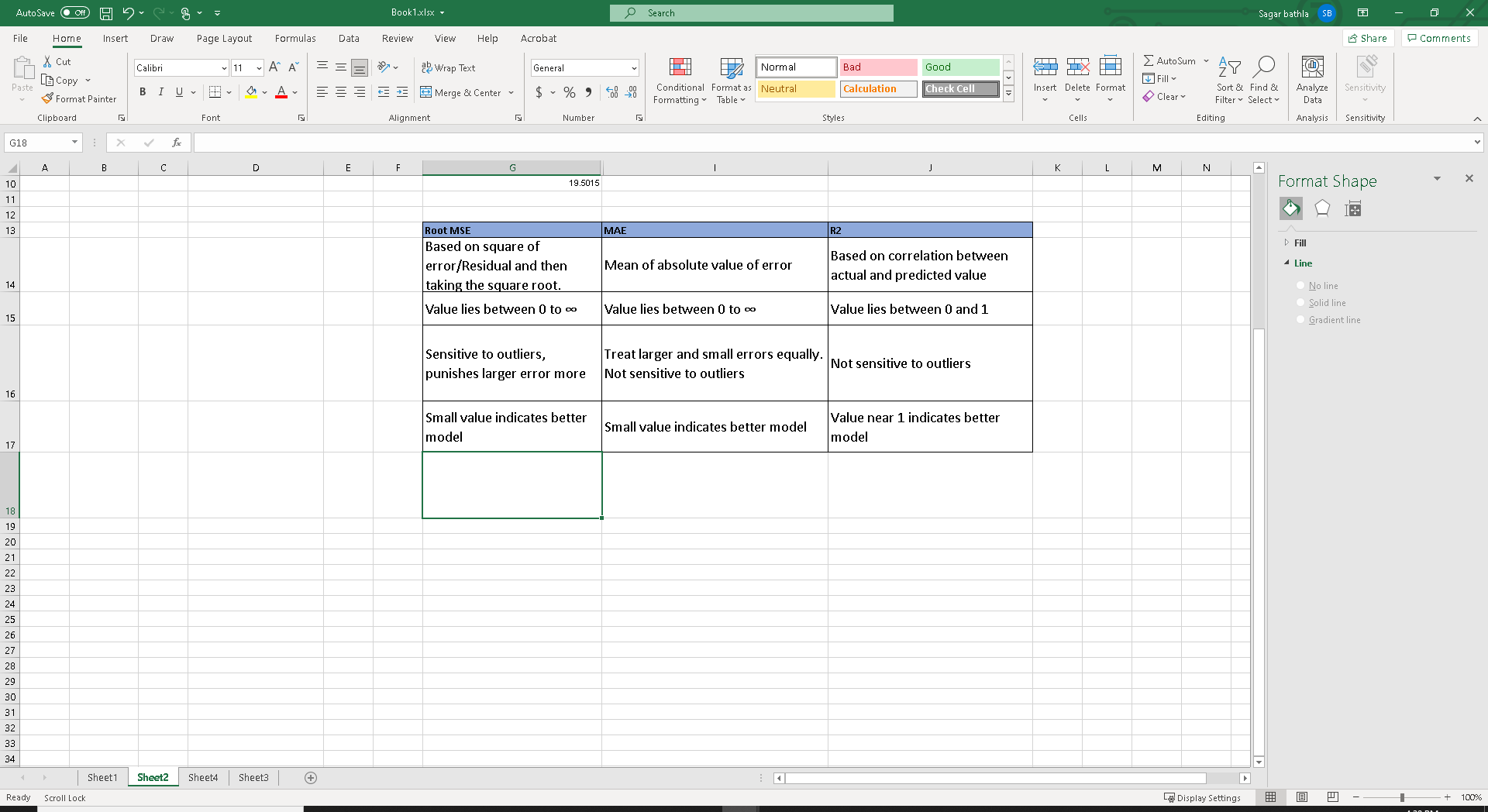


Figure20-Methods for Accuracy

MAE (Mean Absolute Error) for my model is **0.80**

RMSE for my model is **1.08**

R-Squared value is **0.6005**

This means that the average difference between the predicted and observed like value is **0.80**. Since **RMSE> MAE** it means there is some variation in the magnitude of error.

1. **Checking Reliability of Model**

Reliability is a measurement of how well the final model handles the future predictions. Model has been built on training group (2/3 of the data) and will be validating the model on holdout group (1/3 of data) in this step. I will use cross validation method that uses a shrinkage measurement to assess reliability.

Shrinkage=R-Square (training Group)- R-Square (holdout).

We will compare the R2 between the train model and holdout model to see how the shrinkage changes. If the shrinkage < **0.10**, it means out model is reliable.

R-Square for the training group is **0.6005** and when I run the model on holdout group, I got the correlation coefficient(r) as **0.814** which makes R square as **0.66** so the Shrinkage=**0.6005-0.66=-.06 <.10** hence model is reliable.

**ALL THE PROCESS REPEATED ABOVE FOR FEMALE**

Descriptive Statistics has been done in the start of model for female as well.

**Model Building:**

The process is exactly same as male but here our target variable is Likef and other independent variables are Attractivef, sinceref , intelligentf , funf , ambitiousf , sharedinterestsf. We want to predict the likef i.e. rating of male given by female based on other variables like fun , attractive ,sincere,ambitious and shared intrests ratings.

Before adding the polynomial terms and Interaction terms, I checked if they are needed into our model.

**For Interaction**: I used the chunkwwise approach to check if any of the pairwise interaction is useful. I ran the model including all pairwise interactions and then without them and found that at .05 significance level none of them is useful.

H0: None of the pairwise is useful ; Ha:Atleast one of the pairwise interaction is useful.

Test statistics: F(15,162)=MSR/MSE where MSR=(394-365.4)/(21-6) and MSE=1.29 so F=1.47

p-value=0.121 which is greater than significance level . Thus we do not reject the null hypothesis and conclude that none of the pairwise interaction is useful.

**For Polynomial terms**: I checked the significance of the cubic terms above and beyond the quadratic and base variable and found that none of the cubic term is significant for any of the six variables.Therefore I only added the base terms into my model and the quadratic terms.Once this is done, I divided my data into training and Holdout Groups.

On the training group, I started with the variable selection process for model building i.e. All Possible Method, Forward, Backward and Stepwise. Results obtained are in Figure21:

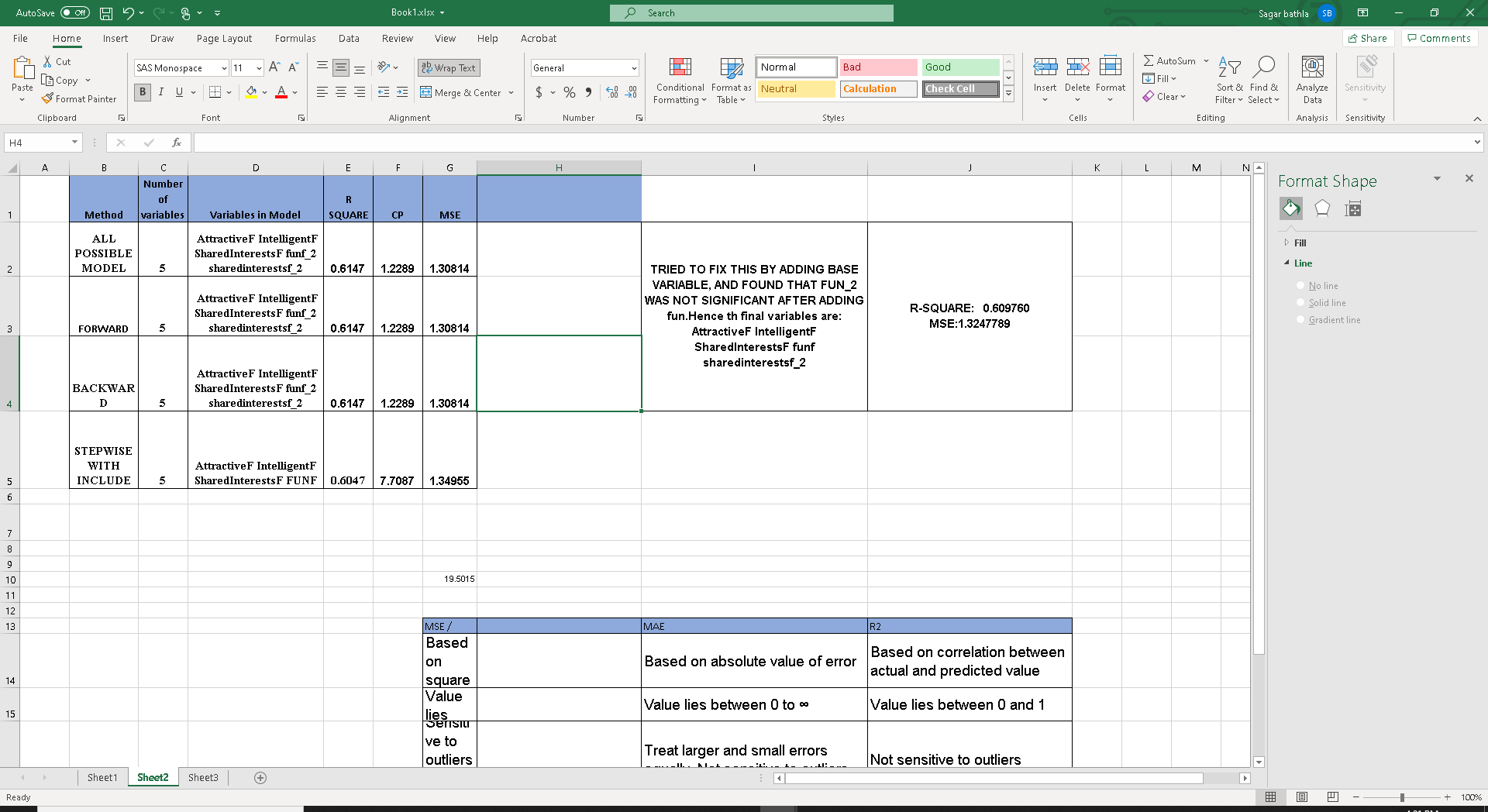


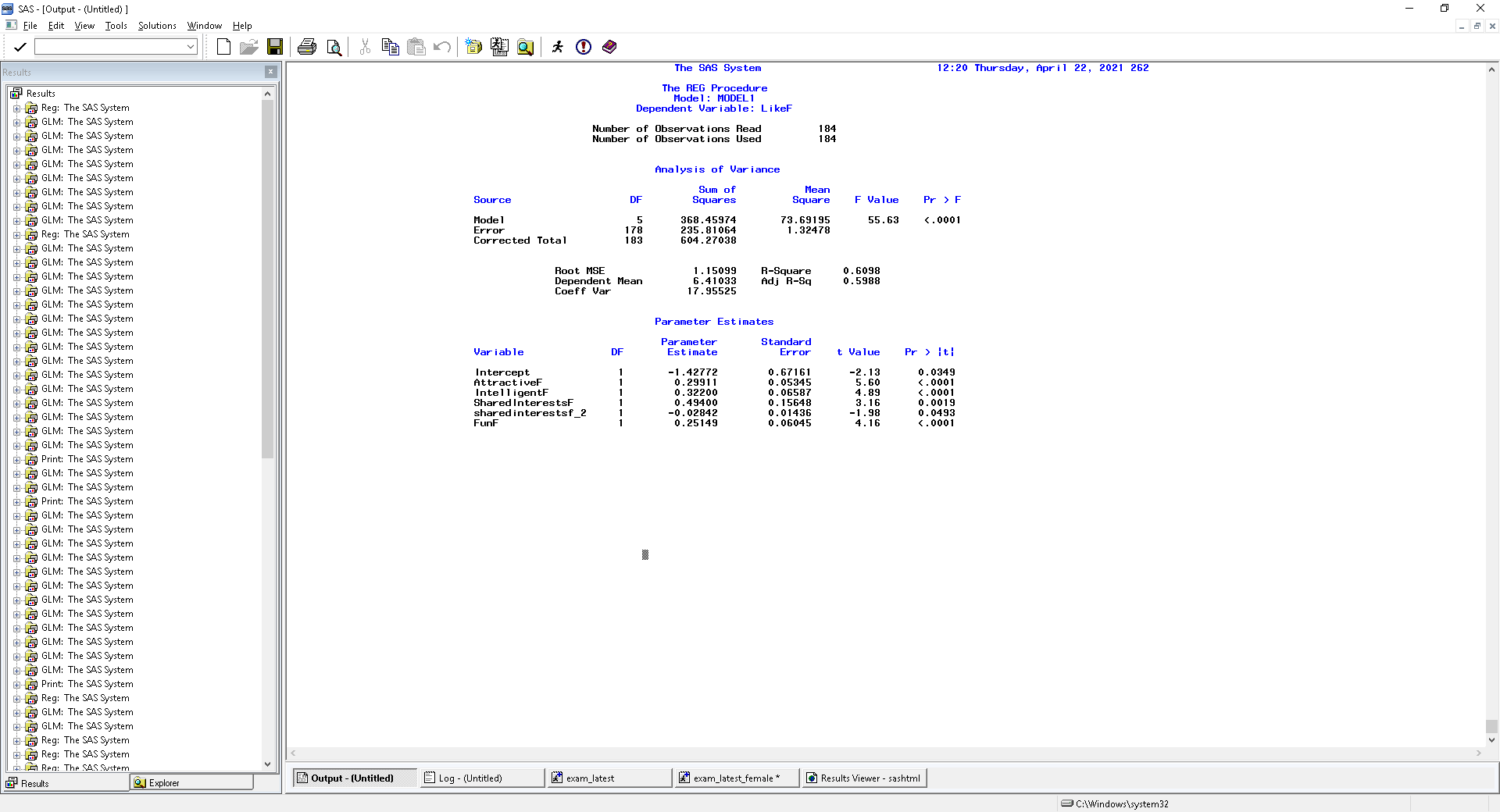
Figure21-Variable selection from the different approaches with the significance level of .05

Now, we can see that the first three approaches have quadratic term without the base term. Stepwise approach is ran with the include statement which tells SAS to include the base variables always and add any other if significant.No other variable was significant. After this I eliminated the variables that were forced by us(base terms) which were insignificant and hence left with Attractive, Fun, Shared Intrests, Intelligentf

I can see that first three has better R-Square and lower MSE then stepwise so I tried to fix that by including base variable.I added variable and funf along with AttractiveF IntelligentF SharedInterestsF funf\_2 sharedinterestsf\_2 and run the model. The results showed that fun\_2 has become insignificant after addition of funf.Thus, I removed that variable and now my final model has 5 variables : Attractivef intelligentf sharedinterestsf sharedinterestsf\_2 funf

R-Square is better than the Stepwise approach so I conclude that these are my final variables and my Model Equation is:

Y=-1.42772 +0.29911 \*attractivef +0.32200 \*intelligentf+0.49400 \* SharedInterestsF -0.02842 \* sharedinterestsf\_2 +0.25149 \* FunF



It means that with one unit increase in the attractive rating, we expect liking of female will increase by .2991 if all other variables are kept constant. Same goes for all other variable in the model.

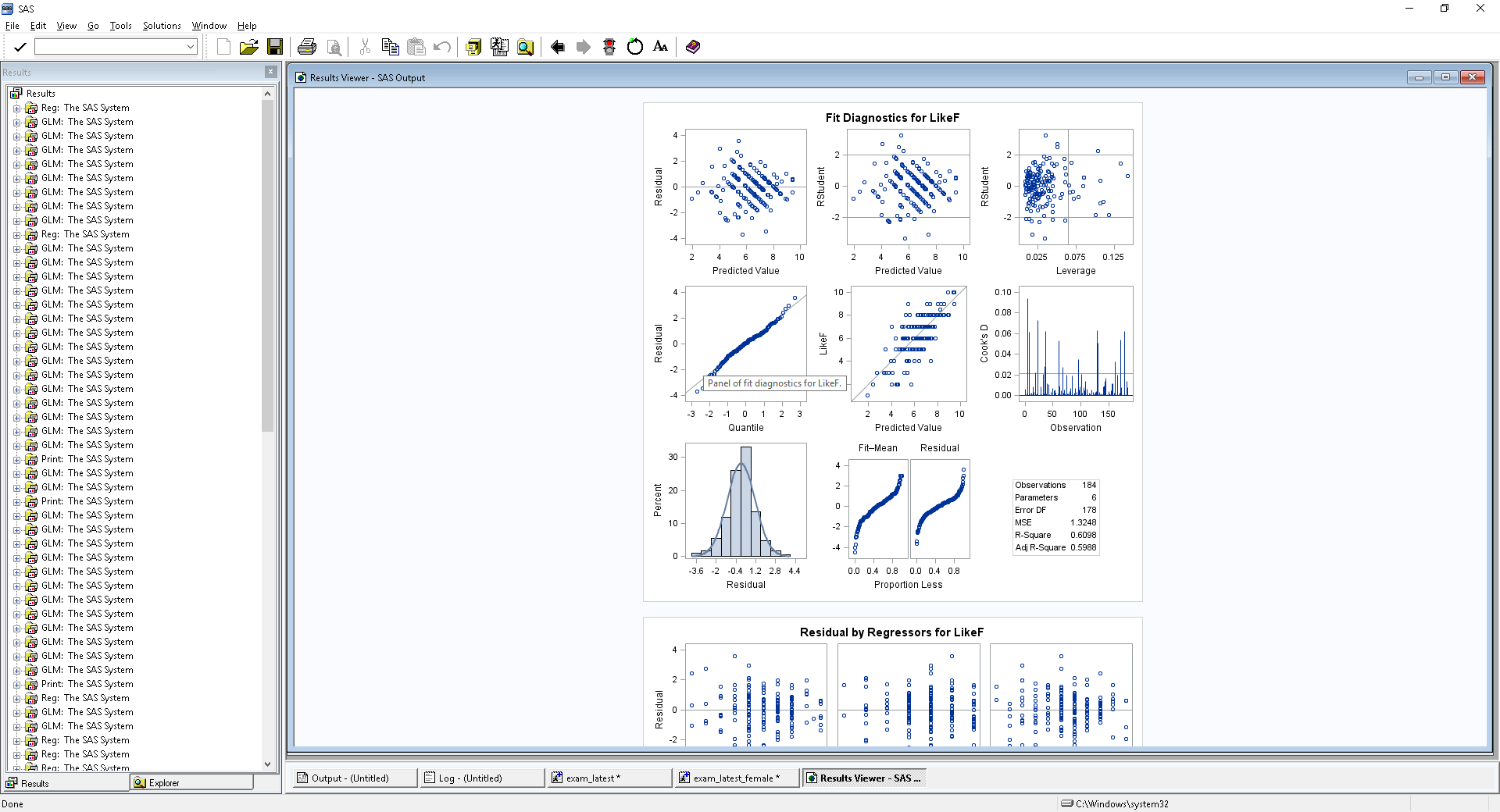
This model has quadratic term for SharedInterest and this was fitting model better than the linear fit.

**REGRESSION DIAGNOSTICS FOR THE FINAL MODEL**

1. **Checking the assumptions:**

Linearity, homoscedasticity, and independence:

To assess the assumption of linearity we want to ensure that the residuals are not too far away from 0 (standardized values less than -2 or greater than 2 are deemed problematic). To assess if the homoscedasticity assumption is met, we look to make sure that there is no pattern in the residuals and that they are equally spread around the y = 0 line.



Based on this plot, the scatter appears to be quite random with no evident funneling or sinuous pattern. Thus, assumptions of homoscedasticity, independence, and linearity are not notably violated.

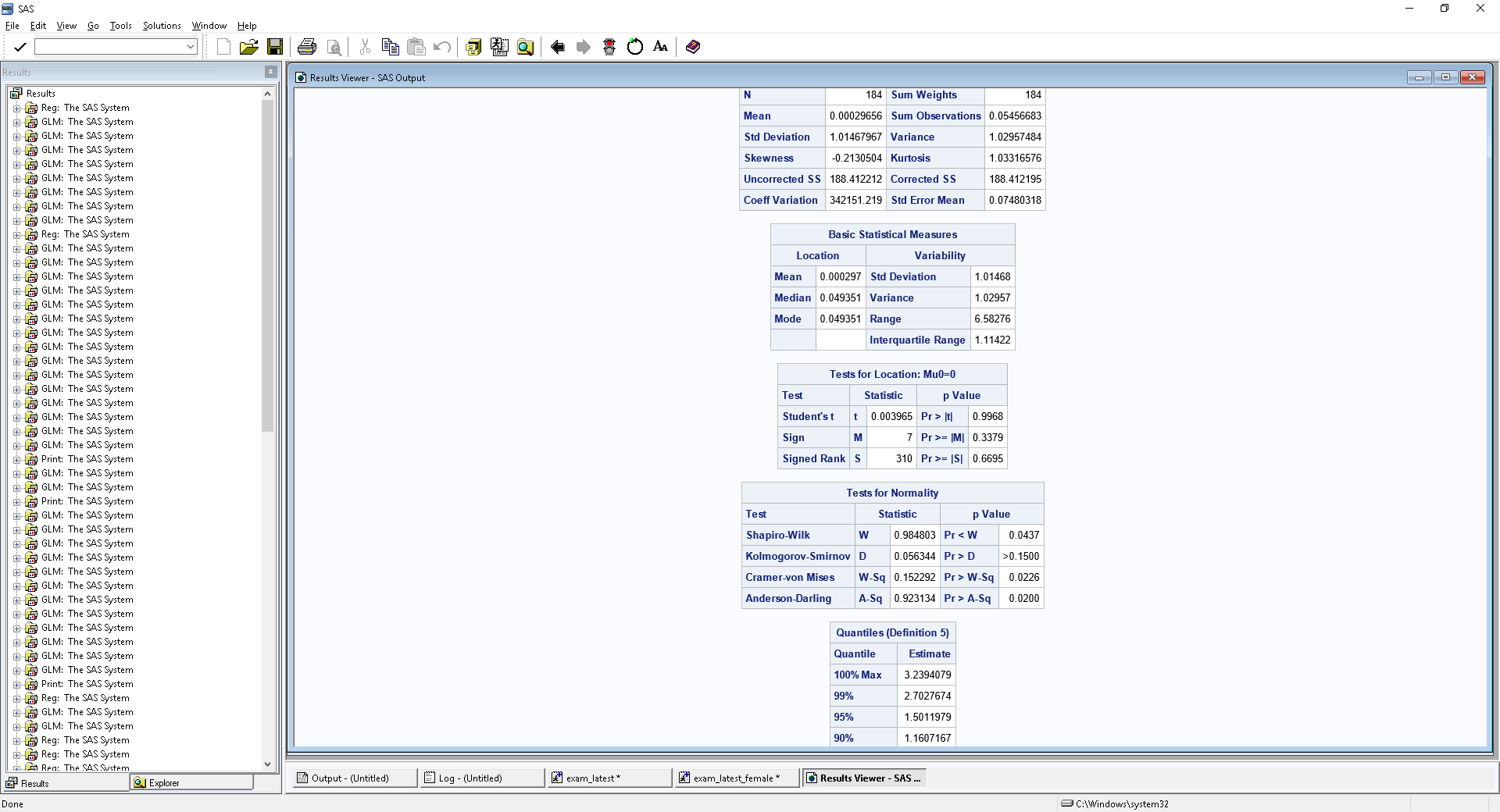
Normality assumption:

Here I used different methods to check the normality assumptions:

1. *Several graphs, plots, and tables can be created via PROC UNIVARIATE*. An investigation into basic descriptive statistics of Jackknife residuals yields a 25% Q1 value of -0.532, a median of 0.049, a 75% value Q3 of 0.5814, a mean of 0.0002, and a standard deviation of 1.01467967. Skewness is -0.2130504 and it measure the degree and direction of asymmetry. Looking at these numbers, it looks like normal because Mean is close to 0, variance is close to 1 and skewness is also close to zero.
2. *Goodness of fit test (Hypothesis Testing):*

H0: It follows the normal distribution

Ha: It do not follow Normal distribution



We will consider the result from Kolmogorov-Smirnov because our sample data is more than 50 otherwise, we would have considered Shapiro-Wilk test. Since the p value is greater than the significance level (.05) it means reject the null hypothesis i.e., the data follows normal distribution.

1. *Normal Quantile Plot (QQ Plot):*

This tells us that if the data is normal then it should be close to straight line. Observations lie well along the 45-degree line in the QQ-plot, so we may assume that normality holds here.

1. *Histogram:*

To confirm the normality of residuals I checked the Histogram, and it looks Normal.

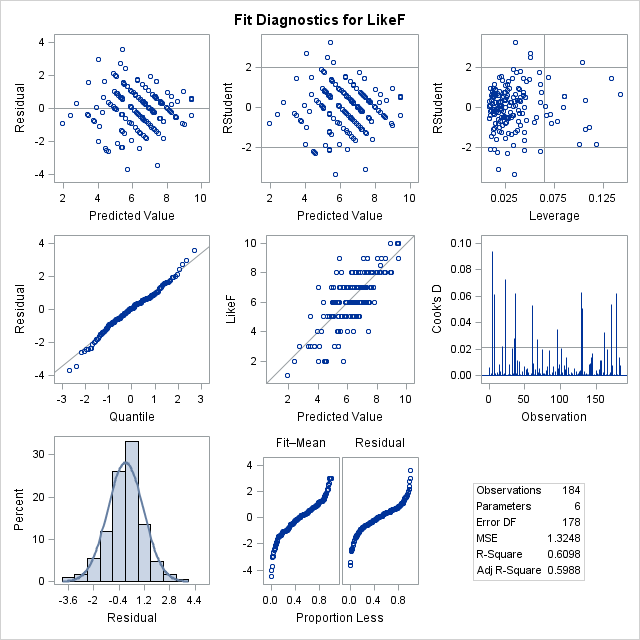
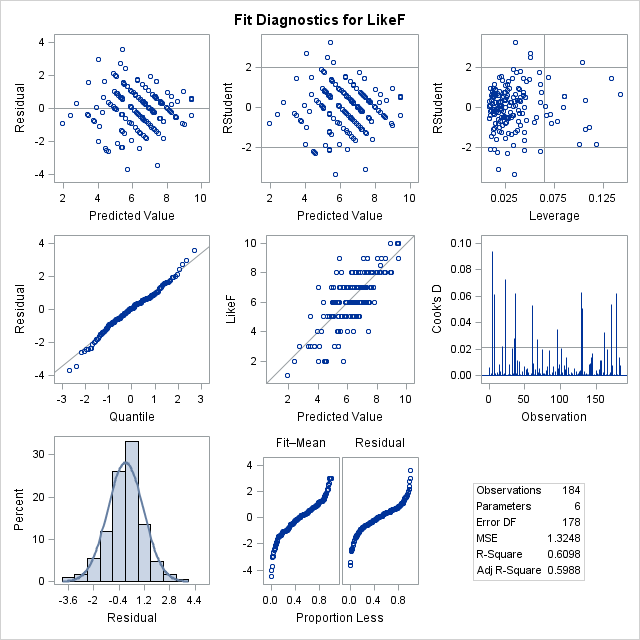


Figure22 QQ Plot Figure23 Histogram

Therefore, all the four assumptions of linear regression model are satisfied.

1. **OUTLIER DETECTION:**

Using the jackknife with absolute value greater than 2, Cook’s distance >1, and leverage >12/276=0.043, following are the outliers observed (Id refers to the observation number in the original dataset).

Figure24- Observations containing outliers identified via a Cook’s distance of greater than one, leverage values greater than 12/276, and Jackknife residuals with absolute values greater than 2. No observations were noted for suspicious di, but observation number255 has jackknife\_res= 3.239 with original like value=9 and predicted=5.4 seems to be suspicious.

1. **COLLINEARITY TEST**

Using VIF and condition Index, the results obtained are shown in Figure25

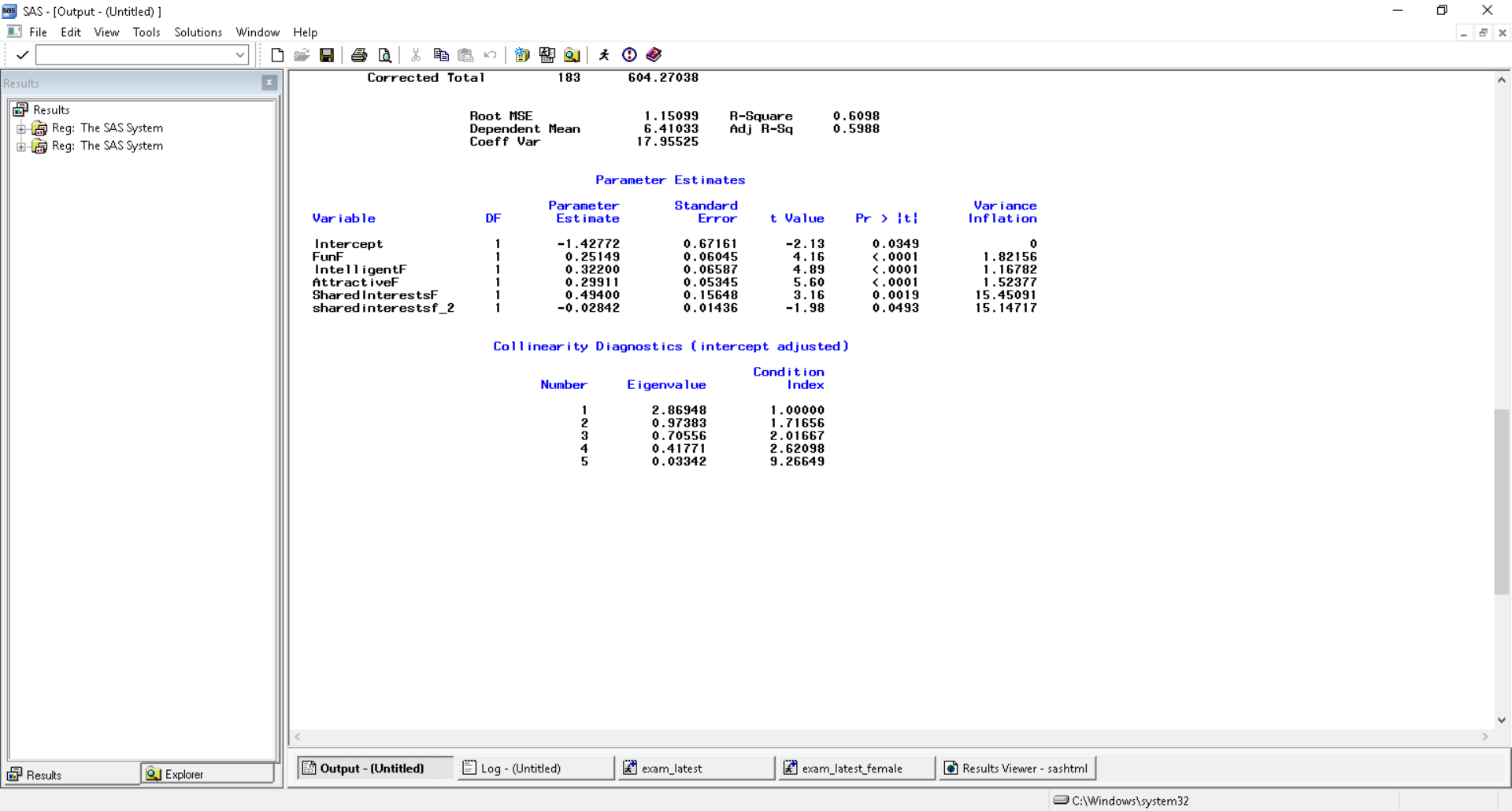
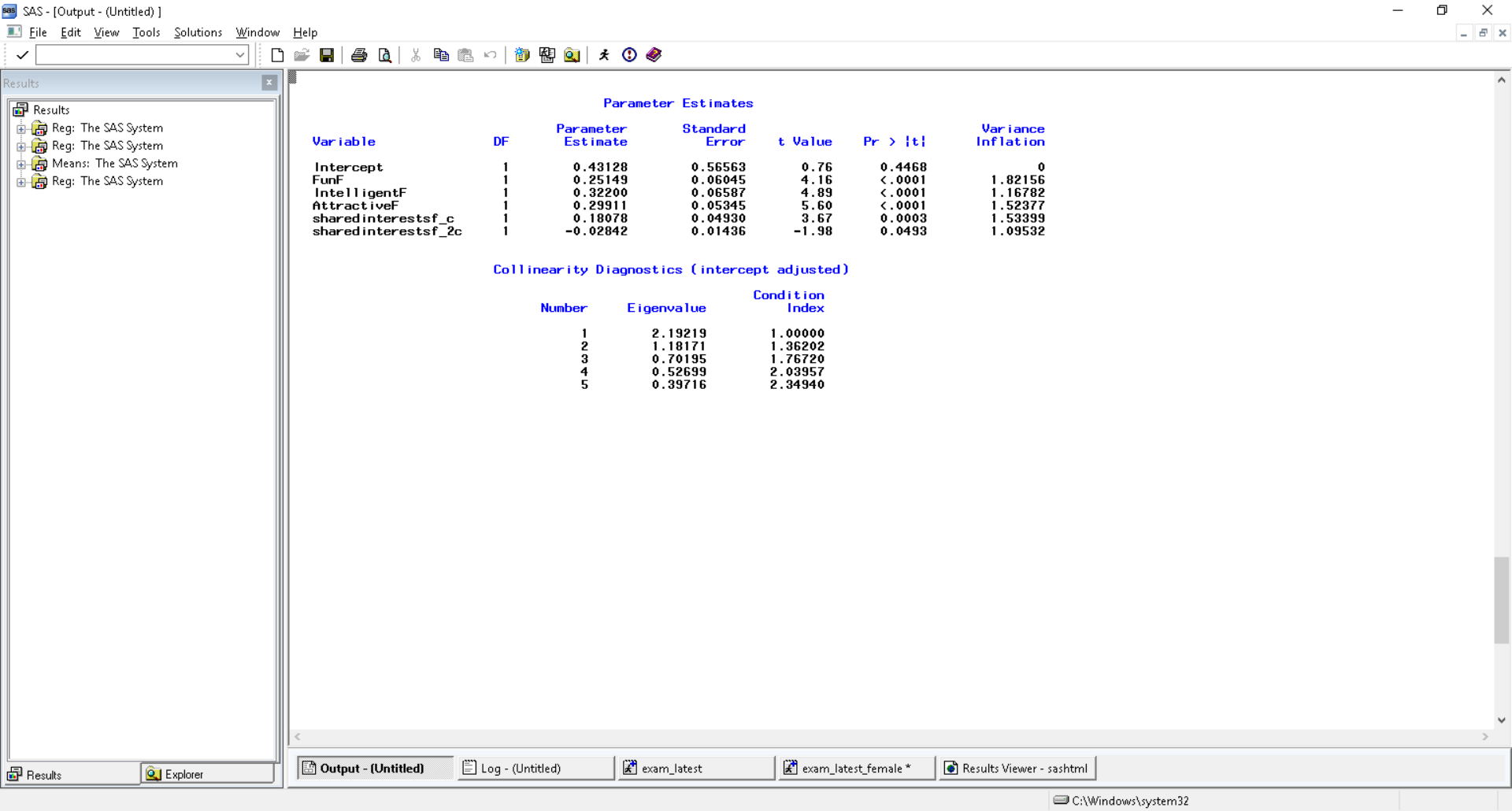
   
 Figure25 Detecting Collinearity Figure26 Resolving Collinearity by centering the variable

Figure 26 shows that after centering the variable shared\_interestsf, the problem has been resolved. Thus, Final Model is:

**ACCURACY OF MODEL:**

I have checked the closeness of the actual and predicted values using Mean Absolute error, Root MSE and R square as discussed above in the male model.

MAE for my model is 0.86

RMSE for my model is 1.15

This means that the average difference between the predicted and observed like is 0.86.

1. **Checking Reliability of Model**

R-Square for the training group is 0.6098 and when I run the model on holdout group, I got the correlation coefficient(r) as 0.78 which makes R square as 0.6084

Shrinkage=R-Square (training)- R-Square (holdout)

Shrinkage=0.6098-0.6084=.0014<.10 hence model is reliable.

**Summary of Findings:**

The analysis was conducted to predict the likelihood of a dater’s opinion of their partner based on different dimensions including attractiveness, sincerity, intelligence, fun, ambitiousness, and shared interests of their partner. The analysis was based on the speed dating event which recorded responses of 276 couples. The study was done separately for males and females based on their corresponding responses.

A linear regression model was built to create an equation between independent and dependent variables, and it is observed that key factors for a male like ratings are attractiveness, Sincere, fun and Shared Interests in the given order.

Model accuracy was measured using Mean Absolute Error, means that the average difference between the predicted and observed Like is 0.80. The model performance is stable when measured using R square on the development (0.60) and hold-out sample (0.66) with a shrinkage of 0.06.

Similarly, for females, the key factors are Intelligent Attractive Fun and Shared Interests With model accuracy of 0.86(MAE) and R-square value of development (.6098) and hold-out (.6084) with a shrinkage of .0014.

As observed, the different dimensions have different weighting for both males and females and hence, it is advised to use different equations for male and female, respectively.

We also noticed that there are 121 couples that belonged to same race and 111 were of close age group(+/-2yrs). To further check if the race or close age effects the like response, I created the dummy variables for both same race and close age (+/- 2yrs) and did hypothesis testing if race or close age have significant relationship with Like outcome and found that for both males and females, the P value is > 0.05 in all cases leading to a conclusion that there is no significant relationship between a partner of same race or close age with like variable. Therefore, it does not matter if the partner belongs to same race or close age.

Lastly, I also checked the integrity of responses, if there are any suspicious cases where ratings have been inflated and based on outlier detection analysis between actual and predicted values (>2 or based on cooks distance or leverage and like ratings with 9 or 10), found that observation number 157 looks suspicious for male model and 255 for female model. And hence, should be further investigated.

**APPENDIX:**

Here is the SAS Code attached for reference:

