

PREDICTION OF CHANCE OF GETTING ADMIT INTO UNIVERSITY USING ADMISSIONS DATASET

|  |
| --- |
|  |
|  |
| Group 19  Abhinav Saluja  Prarthna Vasudevamurthy  Shivank Aggarwal  Sruthi Bandarupalli  Sindhu Nagaraj |

# INTRODUCTION

Problem Statement:

The United States has attracted millions of international students to its universities and colleges because the USA offers so many choices and some of the best faculties in the world. With more than 3,000 universities and colleges in united states, the options are almost limitless. Yet, because the choices are so varied, deciding which university to attend is not an easy choice.

There’s also uncertainty about whether the student will get admitted to a university or not. The price of the admissions process often totals hundreds of dollars, accounting for the costs associated with standardized tests, test-prep resources and application fees. This Project will help students will to guess their capacities and to decide whether to apply for a master's degree in a particular University or not.

Objective:

To provide a quality model for the project beneficiaries in order for them to predict their chances of getting university admits by selecting the most accurate model to predict the probability of admission.

Monetary Benefit:

* Students Often have to spend huge amounts of money while applying to many universities not knowing where they would get accepted, also there are a plethora of University Consulting Advisors who charge a massive amount of money from Students to help them choose where to apply
* This project will help students to narrow down their options, aiding them in applying to schools when their chance of admission is high, ultimately reducing application costs.

Project Beneficiaries:

1. Students Applying to Graduate Schools and Universities
2. University Admission Committee
3. Private University Consulting Advisors helping students with Graduate school applications and admissions process

# DATA EXPLORATION

Dataset:

This dataset is created for prediction of Graduate Admissions from an Indian perspective. 500 applicants have been surveyed as potential students for UCLA. The university weighs certain aspects of a student's education to determine their acceptance.

The dataset is clean, has 500 Observations with no null records and contains 7 parameters considered important during the application for Masters Programs

* Independent Variables

1. GRE Scores (out of 340)
2. TOEFL Scores (out of 120)
3. University Rating (rated from 1 to 5)
4. Statement of Purpose (rated from 1 to 5)
5. Letter of Recommendation Strength (rated from 1 to 5)
6. Undergraduate GPA (out of 10)
7. Research Experience (either 0 or 1, 1 meaning student has Research Experience)

* Dependent Variable

Chance of Admit (ranging from 0 to 1)

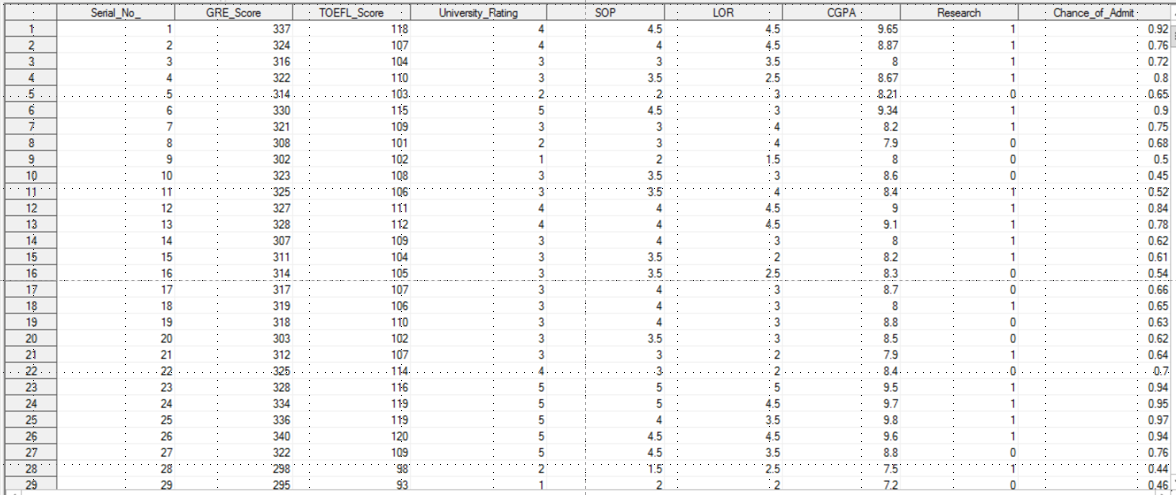


Fig1.1 Screen shot of the clean Data

Correlation Between all the variables:

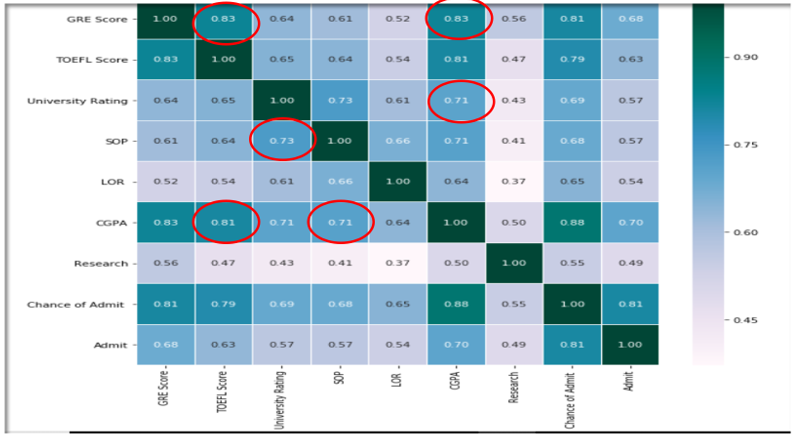


Fig 1.2 Correlation matrix produced in SAS for all variables in the Dataset

* The correlation between GRE Score and TOEFL Score is very high at 0.83
* The Correlation between CGPA with GRE Score and TOEFL Score is also quite high at 0.83 and 0.81 respectively
* We observe that most of the variables in the dataset are highly corelated with each other. This is expected as the dataset is education related dataset and we can assume that all the parameters are correlated to how good a student is academically.
* We use Factorial analysis to understand the variability among these highly correlated variables.

Factor Analysis:

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. For example, it is possible that variations in six observed variables mainly reflect the variations in two unobserved (underlying) variables. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modelled as linear combinations of the potential factors, plus "error" terms. Factor analysis aims to find independent latent variable. The theory behind factor analytic methods is that the information gained about the interdependencies between observed variables can be used later to reduce the set of variables in a dataset.

An important feature of factor analysis is that the axes of the factors can be rotated within the multidimensional variable space. In simple terms the factor analysis program looks first for the strongest correlations between variables and the latent factor, and makes this Factor 1. Visually, one can think of it as axis (Axis 1) then the factor analysis program looks for the second set of correlations and calls it Factor 2, and so on.

Another option/ Criterion for determining the number of factors is the Scree Plot. The Cattell scree test plots the components as the X-axis and the corresponding eigenvalues as the Y-axis. As one moves to the right, toward later components, the eigenvalues drop. When the drop ceases and the curve make an elbow toward less steep decline, Cattell's scree test says to drop all further components after the one starting at the elbow.

The factoring in our analysis is as follows:

Factor 1:

* CGPA
* University Rating
* LOR
* SOP

Factor 2:

* GRE
* TOEFL
* Research

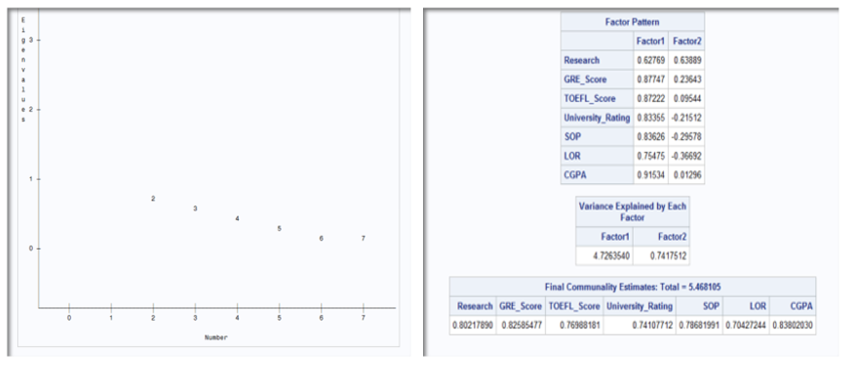


Fig 1.3: 2 factors will be retained by the NFACTOR criterion

Rotation Methods:

The unrotated output maximizes variance accounted for by the first and subsequent factors, and forces the factors to be orthogonal. This data-compression comes at the cost of having most items load on the early factors, and usually, of having many items load substantially on more than one factor. Rotation serves to make the output more understandable, by seeking so-called "Simple Structure": A pattern of loadings where each item loads strongly on only one of the factors, and much weakly on the other factors. Rotations can be orthogonal or oblique (allowing the factors to correlate).

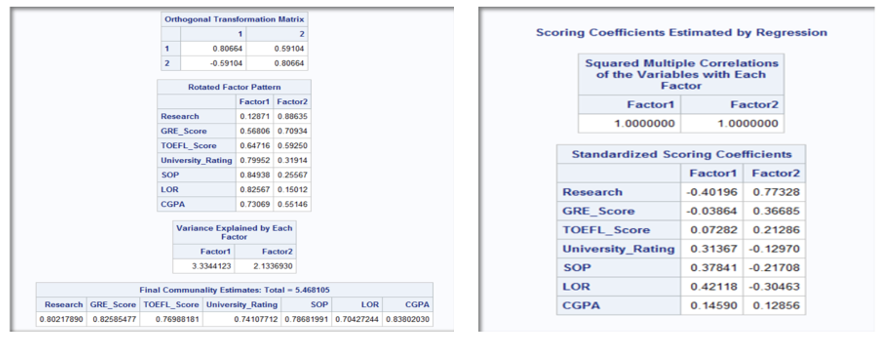
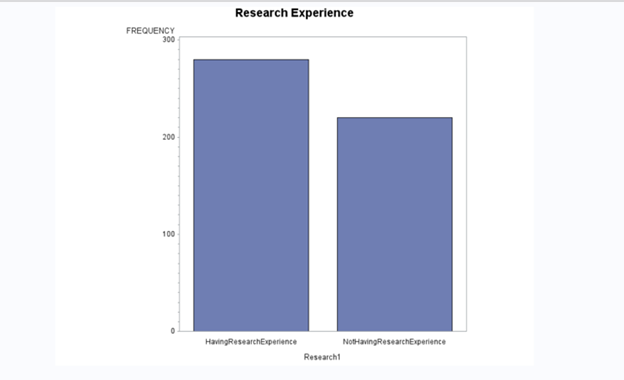


Fig 1.4 Rotation Method using SAS

* The initial solution results in strong correlations of a variable with several factors or in a variable that has no strong correlations with any of the factors.
* In order to make the location of the axes fit the actual data points better, the program rotates the axes. Ideally, the rotation will make the factors more easily interpretable.

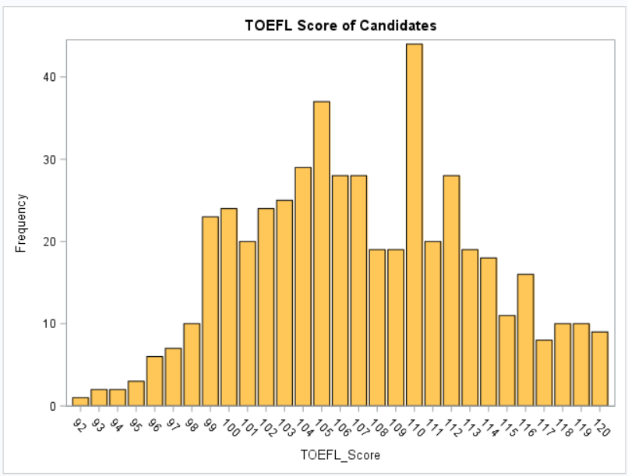
# Data Visualization

Number of Students involved in research



The graph clearly indicates that a greater number of students have been involved or have research experience as compared to students with no prior research experience.

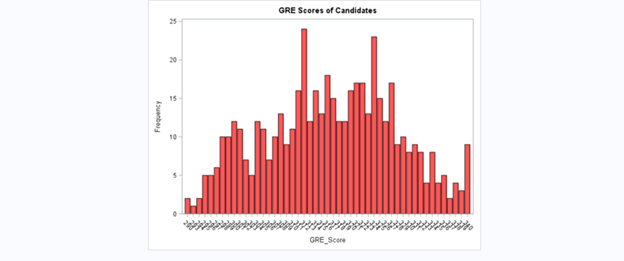
Plot for TOEFL Scores



This histogram shows the frequency for TOEFL Scores.

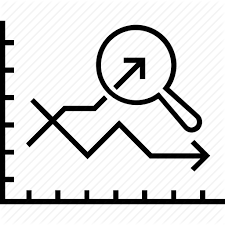
There is a density between 92 and 120. The minimum and maximum scores among university students are reported to be 92 and 120.

Histogram – GRE Scores

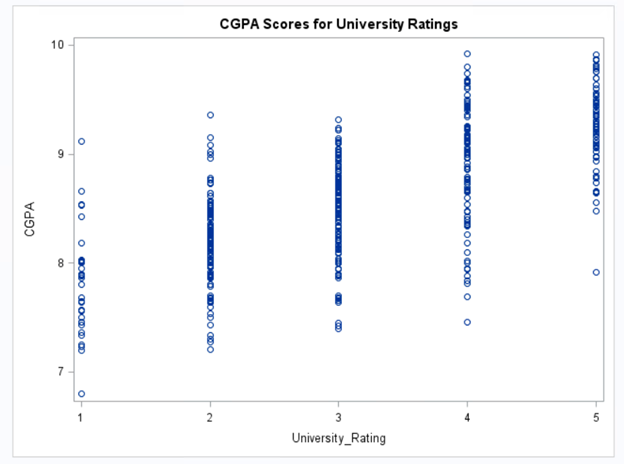


* This histogram shows the frequency for GRE scores.
* There is a density between 290 and 340. Being within this range would be a good feature for a candidate to stand out.
* How many students have good chances for getting an admit?

Let’s have a look at relations between the variables

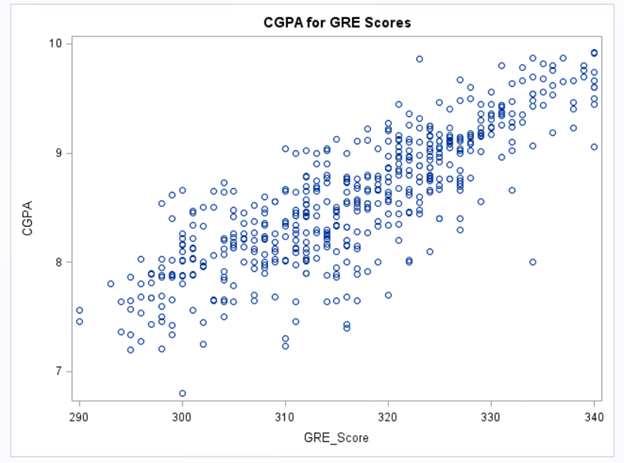


CGPA v/s University Ratings



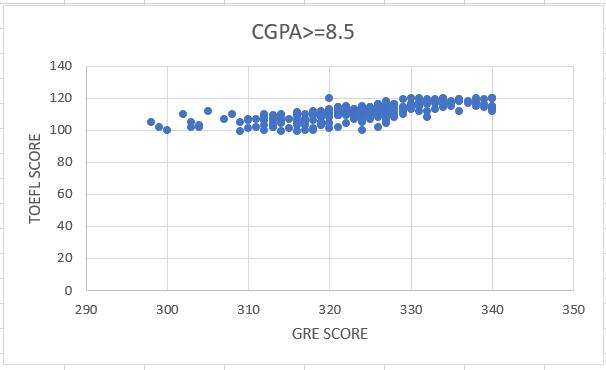
Clearly, high CGPA is associated with higher University Ratings. This is a clear indication that a higher CGPA in your undergraduate college can land you to a good university with a high ranking.

CGPA v/s GRE Scores



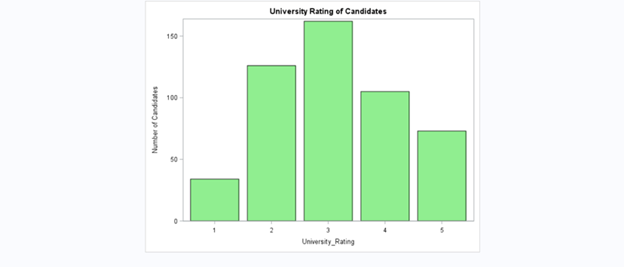
Students who have high CGPA have a tendency to score high score in GRE. We can associate it with high Intelligence Quotient (IQ). Students with high IQ have a capability to score well in exams, be it the university exams or other competitive exams like GRE.

TOEFL v/s GRE (CGPA>=8.5)



University students having high score have a tendency to score high in TOEFL given that their CGPA is higher than 8.5.

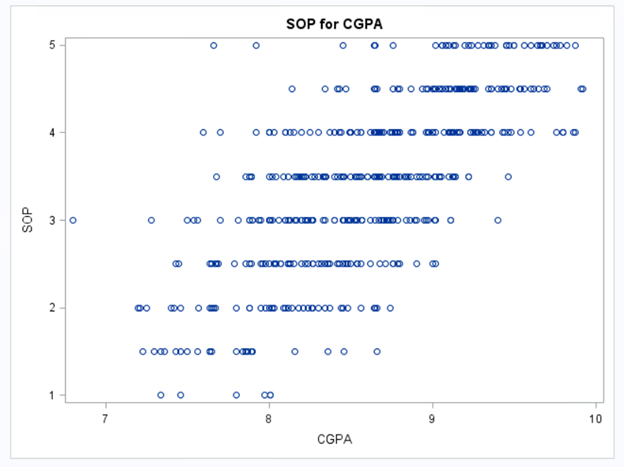
University Rating v/s Number of Students



Most of the students from our data set are from the Average Universities with Rating 3.

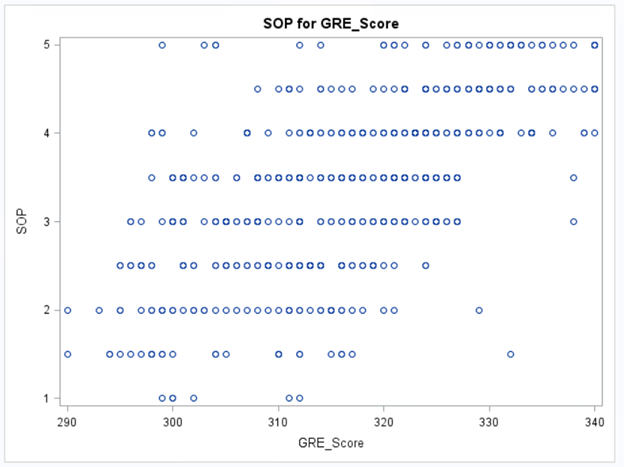
The data set has less number of students from low ranking universities.

SOP v/s CGPA



Candidates with high CGPA usually have high SOP scores.

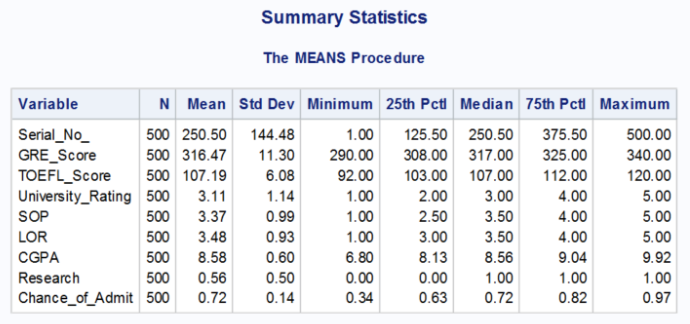
GRE Score v/s SOP Score



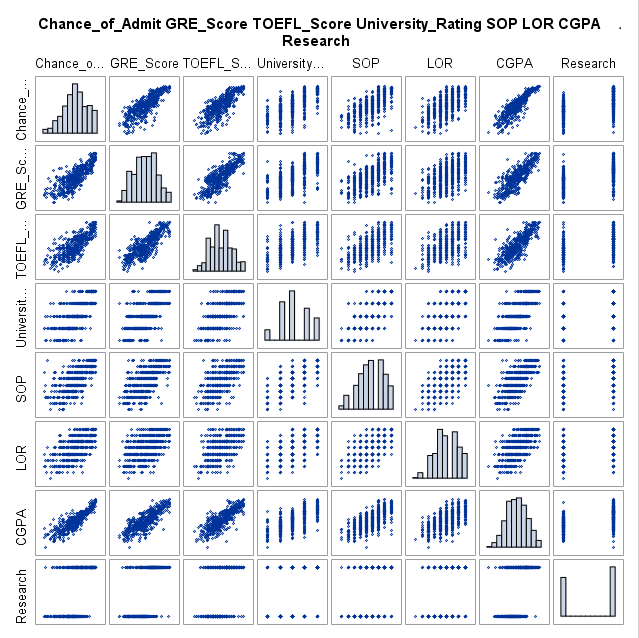
Candidates with high GRE scores are expected of getting high SOP scores too.

Model Selection

We did summary statistics in order to know the statistical properties of the data we have, which we can use for inferring further predictions.



We have also plotted our dependent variable Chance\_of\_Admit against all the other explanatory variables in order to understand the relationship and the trends between them.



This shows us that our dependent variable has positive relationship with all the other independent variables.

Modelling

Simple Linear Regression

* Linear regression is a linear approach to modelling the relationship between a dependent variable and one or more explanatory variables. Since we have more than one explanatory variables, the approach is called as multiple linear regression. Here the relations are modeled using linear predictor functions whose unknown model parameters are estimated from the data. This is more popular because the models which depend linearly on their unknown parameters are easy to fit than models which are non-linearly related to their parameters. Our goal is to explain the variation in dependent variable to variation in the explanatory variables. We can also determine the strength of the relationship between the response and explanatory variables and can identify the subsets of independent variables which contains redundant information about the response.
* Our model is as below:

***Chance\_of \_Admit= GRE\_Score +TOEFL\_Score +University\_Rating+SOP+LOR+CGPA+Research***

Results:



Insights:

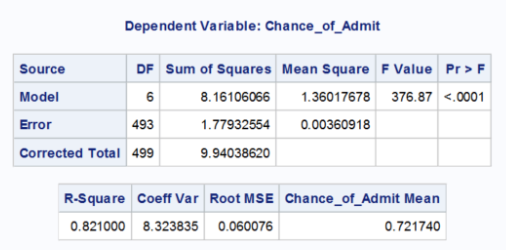
* We got a relatively high R square of 0.82
* An increase in GRE score by 1 unit raises chances of admit by 0.2%.
* Increase in TOEFL by 1 unit raises the chances of admit by 0.28%.
* University ratings and SOP are turned out to be insignificant.

Due to high level of correlation we added interactive term of GRE score and TOEFL score and ran the regression.

*Chance\_of \_Admit= GRE\_Score \* TOEFL\_Score +University\_Rating+SOP+LOR*

*+CGPA+Research.*

Results:

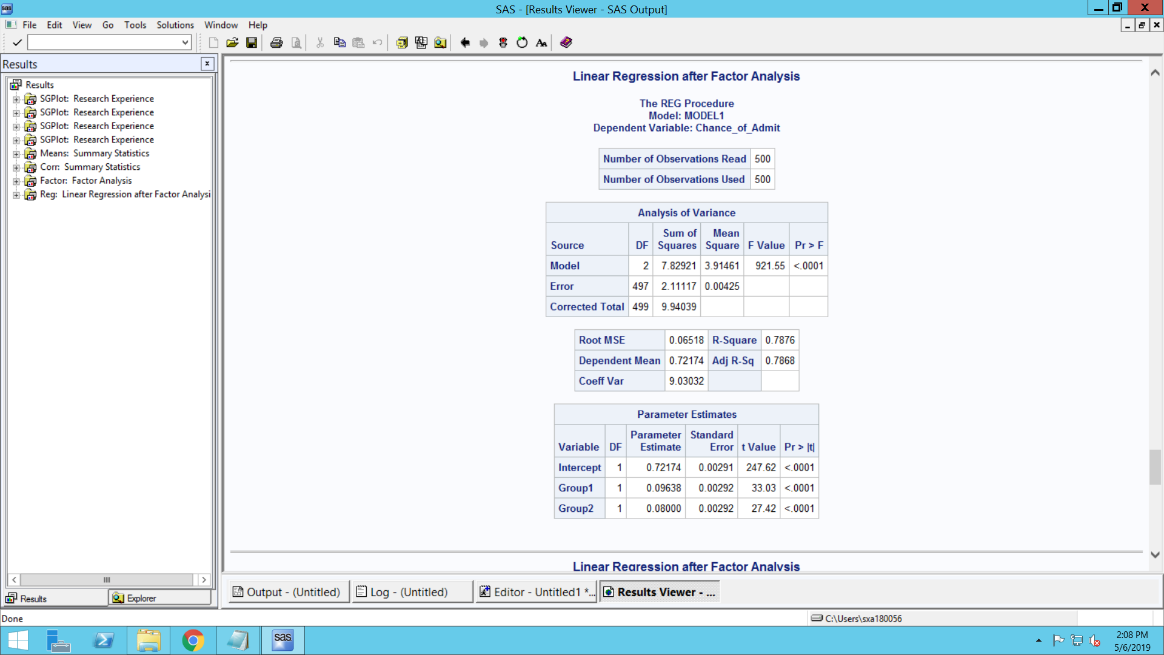
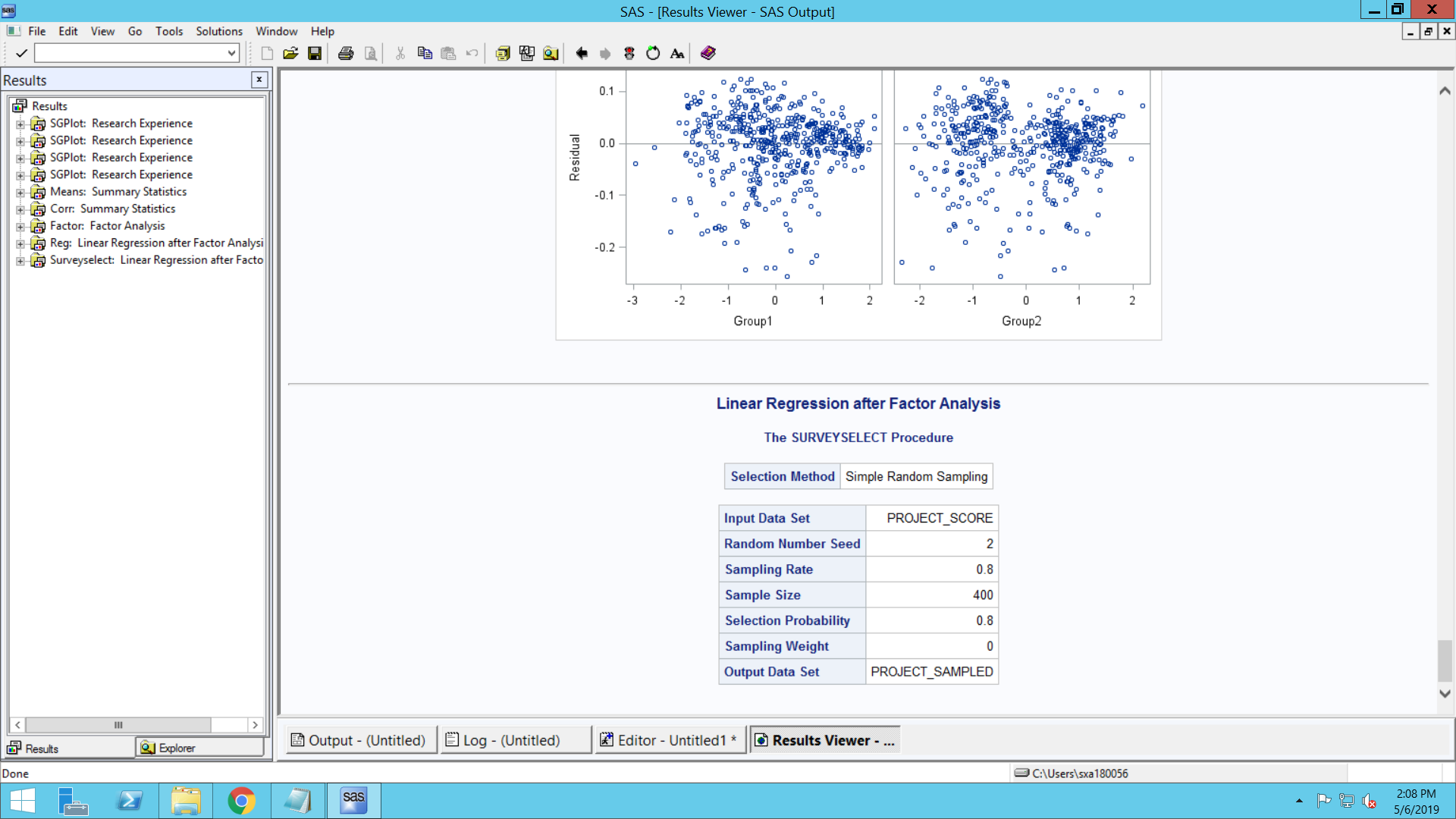




Insights: We see that there is no significant increase in R square at 0.82 from the previous regression.

Linear Regression Using Factor Analysis:

Since the data is highly correlated, we considered doing the factor analysis on the data and suing those factors as our model predictors.

Results:  


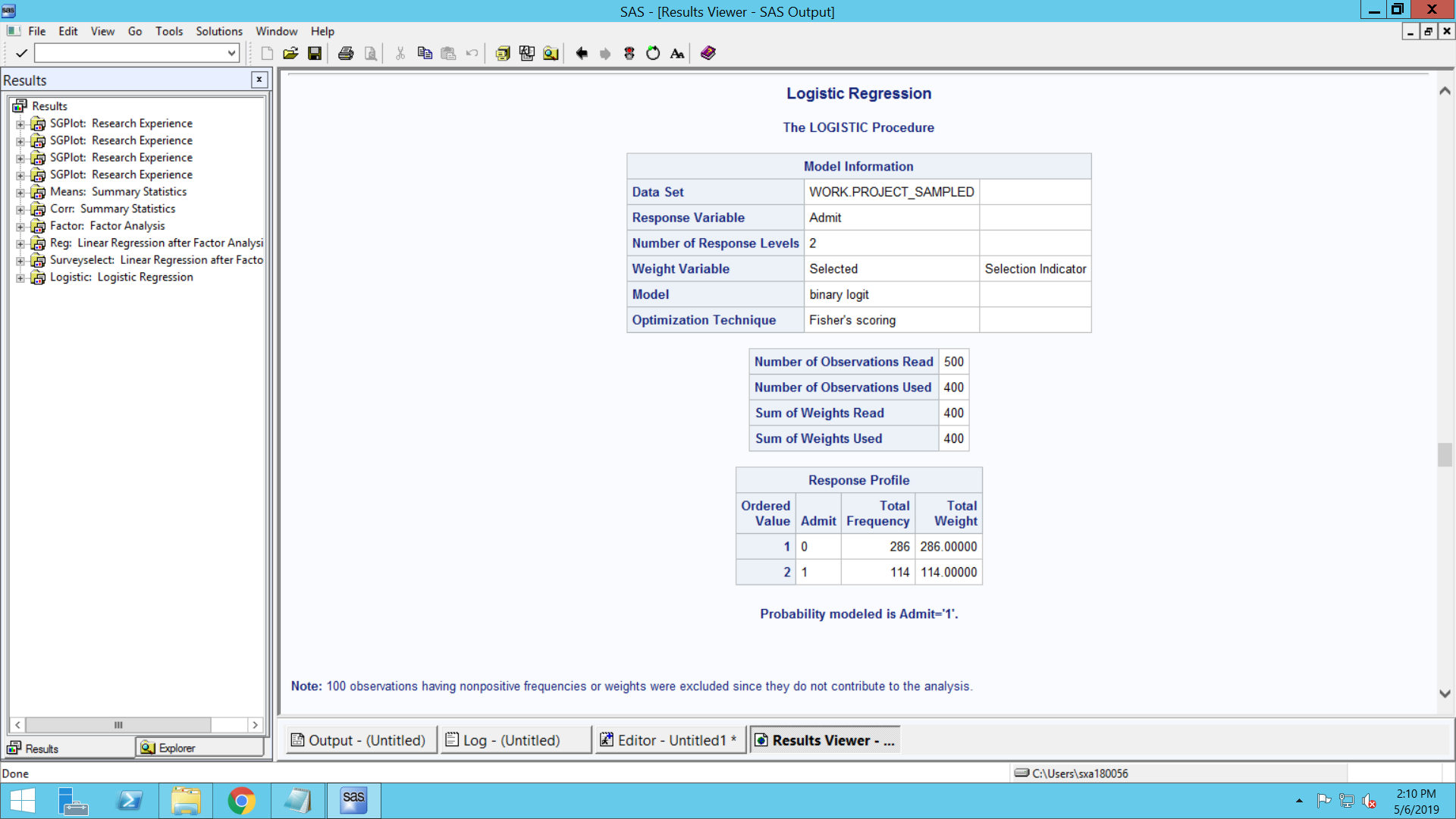
Insights:

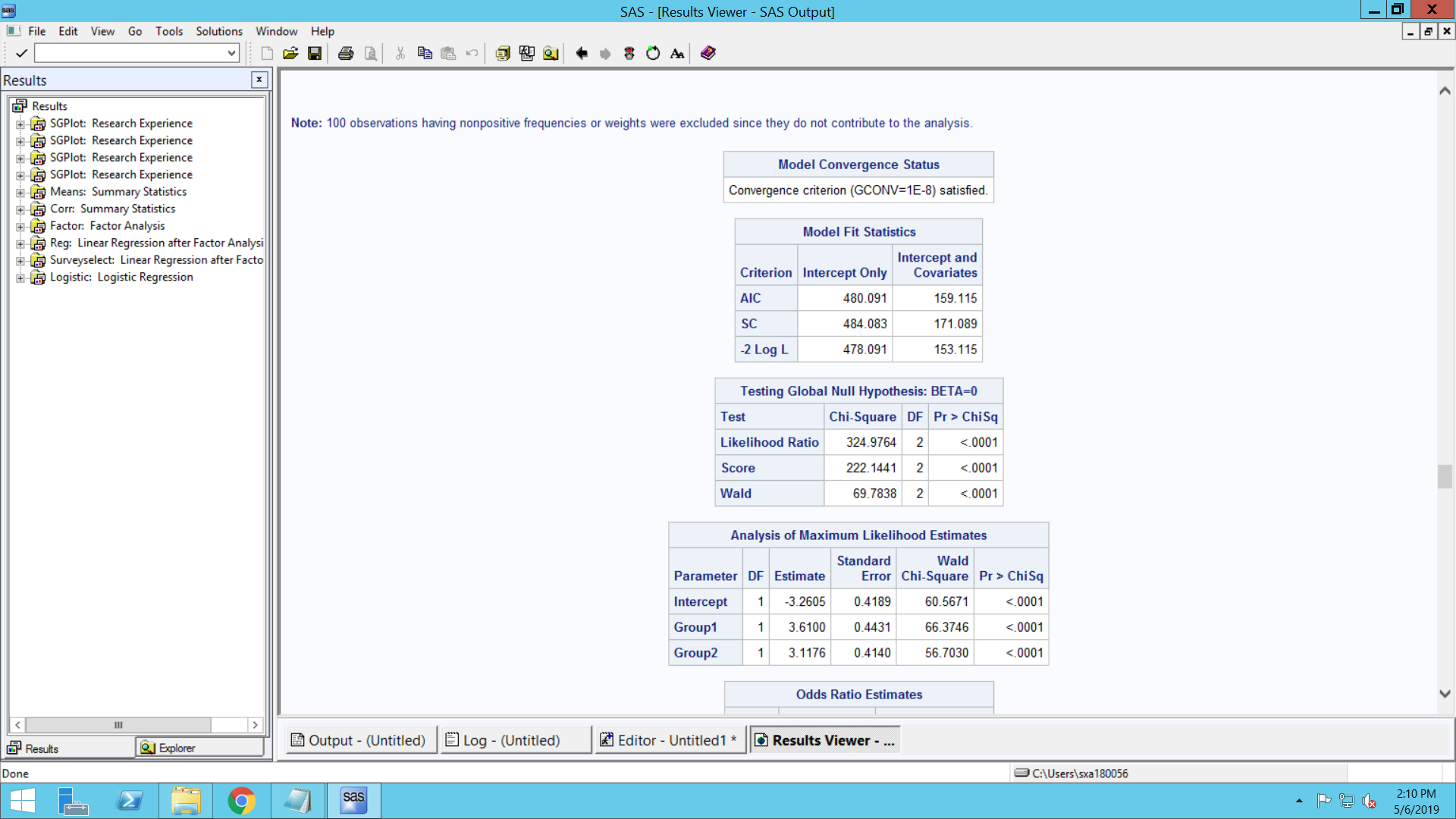
* The model explains 78.76% of variability of the response data around its mean.

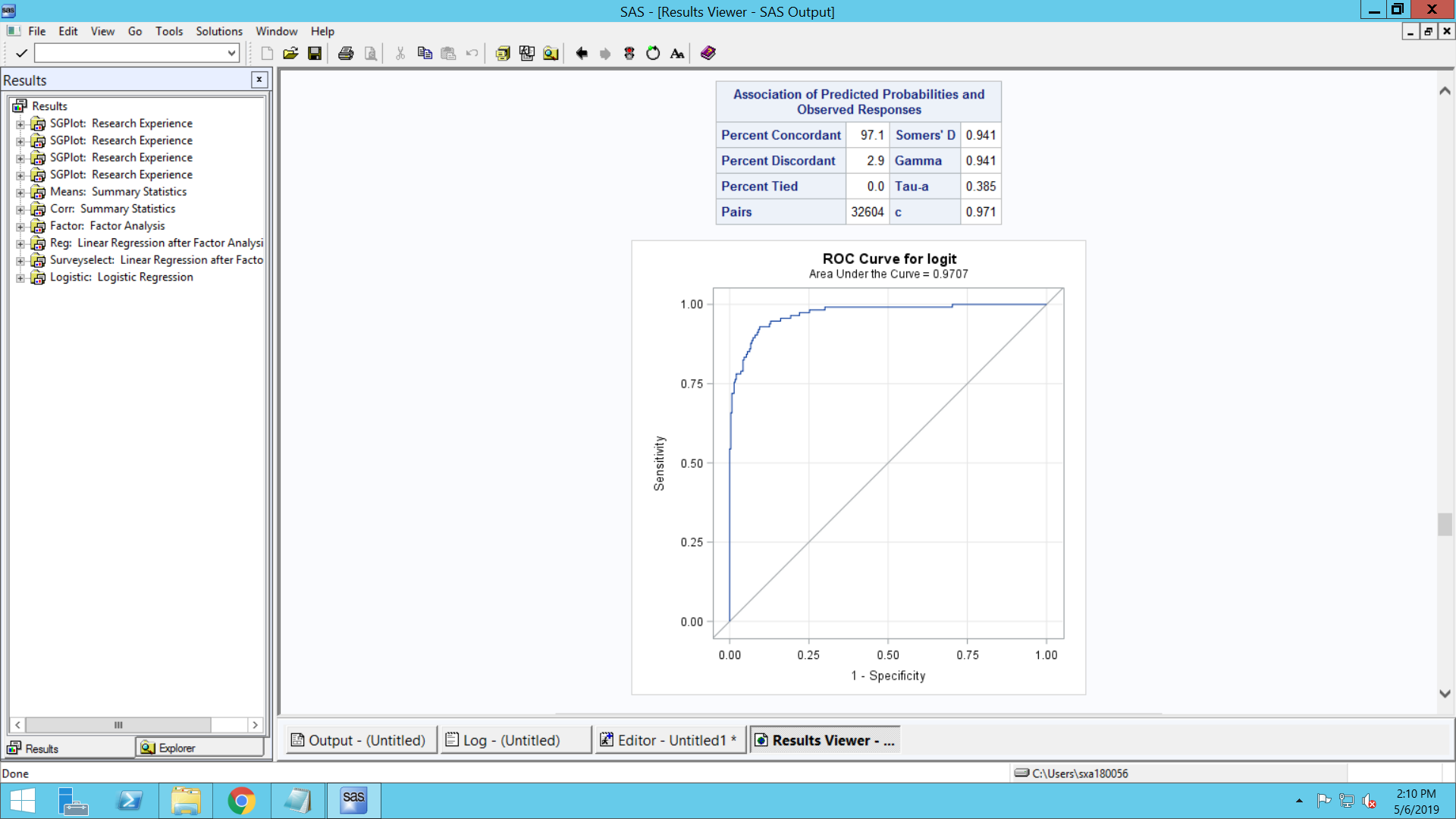
Logit Model Using Factor Analysis:

* Logistic regression estimates the parameters of a logistic model in a form of binary regression when the dependent variable has two level and as multinomial regression when Y has more than 2 levels.
* Since our dependent variable is chance of admit which is a probability, we categorized our dependent variable based on the probability.
* We took, if the chance of admit>0.8, then he/she is admitted. So we divided on scale of 08 & 0.2. we have also tried with 0.9 and 0.1 but the above model gave us more accurate results and hence we considered that model.

Results:







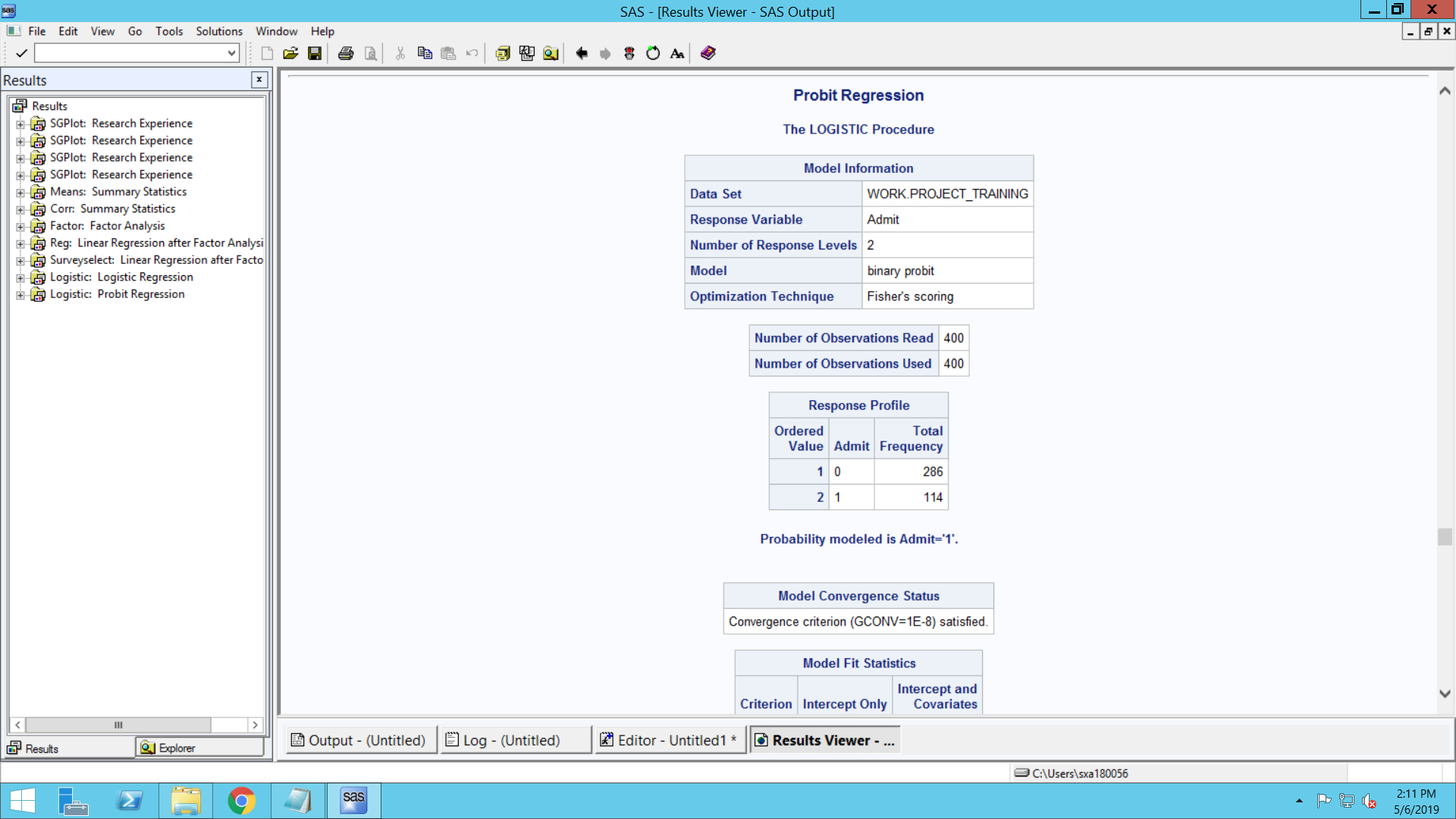
Insights:

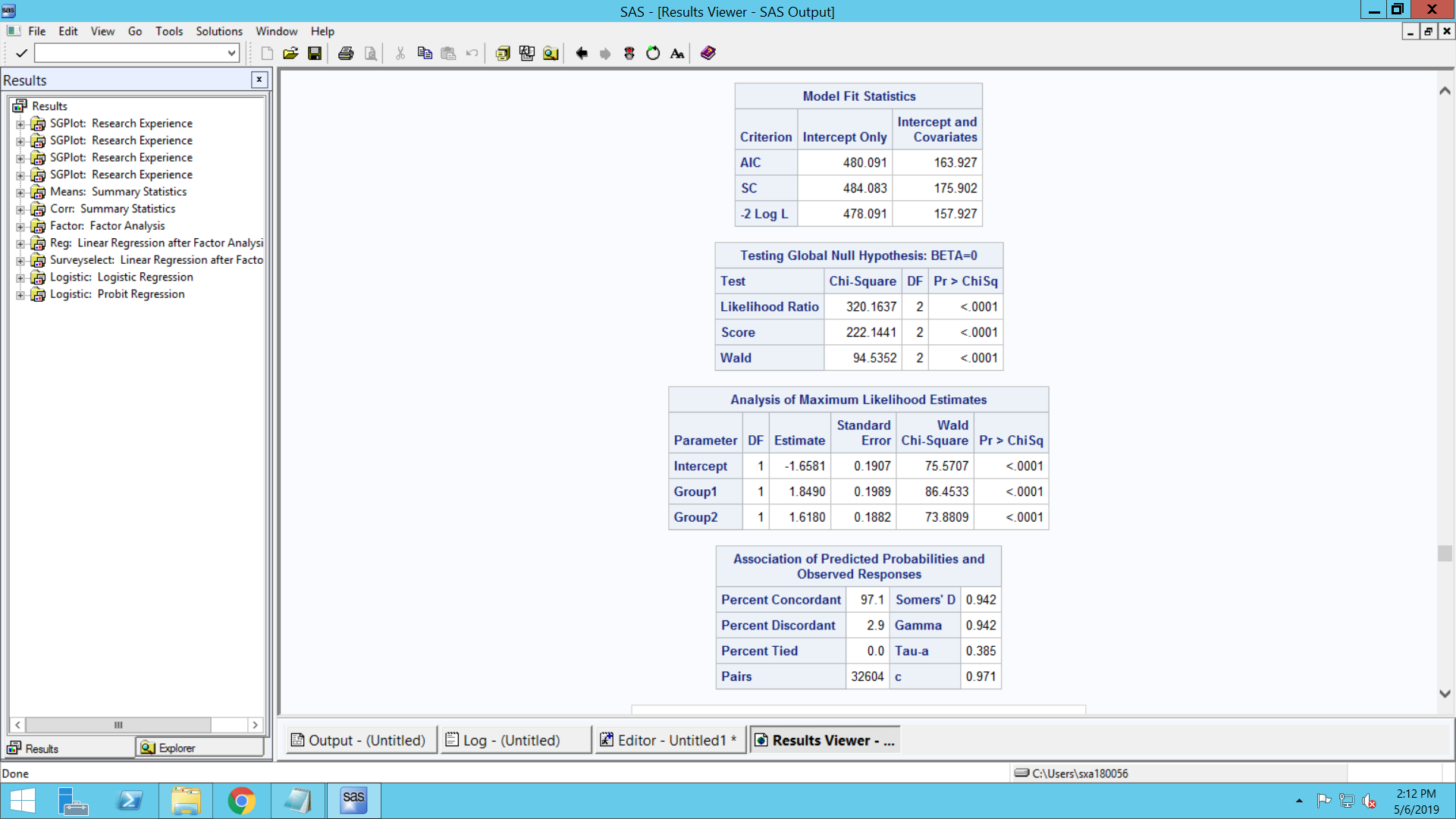
* The AIC we got is 159.115 and likelihood ratio is 324.97.
* We can also see that both group1 and group 2 are statistically significant.
* The Area under curve for logit is 0.9707.

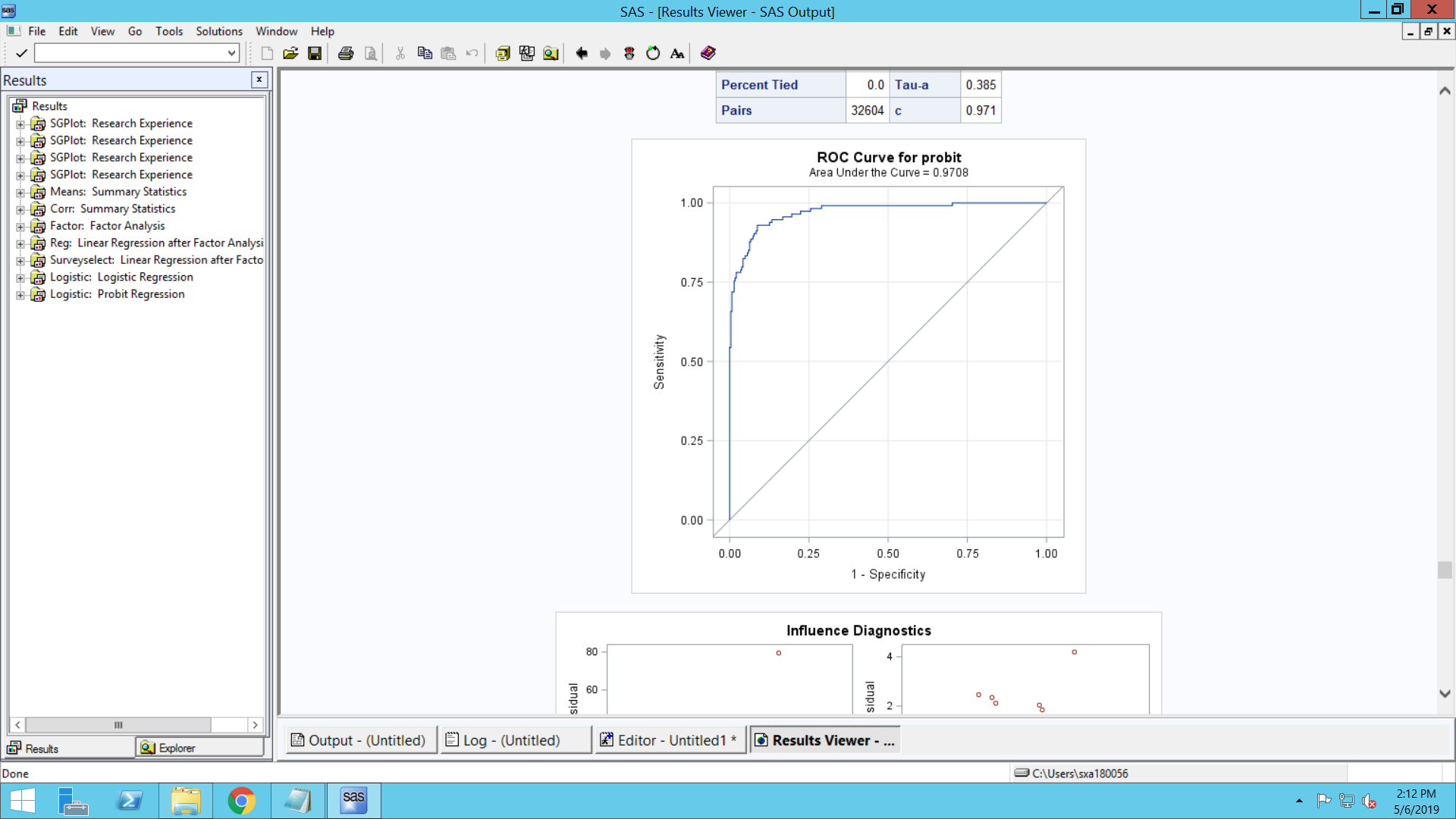
Probit Model Using Factor Analysis

* Probit model is same as logit model but only the difference is in distribution.
* Probit follows normal distribution unlike as for logit model.
* The results will be mostly similar from Probit and Logit.
* We are trying to chose the best model based on the Area under curve for logit and probit.

Results:







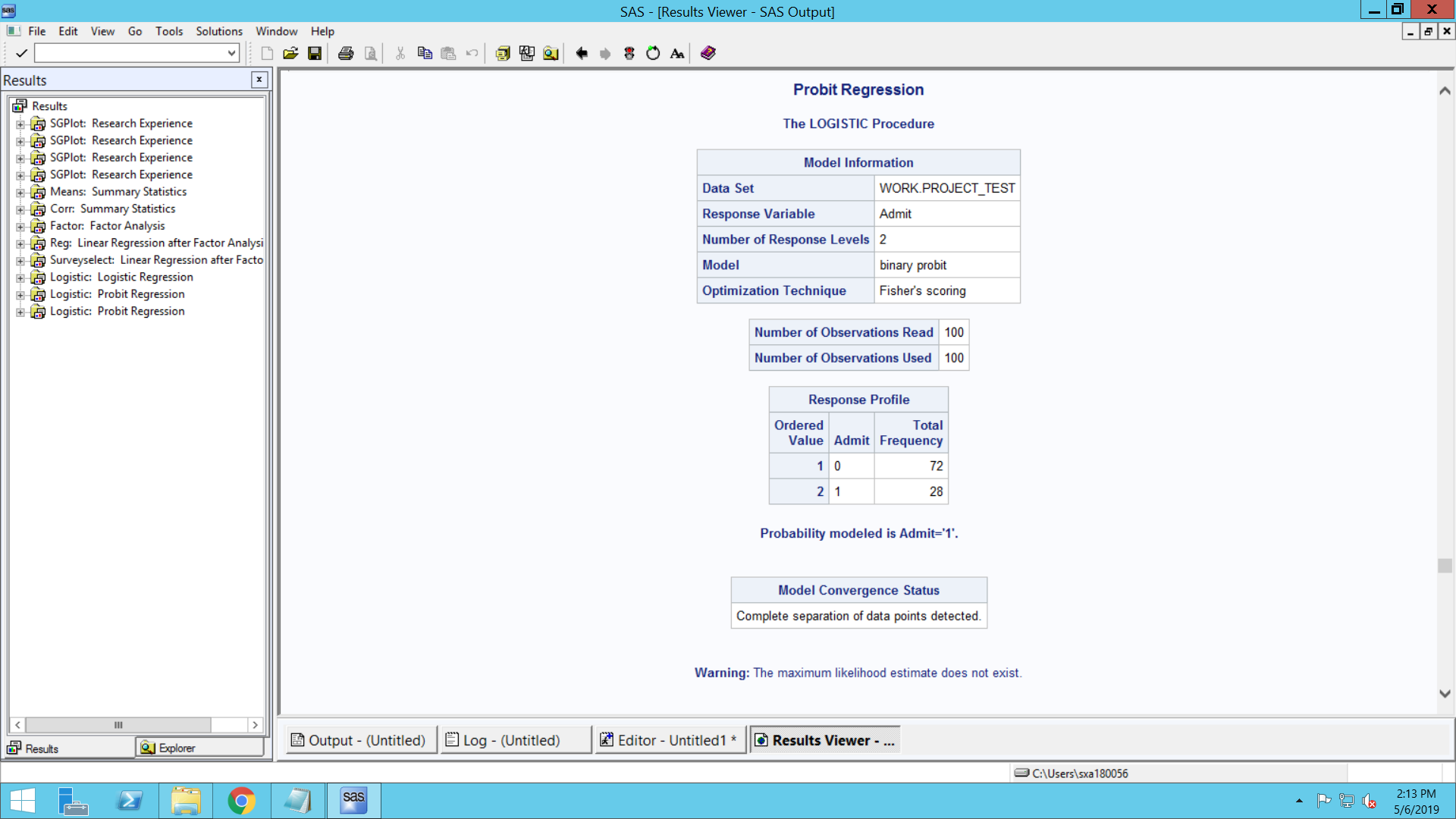
Insights:

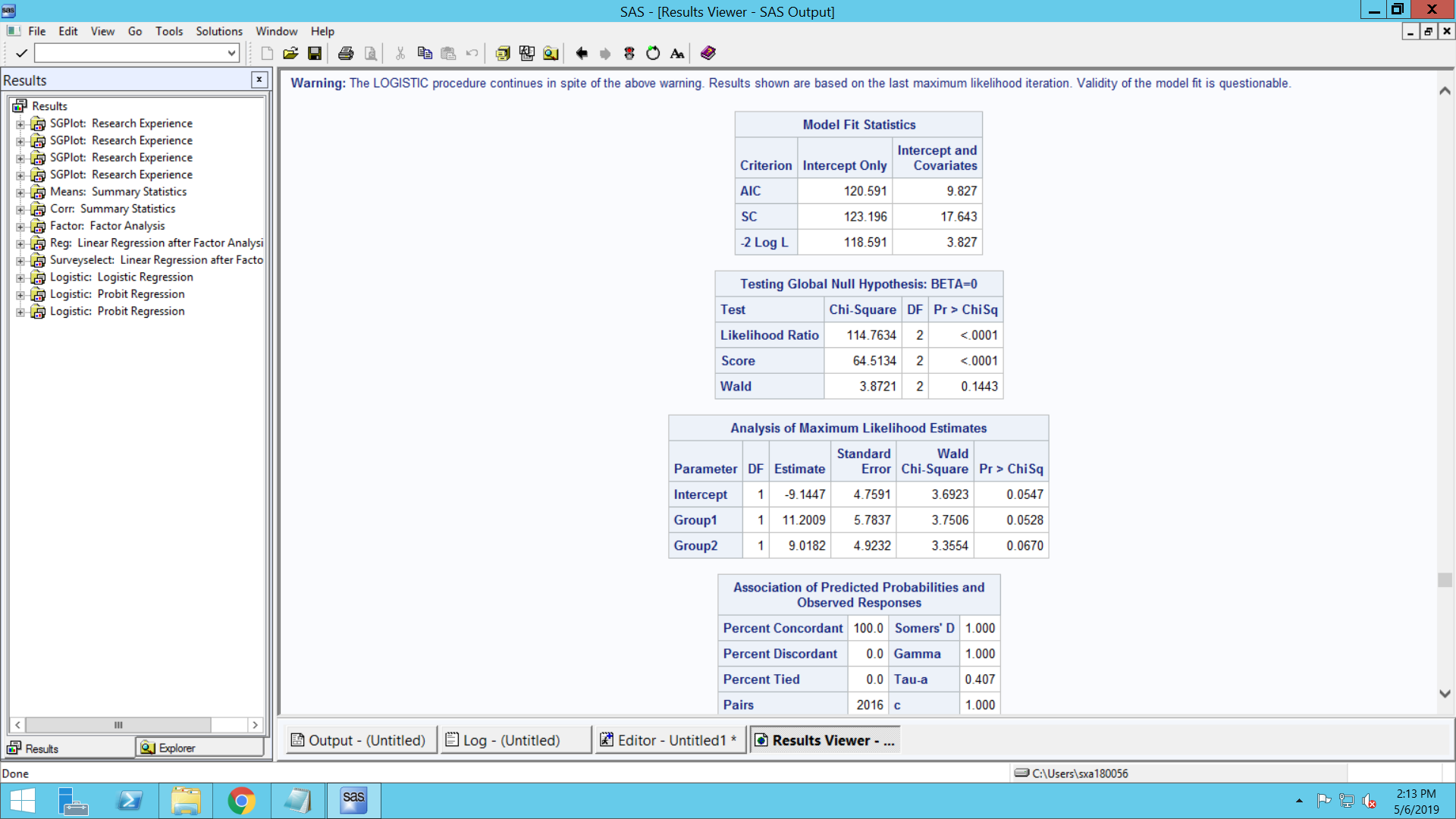
* The AIC we got is 480.091 and likelihood ratio is 320.16.
* We can see that both groups are significant.
* The Area under curve is almost similar to Logit 0.9708.

Probit Model On Test data:

* As we did the above models on training data, we also wanted to run a model on test data to asses the performance of the model.
* We are trying to test the model by making predications against the test data set.

Results:







Insights:

* The AIC we got is 120.29 and likelihood ratio is 114.76.
* We can see that none of the groups are significant.
* But the area under curve is which is slightly greater than 1.

So, for the data we have and among the models we did so far, we prefer Probit was best fit to estimate the data as it’s giving high area under curve under all the considerations we took.

SAS Code:

LIBNAME Project 'E:\Users\sxa180056\Downloads\Project';

/\* This imports the csv dataset into SAS. \*/

/\* You can do it by using the "Import Data" option in File on the main menu \*/

PROC IMPORT OUT= Project.Admission

DATAFILE= "E:\Users\sxa180056\Downloads\Project\Admission.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=2;

RUN;

/\* generating the working dataset in Work library \*/

data Admission;

set Project.Admission;

if Research = 0 then Research1 = "Not Having Experience";

if Research = 1 then Research1 = "Having Experience";

run;

proc sgplot data = Admission;

vbar Research1 / fillattrs = label(color = blue);

title 'Research Experience';

run;

proc sgplot data = Admission;

vbar University\_Rating / fillattrs = label(color = lightgreen);

run;

proc sgplot data = Admission;

vbar GRE\_Score / fillattrs = label(color = lightred);

run;

proc sgplot data = Admission;

vbar TOEFL\_Score / fillattrs = label(color = yellow);

run;

/\* Summary Statistics using PROC means: We can choose our variables \*/

proc means data= Admission n mean stddev min p25 median p75 max maxdec= 2;

var GRE\_Score TOEFL\_Score University\_Rating Research CGPA SOP LOR Chance\_of\_Admit;

title 'Summary Statistics';

run;

proc corr data = Admission out = corr;

var GRE\_Score TOEFL\_Score University\_Rating Research CGPA SOP LOR Chance\_of\_Admit;

run;

/\*\*\*\*\* Factor Analysis \*\*\*\*\*\*\*/

proc Factor data = Admission outstat=Factout nfactor = 2 scree rotate = Varimax out = Project;

var Research GRE\_Score TOEFL\_Score University\_Rating SOP LOR CGPA;

title 'Factor Analysis';

run;

proc score data= Project score=factout out=Project\_score (rename=(Factor1=Group1 Factor2=Group2));

var GRE\_Score TOEFL\_Score University\_Rating Research CGPA SOP LOR;

title 'Naming Factors';

run;

proc reg data = Project\_Score;

model Chance\_of\_Admit = Group1 Group2 ;

title 'Linear Regression';

run;

data Project\_Score;

set Project\_Score;

if Chance\_of\_Admit > 0.80 then Admit = "1";

if Chance\_of\_Admit =< 0.80 then Admit = "0";

run;

/\* Create training and test datasets. 80% of sample in training \*/

proc surveyselect data=Project\_Score out=Project\_sampled outall samprate=0.8 seed=2;

run;

data Project\_training Project\_test;

set Project\_sampled;

if selected then output Project\_training; /\* Tell SAS that only keep the 80% selected one in sample. The rest will be in test data \*/

else output Project\_test;

run;

/\* Logistic Regression \*/

proc logistic data=Project\_sampled plots = All;

logit: model Admit (event='1') = Group1 Group2;

weight selected;

title 'Logistic Regression';

run;

quit;

/\* Probit Regression \*/

proc logistic data=Project\_training outmodel=Probitmodel plots = All;

probit: model Admit (event='1') = Group1 Group2/ link = NORMIT;

title 'Probit Regression';

run;

/\* Probit Regression \*/

proc logistic data=Project\_test outmodel=Probitmodel plots = All;

probit: model Admit (event='1') = Group1 Group2/ link = NORMIT;

title 'Probit Regression';

run;