# Final Project Report

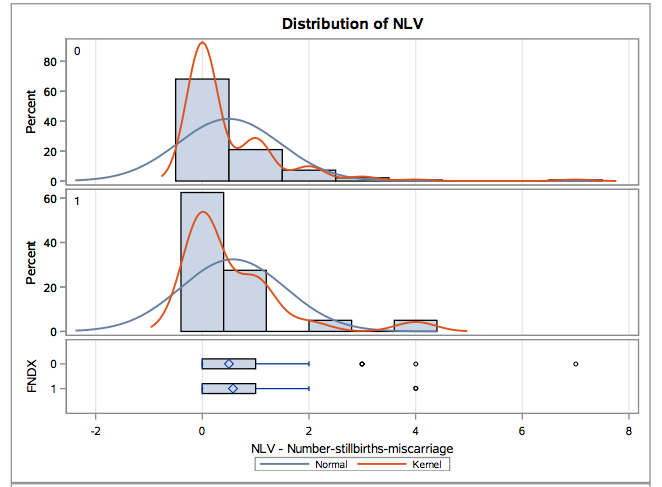
# Data Characteristics:

**The Distribution:** The distribution is a summary of the frequency of individual values or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of persons who had each value.

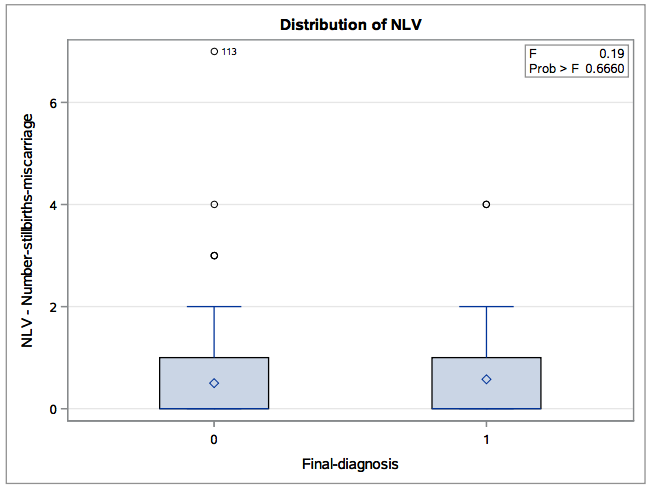
**The Mean** or average is probably the most commonly used method of describing central tendency. To compute the mean all you do is add up all the values and divide by the number of values.

**The Standard Deviation** is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range (as was true in this example where the single outlier value of 36 stands apart from the rest of the values. The Standard Deviation shows the relation that set of scores has to the mean of the sample.

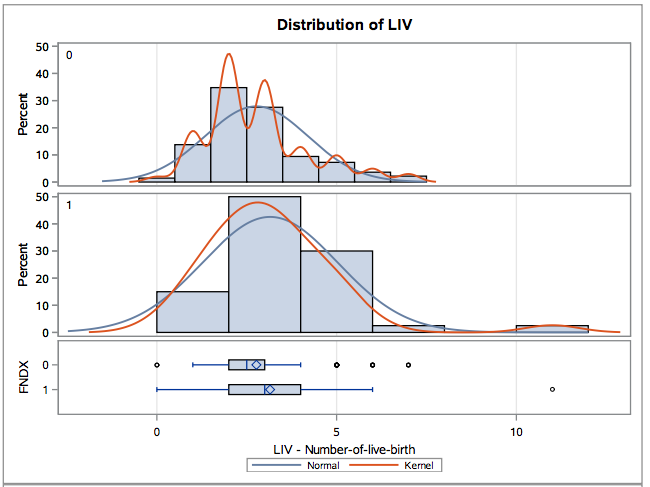
# NO OF OD STILLBIRTHS AND MISCARRIAGES (NLV):



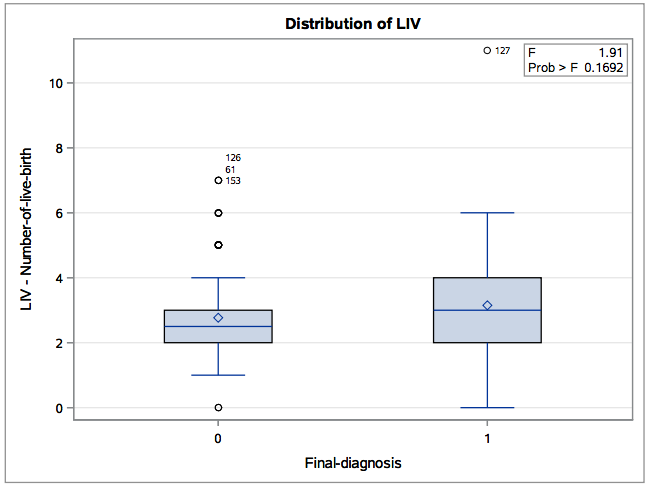
The Univariate test and Ttest was done on this variable and we got the lowest value to be 0 and highest value to be 7. The mean value was found to be 0.51 with standard deviation 0.96. From the above graph we can see that the standard deviation is equal. The red line(kernel) in the diagram is on the leftside for the blue line (normal) and so the skewness is greater than Zero(positive)

  
  
ANOVA test was done on this variable and we found the Pvalue to be 0.666 which is more than 0.05 and so it is not significant

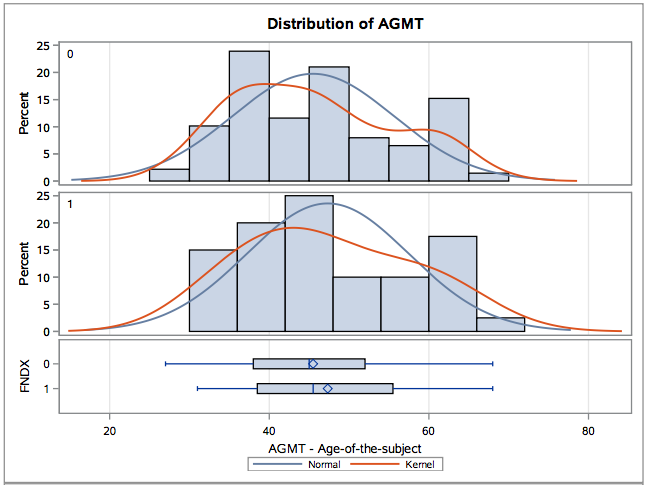
# NUMBER OF LIVE BIRTHS (LIV):



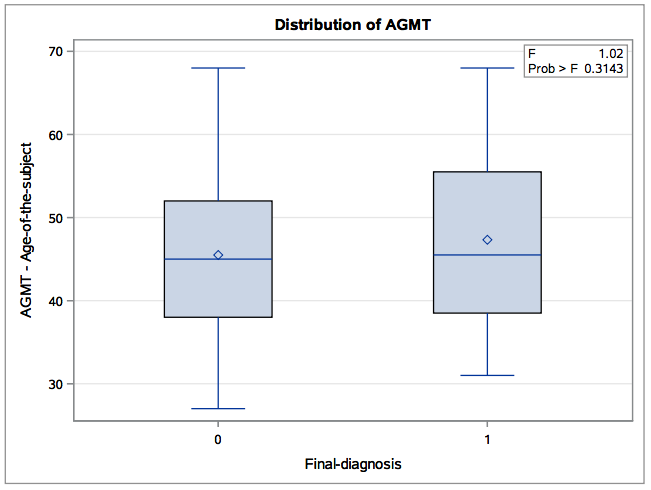
The Univariate test and Ttest was done on this variable and we got the lowest value to be 0 and highest value to be 11. The mean value was found to be 2.85 with standard deviation 1.54. From the above graph we can see that the standard deviation is no equal. The red line(kernel) in the diagram is on the leftside for the blue line (normal) and so the skewness is greater than Zero(positive)

  
  
ANOVA test was done on this variable and we found the Pvalue to be 0.1692 which is more than 0.05 and so it is not significant

# AGE AT INTERVIEW (AGMT):

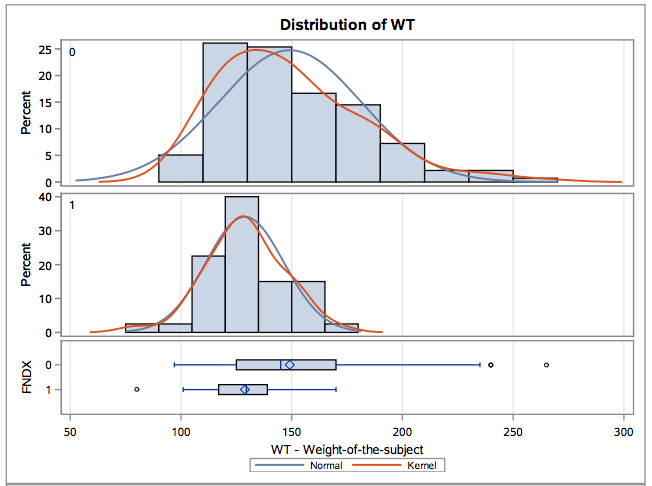


The Univariate test and Ttest was done on this variable and we got the lowest value to be 27 and highest value to be 64. The mean value was found to be 45.9 with standard deviation 10.11. From the above graph we can see that the standard deviation is no equal. The red line(kernel) in the diagram is on the leftside for the blue line (normal) and so the skewness is greater than Zero(positive)

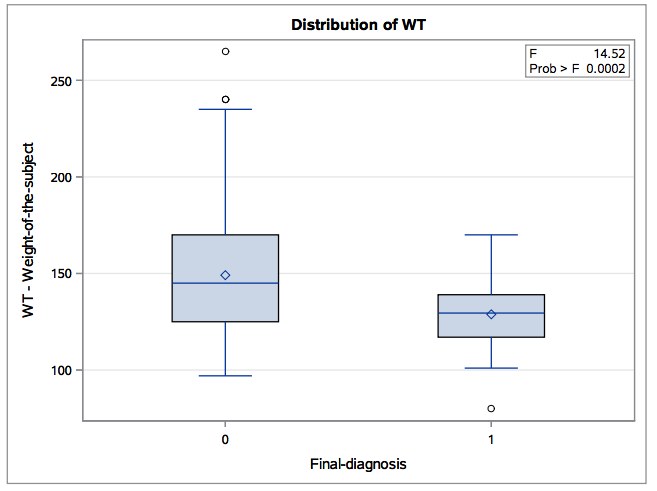


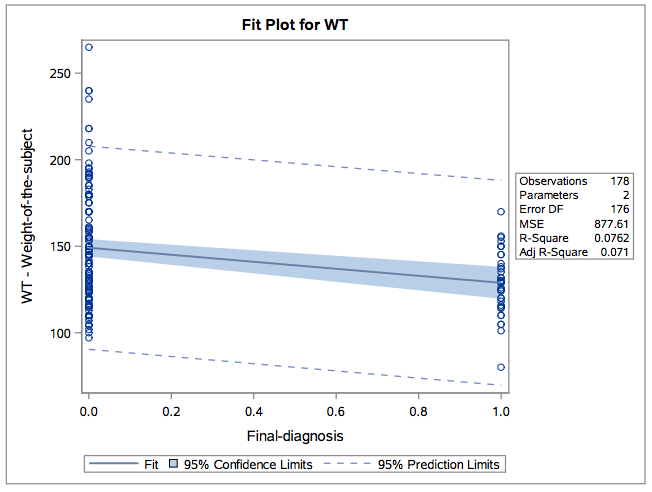
ANOVA test was done on this variable and we found the Pvalue to be 0.3143 which is less than 0.05 and so it is significant

# Weight (WT):



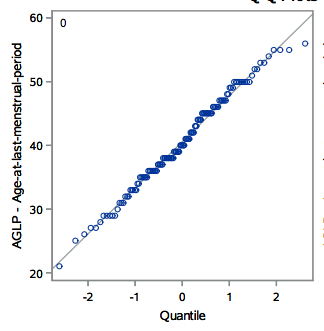
The Univariate test and Ttest was done on this variable and we got the lowest value to be 80 and highest value to be 265. The mean value was found to be 144.5 with standard deviation 30.73. From the above graph we can see that the standard deviation is no equal. The red line(kernel) in the diagram is on the leftside for the blue line (normal) and so the skewness is greater than Zero(positive)

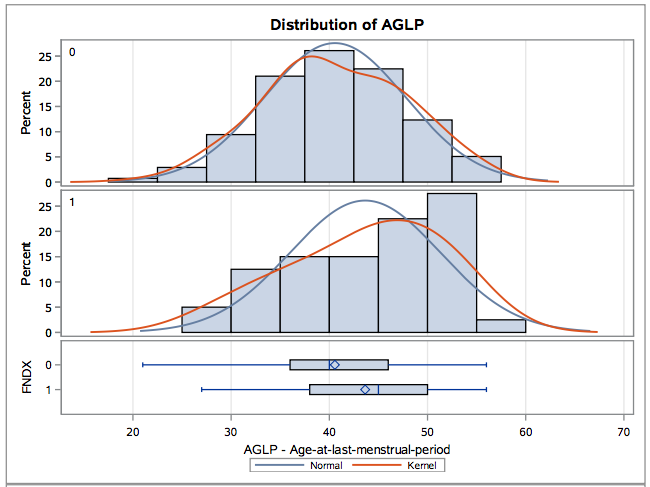
  
  
ANOVA test was done on this variable and we found the Pvalue to be 0.0002 which is less than 0.05 and so it is significant



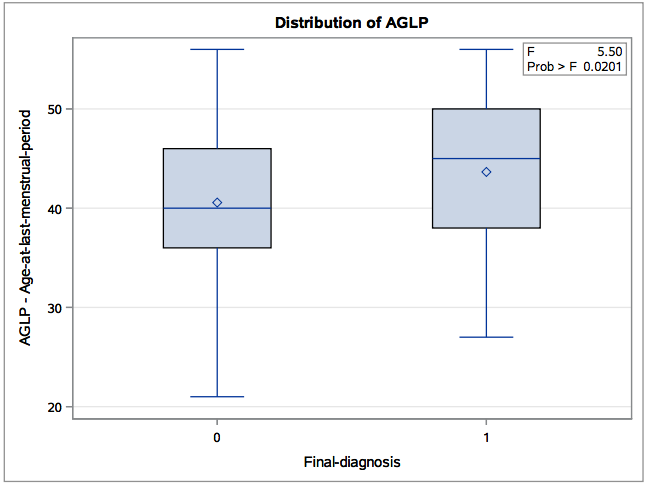
Linear regression was done on this variable and we can see from the diagram that both the intercept and slop have significant P-value but R-square value 0.00762 which is low .

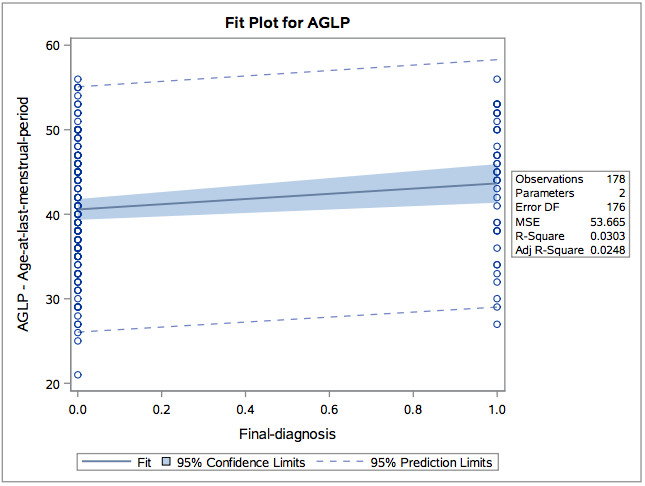
# AGE AT LAST MENSTRUAL PERIOD (AGLP):





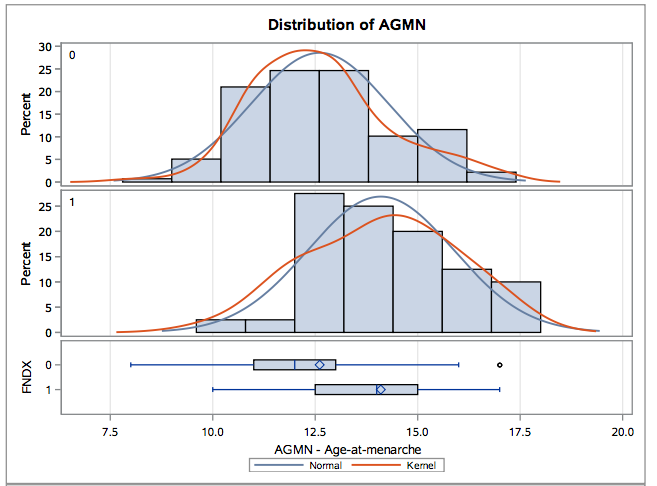
The Univariate test and Ttest was done on this variable and we got the lowest value to be 21 and highest value to be 56. The mean value was found to be 41.25 with standard deviation 7.41. From the above graph we can see that the standard deviation is no equal. The red line(kernel) in the diagram is on the leftside for the blue line (normal) and so the skewness is greater than Zero(positive)

  
  
ANOVA test was done on this variable and we found the Pvalue to be 0.0201 which is less than 0.05 and so it is significant.

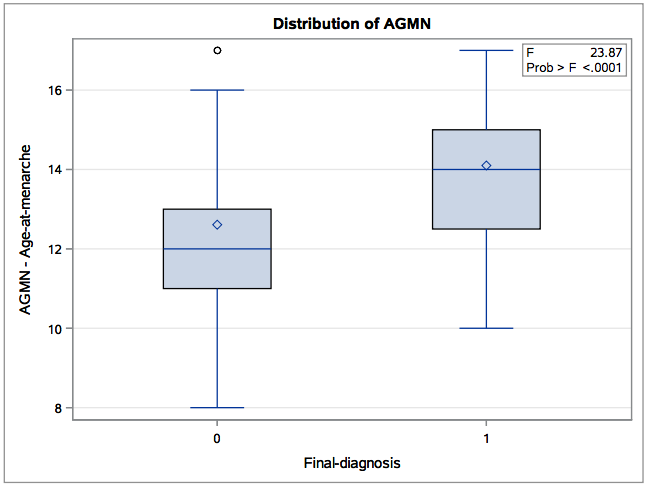


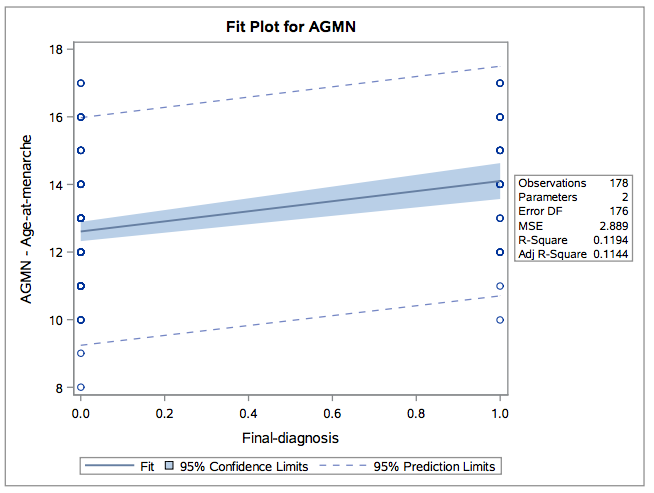
Linear regression was done on this variable and we can see from the diagram that both the intercept and slop have significant P-value but R-square value 0.03 which is low.

# AGE AT MENARCHE (AGMN):



The Univariate test and Ttest was done on this variable and we got the lowest value to be 8 and highest value to be 17. The mean value was found to be 12.94 with standard deviation 1.8. From the above graph we can see that the standard deviation is no equal. The red line(kernel) in the diagram is on the leftside for the blue line (normal) and so the skewness is greater than Zero(positive)

  
  
ANOVA test was done on this variable and we found the Pvalue to be 0.0001 which is less than 0.05 and so it is significant



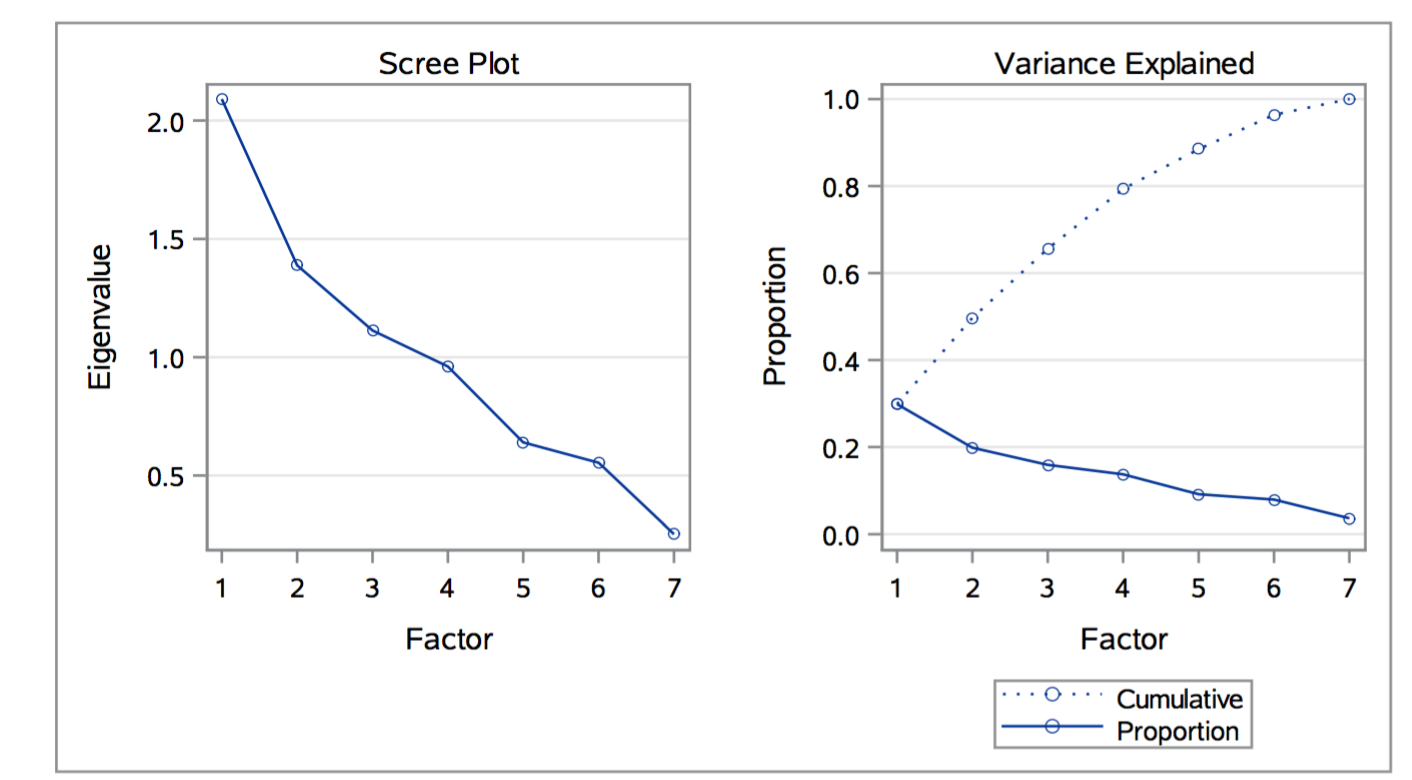
Linear regression was done on this variable and we can see from the diagram that both the intercept and slop have significant P-value but R-square value 0.11 which is low .

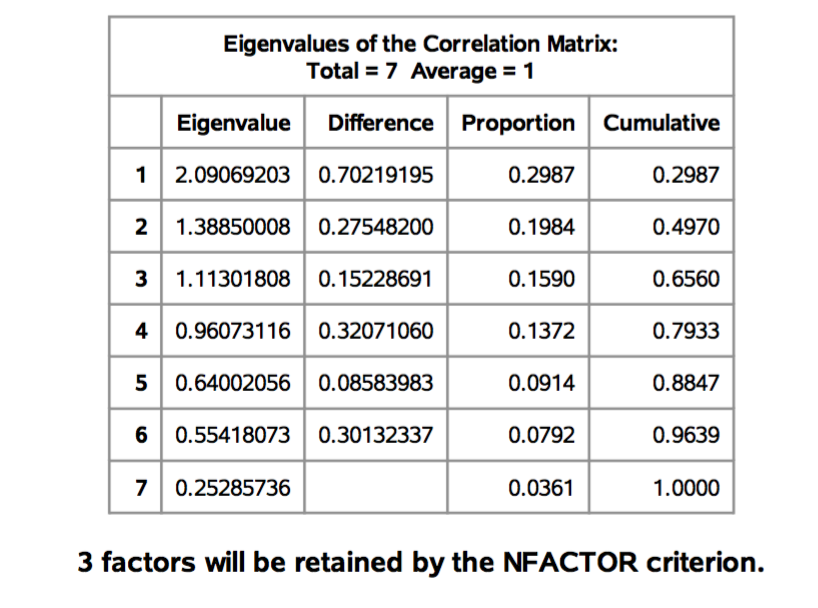
# Factor Analysis:

An exploratory factor analysis was done to understand if there are any underlying patterns between seven variables, which are metric.

The number of observations used is 178, which equals a ratio of 1:18, which is far higher and satisfactory than the minimum of 1:10.

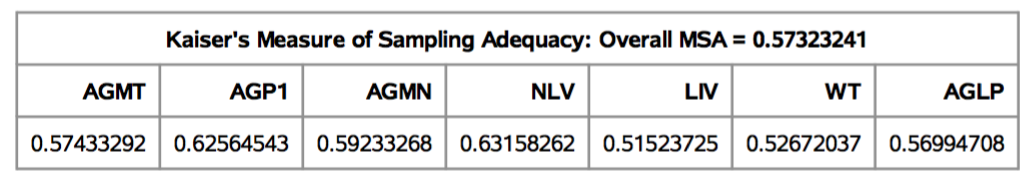
The correlation matrix obtained from the principal component analysis shows around 50% of correlations above ±0.30 indicating there is sufficient correlation in the dataset to justify use of factor analysis. We start with seven principal components or factors i.e. as many as the number of variables and see how much variance in the data each component or factor explains. The below table and plots show this.





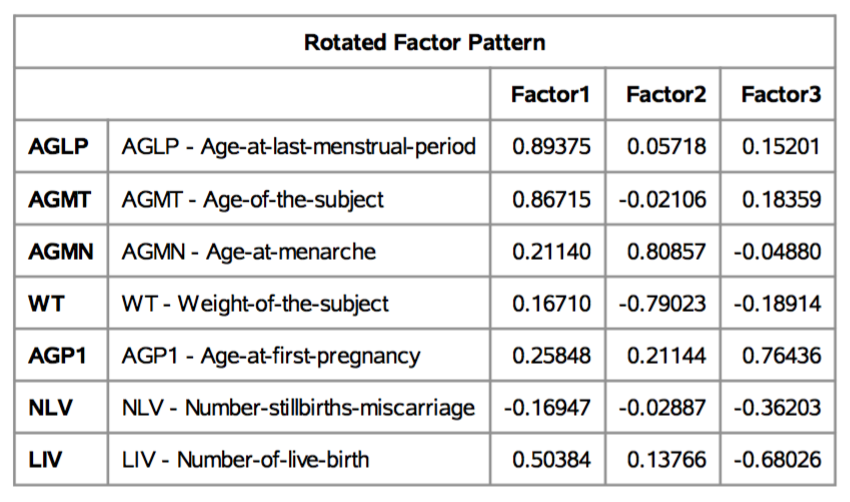
From the table, the variance and scree plots, the first three components combined explain about 66% of variance. They all have eigenvalues greater than 1 and each of them explains substantial amount of variance of about 30%, 20%, 16% respectively. This satisfies our conceptual criteria for selecting the appropriate number of factors.

After extracting the first three factors, the results are examined to check the overall MSA value (degree of correlation) for the data matrix and the MSA value for each variable



The overall MSA (Measure of sampling Adequacy) is calculated to be above the acceptable level of 0.5 at 0.57 which can deemed satisfactory. Similarly, all the seven variables chosen are found to be significant.

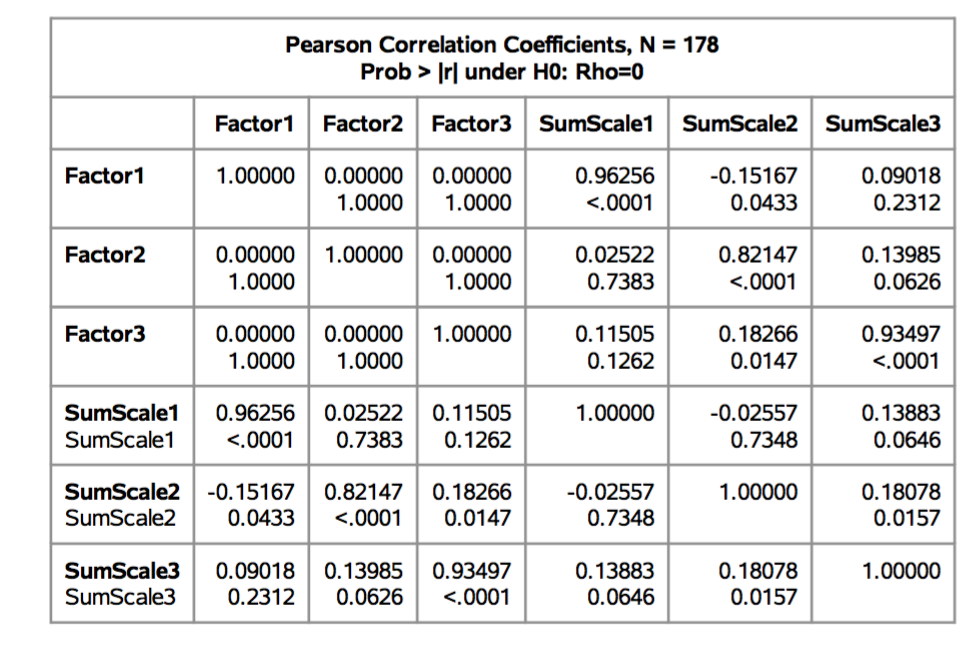
To examine how each variable loads on the factors, we do a Varimax rotation to make the loadings on each factor more significant. The results are shown below:



AGLP and AGMT load heavily on Factor 1, AGMN and WT load heavily on factor 2 and AGP1 and LIV load on Factor 3 significantly. If our objective is data reduction, we can remove LIV, which loads significantly on two factors and NLV, which seems to load not as significantly as we would like. As LIV is loading significantly both positive on Factor 1, also negatively on Factor 3. We choose to remove the variable and run the analysis without it.

.

The correlation matrix for the computed factor scores and summated scales is shown on the next page. The correlation coefficient for the two factors is 0 which shows that the solution is perfectly orthogonal i.e. the two factors are fully independent of each other. Similarly, the summated scales are also independent of each other. They don’t have a 0 correlation coefficient as summated scales are not standardized like factor scores, they are a straight average. Nonetheless their correlation coefficient is very close to 0 indicating orthogonality.



# Logistic Regression

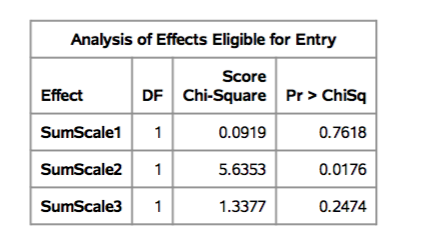
As our objective of factor analysis is create summated scales, which usually reduce measurement error and are a compromise between surrogate variable and factor score options. We use these summated scales to see how they influence the factor Final diagnosis.

Stepwise selection procedure is used. The dataset is divided and 90 observations are used.

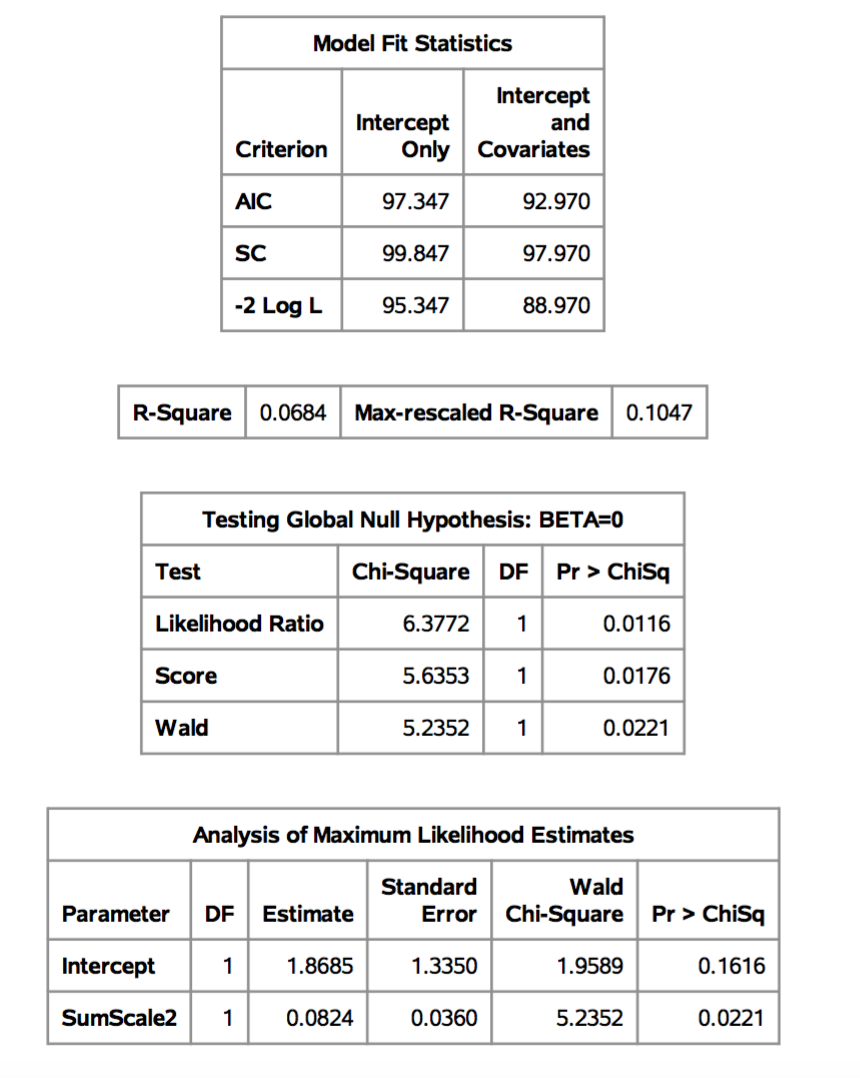
FNX(Final diagnosis) is designated as the response variable.

In the three variables, Sumscale2 is the most eligible for entry

Probability modeled is FNDX=1.

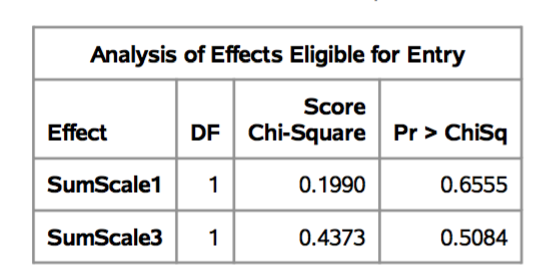


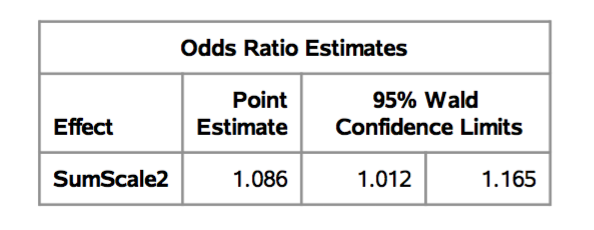
After entering X7 to the model, AIC, Schwarz Criterion and log likelihood are 92.97, 97.97 and 88.97. The Nagelkerke R-Square is 0.0684. The testing global Null hypothesis p-value for likelihood ratio, score and Wald is around zero showing that at least one variable is not equal to zero in the overall model.



The above table of Maximum Likelihood estimates show the coefficients, standard error , Wald Chi-square statistic and associated p-values. The coefficient of Sumscale2 is statistically significant with a p-value less than 0.05

The analysis of eligible for entry shows no other variable is eligible for entry.





An odds ratio is the exponentiated coefficient, and can be interpreted as the multiplicative change in the odds for a one-unit change in the predictor variable.

Here, one unit change in Sumscale2 increases the odds of Final diagnosis being 1 by 1.086 times.

Finally the percentage of 0 and 1 for the holdout sample of 88 is calculated and displayed in a table