**Individual project report**

**Predicting the somatotype of children in age 18**

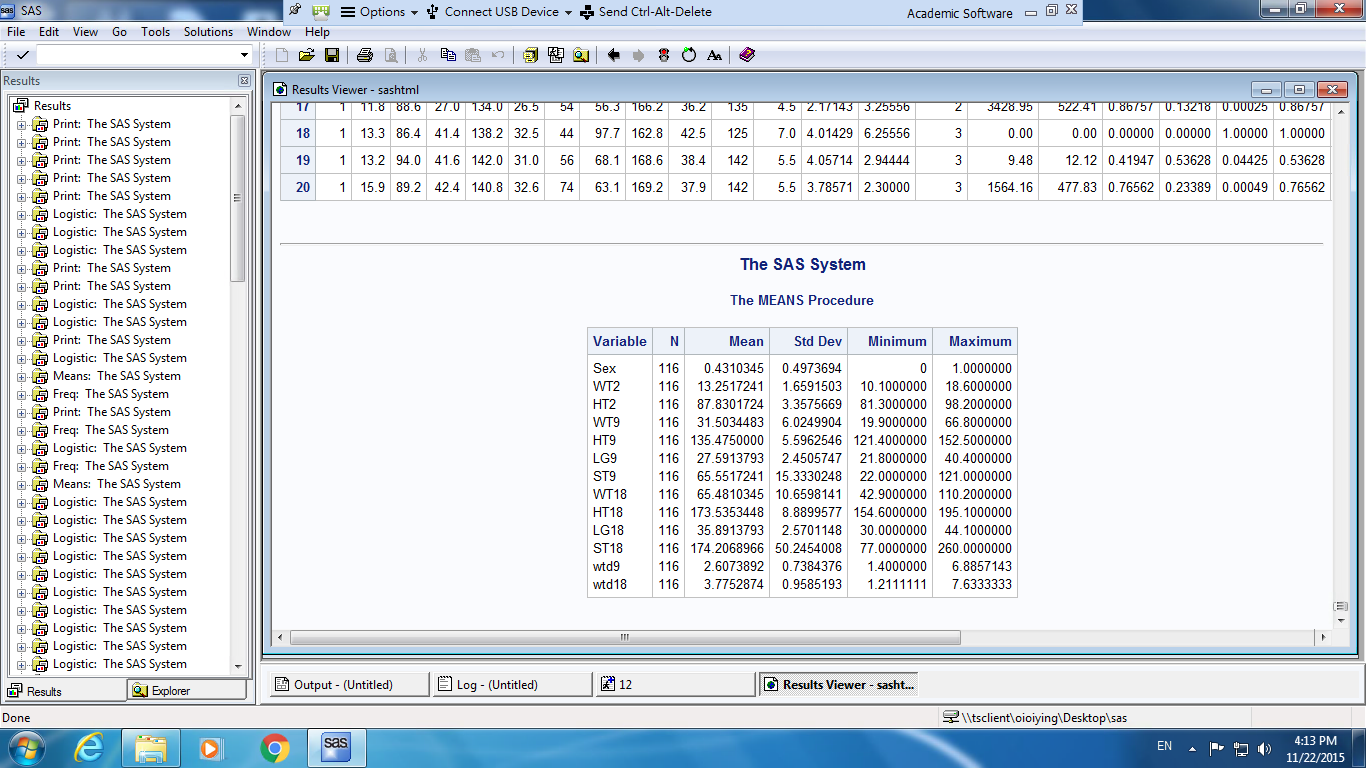
Name: Wong Choi Yan Heki Student ID : 20040850 Data set: 20

**Summary**

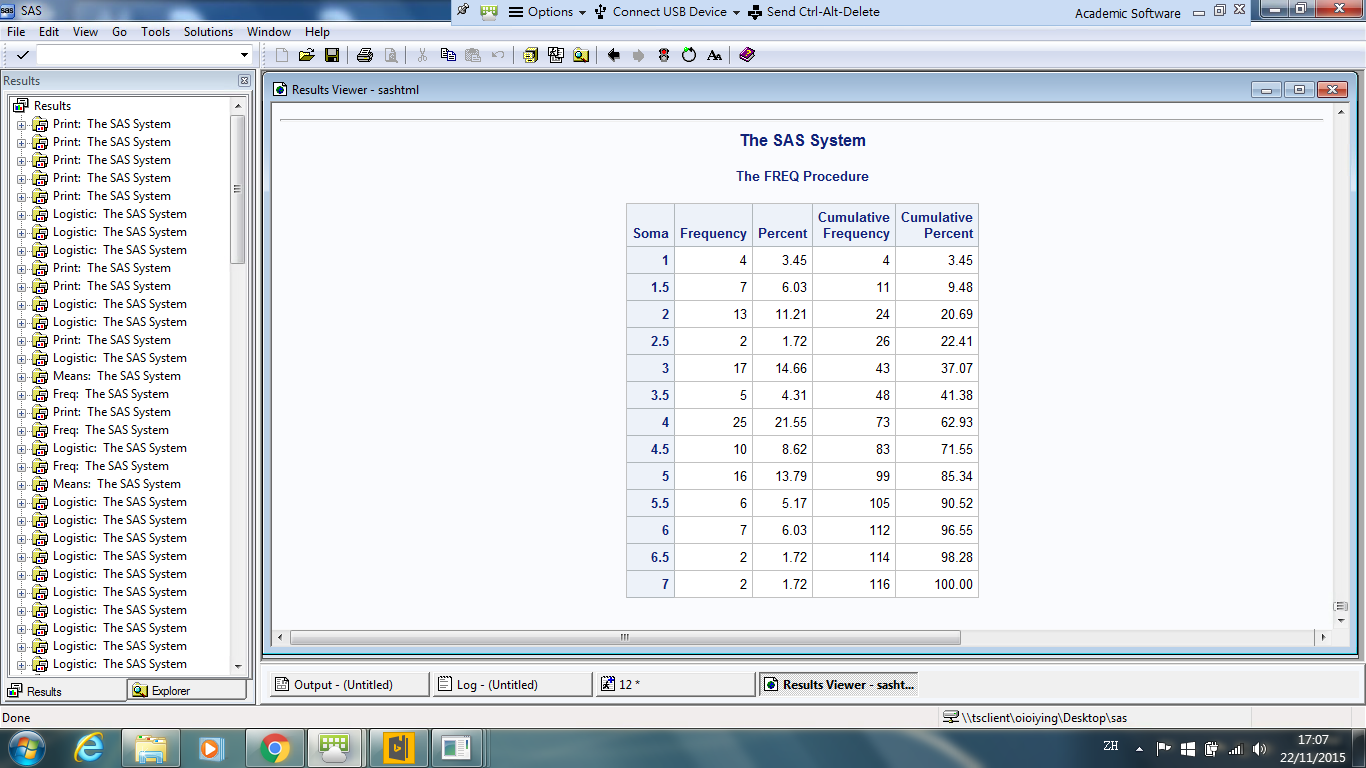
The project is to analyze the data set of 136 children born in 1928-29 in Berkeley in order to identify their somatotype, which is the body types based on their body shape, at age 18, so that they can then match with the most suitable sport or activities according to their somatotype. The objective is developing a new multiple regression model for predicting the somatotype of children, at age 18. Gender, weight gain from age 2 to 9, weight gain from 9 to 18, and height in age 18 are found to be the significant predictors for the somatotype at age 18. Results come out that children who increase their weight gain from age 2 to age 9 and from age 9 to age 18, may have less chance of being thin, and being normal, at age 18. In addition, children who is taller at age 18 and who is a boy, have higher chance of being thin or being normal. Apart from these, height of children at age18 may have same effect on predicting children being thin and children being normal, but other variables may have different effect on predicting.

**Background information**

The data set contains 12 variables on 136 children born in 1928-29 in Berkeley. The outcomes variables is **soma**, somatotype, a scale from 1, very thin, to 7, obese, of body type. The variables includes **sex**(0=male, 1=female), a binary categorical variables, and nine continuous variables: age 2 weight(**wt2**), age 2 height(**ht2**), age9 weight(**wt9**), age9 height(**ht9**), age9 leg circumference(**lg9**), age9 strength(**st9**), age18 weight( **wt18**), age18 height(**ht18**), age18 leg circumference(**lg18**), age18 strength(**st18**).



*Table1: Weighted means of all of the independent variables*



*Table2: Weighted frequencies for the dependent variables*

**Objective**

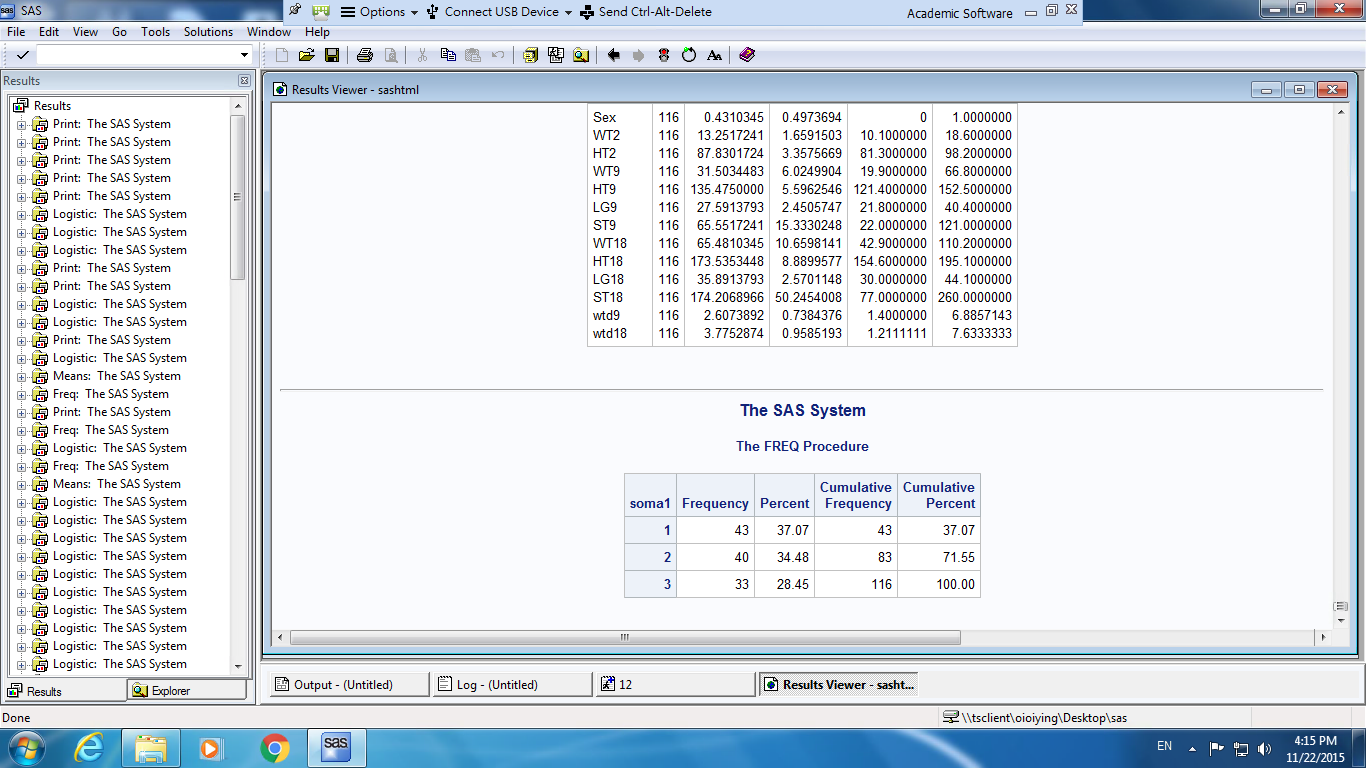
The objective is developing a new multiple regression model for predicting the somatotype of children, at age 18. We will do regression analysis in order to determine which predictors are significant to create model to identify the somatotype of children. To this end, a new multiple regression model will be proposed based on the analysis results.

**Methodoloy and Results**

Since there are 13 possible outcomes variables, which are somatotype, a scale from 1, 1.5, 2, 2.5, 3, … , 7 (from very thin to obese), we will use either proportional odds model or multinomial logit model. It is checked that the proportional odds assumption is not satisfied ( P-value of Score test <0.0001, reject the null), which means the slope coefficients depend on the level of the response variables, multinomial logistic regression is thus chosen for modeling, designating last level as reference level.

***Step1 :Regroup the responses***

As some of the responses have only few observations, we need to regroup the responses in order to get a more precise model. By considering the means of independent variables, we will group the response if the corresponding independent variables have similar mean. The 13 responses are thus regrouped into 3 main responses, which is somatotype, a scale from 1(thin), 2(normal), 3(obese).



*Table3: Weighted frequencies for the dependent variables after grouping*

In place of the preceding variables, we also consider the following variables:

WTD9 = = average weight gain from age 2 to 9

WTD18 = = average weight gain from age 9 to 18

***Step2: Model selection***

|  |  |  |  |
| --- | --- | --- | --- |
|  | Model | AIC | BIC |
| 1 | **Intercept, wtd18** | 253.991 | 265.005 |
| 2 | **Intercept, sex, wtd18** | 170.682 | 187.203 |
| 3 | **Intercept, sex, wtd18, wtd9** | 153.911 | 175.940 |
| 4 | **Intercept, sex, wtd18, wtd9, ht18** | 138.633 | 166.169 |
| 5 | **Intercept, sex, wtd18, wtd9, ht18, ht9** | 138.349 | 171.392 |
| 6 | **Intercept, sex, wtd18, wtd9, ht18, ht9, ht2** | 141.525 | 180.075 |
| 7 | **Intercept, sex, wtd18, wtd9, ht18, ht9, ht2, wt2** | 144.977 | 189.035 |

*Table4: AIC and BIC of models*

We input the variables from the most significant one, to the least significant one, to compare the corresponding AIC and BIC. All the interaction terms are found to be insignificant. From table 4, the have the smallest BIC and model have the smallest AIC. The model with the smallest Akaike Information Criterion (AIC) represents the best approximation to the true model, and BIC is closely related to AIC. For the adjusted value, which derived to reflect the extent of how the independent variables contribute to the variation in the dependent variable.model have adjusted =0.4925 and model have adjusted =0.50156. As model only have slightly lower adjusted and slightly higher AIC than model, we use the sequential methods to decide the finalize model.

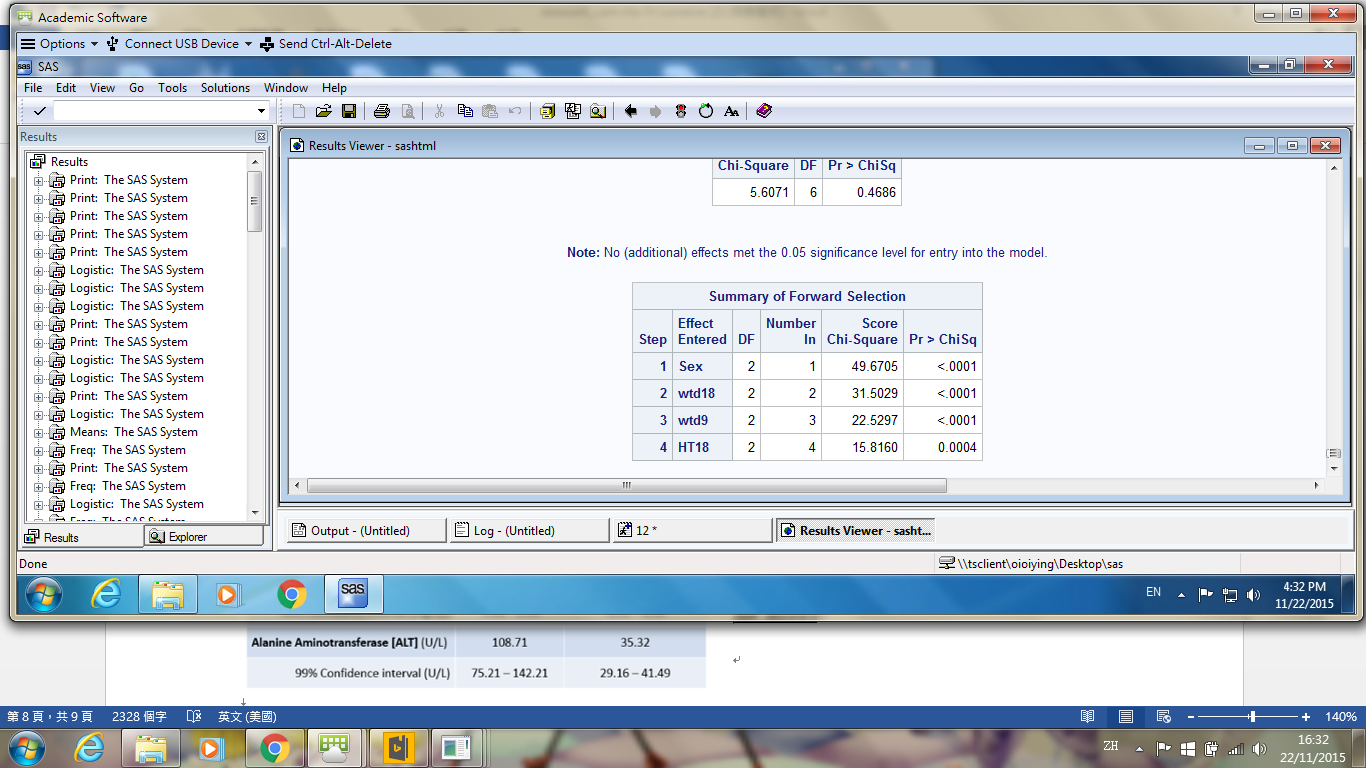
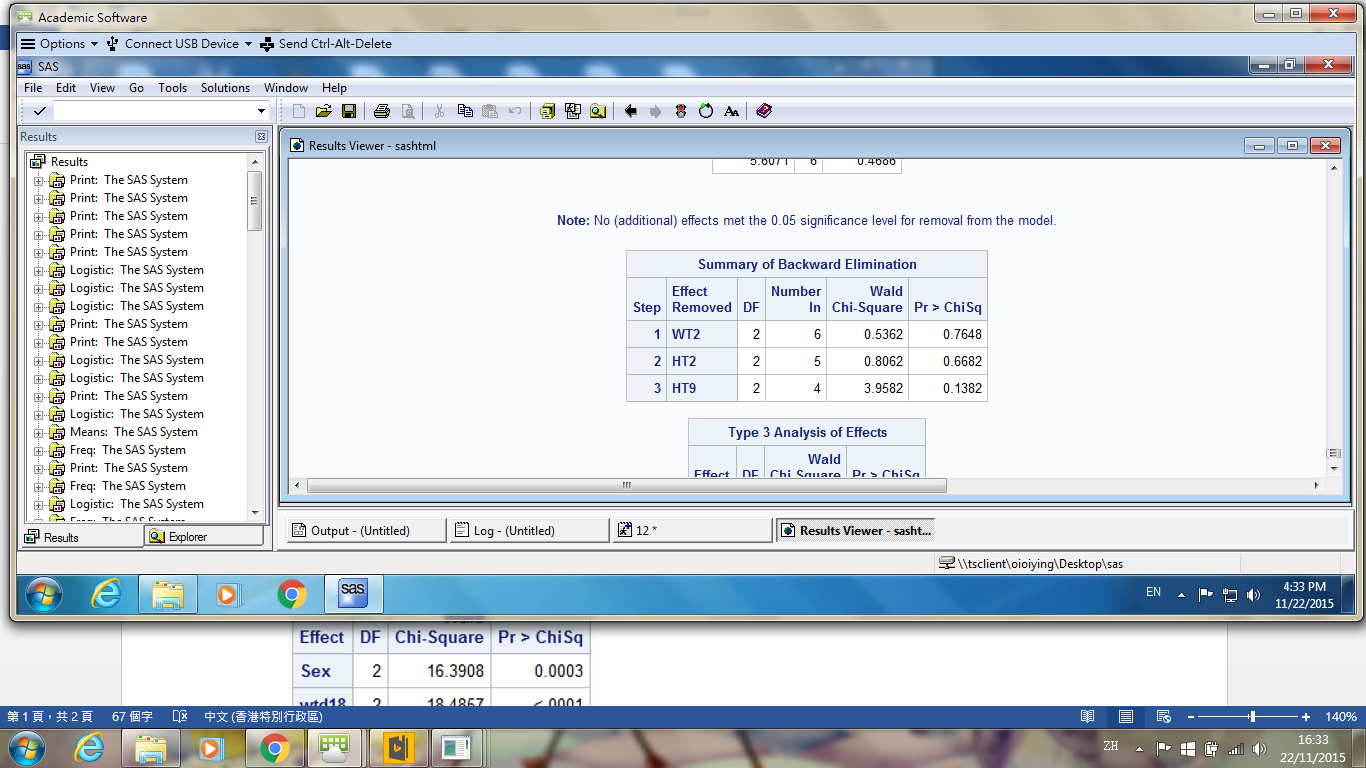
We can get the follow result from the SAS,

Forward selection: **Intercept, sex, wtd18, wtd9 and ht18**

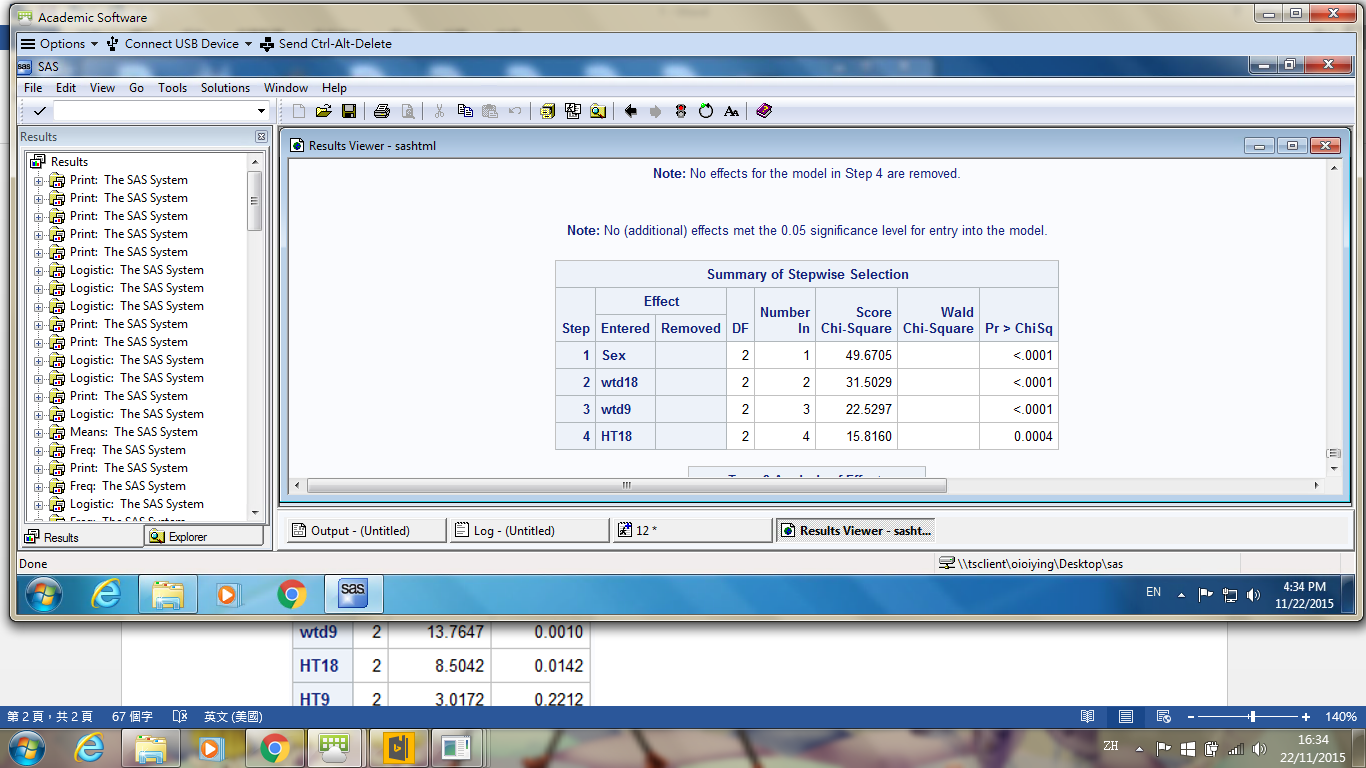
Backward elimination: **Intercept, sex, wtd18, wtd9 and ht18**

Stepwise selection: **Intercept, sex, wtd18, wtd9 and ht18**

The all results are consistency. Thus, variables (**sex, wtd18, wtd9 and ht18**) have been entered into the model.

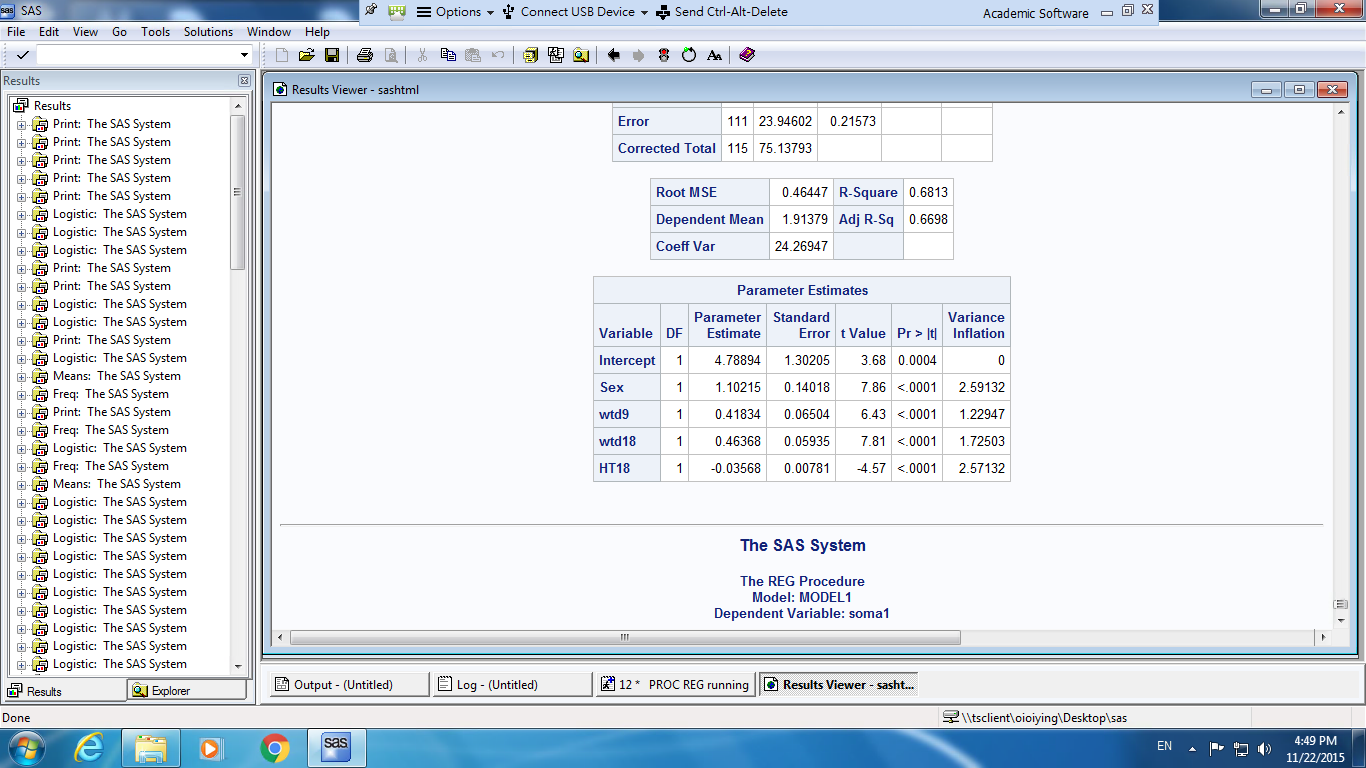
*Table5: Forward selection Table6: Backward elimination*



*Table7: Stepwise selection*

***Step3: Measure the multicollinearity of the independent variables***

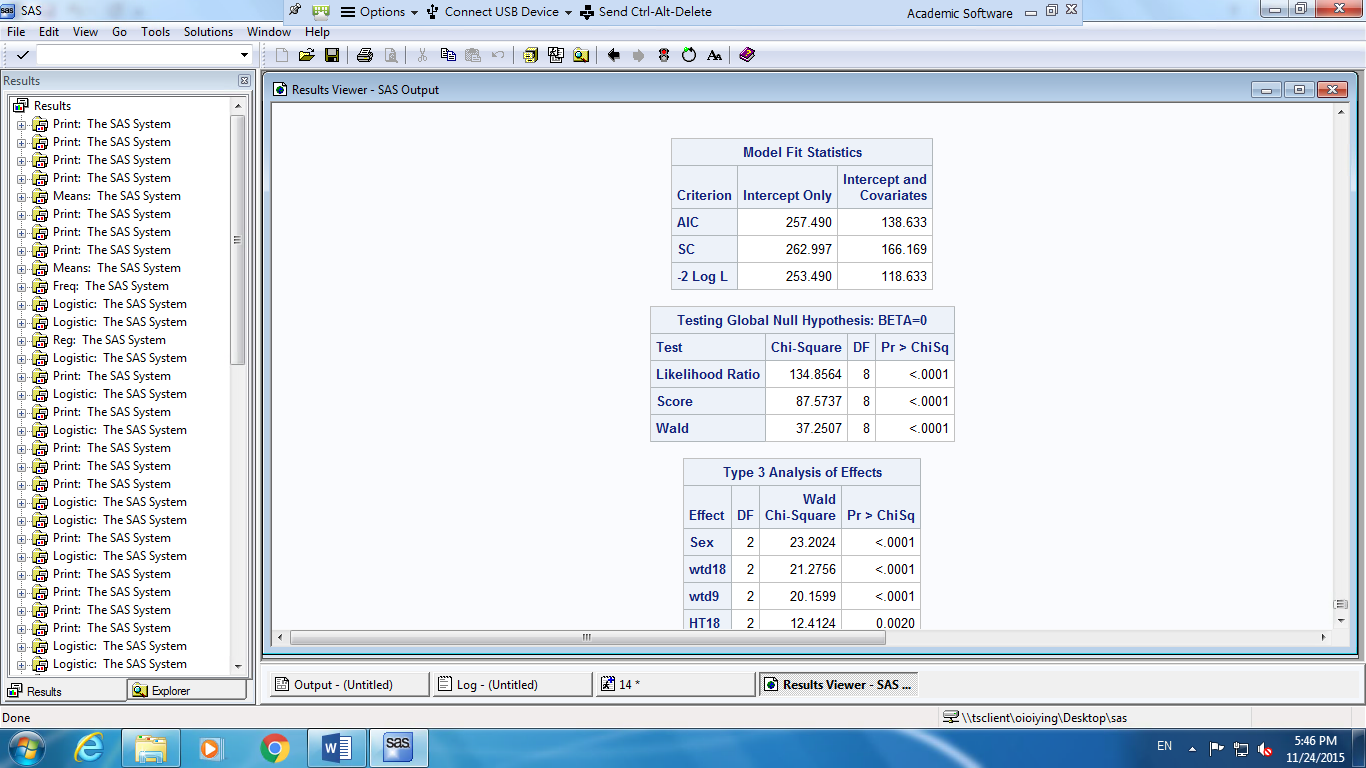
The variance inflation factor (VIF) measures the multicollinearity of the independent variables in a multiple linear regression model. The higher the VIF, the lower the precision in the estimate of the parameter. In the table below, we can see that the VIF of all the independent variables is low (all <5), so the multicollinearity of the variables is low, and thus the parameter estimate in the model is precise.



*Table8: VIF of model*

***Step4: Goodness of fit test***

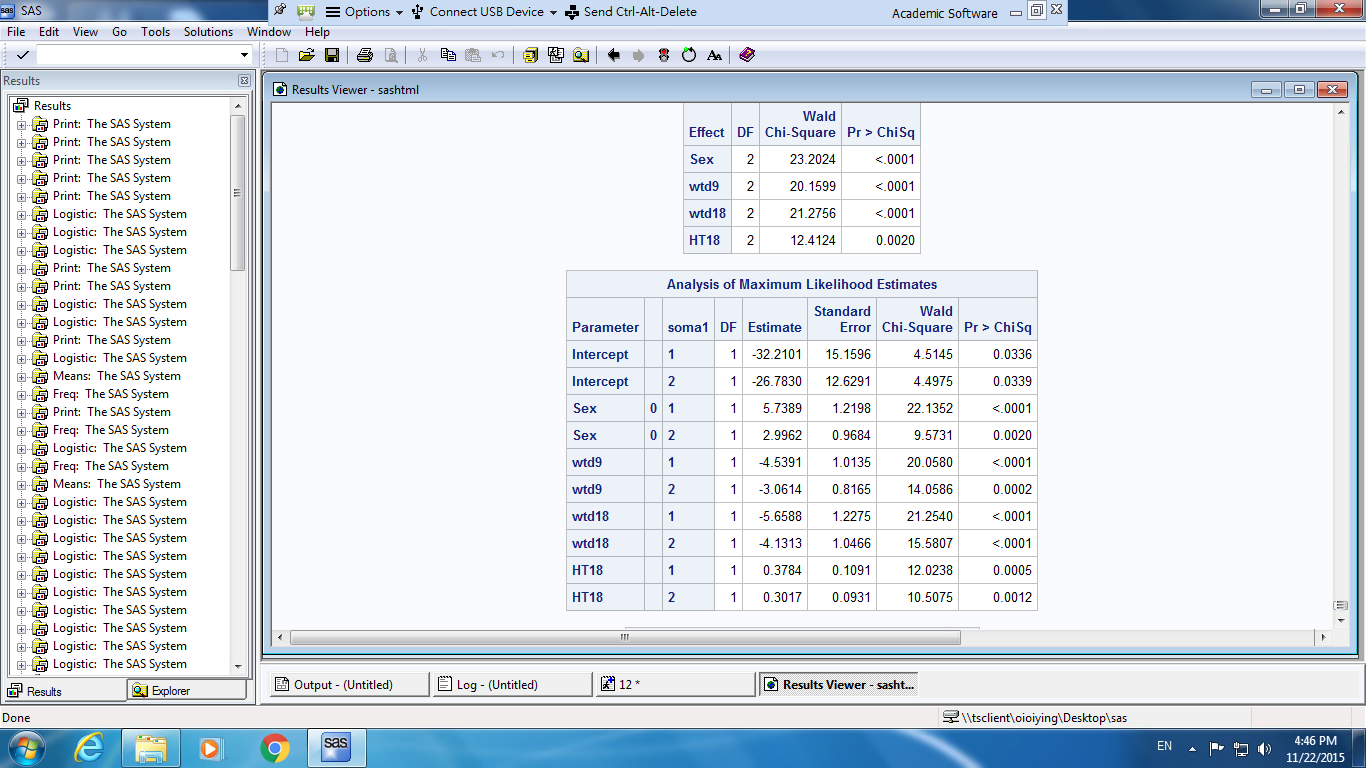
The p-value of Deviance and Pearson Goodness-of-fit statistic >0.05, so we do not reject the null hypothesis and thus the model fitted well at significant level 0.05. In addition, we reject the null hypothesis in Likelihood Ratio test, which means the model provides a good fit to the dependent and independent variables.



*Table9: Deviance and Pearson Goodness-of-fit statistics Table10: Likelihood Ratio Statistic*

***Step5: Fit the model***

Hence, the fitted model is:



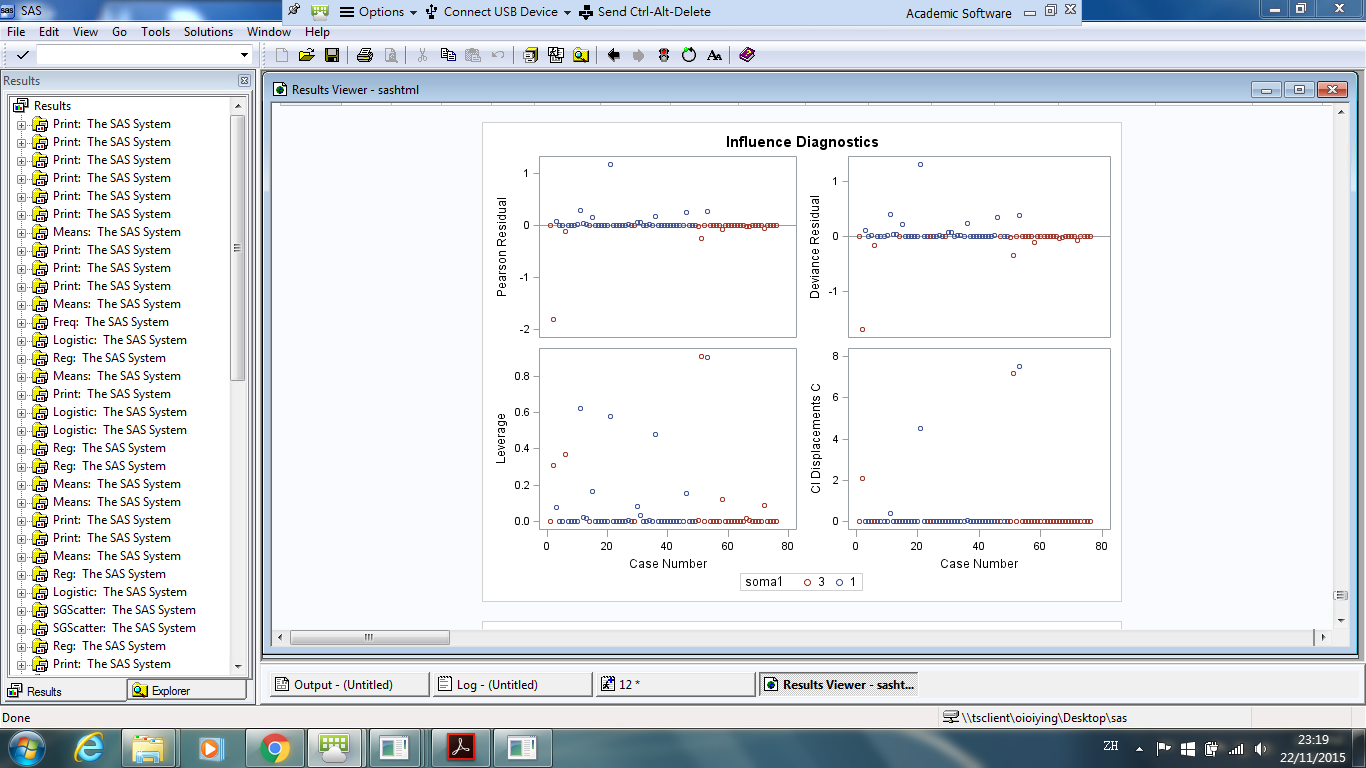
*Table11: Parameter Estimates of model*

***Step 6: Outliers or influential points***

Outliers and influential points in the model can be caused by measurement errors or may be the result of inherent variability of the data. Thus, we should find out those data, which are considerably exceptional or inconsistent with the rest of the data after we fitted the model. As multinomial logistic regression in SAS does not compute any diagnostic statistics, we use the Logistic Regression procedure to calculate and examine diagnostic measures. We run two binary logistic regressions (somatotype level1 to level 3, somatotype level2 to level3).

*For somatotype level 1 and level 3:*

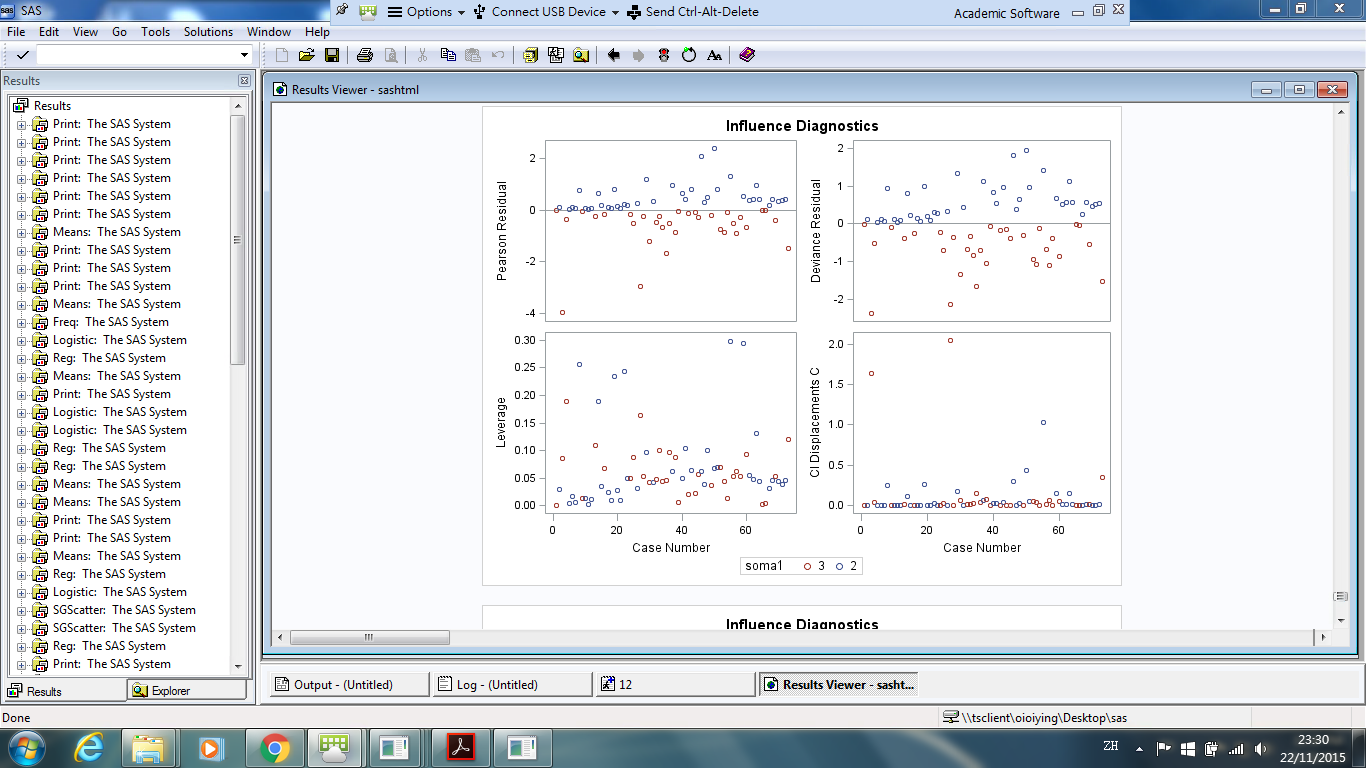
Pearson residual is used as an indicator to determine outlier. We use 2 as the cutoff value. From the graph 1, we can see that the highest Pearson Residual is (1.79977)and (1.16933)data, but their value is not big. So they are not regarded as an outlier. For influential point, confidence interval displacement diagnostics measures the influence of individual observations on the regression estimates, and we use 1 as the cutoff value. From the graph, we find (2.06476), , (7.49434) are greater than 1. However, after deleting the individual observation, we cannot find significant change in the regression estimates, so these four datum are not influential point.



*Graph 1: Influence Diagnostics for somatotype 1 and 3*

*For somatotype level 2 and level 3:*

From the graph 2, we can see that data have large Pearson residual and so these four datum are outlier. For influential point, we find and also have confidence interval displacement C greater than 1. After deleting the data, we find significant change in the regression estimates, but we cannot get the significant change of parameter estimates after deletingdata. So, only data are influential point. We thus delete observations and fit the model again.



*Graph 2: Influence Diagnostics for somatotype 2 and 3*

***Step 7:* After remove the outlier**

Repeat Step2 to step6 again. By sequential method, it output the following results:

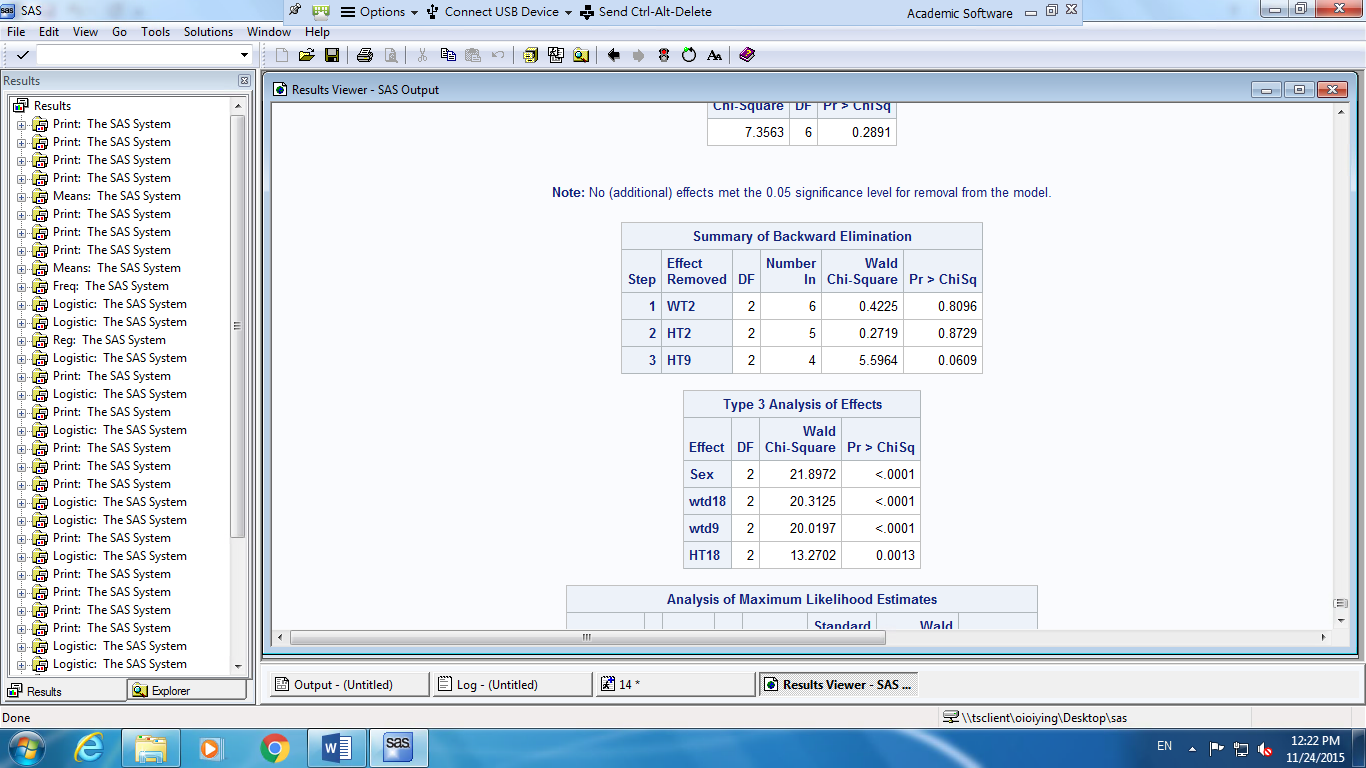
Forward selection: **Intercept, sex, wtd18, wtd9, ht9 and ht18**

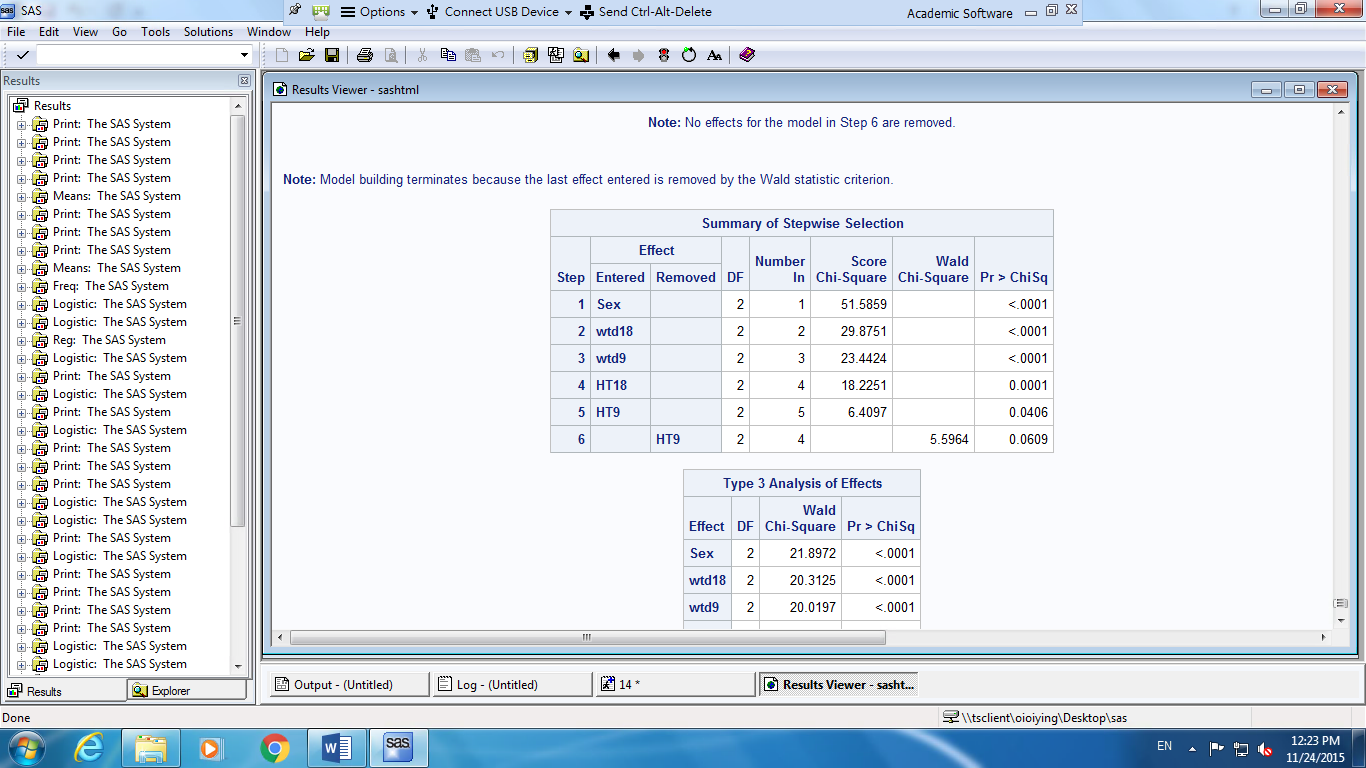
Backward elimination: **Intercept, sex, wtd18, wtd9 and ht18**

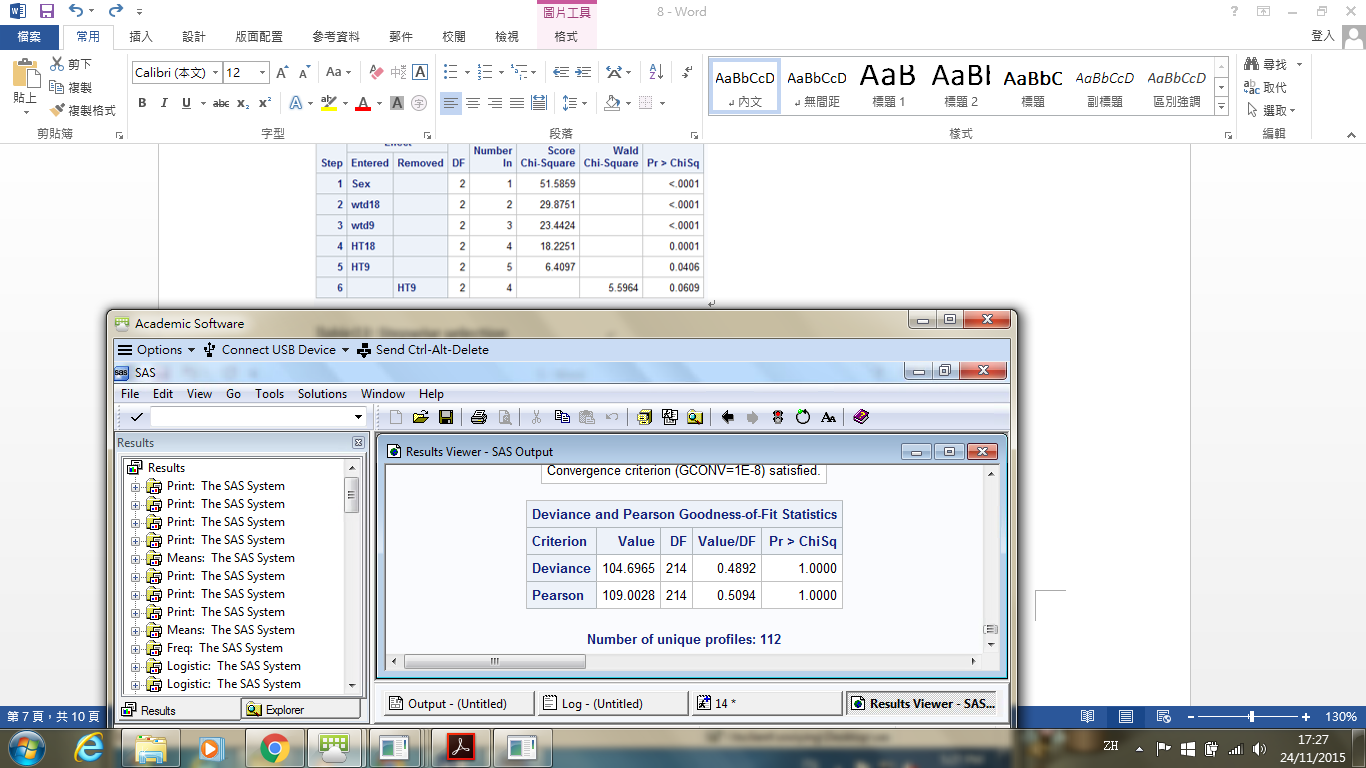
Stepwise selection: **Intercept, sex, wtd18, wtd9 and ht18**

As both backward and stepwise selection agree with the same model, thus variables (**sex, wtd18, wtd9 and ht18)** is entered into the model. In addition, from table 15, the p-value of Deviance and Pearson Goodness-of-fit statistic >0.05, so we do not reject the null hypothesis and thus the model fitted well at significant level 0.05. In addition, we reject the null hypothesis in Likelihood Ratio test, which means the model provides a good fit to the dependent and independent variables. From table 17, the overall effects of sex, wtd18, wtd9 and ht18 are listed in this tables. Since P-value of Wald test <0.05, the null is rejected and all of the variables are significant.

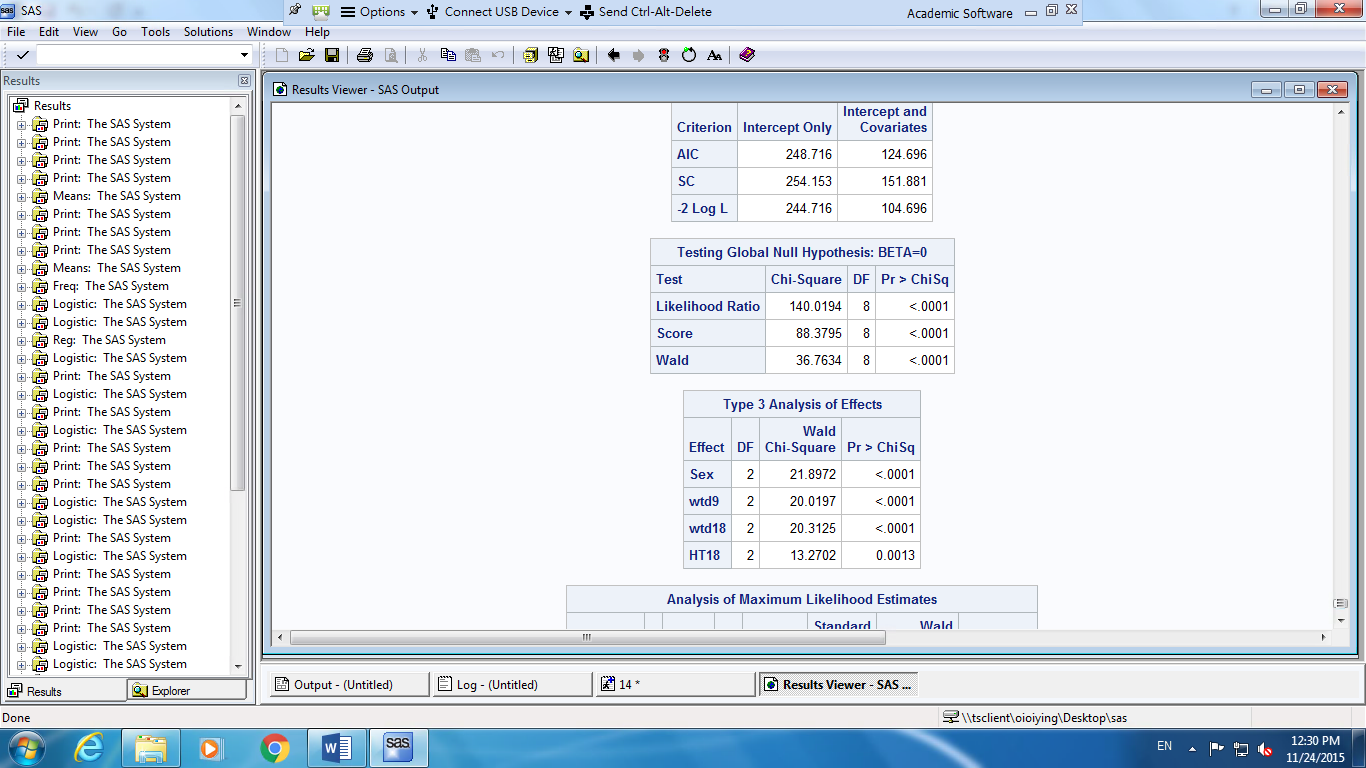
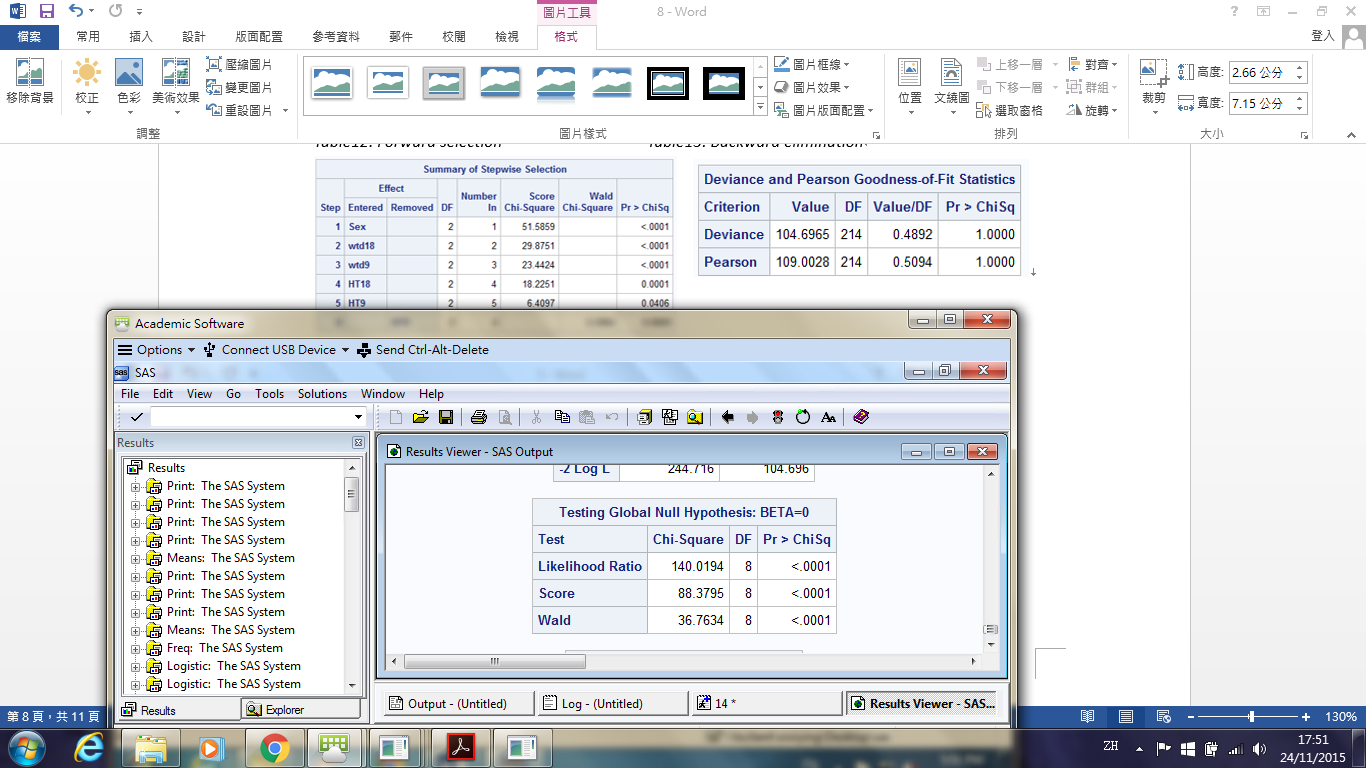
For this model, the adjusted = 0.5313, AIC =124.696 and BIC=151.881. The adjusted is larger and the AIC and BIC is smaller, so the new model is better.



*Table12: Forward selection Table13: Backward elimination*

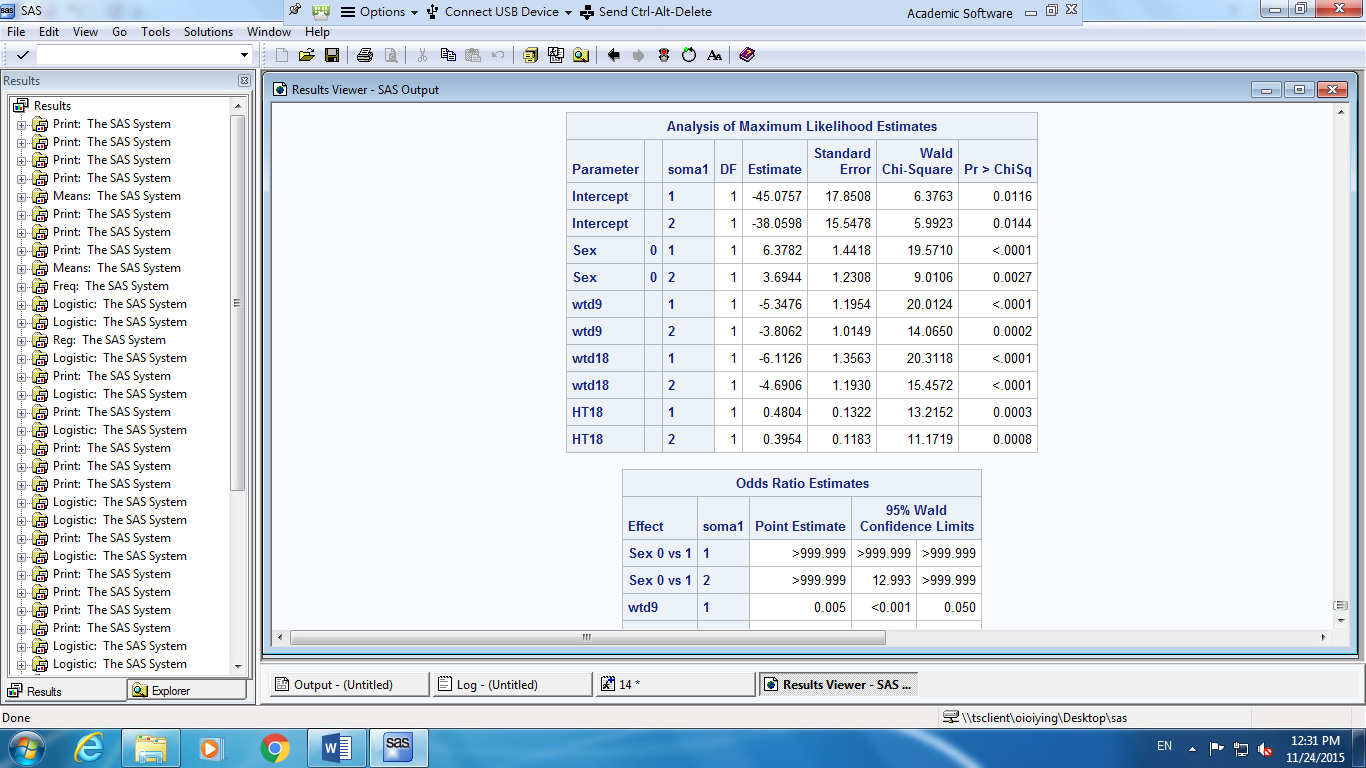


*Table14: Stepwise selection Table15: Goodness-of fit statistic*



*Table16: Likelihood Ratio Statistic Table17: Type 3 Analysis of Effects*

From the SAS result, the new fitted model is



*Table18: Parameter Estimates of model*

**Step 7: Odds**

1. A one unit increase in **wtd9** multiplies the odds of being in somatotype 1 vs somatotype 3 by 0.005(. Because 100(0.005-1)% = -99.5%, the odds are expected to decrease by about 99.5%. Also, a one unit increase in **wtd9** also multiplies the odds of being in somatotype 2 vs somatotype 3 by 0.022(. Because 100(0.022-1)% = -97.8%, the odds are expected to decrease by about 97.8%.

**In other words, children who increase gain weight from age 2 to age 9, may have less chance of being thin, in somatotype1 relative to in somatotype3 (chance decrease by 99.5%) and being normal, in somatotype2 relative to in somatotype 3(chance decrease by 97.8%), at age18.**

1. Similar to the above result, a one unit increase in **wtd18** multiplies the odds of being in somatotype 1 vs somatotype 3 by 0.002(. Because 100(0.002-1)%=-99.8%, the odds are expected to decrease by about 99.8%. Also, a one unit increase in **wtd18** also multiplies the odds of being in somatotype 2 vs somatotype 3 by 0.009(. Because 100(0.009-1)%=-99.1%, the odds are expected to decrease by about 99.1%.

**In other words, children who increase gain weight from age 9 to age 18, may have less chance of being thin, in somatotype1 relative to somatotype 3(chance decrease by 99.8%) and being normal, in somatotype2 relative to somatotype 3(chance decrease by 99.1%), at age18.**

1. A one unit increase in **ht18** multiplies the odds of being in somatotype 1 vs somatotype 3 by 1.617. As 100(1.617-1)%= +61.7%, the odds are expected to increase by about 61.7%. A one unit increase in **ht18** also multiplies the odds of being in somatotype 2 vs somatotype 3 by 1.485. Because 100(1.485-1)%= +48.5%, the odds are expected to increase by about 48.5%.

**In other words, children who is taller in age 18, may have higher chance of being thin, in somatotype1 relative to somatotype 3(chance increase by 61.7%) and being normal, in somatotype2 relative to somatotype 3(chance increase by 48.5%), at age18.**

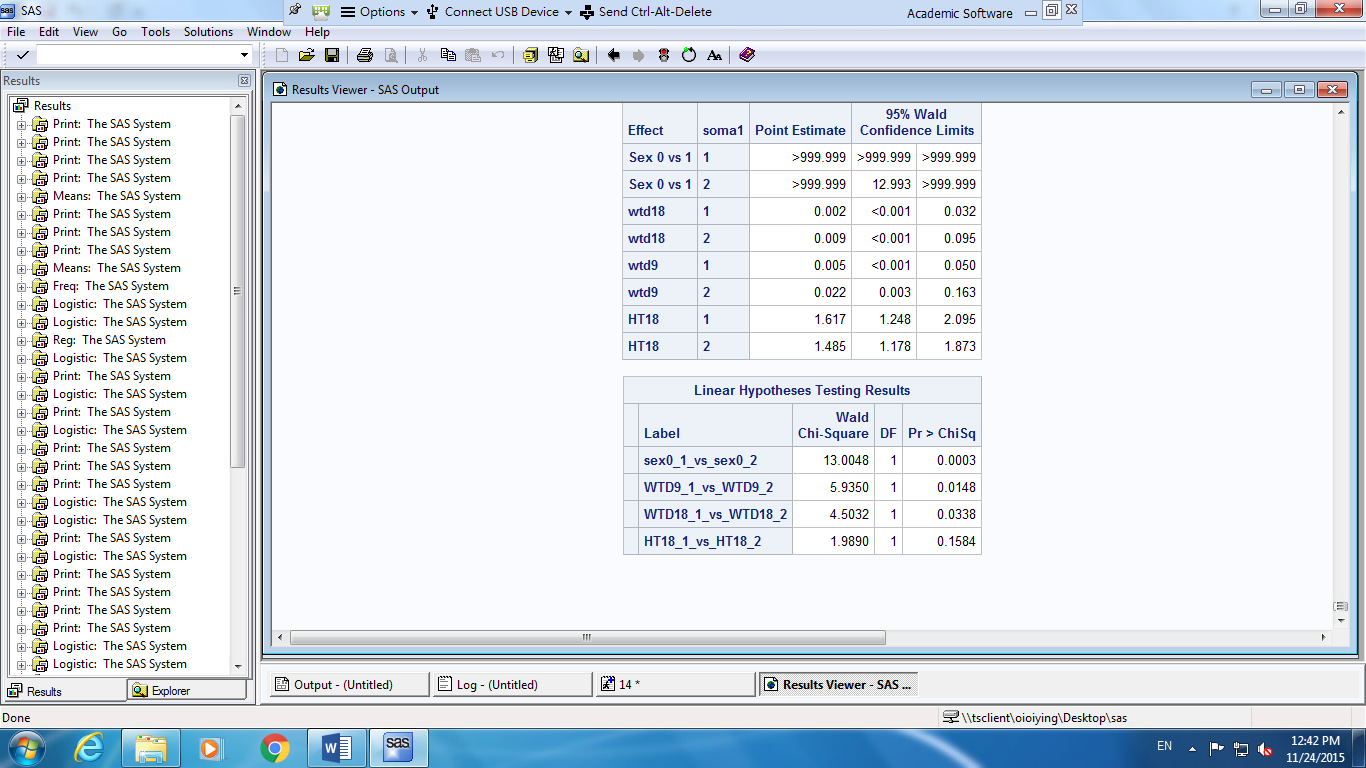
1. The relative log odds of being in somatotype 1 vs. in somatotype 3 will increase by 6.3782 if he is a boy instead of a girl. Similarly, the relative log odds of being in somatotype 2 vs. in somatotype 3 will increase by 3.6944 if he is a boy instead of a girl.

**In the other words, boy will appear to be thinner in age 18 than girl. And the chance for boy to be thin( in somatotype 1 relative to somatotype 3) is higher than to be normal ( in somatotype 2 relative to somatotype 3) at age18.**

**Step 8: Test the equality of parameter estimate for each variables in two models**

To test whether effect of the four variables for predicting somatotype 1 to 3 and for predicting somatotype 2 to 3 is equal or not, we do the following hypotheses testing. The null hypothesis is that the parameter estimate of variables for predicting somatotype 1 to 3 is equal to that for predicting somatotype 2 to 3. The results come out that the effect of **ht18** for predicting somatotype 1 to 3 is not different from that of predicting somatotype 2 to 3 (P-value >0.05). But for the other variables **sex, wtd9 and wtd18**, the effect for predicting is different (P-value <0.05).

**In other words, the height of children at age 18, may have same effect on predicting whether he or she is thin (in somatotype1 relative to somatotype 3) or normal ( in somatotype2 relative to somatotype 3). However, gender of children, weight gain from age2 to 9, and weight gain from age 9 to 18, may have different effect on predicting whether he or she is thin or normal.**



*Table19: Hypotheses testing of equality of parameter estimate in both model*

**Prediction**

For new observation, we need to identify whether he/she is in somatotype 1, 2, or 3. For each observation, we put the value of sex, wtd9, wtd18, and ht18 into both of the model, in order to get the probability of he/she is in somatotype1, 2 and 3. We thus compare three of the probability and use the highest probability for our prediction result. To measure the accuracy of our model, we use the prediction results of 20 new observations to compare with their original results. It comes out that 14 out of 20 datum get the same result as the original. Our model thus have 70% accuracy.

**Conclusion**

Significant predictors for the somatotype predicting model will be gender, weight gain from age 2 to 9, weight gain from 9 to 18, and height in age 18. Results appear that children who increase their weight gain from age 2 to age 9 and from age 9 to age 18, may have less chance of being thin, and being normal, at age 18. In addition, children who is taller at age 18 and who is a boy, have higher chance of being thin or being normal. Moreover, height of children at age18 may have same effect on predicting whether he or she is thin or normal, but other variables may have different effect on predicting whether or she is thin or normal. These model can help to predict the somatotype of children at age 18, so that they can match with the most suitable sports and activities according to their somatotype.

**Limitation on the interpretation and application**

All the data is from children born in 1928-29. As the data were taken long time ago, the model may not suitable for measure the somatotype of children anymore. Moreover, there are only 136 data in the data file and only 112 data left for creating the model after deduce the outliers and prediction data, so more data should be collected in order to get a more accuracy model.