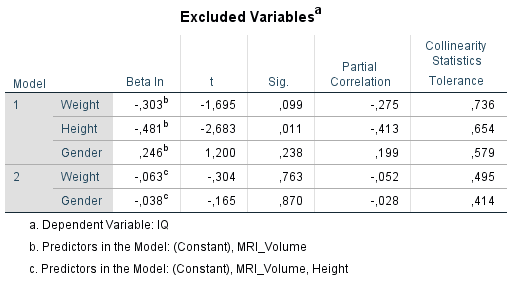
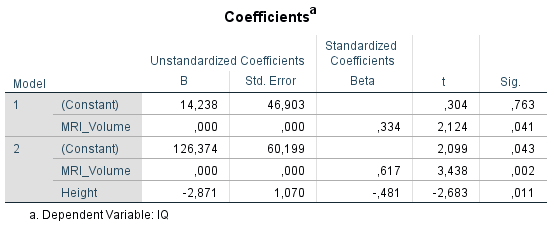
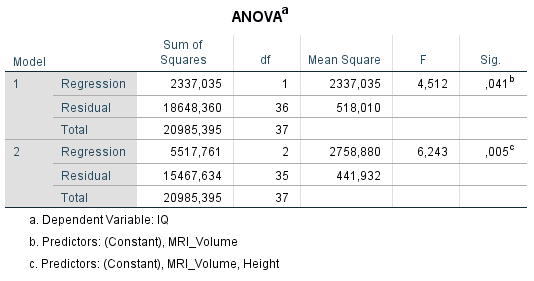
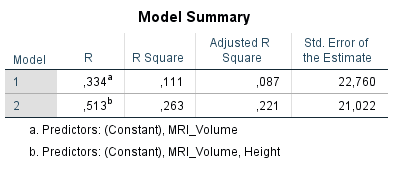
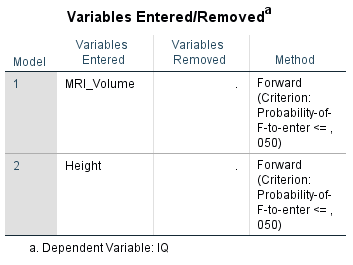
**#this assignment, a class assignment I took from the cognitive science department ("RESEARCH METHODS AND STATISTICS FOR COGNITIVE SCIENCE"), the approach to statistical problems using SPSS was explained.**

**COGS 536**

**Homework 4**

**Leman Nur Erkan**

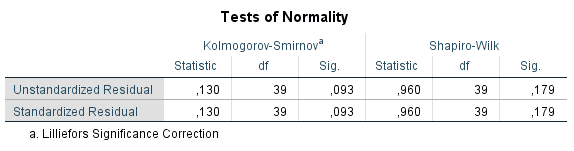
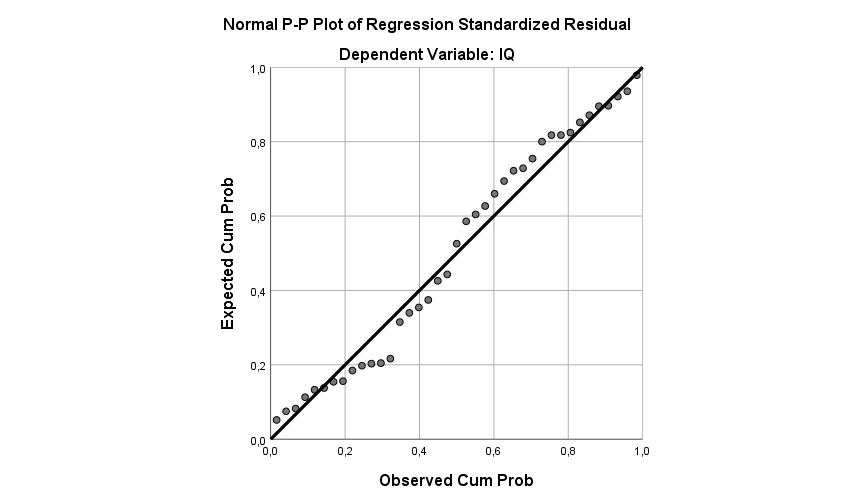
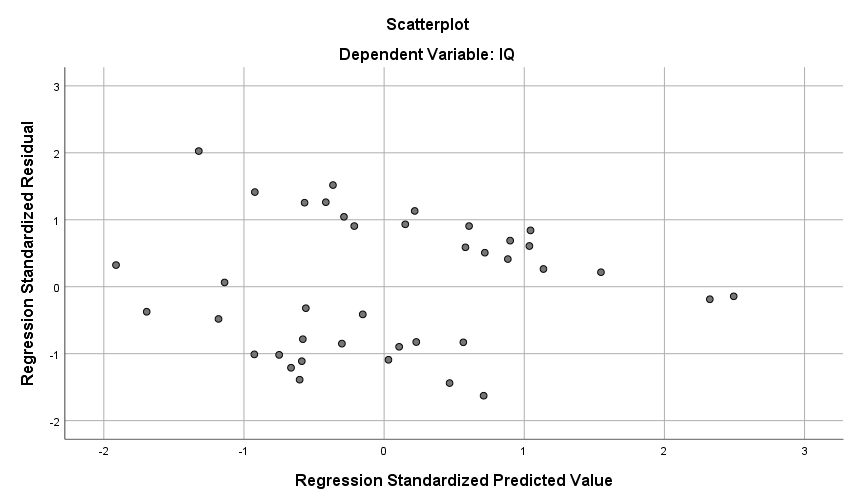
**1. a) **

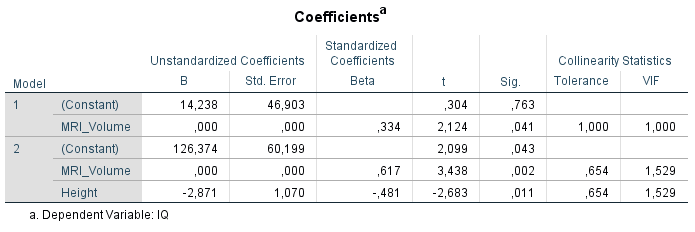
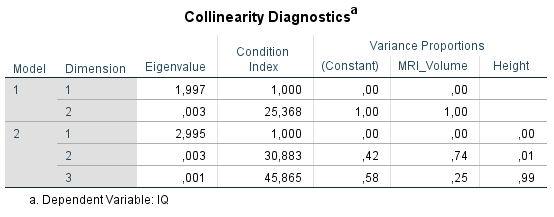
Forward selection is a type of stepwise regression that starts with an empty model and adds variables one at a time. At each further step, a single variable is added to the model that gives the best improvement—the one that gives the best regression equation. This method prioritizes the way that the variables can make the best partial contribution to the equation, not to get the best equation between the variables in the regression. In our model, SPSS generated 2 models to find the best equation and ended the process by choosing the most appropriate regression equation between them. SPSS chose the model containing the MRI\_volume and height variables as the best model. Because the regression equation obtained by using the partial contributions of these variables gave the best adjusted r square value when compared to model 1. Weight and gender were excluded from the model based on their partial contributions to the regression equation. Adjusted r square (which represents the proportion of the variance for the dependent variable) from .087 in model-1; it went up to 221 when switching from 1st model to 2nd model. Since the model is multiple linear regression, the adjusted r square is not looked at directly, but rather at the adjusted r square.

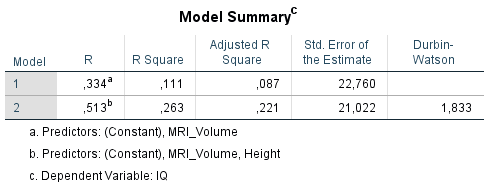
**b)** SPSS chose the model containing the MRI\_volume and height variables as the best model and they are preserved in the final model. They are much more significant predictors in the final regression equation to contribute partially in the equation. Height is the more significant predictor in the equation since it has the slope of -2.871 and comparing to the MRI volume, its contribution to the slope of line is higher. (in coefficients table, looking at unstandardized B part, we can see constants as the constant part of the regression equation -y=ax+b; constant is b- and we can see the a in the equation as the corresponding variable, -2,871 for height). Each variable makes a partial contribution equation, y=ax1+cx2+b; aspect, and the outcome variable will be in the y part of the equation.

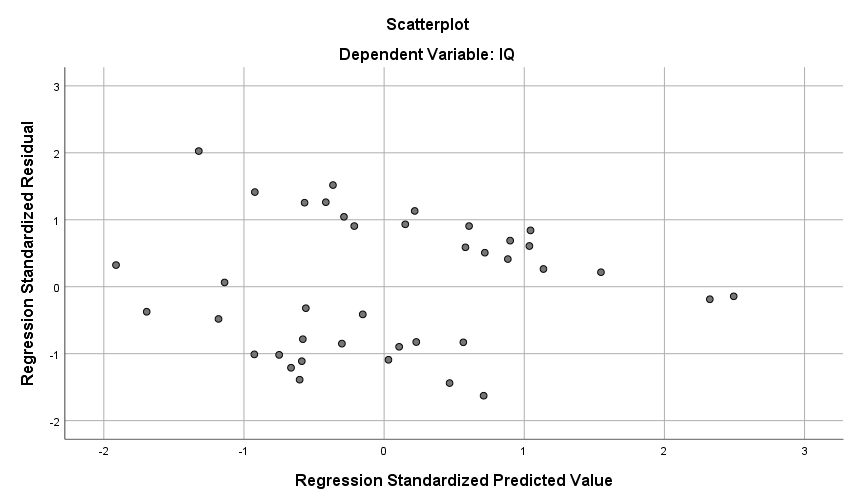
**c)** as y=ax1+cx2+b; therefore; 126.374 +000(mrı\_volume)-2.871(height)

**d)** 126.374 +000(mrı\_volume)-2.871(height) **= IQ --->** in this equation, the IQ value is negative. residual is the difference between the predicted value and real value, since we don't know the real value of IQ in this question, we cannot calculate the residual.

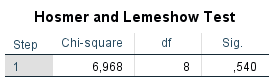
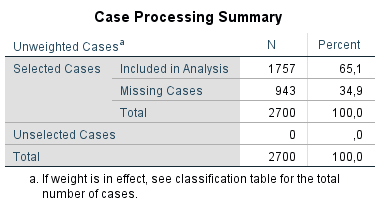
**e)** 

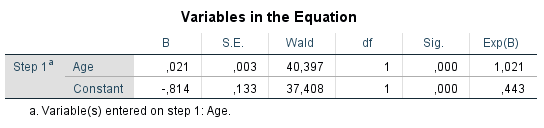
No difference in standardized and unstandardized in the above, we assume they are normally distributed. and significance is greater than 0.005. collinearity statistics is not less than one in the table above, therefore there are no collinearity -variables not highly correlated to each other . -trustable model-



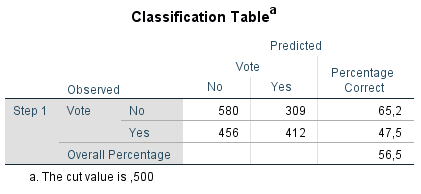
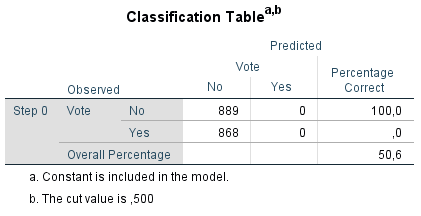
Durbin-watson is close to 2, therefore residuals are uncorrelated. rectangular plot==represents homoscedasticity, → no violation of assumption which is non linearity

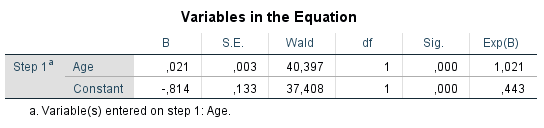
**2.i**

**a) **The Hosmer-Lemeshow statistic indicates a poor fit if the significance value is less than 0.05, this table above indicates a good fit to model therefore.

**b) **

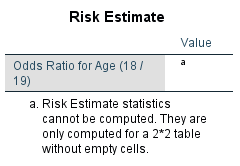
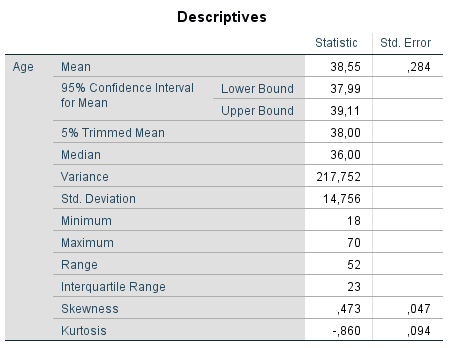
log(p/1-p) = b0 + b1\*x1, therefore, log(p/1-p)=-.814+age\*.021

**c)**  The overall percentage is bigger in the B1 model (56,5) than null model (50,6), therefore prediction power is increased in the model comparing to the baseline model.

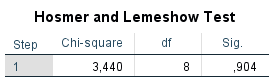
**d)**  for the B part, the independent variables which are not significant, the coefficients are not significantly different from 0, which means age is not a significant variable to predict the outcome.

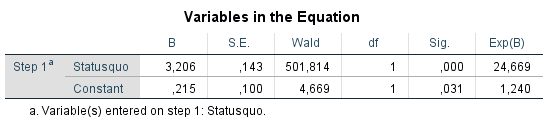
**e)** Exp(B) is the odd ratio for the predictions, which is 1,021 in the above table. In this model -logistic regression- the log-odd ratios converted to the odd ratios to facilitate the model explanations. Odd ratios are the exponentiation of the coefficients. An odd ratio that is close to the 1 means that variables do not affect the odds of property of the dependent variable.

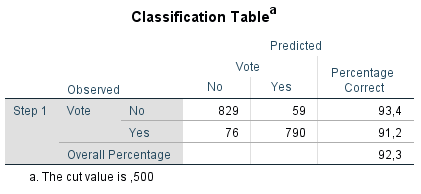
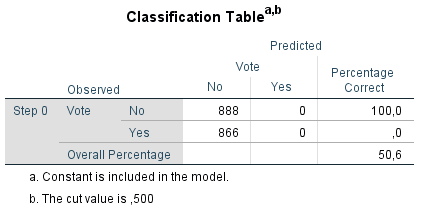
**f)** The confidence interval indicates the level of uncertainty around the measure of effect, if the confidence interval is 1, there is no difference btw the arms of the study. If we exponentiate 0, we get 1 (exp(0) = 1). 95% confidence interval lower bound is closer to the 1, and in there the p-value is very close to .05 and odds ratio is not significant.



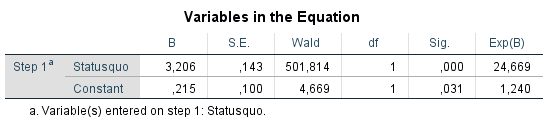
**2.ii**

**a) **The Hosmer-Lemeshow statistic indicates a poor fit if the significance value is less than 0.05, this table above indicates a good fit to model therefore.

**b) **log(p/1-p) = b0 + b1\*x1, therefore, log(p/1-p)=-3.206\*statusquo+.215

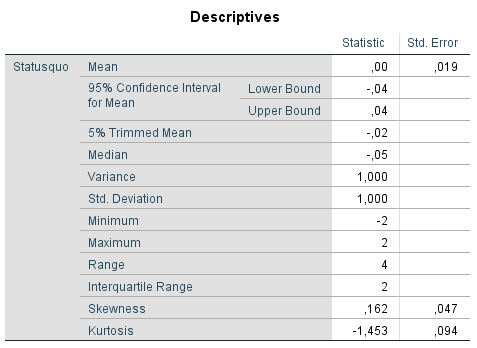
**c) **

improved from 50.6 to 92.3

**d)** 

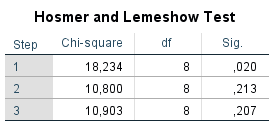
significant contribution to the equation, as above 3

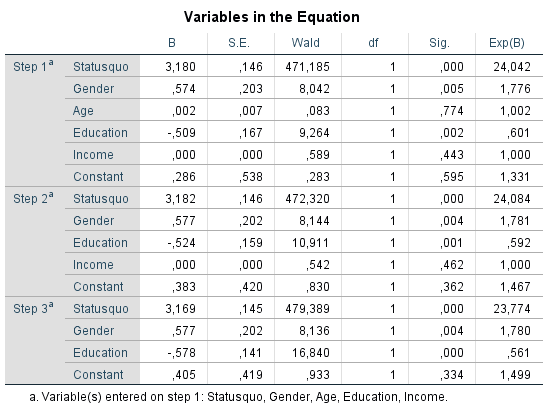
**e)** odds ratio 24→ strong association, since odds ratios measure how many times bigger the odds of one outcome is for one value of a value compared to another value

**f)** 

confidence interval not include 1, which is not significant as mentioned in part 2.i.f.

**2.iii**

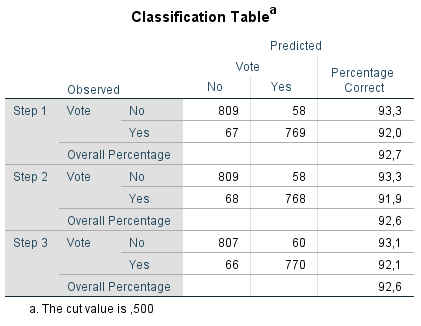
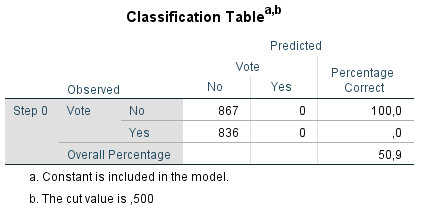
**a)**  the fitness of the 3 different model, increasing through to the third model

**b)** 

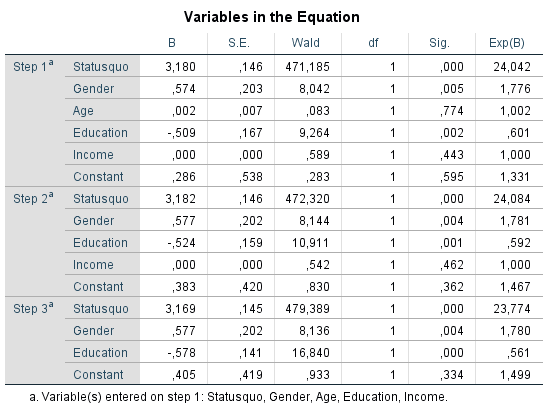
equation of the 3st model: log(p/1-p) = b0 + b1\*x1, therefore, log(p/1-p)=3.169\*statusquo+.405 +gender\*.577+education\*-578

equation of the 2st model: log(p/1-p) = b0 + b1\*x1, therefore, log(p/1-p)=3.182\*statusquo+.383 +gender\*.577+education\*-524+income\*000

equation of the 1st model: log(p/1-p) = b0 + b1\*x1, therefore, log(p/1-p)=3.180\*statusquo+.383 +gender\*.574+education\*-509+income\*000+age\*.002

**c) **

increased among the 3 models compared to the baseline model from 50 percent to higher than 90 percent.

**d)** 

In all models statusqua is the most significant contributor, after that gender and education are contributing significantly and least ones are income and age , based on their B values in the equations.

**e)** odds ratios are the Exp(B) values for the variables in the table above. odds ratio shows the association strength between the variables, statusquo has the most strong relation between the dependent variable, gender, age and income is associated btw dependent variable with odds ratio 1, and least associated is education as near the value of .5

**f)** The confidence interval indicates the level of uncertainty around the measure of effect, if the confidence interval is 1, there is no difference btw the arms of the study. If we exponentiate 0, we get 1 (exp(0) = 1). 95% confidence interval lower bound is closer to the 1, and in there the p-value is very close to .05 and odds ratio is not significant.confidence intervals can be calculated from the odd ratios for each of the variables.

**g)** in step 1, all the variables is entered the model, in step 2, age is removed from the model to make the regression equation better, in step 3, the income is removed from the equation to make the regression equation better and income and age have the lowest significances among the independent variables to predict voting behavior of the individuals.