**Statistical Learning Lab**

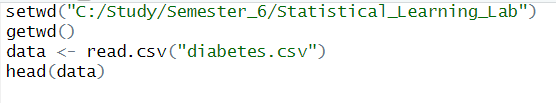
**Assignment – 2: Logistic Regression**

**Name: Semanti Ghosh**

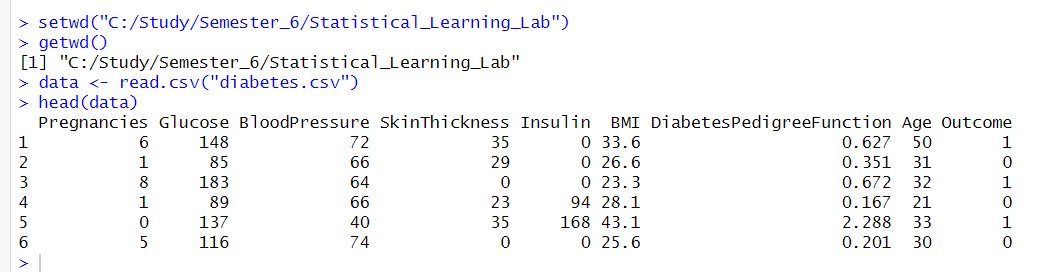
**Roll No.: 22IM10036**

**Loading the dataset**

Code snippet

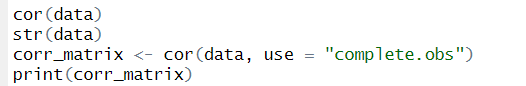


Output

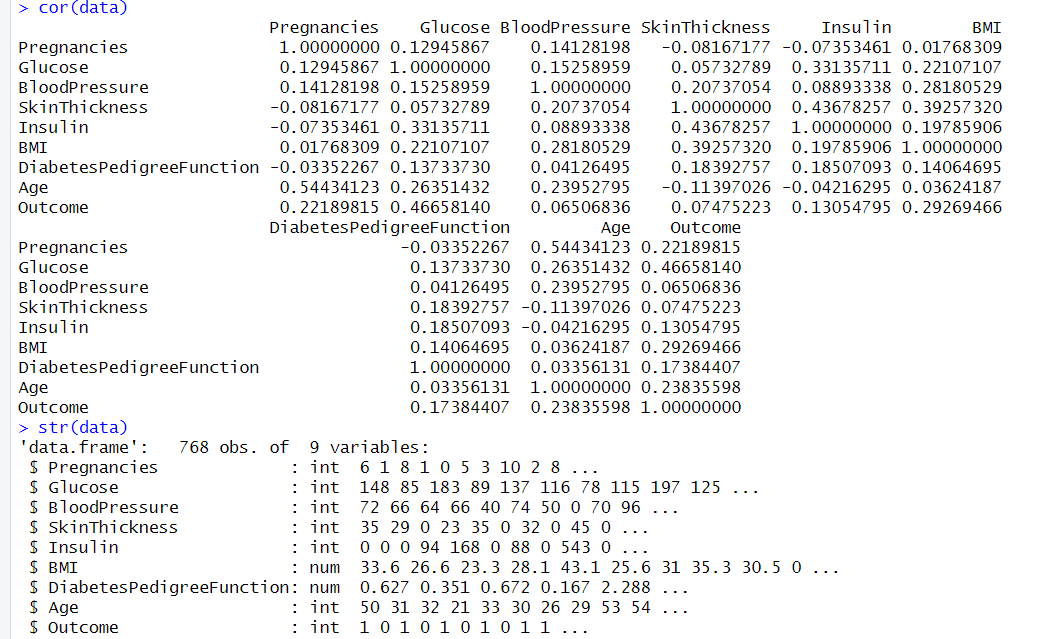


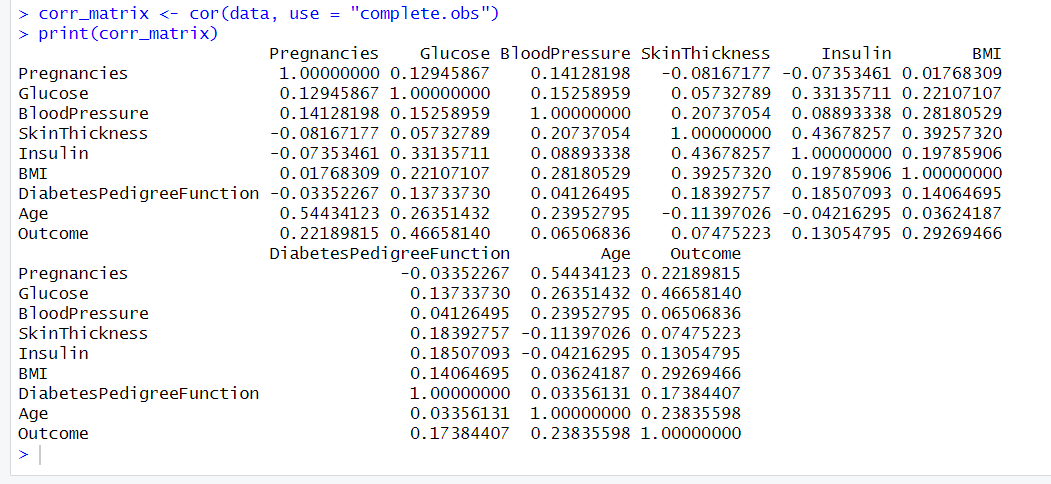
**Preliminary Analysis**

Obtaining the correlation between variables

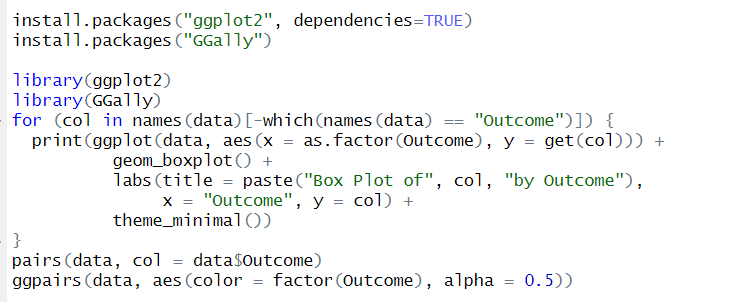


Output

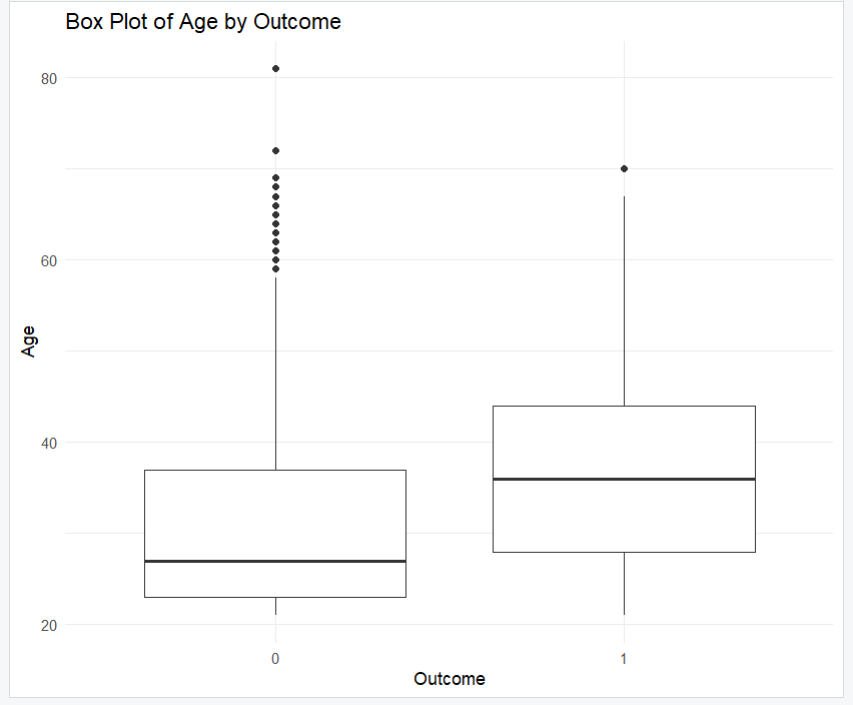




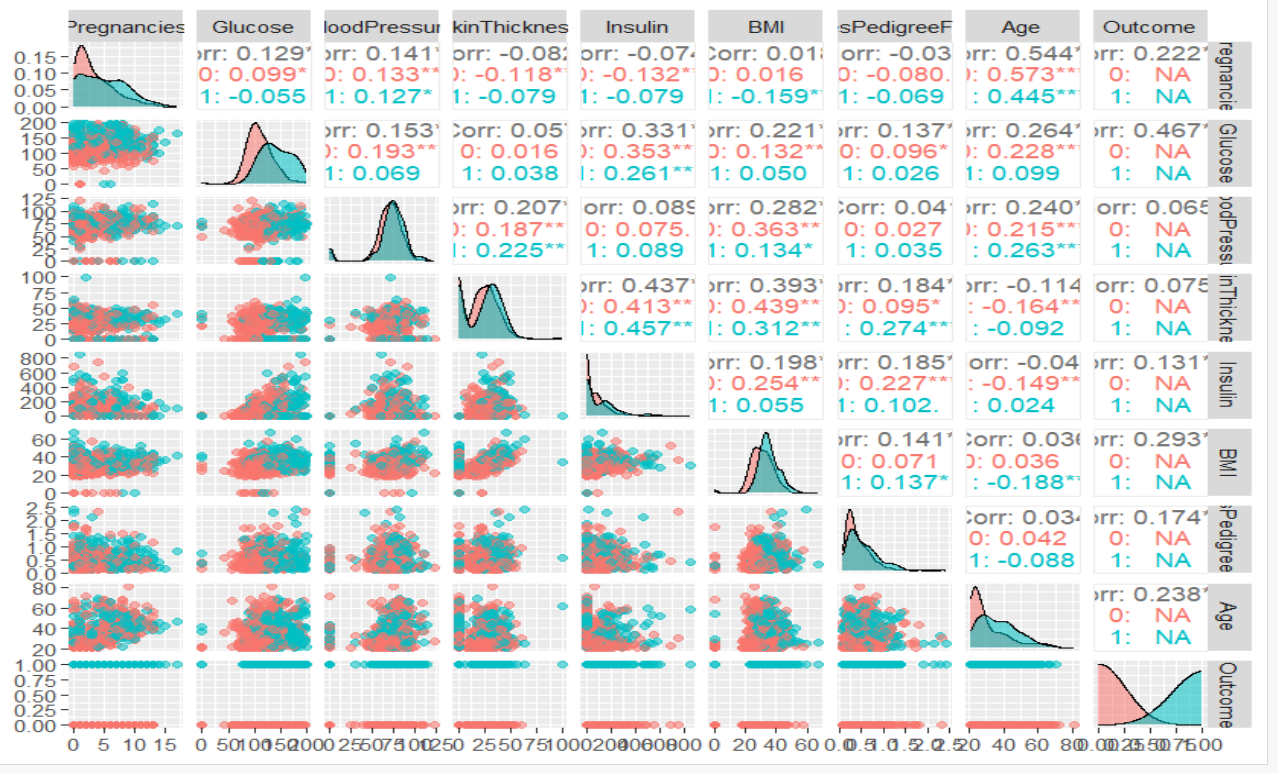
Code snippets – obtaining the box plot and scatter plot



Box plot

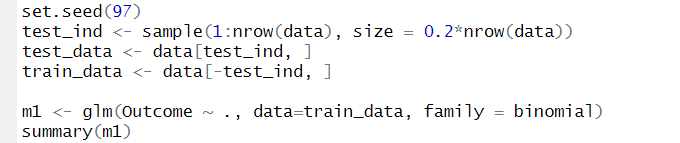
****

Scatter plot

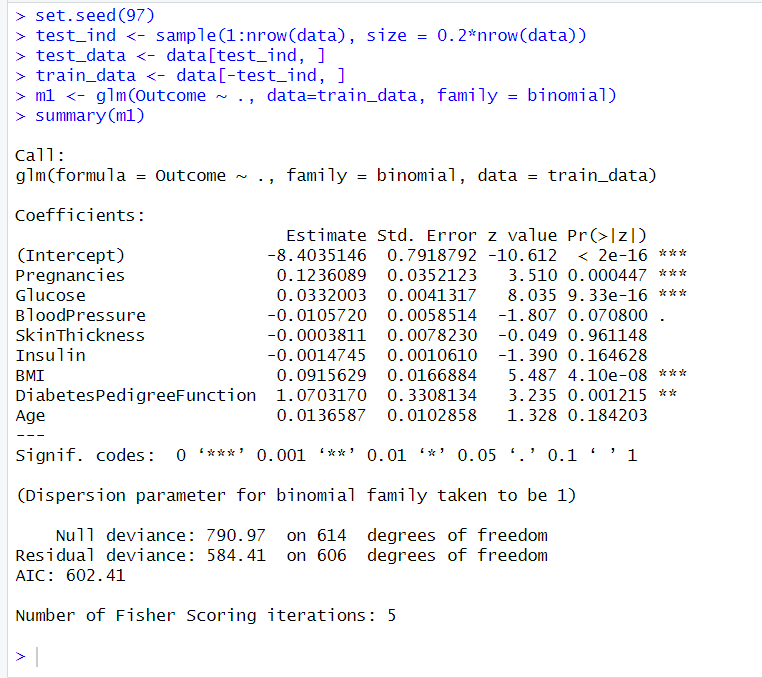
****

**Random Sampling of data and fitting a Logistic Regression model**

Code snippet



Output

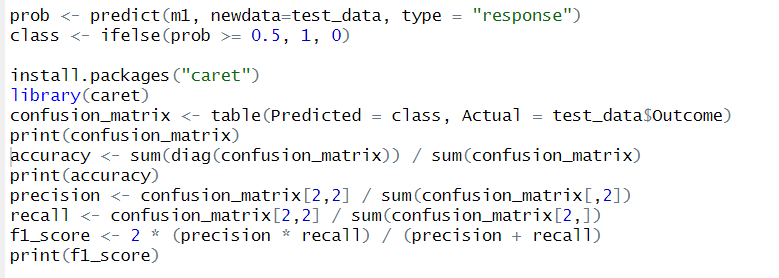


From the summary obtained, the number of pregnancies, the glucose level and the BMI are the most significant parameters (and the intercept is significant too). The Diabetes Pedigree function is also significant, but not as significant as the ones mentioned.

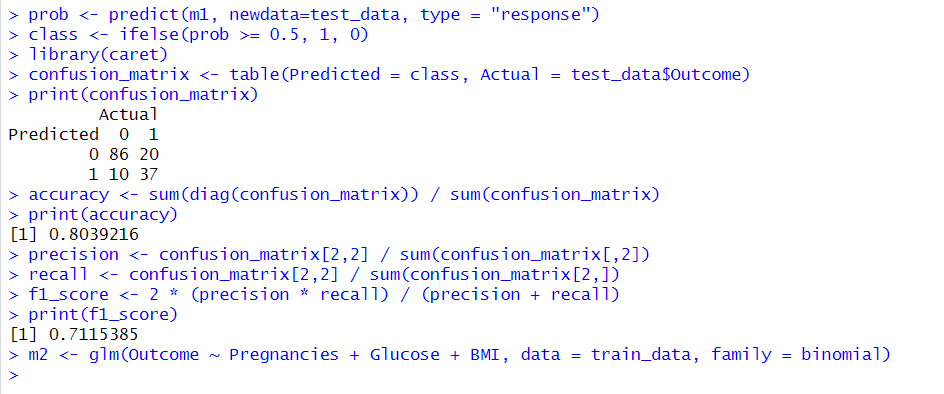
The coefficients of the parameters or the predictors measure the change in log-odds ratio for a unit change in the given parameter. If the coefficient is positive, the log-odds ratio increases with increase in the parameter, and if it is negative, the log-odds ratio decreases.

**Confusion Matrix, Accuracy, F1 Score**

Code snippet

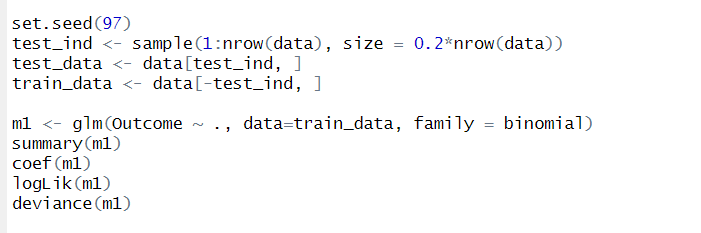


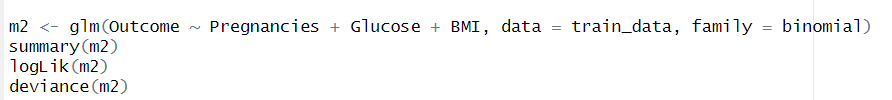
Output



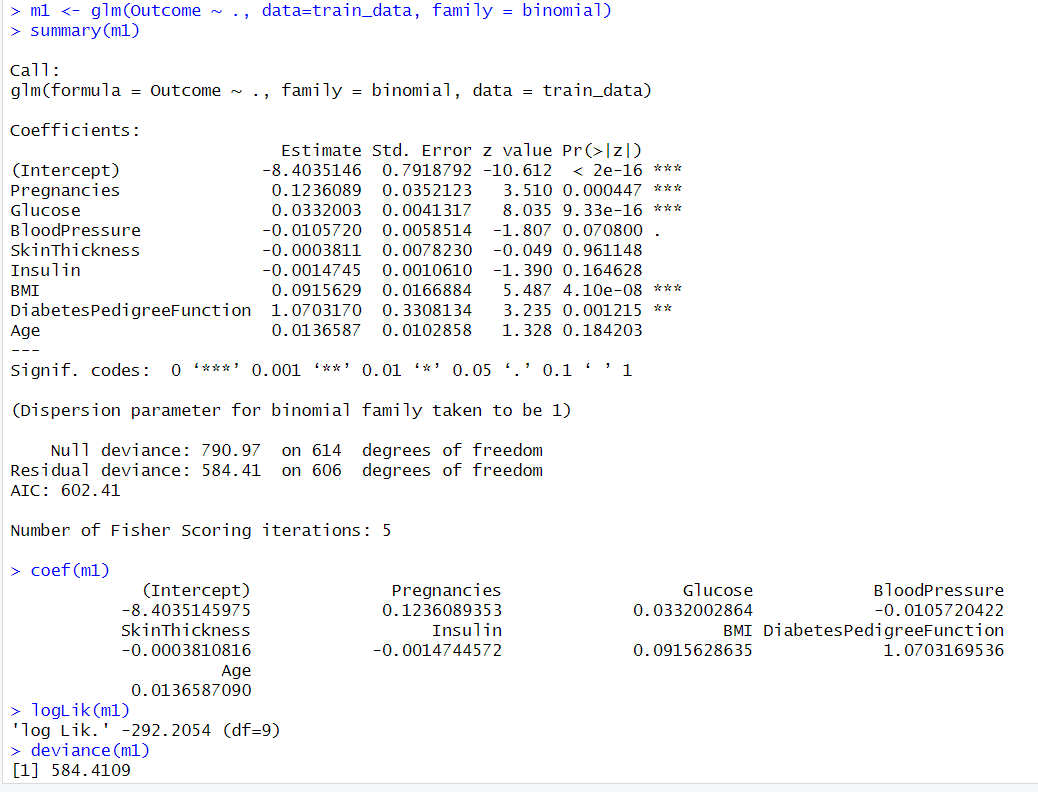
**Fitting and comparing the two models**

Code snippets

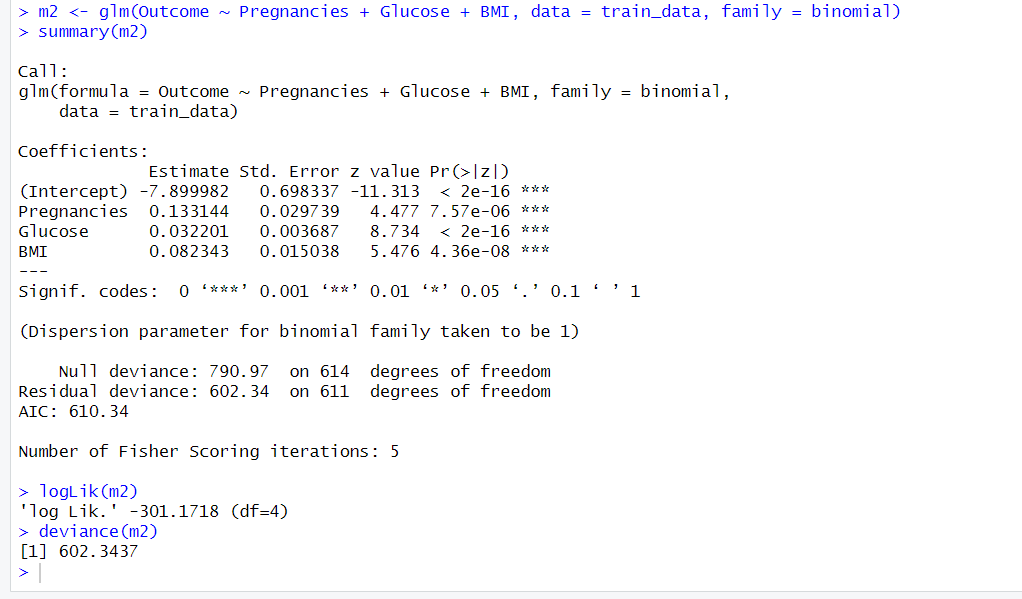




Outputs for model 1



Outputs for model 2



Deviance of model 1 = 584.4109

Deviance of model 2 = 602.3437

Model 1 has lower deviance than model 2, which means that model 1 fits the training data better than model. This is expected considering the fact that model 1 is trained taking greater number of parameters into consideration. However, we also try to reduce the number of parameters taken into consideration since too many parameters or too high powers since then the model becomes prone to overfitting. This might be the problem with model 1.