Lab 12 – Report – Jack Cleary 19333982

Introduction

The purpose of this analysis of the Pima Indians Diabetes Dataset is to firstly fit a logistic regression model to the dataset, then assess the effectiveness of the model using residuals or variable selection methods etc. Then, using the results of this assessment I will fit a new logistic regression model to the dataset that may better fit the data. I will then assess this new model, and compare it to the original in order to make a judgement on which is a better fit to the data. I will carry out this exploration of the dataset using R, and all graphs attached in this report will be created using the same. I will attach the R file which I used.

Methods

I fit a logistic regression model to the dataset using the "glm" function in R. In order to assess this model, I used three different methods. Firstly, I plotted the Pearson residuals for each variable, and also the Deviance residuals for each variable. Lastly, I plotted the Binned Residual Plot for each variable and also for the entire model. An example of the three resulting graphs can be seen below in Figure 1.1, which shows the Pearson, Deviance and binned residual plots for the "Pregnancies" variable.

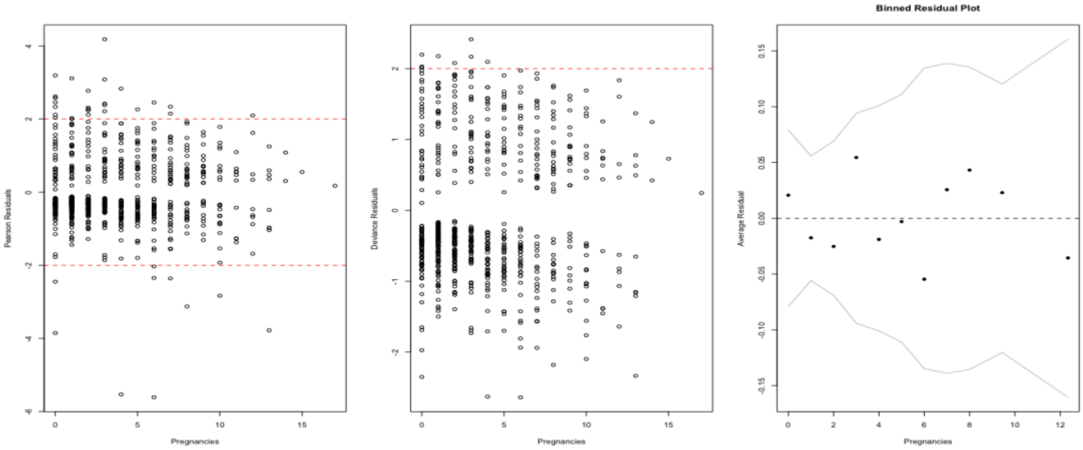

Figure 1.1

This analysis told me a lot about the model. Firstly, the binned residual plot for the entire model, seen below in Figure 1.2, demonstrated to me that our model is flawed. This can be concluded as there are 4 outliers, which is more than the 5% that we might expect from chance alone. Also, the fact that three of the outliers, marked with red, are bunched so close together demonstrates that the model probably doesn't work well for fitted values under 0.1, predicting a higher rate of diabetes than is actually the case. Secondly, the Pearson and Deviance residual plots for each of the different variables in question also showed that there is a flaw in our model. The Pearson residual plot should show values between -2 and 2, and the Deviance residual plot should show values below 2. In a lot of cases however, an example being in Figure 1.1 above, it was evident that a number of residuals were outside of these bounds indicating that our model has potential to be improved.

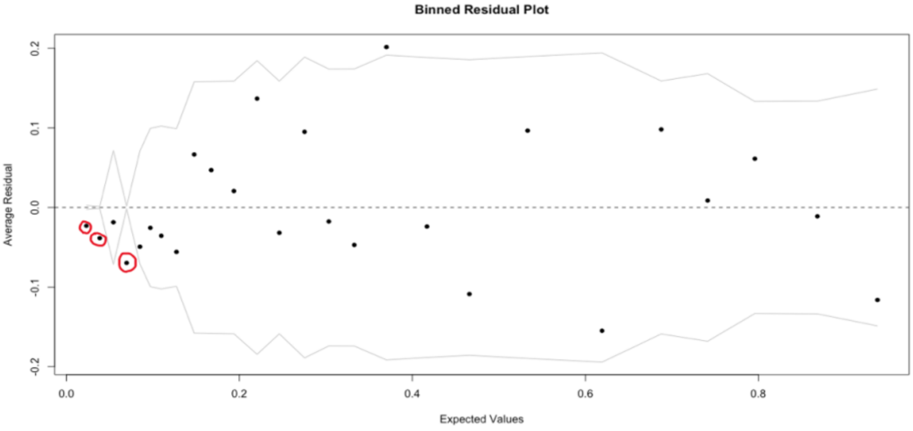

Figure 1.2

On inspection of the binned residual plots, I noticed one variable in particular that stood out. For the BMI, there were 4 outliers, as shown below in Figure 1.3. This demonstrates the fact that in these bins, the model is predicting a higher rate of diabetes than is accurate, which may indicate that the relationship between BMI and the outcome is non-linear. For this reason I chose to change the model and transform this variable by finding the log of the BMI values.

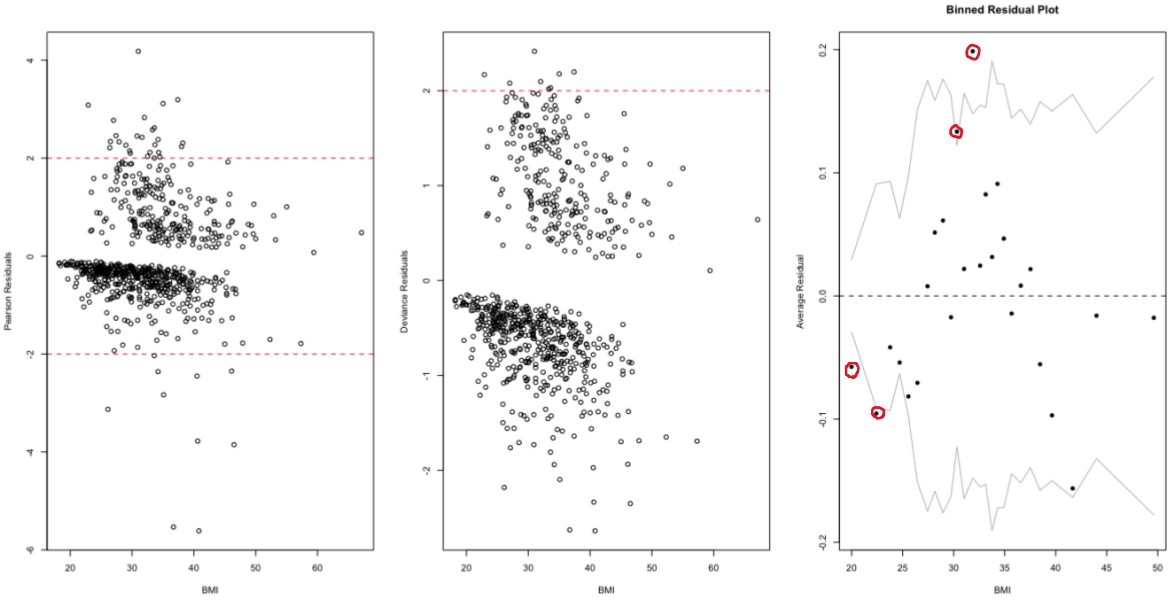

Figure 1.3

Results

The comparison of these two models showed that my initial idea was not as successful as I had hoped it would be. As we can see from the binned residual plot for the second model below, Figure 1.4, very little has changed in terms of accuracy for the 0.1 and under fitted values. However, on a more positive note, the AIC, BIC and Residual Deviance values all decreased from model one to model two. Overall, it is evident that logging the BMI value was not very successful in improving this model. To further try to perfect it, we could try either a different transformation of the BMI values, or we could exclude this variable altogether, focusing solely on Pregnancies, Glucose, Blood Pressure, Insulin, Diabetes Pedigree Function and Age. In these instances we may see an improvement in the model’s accuracy.

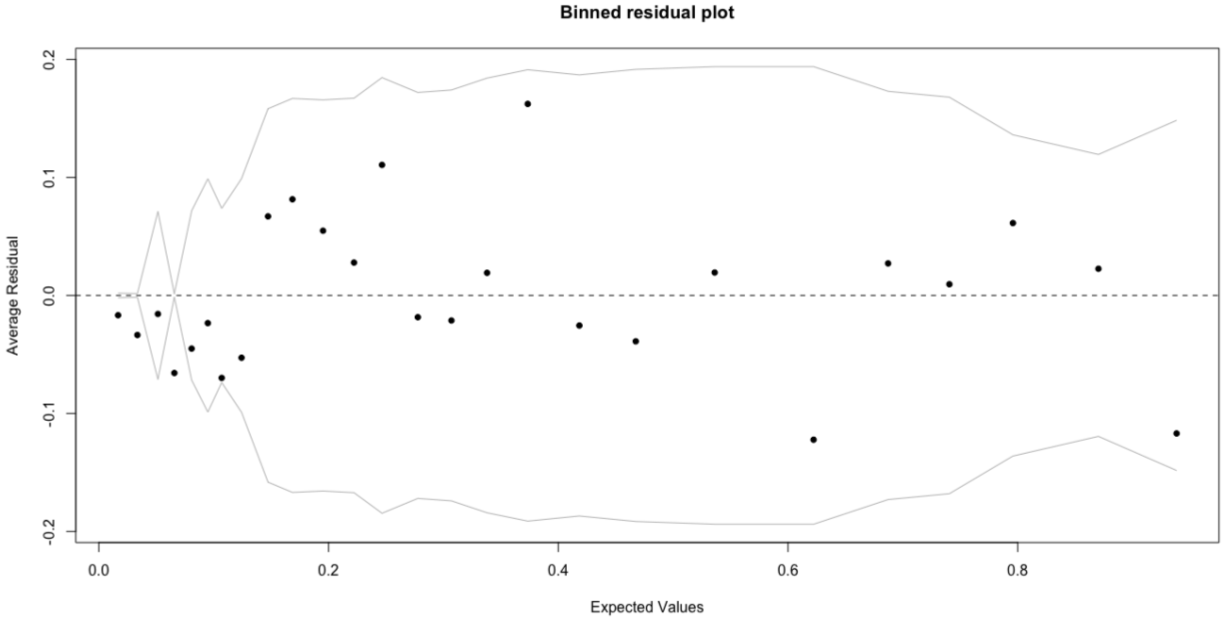

Figure 1.4