Final Project:
Diabetes Dataset

Generalized Linear Models

Rose Gogliotti

East

Contents

Introduction

The goal of this project was to use the diabetes dataset from the faraway package to develop a generalized linear model(glm) for the prediction of a diabetes diagnosis. The diabetes dataset is comprised of 403 African American individuals' test results from a study to understand the prevalence of obesity, diabetes and other cardiovascular risk factors in central Virginia. Although this data was not collected specifically to understand the just prevalence of diabetes, this data will be used to predict when a diabetes diagnosis occurs. A positive diagnosis in the dataset is assumed to be a glycosolated hemoglobin(glyhb) result of greater than 7.

The dataset has 19 variables in total: id, cholesterol (chol), stabilized glucose (stab.glu), high density lipoprotein (hdl), chol/hdl ratio, glyhb, location, age, gender, height, weight, frame, first systolic blood pressure (bp.1s), first diastolic blood pressure (bp.1d), second systolic blood pressure(bd.2s), second diastolic blood pressure (bp.2d), waist, hip, and postprandial time in minutes (time.ppn). As additional factor variable of diagnosis was added to the data using the formula specified above. It is expected that most of these variables will not be useful in the glm. Prior to modeling the variables id, location, glyhb, and time.ppn were removed immediately from the data because it was determined that they have no predictive ability or could not be used to generalize to a new population. Additionally, the variable bp.1s and bp.1d were also removed because the majority of their values were NAs. With these variables removed, all observations with NAs were dropped from the model. This resulted in 367 observations remaining in the dataset, a diagnosis variable and 13 predictor variables. The modified diagnosis variable has 311 negative values and 56 positive values.

All remaining data will be used to fit a model. With an initial model fit, we will use backward selection to reduce the number of variables in the dataset. The reduced model will then be tested to see if it meets the glm assumptions using loess plots. If the model fails to meet these assumptions, variables will be modified using splines. A new model will be split with the splines if necessary then the model will then be tested for goodness of fit. The model will be explored to see if outliers exist in the data that are impacting our results. Then finally we will evaluate the developed model and see how well it is able to discriminate positive and negative diagnosis.

Step 1: Exploratory Data Analysis

- Interpretation of a scatterplot: Scatterplots show the relationship between two different variables. A dot represents a single data point.
    - Diagnosis vs stab.glu: This plot shows that stab.glu negative diagnoses tend to have values under 150 while positive diagnoses seem to be more evenly spread for the stab.glu variable. Jitter or variation is added to this graph to visualize points that may appear more than one time.

    - Diagnosis vs waist: This plot shows that positive diagnoses tend to not have waists of less than about 33. Similarly, all waist values above 53 have positive diagnoses. In between those values I do not see a clear trend. Jitter is also added to this graph
- Interpretation of an interleaved histogram: A histogram is a common way of visualizing the range and frequency of occurrence of a variable. Count is shown on the y-axis and the variable of interest is shown on the x-axis. The shape of a histogram helps to show the distribution that the data follows. An interleaved histogram adds an additional variable and shows frequencies side by side. This type of visualization can identify trends in the data.
    - Stab.glu: This plot shows that negative diagnoses in stab.glu variable tend to follow a fairly normal distribution with a median of around 80. Conversely the positive diagnoses do no show the same normality trend and appear largely undefined.
    - Waist: This plot shows that negative diagnoses in the waist variable tend to follow a normal distribution with a median around 35. It is difficult to tell based on the small number of positives in the data, but it appears as though positive diagnosis also follow a normal distribution with the median centered around 40.

Step 2: Variable Selection

- Variables with more than 5% of the data missing were removed from the data. These variables were bp.2s and bp.2d. They each ended up having 262 NA values. With a large amount of missing data, they are not useful predictors and would cause substantial data loss when NAs are dropped.
- The variables glyhb, id, time.ppn and location were removed from the dataset prior to modeling. Glyhb can no longer be used because it was directly used for prediction. The other variables were removed because they were determined to not have predictive abilities. For instance, location is not useful in the model because it can’t be applied to a bigger population
- The variables gender and frame were converted to factors from numeric variables. In order to ensure they are treated correctly in modeling, this was an important step. Additionally, NA values were dropped.
- A binary generalized linear model was fit to the data using diagnosis as a response and all remaining variables as predictors. A binary model was selected because all covariate classes had less than 5 members. The initial model violates the rule of 5, which for this dataset with 56 remaining positives in the data can only use 11 predictors. This will be addressed in the next step.
- The step() function was used to reduce the predictors used in the model and select a more powerful model. The function works by systematically removing variables one at a time, if removing the variable lowers the AIC, the variable is removed. This process continues until removing an additional variable will not reduce the AIC further.
- The step function selected using the variables chol, waist, age, and stab.glu as predictors. I find that the variables chol, waist, age, and stab.glu make a lot of sense that they are

predictive in that they are commonly given as risk factors for diabetes. I was surprised to see that weight was removed because being overweight can lead to diabetes. I suspect that this may be because your weight is not the only determinant of being overweight (ie. Gender, height and frame impact the determination of being overweight).

Step 3: Assess Model Fit

- Loess Plots Interpretation: Loess plots can be used to verify that continuous predictors have a linear relation with the logit. For this graph, a prediction was run on the loess function which helps smooth data for visualization. The values of this prediction that fell between 0 and 1 were plotted on the y-axis as a log odds. The corresponding x-values were plotted along the x-axis. If a linear trend is observed the variable is interpreted to meet the logistic regression assumptions. If a linear trend is not apparent, then the variable does not meet the logistic regression assumptions.
    - Linear trend: The chol and waist variables loess plots have fairly linear trends apparent. Notably, they are not perfectly straight lines but a linear trend can be inferred. These variables meet the logistic model assumption that the x and the logit have a linear relationship
    - No linear trend: The stab.glu and age loess plots do not follow a linear trend. The age variable has a more sigmoidal plot and the stab.glu variable plot has a more arched trend. Neither variable meets the assumptions of the logistic model.
- Splines: Splines can be used to address the failure of the age and stab.glu variables to meet the logic assumptions. They are line segments that help to approximate the curve. Knots are the points that these segments cross the line. As a rule of thumb, knots are located at the 10th, 50th and 90th percentiles so that was done for both the age and stab.glu variables.
- From these 3 knots , 4 splines are created for each variable. The splined variables were then fit back into the model in place of the original variable. This new model has an AIC of 168.72 which is lower than the reduced model of 173.4 and indicates an improved fit.
- The Hosmer-Lemeshow (HL) test was used to check for goodness of fit in the model fit in the last step. This test had a p-value of 0.68 which indicates that the model fit is adequate.

Step 4: Model Inferences

- Checking for influential observations: An influential point is defined as a point that's presence impacts the model. In other words, its removal will impact the model slopes. Checking for influential points can be done using Cook's Distance.
- No points were identified by Cooks distance as outliers using the assumption that absolute values of the Cook's distance are greater than 1 are potentially influential. Since no points were identified, no points were removed from the model.
- It was noticed that the p-value of the age splines and waist variables were high so these variables were removed and a nested models test was run on the larger and smaller models. This test resulted in a p-value of 0.07 so at a alpha of 0.05 it was determined that

these variables needed to remain in the model though a modestly higher alpha would cause a different conclusion.

- For splines, confidence internal don’t have much meaning so they are not reported for the spline variables.
- For the chol variable a confidence interval of (0.0015, 0.0215) with a p-value of 0.06 was found for the slope. This can be interpreted to mean, “we are 95% confident that the true population value of chol slope lies within the interval (0.0015, 0.0215)”.
- For the waist variable, a confidence interval of (0.0321, 0.1173) with a p-value of 0.26 was calculated for the slope. This can be interpreted to mean, “we are 95% confident that the true population value of waist slope lies within the interval (0.0321, 0.1173)”.
- Confidence intervals and p-values for the spline slopes are not meaningful so they are not reported.

Step 5: Assess Predictive Power

- ROC Curve: the ROC curve can be used to identify how well a model is able to discriminate between the positive and negative diagnosis classes. In this plot the true positive rate is plotted on the y-axis and the false positive rate is assigned to the x-axis. Curves that are closer to the upper left corner of the unit square indicate better discrimination.
  - In our plot, we see that the line is very close to the left corner. This indicates that our model has good discrimination. That is to say that it can discriminate between positive and negatives diagnosis never well.
- Area Under the Curve(AUC): To numerically quantify the ROC curve, the area under the curve can be tabulated. The lowest AUC possible is 0.5 and indicates poor discrimination while the closer the AUC gets to 1 the better the discrimination.
  - Our graph produced an AUC of 0.95 which indicates that the model has excellent discrimination. A high AUC was expected based on visual inspection of the ROC Curve.
- Positive and Negative frequency graphs: These plots show the positive and negative classification based on the predicted probabilities. In the negative plot we see that almost all zeros were predicted zeros with very few predicted as positives. In the positive plot we see that most positive values are predicted correctly however positives are identified correctly at a lower rate the negatives. Notes that the y-scales on these plots are different orders of magnitude.
- Fitted values were also plotted against observed values as a jittered scatterplot. From the scatterplot we see a different visualization that the above histograms of the same data. The large cluster of points close to (0,0) are correctly identified negative values. The smaller congregated points in the upper right corner are correctly identified positive values. The remaining more dispersed points correspond to points where the observed and fitted values do not match.

Conclusions

The final glm used waist, cholesterol, and a 4 knotted spline for age and stab.glu to predict diagnosis. It has an AIC of 168.72 and a model deviance of 146.72 on 356 degrees of freedom. Overall the modeled glm performed very well. The ROC curve and AUC values show that the model has excellent discrimination, which means it does well in discriminating between positive and negative diagnosis. Additionally, when we look at the frequency histograms, we see that most positive values are correctly identified.

This model could be improved with the addition of more data with positive diagnosis. As can be seen in the table summary of the response variable, the data is highly imbalanced with positive diagnoses. Imbalanced data can be hard to predict and model with. The impact of this can be seen clearly in the fitted values vs observed values scatterplot. We seen that despite having fewer positive diagnosis, more positive than negative points are incorrectly predicted.

Additionally, I am somewhat concerned the data model overfit the data. This means that the model predicts well on the data it was trained with but performs poorly with new data. This could have been quantified if I had split the data into testing and training datasets before modeling. New data could also be used to test if overfitting occurred.

With additional time I would like to explore the effects of using all variables and the splined variables with forward selection. Backward selected the variables as significant prior to them having splines. I am interested in how splines would impact those results. Many of the p-values for the spline knots are low, and I suspect the age spline would be removed through the step function. I did test if the variables age and waist could be removed from the model using a nested model hypothesis test but the p-value of 0.07 does not suggest they are that important and they would be removed at a different alpha.

Finally, I would like to talk to as subject matter expert on other variables and variable combinations that might be important predictors. For instance, I suspect creating a new variable that combines other variables deemed insignificant such as BMI might be a useful predictor. Also, I do not know enough about several of the variables to make informed decisions on which others might be meaning to be combined.