Important predictors in stroke prediction

Saji Nammari and Markos Meytarjyan

Introduction

Stroke is a leading global health issue, with over 13 million people affected each year, according to the World Health Organization (WHO). It is the second leading cause of death worldwide, and survivors often experience long-term disability and impaired quality of life. Stroke also affects different populations differently, as highlighted by research on the impact of sex and gender on stroke occurrence and outcomes (Rexrode et al., 2022). Prevention and early detection are crucial to reduce the economic and social burden of stroke. Statistical modeling can aid in this effort by providing valuable insights into the risk factors associated with stroke occurrence and the development of targeted prevention strategies.

Recent studies have shown the potential of machine learning algorithms to predict stroke mortality accurately. Schwartz et al. (2023) conducted a systematic review, demonstrating the ability of these models to identify individuals at high risk of stroke and implement targeted interventions. Such data-driven approaches can help healthcare professionals identify high-risk patients who may benefit from early interventions and preventive measures.

For example, a study by Rexrode et al. (2022) investigated gender-specific risk factors for stroke, such as pregnancy, menopause, and hormone therapy use in women. Identifying such risk factors can help healthcare professionals tailor preventive strategies to specific patient groups and improve stroke prevention efforts.

Our project aims to build on this research by developing a predictive model for stroke occurrence based on multiple risk factors, including age, hypertension, heart disease, smoking, gender, work type, residence type, average glucose level, and BMI. We will analyze a dataset that includes information about a patient's medical history, lifestyle, and demographic characteristics using statistical techniques, including logistic regression and lasso regression. Our goal is to identify the most significant risk factors for stroke occurrence and advise targeted prevention strategies to improve patient outcomes. We predict that our statistical analysis will reveal several risk factors that are strongly associated with stroke occurrence. Specifically we anticipate that older age, hypertension, heart disease, smoking, and higher average glucose levels will be significant risk factors.

Matrials and Methods

The "Stroke Prediction Dataset" was used in this project, which is publicly available on Kaggle.[1] The dataset comprises demographic characteristics, lifestyle factors, and medical history of 5,110 patients. The source of the data is confidentially collected, and there is no information regarding the data source.

The dataset includes several important variables for our project. The Gender variable is a categorical variable indicating the gender of the patient, which can be used to investigate the impact of sex and gender on stroke occurrence and outcomes. The Age variable is a quantitative variable indicating the age of the patient in years, which is an important risk factor for stroke. The Hypertension and Heart disease variables are categorical variables indicating whether the patient has hypertension or a history of heart disease, respectively, which are two significant risk factors for stroke. The Average glucose level and BMI variables are quantitative variables that can provide information about the patient's overall health status, which can be used to investigate the impact of lifestyle factors on stroke occurrence. Finally, the Smoking status variable is a categorical variable indicating the patient's smoking status, which is a well-known risk factor for stroke.

The dataset was cleaned by removing the multiple NAs found in the BMI column, which was initially set as character type. After changing the BMI column to double type, we found that 201 rows had missing data in the BMI column. We decided not to drop the rows instantly as we won't be using the BMI column the most, and we will only drop the rows when we are using the BMI column during our modeling. Additionally, we changed the ever_married column to a 1, 0 binary instead of Yes/No, and we removed one row from the gender column that was listed as "other" for consistency purposes. [2]

When building the models, several strategies were employed to create a strong model with the best selected variable. We aimed to develop a binary classification model to predict occurrence of strokes based on various demographic and health related factors. Initially we separated the predictor variables and the response variable, stroke. Then we fit the data into a lasso regression model in order to select the best variables, then plotted the coefficients to visualize the shrinkage of coefficients in the lasso model. Additionally cross validation was done to ensure the accuracy of the lasso model. After the variable selection process, we fit a multitude of logistic regression with the different selected variable. This process ensured us that we can compare the logistic regression model with all the different combination of the selected predictor variables. Along with the logistic regression models that only took into account the predictor variables, we built multiple logistic models with the interaction between the selected predictor variables to capture potential interaction effects. When comparing the models in order to build a final model, we performed anova tests to compare the fitted models and assess their statistical significance to select the final model. After that we created empirical logit plots to visualize the relationship between the predictor variables and the log odds of stroke occurrence to ensure that the final model was the strongest fit. Finally we performed several assessment techniques to ensure the accuracy of our final fit.

Results

Our exploratory data analysis reveled that the majority of patients in our data set were female, above 40 years old, married, did not have hypertension or heart disease, and did not have a stroke. The lasso regularization method was used to select the optimal predictors/risk factors to predict the occurrence of a stroke. The results of the variable selection indicate the significance of the coefficients for the predictor variables. The variables that had a positive association with the likelihood of stroke occurrence are age, at 0.045, hypertension, at 0.0343, heart_disease, at 0.0423, and avg glucose level, at 0.0014, the rest of the variables where excluded from the selected model.

After the variable selection process, we fit four different models, each including different combinations of predictor variables: age, hypertension, heart disease, and average glucose level. The aim was to identify the most informative variables for predicting stroke. Using the anova test we were able to analyze the results of comparing the models. When comparing a logistic model with all the selected variables to one where hypertension is excluded the results shows a deviance of -5.28 with 1 degree of freedom and a PValue of 0.02157. Since the PValue is below the significance level it is understood that excluding the hypertension variable significantly worsens the model fit. When comparing the full logistic model to one where heart disease was excluded, the results shows a deviance of -2.97 with 1 degree of freedom and a PValue of 0.085. Since the PValue is above the significance level it is understood that excluding the heart disease variable doesn't affect the model fit much. When comparing the full logistic model to one where average glucose level was excluded, the results shows a deviance of -12.144 with 1 degree of freedom and a PValue of 0.00049. Since the PValue is extremely small it is understood that excluding the average glucose level variable significantly worsens the model fit. When comparing the full logistic model to one where age was excluded, the results shows a deviance of -255.93 with 1 degree of freedom and a PValue < 2.2e-16. Since the PValue is extremely small it is understood that excluding the age variable significantly worsens the model fit.

Subsequently we fit different logistic models, each including different combinations of predictor variables and their interactions. When comparing a model with the interaction between age and hypertension to one with with only selected variables, it was observed that excluding the interaction term significantly worsened the model fit (deviance = -286.52, $p < 2.2e\text{-}16$). Similarly, the comparison between the model with the interaction between average glucose level and hypertension and to the selected variable model revealed that excluding the interaction term significantly deteriorated the model fit (deviance = -287, $p < 2.2e\text{-}16$).

Additionally, the comparison between the model with the interaction between average glucose level and age and the selected variable model showed that excluding the interaction term significantly worsened the model fit (deviance = -285.81, p < 2.2e-16). On the other hand, when comparing a model with all the interaction terms to one with only the age and hypertension interaction, excluding the interaction terms age:avg_glucose_level and avg_glucose_level:hypertension did not have a significant impact on the model fit (deviance = -1.6053, p = 0.4481). Similarly, the comparison between a model with all the interaction terms to one with only the average glucose level and hypertension interaction, indicated that excluding the interaction terms age:avg_glucose_level and age:hypertension did not significantly deteriorate the model fit (deviance = -1.1312, p = 0.568). Additionally, the comparison between a model with all the interaction terms to one with only the average glucose level and age interaction revealed that excluding the interaction terms avg_glucose_level:age and age:hypertension did not significantly affect the model fit (deviance = -2.3224, p = 0.3131). In the comparison between a model with the interaction terms avg_glucose_level:age and avg_glucose_level:hypertesnion to a model with only the avg_glucose_level:hypertesnion interaction, excluding the interaction term avg_glucose_level:age did not significantly impact the model fit (deviance = -0.25182, p = 0.6158). Similarly, the comparison between a model with avg_glucose_level:age and age:hypertension interaction to to a model with only the avg_glucose_level:hypertesnion interaction showed that excluding the interaction term age:hypertension did not significantly affect the model fit (deviance = 0.26089, p = 0.4915).

Finally after the our model comparison process the final logistic regression model was developed to predict the likelihood of stroke using age, hypertension, and average glucose level, including their interaction term between average glucose level and hypertensiopn. Empirical logit plots were generated to explore the relationships between these variables and stroke occurrence. The plots revealed important statistical findings. Firstly, higher average glucose levels were associated with an increased risk of stroke, especially when hypertension was present. the empirical logit plot showed a clear relationship between higher glucose levels and increased stroke risk. The estimated logit values further confirmed this association, with a negative logit value of -3.231 for the lowest glucose level group (Group 1) and a progressively higher logit value of -1.857 for the highest glucose level group (Group 20). This indicates a significant positive relationship between average glucose level and the likelihood of stroke. Secondly, advancing age was linked to a higher probability of stroke. For age, the empirical logit plot demonstrated a consistent trend of increasing stroke risk with advancing age. The estimated logit values supported this finding, with a negative logit value of -6.212 for the youngest age group (Group 1) and a progressively higher logit value of -1.443 for the oldest age group (Group 20). These results indicate a significant positive association between age and the probability of stroke.

The performance of the final logistic regression model was evaluated through measures such as accuracy, sensitivity, specificity, and Area Under the Curve (AUC). By applying a prediction threshold of 0.2, the model demonstrated an accuracy rate of 79.7%. It was observed that all predicted probabilities using this model were below 0.5. This implies that using a conventional prediction cutoff, all patients would be classified as non-stroke victims. Therefore, in an effort to augment the correctly predicted stroke cases without jeopardizing the accurate identification of non-stroke patients, a significantly lower threshold was chosen. This decision led to a sensitivity of 80.2% and a specificity of 71.1% at a cutoff of 0.2. The Receiver Operating Characteristic (ROC) curve [7] revealed that, while the model's results were not optimal, its performance was substantially superior to a random predictor, as evidenced by the ROC curve's substantial deviation from the diagonal line extending from (0,0) to (1,1). The model's performance was further reinforced by the AUC value of 0.844, indicating that the logistic regression model exhibited notable effectiveness.

Discussion

Our findings provide valuable insights into the relationship between several risk factors and stroke occurrence. With age, average glucose level, and hypertension standing out as significant predictors, our results indicate the need for increased attention to these factors in stroke prevention strategies. Also, the exploration of interaction terms revealed potential moderation effects, suggesting that risk factors may interact to influence stroke occurrence. Regarding the generalization of our results, it is important to note that our study was based on a specific dataset, which may not represent the global population. For instance, the dataset had more female participants and patients above 40 years old. However, it is important to note that this sample may not be representative of the general population, as the source of the data is not specified, and it is listed as being confidentially collected. Furthermore, the data collection process was not disclosed, limiting our ability to conclude the representatives of our sample to any population.

One potential confounding variable in our study could be socioeconomic status, which is not included in our analysis. Socioeconomic status can impact access to healthcare, diet, and stress levels, all of which can influence health outcomes, including stroke occurrence. Another confounding variable could be the presence of other health conditions that were not recorded in the dataset, such as diabetes or high cholesterol, which are known to be related to stroke.

The main limitation of our study is the missing data in the BMI column. This limitation might have influenced our results, as BMI is an important factor associated with stroke occurrence. The strengths of our analysis include a comprehensive exploration of numerous factors and their interactions. The application of logistic regression models allowed us to evaluate different types of relationships between the predictors and the outcome.

Future research could build on our work by including additional variables such as socioeconomic status, physical activity level, or dietary habits. Researchers could also further explore the interaction effects among the variables, as our results suggest that these can significantly influence the risk of stroke.

In conclusion, our study demonstrates the potential of statistical modeling in understanding and predicting stroke occurrence. Our findings underline the importance of age, average glucose level, and hypertension as significant predictors of stroke, providing valuable insights for targeted prevention strategies. We have been able to confirm risk factors that studies such as Rexrode et al. (2022) confirmed that age is a major risk factor for strokes. Additionally in the study conducted by Scrutinio et al. (2020) found that age is a significant risk factor for the likelihood. Although further research is needed to confirm and expand upon our findings, our study provides hope for future investigations into stroke prediction and possible prevention.

Appendix

Schwartz, Anteby, R., Klang, E., & Soffer, S. (2023). Stroke mortality prediction using machine learning: systematic review. Journal of the Neurological Sciences, 444, 120529–120529. https://doi.org/10.1016/j.jns.2022.120529

Rexrode, Madsen, T. E., Yu, A. Y. X., Carcel, C., Lichtman, J. H., & Miller, E. C. (2022). The Impact of Sex and Gender on Stroke. Circulation Research, 130(4), 512–528. https://doi.org/10.1161/CIRCRESAHA.121.319915

Scrutinio D. Ricciardi C. Donisi L. Losavio E. Battista P. Guida P. et al. Machine learning to predict mortality after rehabilitation among patients with severe stroke. Sci Rep. 2020; (0123456789. Available from:): 1-10 https://doi.org/10.1038/s41598-020-77243-3

[1]

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?select=healthcare-dataset-stroke-data.csv

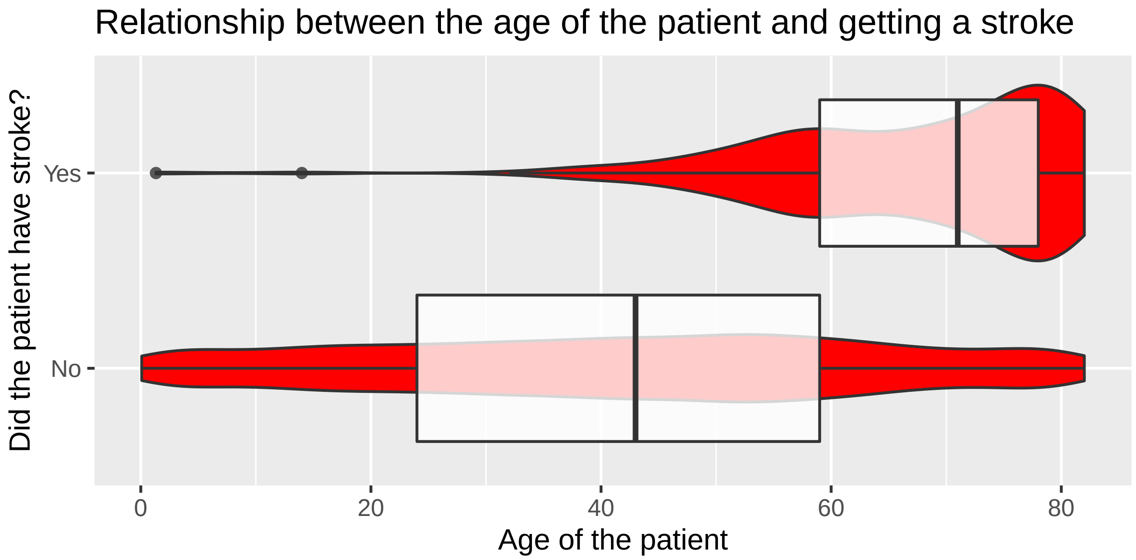
[2]

```
## [1] 201
```

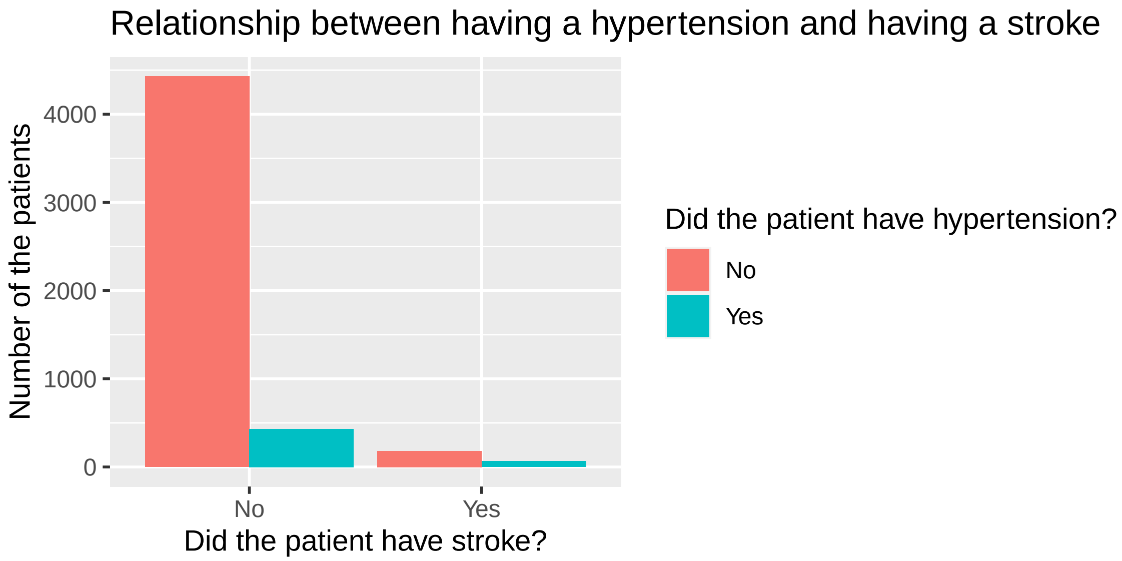
[3]

```
##                   gender    age hypertension heart_disease ever_married Residence_type avg_glucose_level    bmi stroke
## gender             1.000 -0.030        0.022         0.083       -0.036         -0.004             0.053 -0.026  0.007
## age               -0.030  1.000        0.274         0.257        0.681          0.011             0.236  0.333  0.232
## hypertension       0.022  0.274        1.000         0.116        0.162         -0.001             0.181  0.168  0.143
## heart_disease      0.083  0.257        0.116         1.000        0.111         -0.002             0.155  0.041  0.138
## ever_married      -0.036  0.681        0.162         0.111        1.000          0.005             0.152  0.342  0.105
## Residence_type    -0.004  0.011       -0.001        -0.002        0.005          1.000            -0.007  0.000  0.006
## avg_glucose_level  0.053  0.236        0.181         0.155        0.152         -0.007             1.000  0.176  0.139
## bmi               -0.026  0.333        0.168         0.041        0.342          0.000             0.176  1.000  0.042
## stroke             0.007  0.232        0.143         0.138        0.105          0.006             0.139  0.042  1.000
```

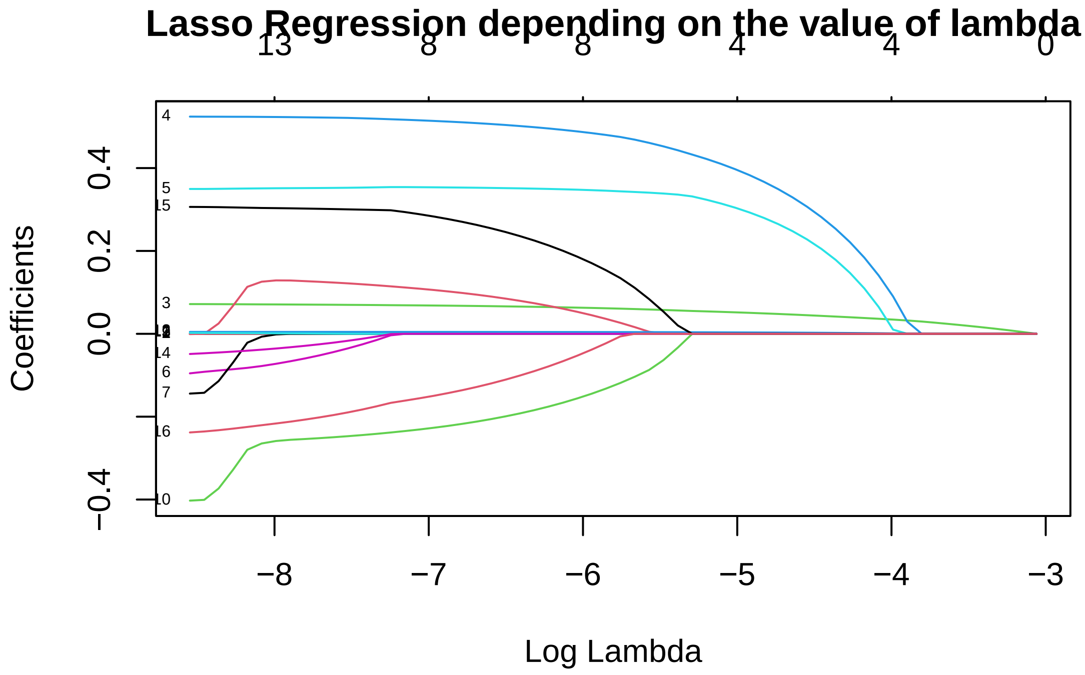
[4]

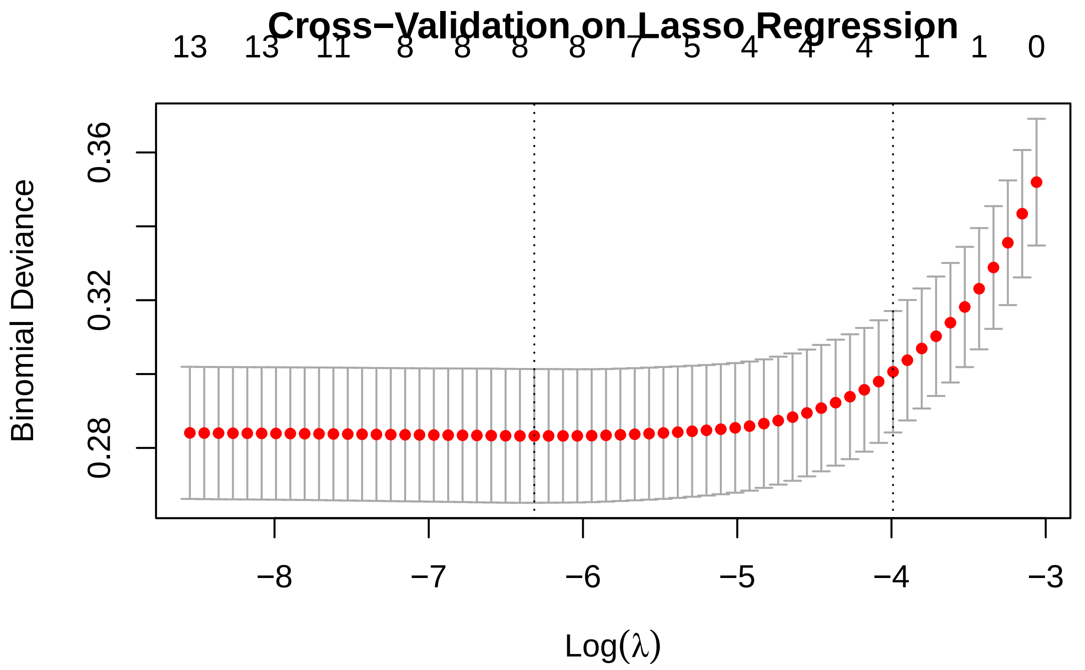


[5]

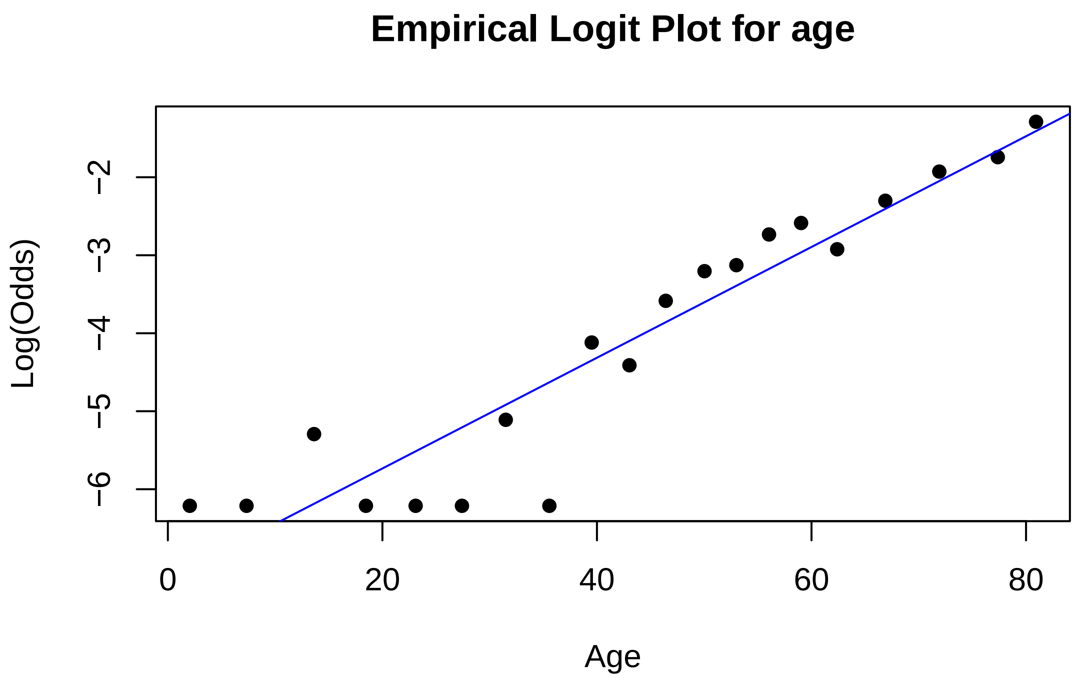


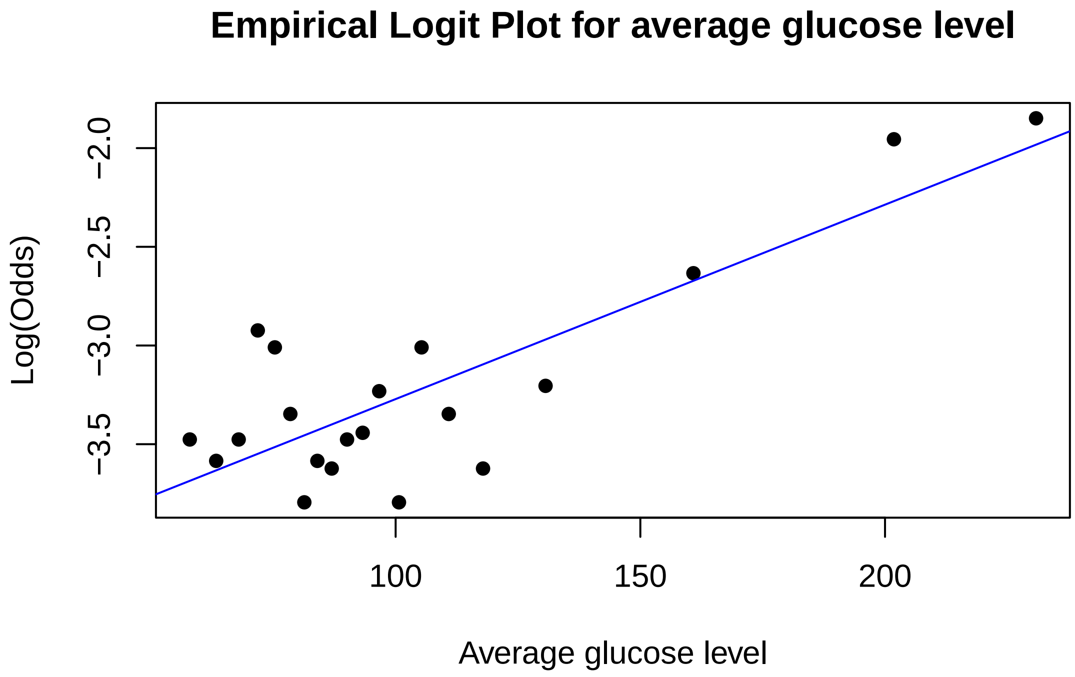
[6] Multiple Logistic Regression model creation

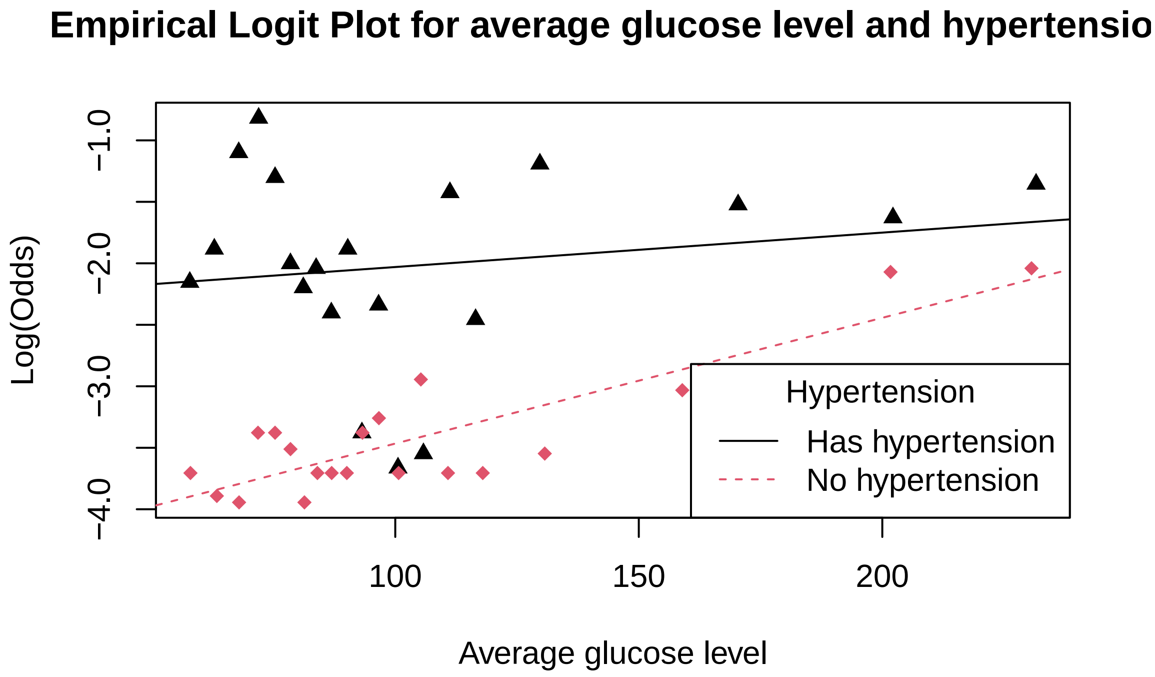




```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                                        s1
## (Intercept)                -4.9592565171
## id                          .
## genderMale                  .
## age                         0.0343513119
## hypertension1               0.0904100780
## heart_disease               0.0101122869
## ever_married                .
## work_typeGovt_job           .
## work_typeNever_worked       .
## work_typePrivate            .
## work_typeSelf-employed      .
## Residence_typeUrban         .
## avg_glucose_level           0.0009092424
## bmi                         .
## smoking_statusnever smoked  .
## smoking_statussmokes        .
## smoking_statusUnknown       .
```




Empirical Logit Plot for average glucose level

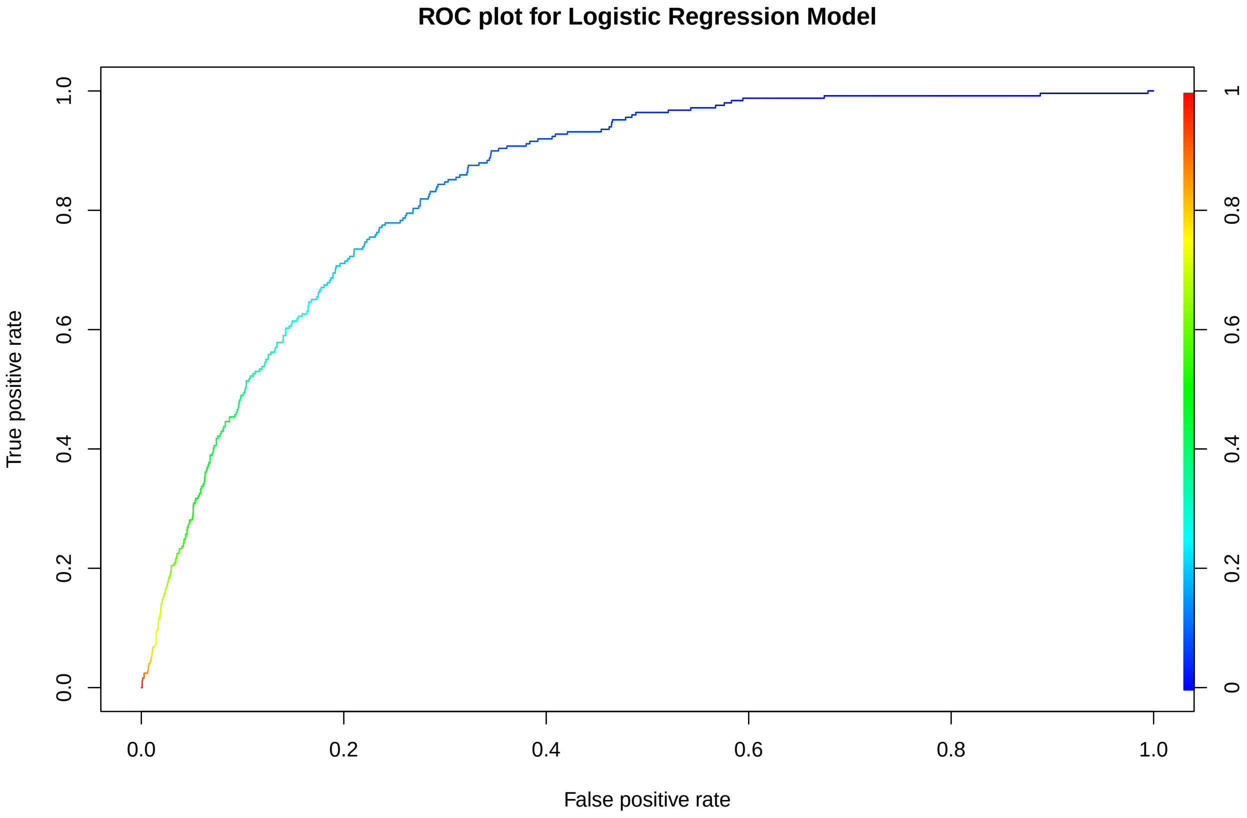

Empirical Logit Plot for average glucose level and hypertensio

[7] Multiple Logistic Regression fit assessment

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.797
```

```
## # A tibble: 1 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 sensitivity binary         0.802
```

```
## # A tibble: 1 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 specificity binary         0.711
```



```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.844
```