A Study of Storke Prediction with GLM

1. Introduction

Background

Stroke is correlated with many factors. It is commonly considered that stroke is highly correlated with age, health status, and lifestyle. Understanding what factors correlate with stroke and how they correlate with stroke help identify at-risk populations and prevent stroke. In the article *Influence of Age and Health Behaviors on Stroke Risk: Lessons from Longitudinal Studies*, the author investigated how age affects stroke by analyzing trends across different populations over time and concluded that stroke risk increases significantly with age (Kelly-Hayes, 2010). In another article *Body Mass Index Measured Repeatedly over 42 Years as a Risk Factor for Ischemic Stroke: The HUNT Study*, researchers calculated average BMI values and related these to the risk of stroke and observed a higher risk for stroke among participants with obesity or overweight over adulthood (Horn et al., 2023). Further in the article *Glucose and Acute Stroke : Evidence for an Interlude*, by analyzing clinical data and categorizing them by blood glucose levels, the researchers concludes that both hyperglycemia and hypoglycemia significantly affect stroke risk (Helgason, 1988).

Research Question

We want to combine all predictors in the existing literature in one comprehensive stroke prediction model. Our goal is to predict the probability of getting a stroke based on age, BMI, average glucose level, which are already discussed in the literature, as well as gender, smoking status and hypertension, three other candidate predictors. We want to identify significant predictors and find the best stroke prediction model.

2.Method

Model

We will use a GLM model for prediction in this study. The outcome 'stroke' in our study is a binary variable, thus the assumption of homoscedasticity does not hold if we use a linear regression model. Our model is:

$$log\frac{E(Y|X)}{1-E(Y|X)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n$$

where Y represents 'stroke'; X_i represents a set of selected predictors. The logit link is log-odds in our case.

Variable Selection

We've first narrow our analysis to health-related predictors, excluding work type, marital status, and residence type in the dataset. Then we use stepwise AIC selection to balance model complexity with fit, considering factors including gender, age, hypertension, heart disease, glucose levels, BMI, and smoking. AIC is used since our dataset is small, we want less penalty on complexity. In the stepwise AIC selection, we add the variable that provides the lowest AIC in forward step while we remove the variable that decreases AIC in the backward step.

Model Diagnostics. and Validation

1. Dfbetas:

We will use dfbetas to identify influential observations by measuring how each one affects the regression coefficients. Observations with a dfbeta over $\frac{2}{\sqrt{(n)}}$ are flagged as influential, and we'll consider their context to decide whether to keep them.

2. Deviance Residuals:

We will plot residuals vs covariates to see whether residuals are independent of covariates. Residuals independent of covariates are a sign of a good model. Having systematic patterns might indicate misspecification of model, where transformation or adjustments are required.

3. Cross-Validation

We'll employ cross-validation to assess our model's predictive accuracy. We fit the model on n-1 data parts and test on the remaining part. A good model's calibration plot will align closely with a 45-degree line, with predicted probabilities equal to actual probabilities.

4.Discrimination With The ROC Curve

We'll assess model performance using ROC curves, plotting sensitivity versus (1-specificity) across thresholds to determine the true positive and false positive rates. A higher AUC indicates better model discrimination.

Results

Description of Data

gender	age	hypertension	heart disease	ever married	stroke
Min. :0.0000	Min. :10.00	Min. :0.0000	Min. :0.00000	Min. :0.0000	Min. :0.00000
1st Qu.:0.0000	1st Qu.:34.00	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:1.0000	1st Qu.:0.00000
Median :1.0000	Median :50.00	Median :0.0000	Median :0.00000	Median :1.0000	Median :0.00000
Mean :0.6091	Mean :48.65	Mean :0.1191	Mean :0.06015	Mean :0.7588	Mean :0.05255
3rd Qu.:1.0000	3rd Qu.:63.00	3rd Qu.:0.0000	3rd Qu.:0.00000	3rd Qu.:1.0000	3rd Qu.:0.00000
Max. :1.0000	Max. :82.00	Max. :1.0000	Max. :1.00000	Max. :1.0000	Max. :1.00000

Residence type	avg glucose level	bmi	smoking status	stroke
Min. :0.0000	Min. : 55.12	Min. :11.50	Min. :0.0000	Min. :0.00000
1st Qu.:0.0000	1st Qu.: 77.23	1st Qu.:25.30	1st Qu.:0.0000	1st Qu.:0.00000
Median :0.0000	Median : 92.35	Median :29.10	Median :0.0000	Median :0.00000
Mean :0.4905	Mean :108.31	Mean :30.29	Mean :0.4593	Mean :0.05255
3rd Qu.:1.0000	3rd Qu.:116.20	3rd Qu.:34.10	3rd Qu.:1.0000	3rd Qu.:0.00000
Max. :1.0000	Max. :271.74	Max. :92.00	Max. :1.0000	Max. :1.00000

Boxplots of stroke each numerical variable against stroke and Proportion of stroke within each binary vabiable are shown in the appendix section (fig.5 & fig.6).

Model

Based on the stepwise AIC results, age, average glucose level, hypertension and heart disease are selected as predictors for the model. Fitting the model with these four variables, we have our model as:

$$log\frac{p(Stroke)}{1-p(Stroke)} = -7.632 + 0.067Age + 0.005AvgGlucoseLevel + 0.568Hypertension + 0.454HeartDisease$$

Godness of Model

1. Dfbetas:

-Age: No influential points exceed the threshold; however, a fanning pattern in age-related dfbetas suggests observations at older ages affect the estimates of age more. This can be due to the non-linearity of the relationship between age and stroke or the skewness of data in age. We will discuss this further in the limitation section.

-Average Glucose Level: Dfbetas values vary around the zero and there are no systematic patterns. The blue lines are relatively flat around zero. No influential observations are observed.

-Hypertension and **Heart Disease**: There are no observations within each group (0 and 1) that appear to have a significant influence on the estimates for hypertension and heart disease.

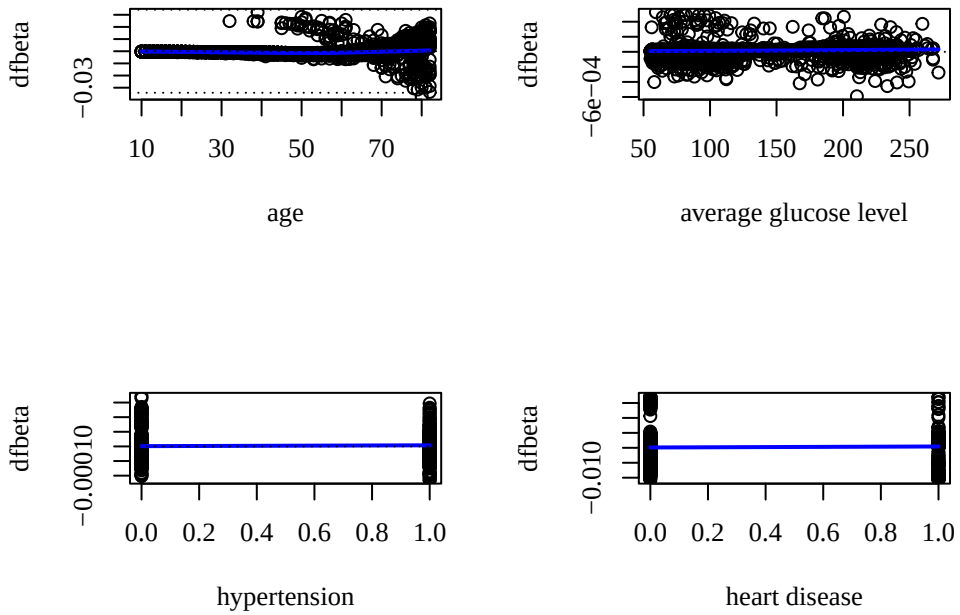

Figure 1: Dfbetas Plots for each Covaraite

2. Deviance Residuals

-Age: The residual plot shows a separation pattern, with one group having significantly higher deviance residuals. This suggests there are missing variables or missing interactions. We will further discuss this in the limitation section.

-Average Glucose Level: Residuals are mostly centered around zero but show some separation. One group have average higher deviance residuals then the other. Reasons might be missing predictors, interactions, or subgroups. We will discuss this in the limitations.

-Hypertension and **Heart Disease**: Residuals are centered slightly below zero with a similar spread in each group, suggesting no systematic bias.The flat blue lines indicate no trends.

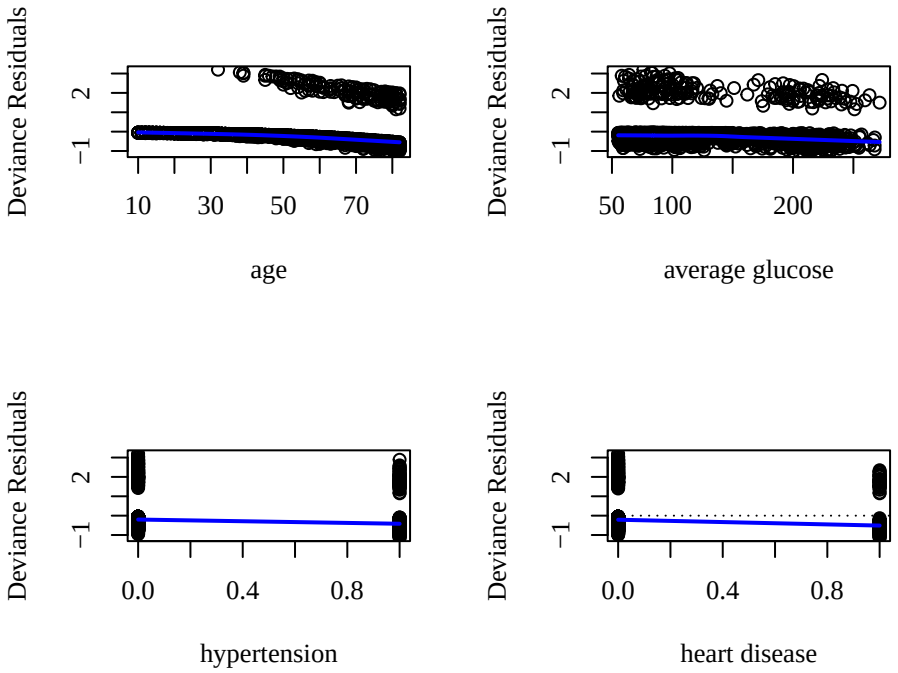

Figure 2: Diviance Residuals Plots for each Covariate

3.Cross Validation

The small deviation of the Bias-corrected line from the off-diagonal line indicates overestimation at high probabilities by the model. The mean absolute error (MAE) is 0.004. That is, on average, the absolute difference between the predicted probabilities and the actual outcomes is 0.004. The model is predicting well for low probabilities but not good enough for high probabilities.

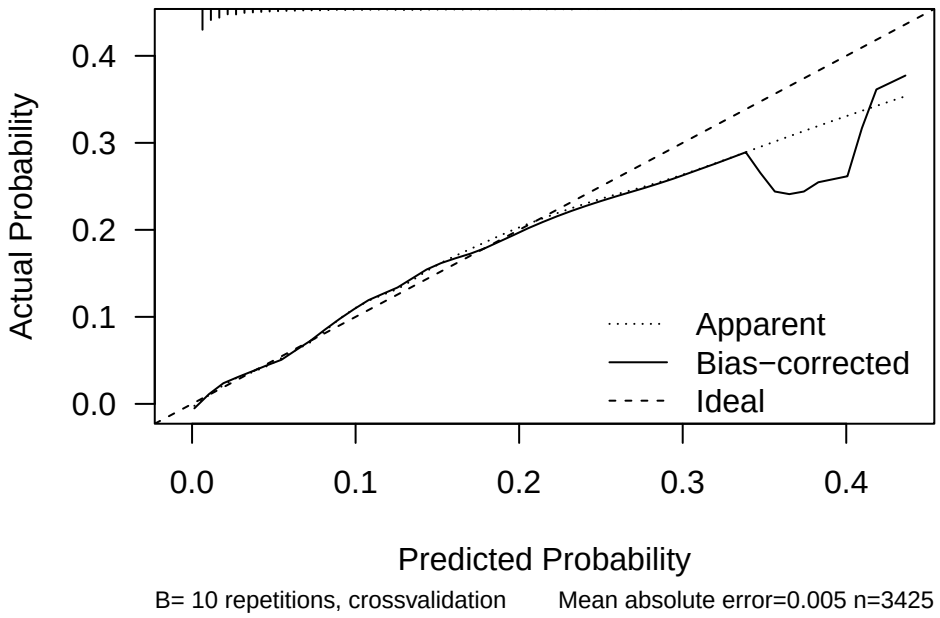

Figure 3: The Calibration Plot

4. Discrimination with ROC curve

The ROC is displayed as the red line with AUC being 0.83. It shows that the model can discriminate 83% of stroke and not stroke. The AUC is high, thus the model has good discrimination ability.

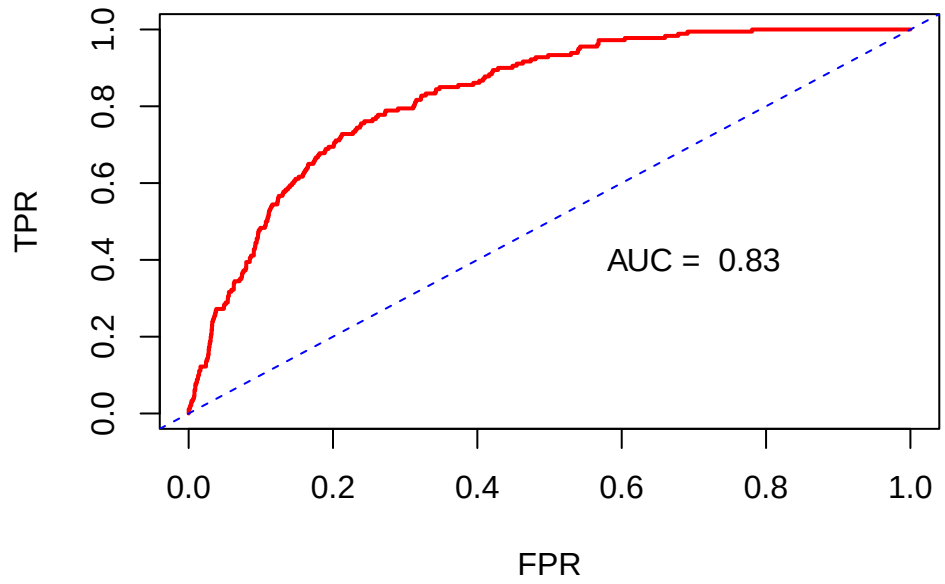

Figure 4: The ROC Curve

Discussion

Final Model Interpretation and Importance

1. Final Model Our final model is

$$log\frac{p(Stroke)}{1-p(Stroke)} = -7.632 + 0.067Age + 0.005AvgGlucoseLevel + 0.568Hypertension + 0.454HeartDisease$$

2. Interpretation (550)

Intercept:The log-odds of having a stroke when age and average glucose level is 0 with no hypertension and no heart disease is -7.632. This is not possible in real life, thus we ignore its interpretation.

Age: Keeping all other variables constant, with each additional year of age, the odds of having a stroke increase by a multiplicative factor of exp(0.67).

Average Glucose Level: Keeping all other variables constant, for each additional one unit increase in glucose levels, the odds of having stroke increase by a multiplicative factor of exp(0.005).

Hypertension: Assuming other variables are held constant, the odds of having a stroke for individuals with hypertension is a multiplicative factor of exp(0.005) of those without hypertension.

Heart Disease: Assuming other variables are held constant, the odds of having a stroke for individuals with heart disease is a multiplicative factor of exp(0.454) of those without heart disease.

We conclude that stroke is positively correlated with older ages, higher average glucose level, having hypertension and having heart disease. This can be used to identify individuals at higher risk and potentially lead to earlier interventions. We can target older people with higher average glucose level, hypertension and heart disease and intervene earlier to prevent stroke.

Limitations

1.Fanning Dfbetas

The fanning pattern indicates bias in estimating the age-stroke relationship, with higher uncertainties for older individuals. We attempted to address this with a polynomial transformation. It slightly reduced the fanning pattern (see fig7) but significantly increased dfbetas for glucose level and hypertension. Given the trade-off, we retained the original model for simplicity and interpretability.

2. Separated Deviance Residuals

The observed separation pattern could result in increased prediction and estimation biases. We attempted to incorporate additional variables, interactions, and applying polynomial and spline transformations, but all adjustments failed to enhance the pattern. Further investigations with other candidate predictors are necessary to improve the model.

Appendix

1. Boxplots of Stroke within each Numerical Variable

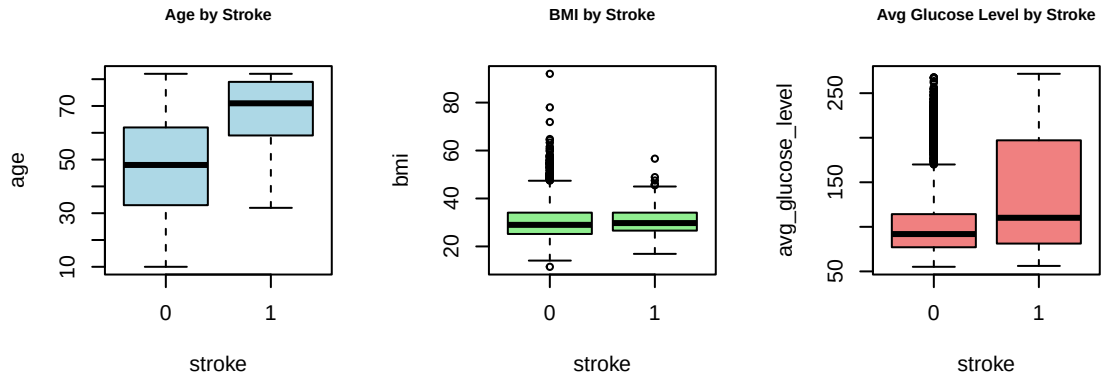

Figure 5: Boxplots of Stroke within each Numerical Covariate

2. Proportion of Stroke within each Binary Variable

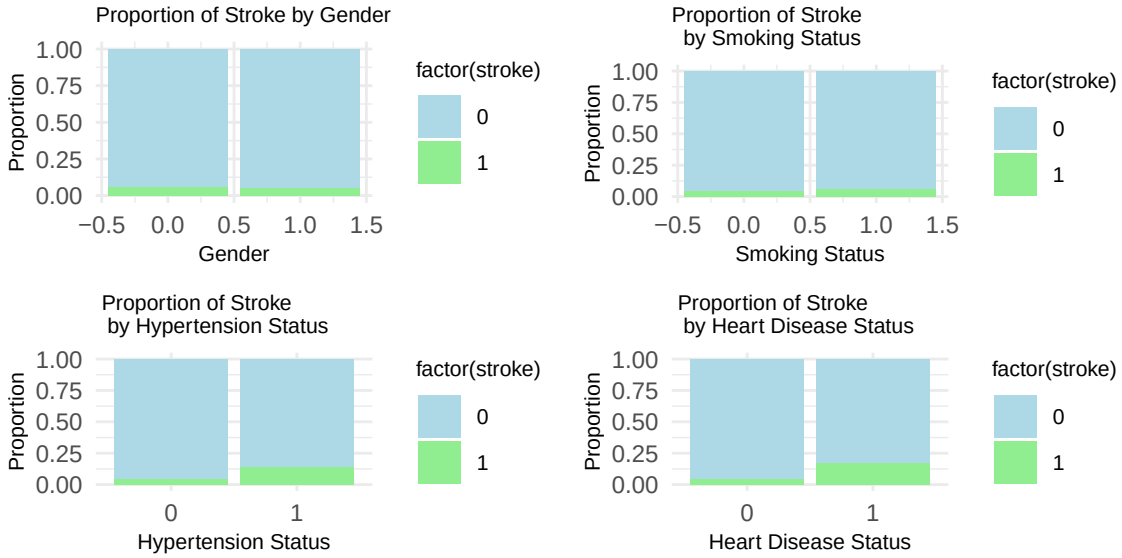

Figure 6: Proportion of stroke within each Binary Covariate

3. Dfbetas within Age after Polynomial Transformation

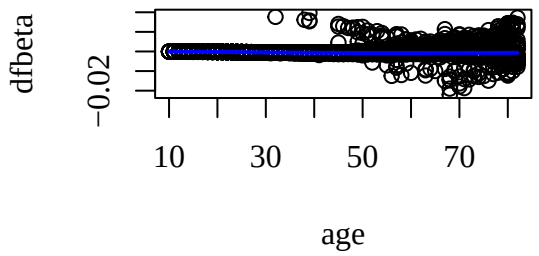

Figure 7: Dfbetas Plot within Age with Second-order Polynomial Transformation

Reference

Helgason, C. M. (1988). Blood glucose and stroke. *Stroke*, 19(8), 1049. https://doi.org/10.1161/01.STR.19.8.1049

Horn, J. W., Feng, T., Mørkedal, B., Aune, D., Strand, L. B., Horn, J., Mukamal, K. J., & Janszky, I. (2023). Body Mass Index Measured Repeatedly over 42 Years as a Risk Factor for Ischemic Stroke: The HUNT Study.*Nutrients*, 15(5), 1232. https://doi.org/10.3390/nu15051232

Kelly-Hayes M. (2010). Influence of age and health behaviors on stroke risk: lessons from longitudinal studies.*Journal of the American Geriatrics Society*,58 Suppl 2(Suppl 2), S325–S328. https://doi.org/10.1111/j.1532-5415.2010.02915.x