Stroke Prediction with Regression based Analysis

Shinjon Ghosh

Department of Mathematics, Illinois State University

Email: sghos10@ilstu.edu

Abstract:

Stroke is one of the leading causes of morbidity and mortality worldwide. It has significant health and economic impacts on individuals and healthcare systems. This study proposed the application of Regression Models to predict and identify factors influencing stroke risk. Developing a robust model that can anticipate risk and prevention strategies. The analysis involves building Generalized Linear Models (GLMs) with a binomial probit link function to predict binary outcomes—whether a patient faced stroke or not. By examining a range of features, the aim is to uncover patterns and relationships that influence patient stroke risk. The results highlight significant predictors (age, gender, hypertension, heart disease, average glucose level, smoke status) of stroke and offer actionable insights for developing targeted prevention strategies. Moreover, the model diagnostics suggest that the proposed GLMs fitted well and adequate according to Patient Stroke dataset. This project provides the value of predictive modeling in driving stroke decisions and optimizing stroke risk of patients.

Keywords: regression models, generalized linear models, stroke risk factors, health details, binomial GLMs, probit link function, interaction variable, odds ratio, quantile residuals, score test, likelihood ratio test, multicollinearity check, cook's distance.

Introduction:

The models used were from the Japan Public Health Center-based Prospective Study (JPHC Study), the Japan Arteriosclerosis Longitudinal Study (JALS), and the Suita Study. The researchers calculated the 10-year stroke risk for each employee and used the mean predicted risk as a representation of each worksite's health status. They found that the three models predicted similar stroke risks, and there were differences in risk among the worksites, even after adjusting for age [3]. This research paper focuses on building a prediction model for ischemic stroke (IS) recurrence using Backpropagation Neural Network (BPNN) and comparing its effectiveness to the traditional multivariate logistic regression model. The results demonstrated that the BPNN model outperformed the logistic regression model [2]. Storke is one of the vital health issues in the worldwide. The GLMs provides insights into factors that influence risk of stroke and guides strategies to reduce risk. Generalized Linear Models (GLMs) provide a suitable approach to predicting binary outcomes, such as whether a patient will face risk of stroke or not. By using GLMs with a probit link function, this study aims to identify key factors that influence patient risk factors of stroke. The purpose of this study is to develop a predictive model to estimate the probability of stroke, enabling health details to understand stroke risk, and improve the treatment. The dataset used for this predictor analysis based on demographic variables, and health details. The results of this project indicate predictor variables (age, gender, hypertension, heart disease, average glucose level, smoke status) are significant. According to the model diagnostics, the proposed GLM is fitted well. The odds ratio suggests the outcome of response variables with comparison between categorized predictor variables.

Background & Literature Review:

Deep Neural Network (DNN) significantly outperformed other models with an AUC (Area Under the Curve) of 82% before feature selection and 79% after. This suggests that DNN is more effective at handling imbalanced data. It also notes that while feature selection generally improves performance, it slightly decreased the AUC for DNN in this case [4]. Recurrent strokes lead to increased healthcare costs and a higher mortality rate, highlighting the importance of preventing recurrence. Model 1 used univariate analysis and multivariate COX regression. Model 2 employed bootstrap sampling, LASSO regression, and random forest algorithms to identify key variables and build a more robust model. Model 2 outperformed Model 1, showing a higher C-index, indicating better predictive ability [1]. The Stacking model outperformed all individual models, achieving an AUC of 0.878 in the training set, 0.871 in the internal validation set, and 0.809 in the external validation set. The Decision Curve Analysis (DCA) showed that the Stacking model had the highest net benefit, providing positive clinical value when the predicted probability threshold was greater than 0.110. Key risk factors identified included atrial fibrillation, pulmonary infection, coma, high creatinine, INR, serum sodium, neutrophil count, and low platelet count [6]. The author proposes a new type of residual, called randomized quantile residuals, address these limitations inverting the fitted distribution function for each response value and finding the equivalent standard normal quantile. This approach ensures that the residuals are exactly standard normal, facilitating more reliable residual analysis [7].

Research Questions:

1. What are the key factors influencing stroke?
2. Can we build a predictive model to forecast the stroke based on these factors?
3. How to measure performance of fitted model?

Objectives:

- Build a regression model to forecast strokes based on patient demographic and health data.
- Identify key factors influencing stroke risk.
- Evaluate model performance to identify early strokes and enhance patient treatment.

Methods & Materials:

Data Source & Overview

The Stroke dataset collects from Kaggle (secondary data resources) which provides demographics and health details of patients. The dataset contains 5110 observations and 12 features. Among 12 variables gender (Male, Female, Others), work_type (children, Govt. Job, Private, Self Employed, Never Worked), smoking_status (formerly smoked, never smoked, smokes), Residency_status (Rural & Urban), ever_married (yes & no), heart_diseases (yes & no), hypertension (yes & no), stroke (yes & no) are categorical features and id, age, bmi (Body Mass Index), average glucose level are numerical variables. Each row corresponds to a specific user, identifiable by their unique patient id. This dataset allows us to explore and analyze stroke base in terms of their health details and demographics measurements. It is a useful resource for statistical modelling, hypothesis testing and developing strategies for predicting strokes and preventing the risk.

Data Preparation

The aim of our project is to develop Regression Models to predict stroke. In the dataset 'bmi' & 'smoking_status' variables have missing values. I removed the missing values from the dataset. After deleting empty values, I found the total observations was 3426. Then the categorical variables are factorized by label 0 & 1 where stroke (no = 0 & yes = 1), gender (Male = 0 & Female = 1), work_type (Private = 0 & others = 1), smoking_status (smokes = 0 & others = 1), Residency_status (Rural = 0 & Urban = 1), ever_married (yes =1 & no = 0), heart_diseases (yes = 1 & no = 0), hypertension (yes = 1 & no = 0),. Later, I define stroke status (yes & no) as a response variable. On the other hand, I set categorical and numerical variable as predictor or explanatory variables according to dataset.

Result & Discussion:

Exploratory Data Analysis

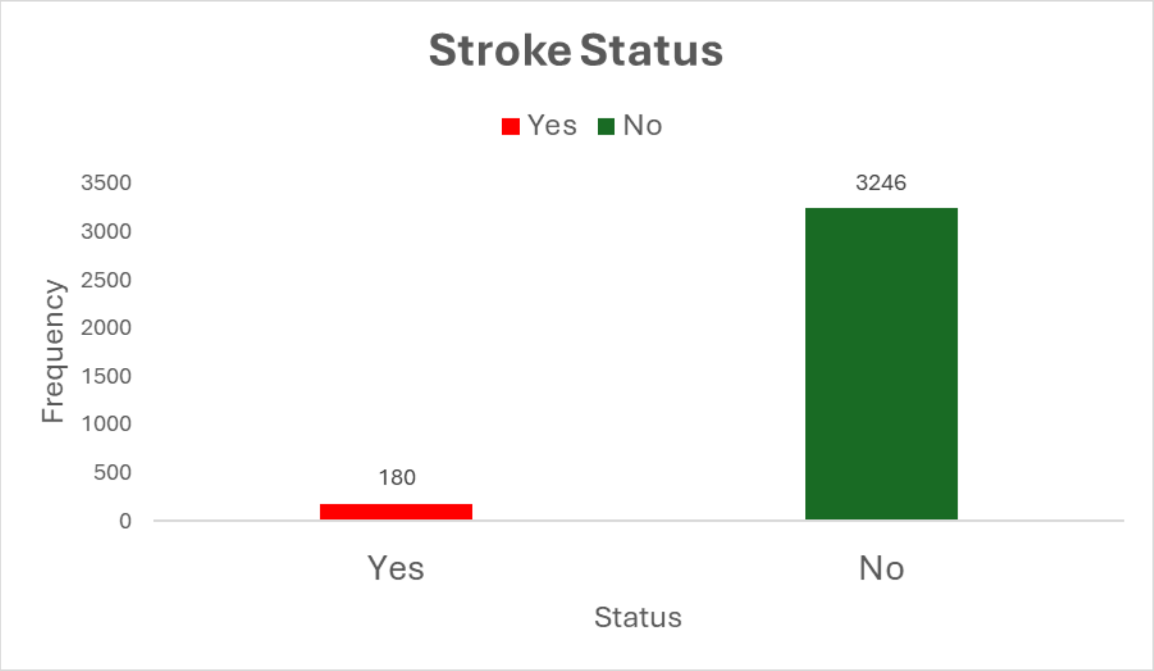


Fig 1: Stroke Status

Information in Fig 1 illustrates the comparison of stroke status values which indicates patients faced stroke or not. From the bar diagram, we can say that stroke patients' numbers (180) and patient did not face stroke those individual values (3246). That means the response variable data is highly imbalanced.

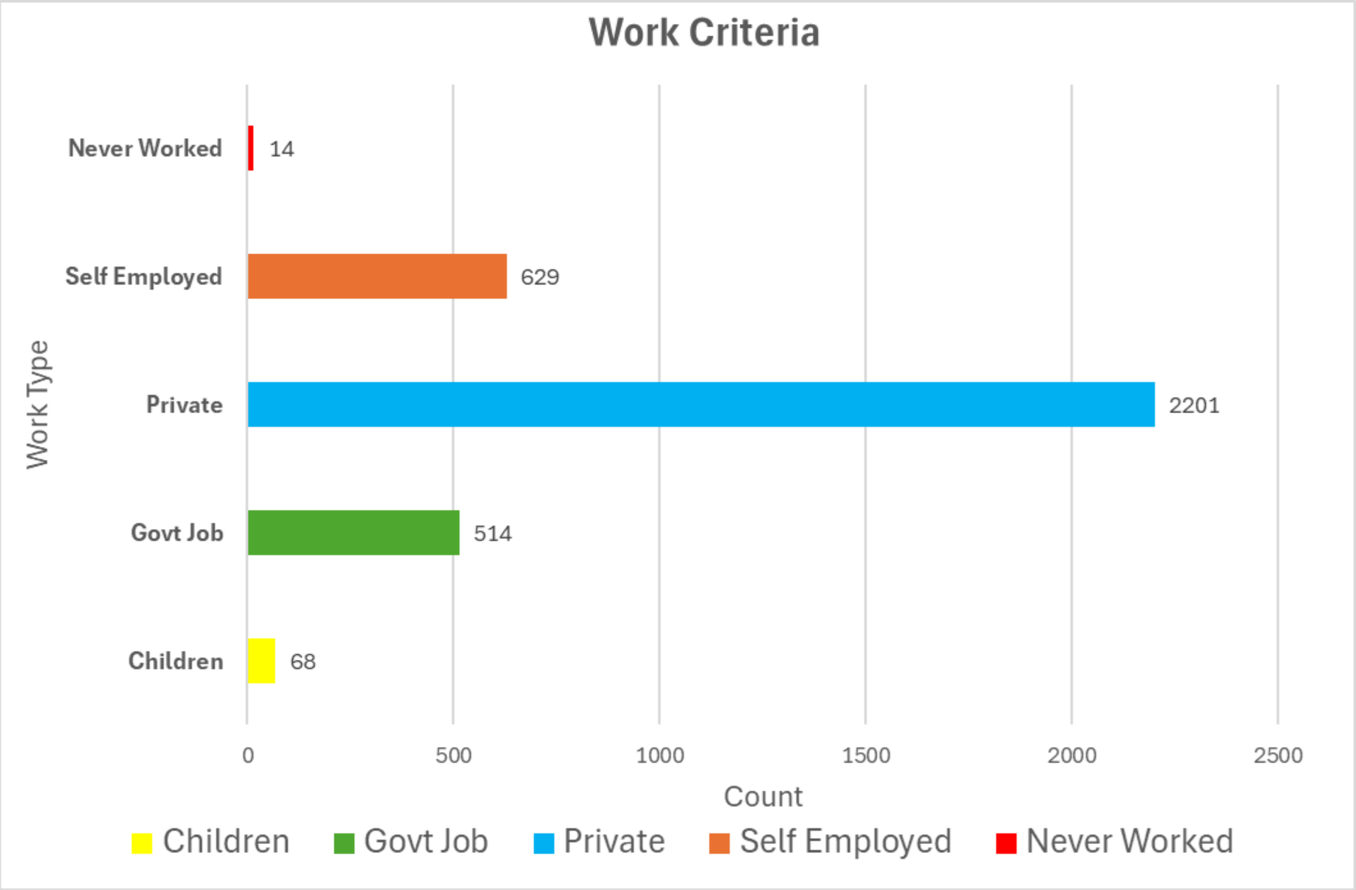


Fig 2: Work Criteria

Fig 2 shows the comparison between patients' work sites. From the horizontal bar plot, we found the maximum number of patients were doing private jobs and the number was 2201. On the other hand, 14 patients were never working, which is the minimum numbers. Moreover, more than 620 patients were self-employed and 514 patients' job status was Govt. job. Furthermore, 68 patients were children.

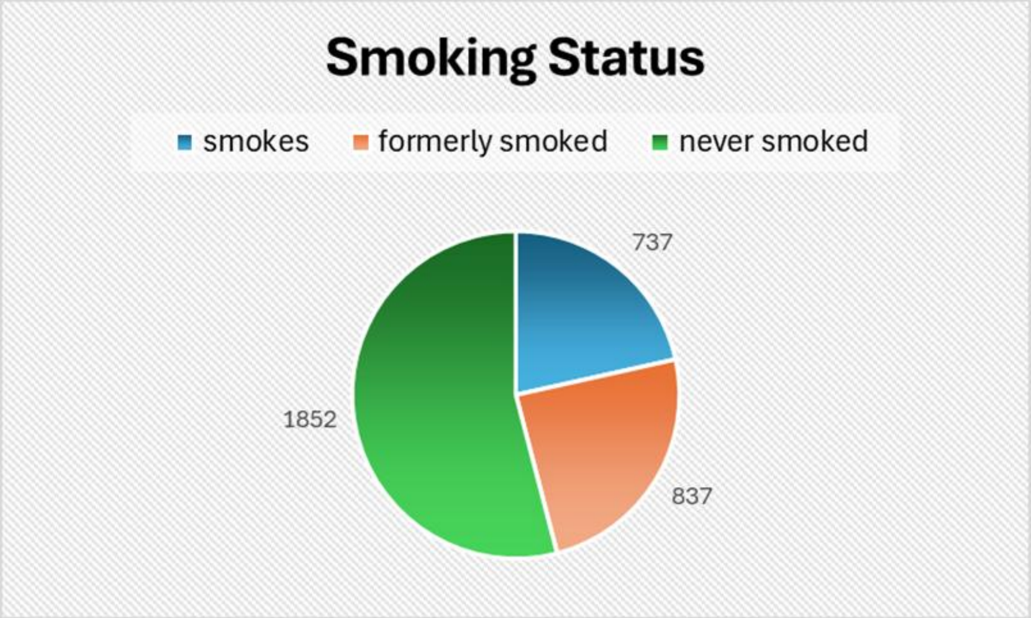


Fig 3: Patients' Smoking Status

This pie chart depicts the overall patient's smoking status. From Fig 3 we can say that more than 50 percent patients did not smoke and 24% patients holds formerly smoked. In addition, only 21 percent of patient were smoker.

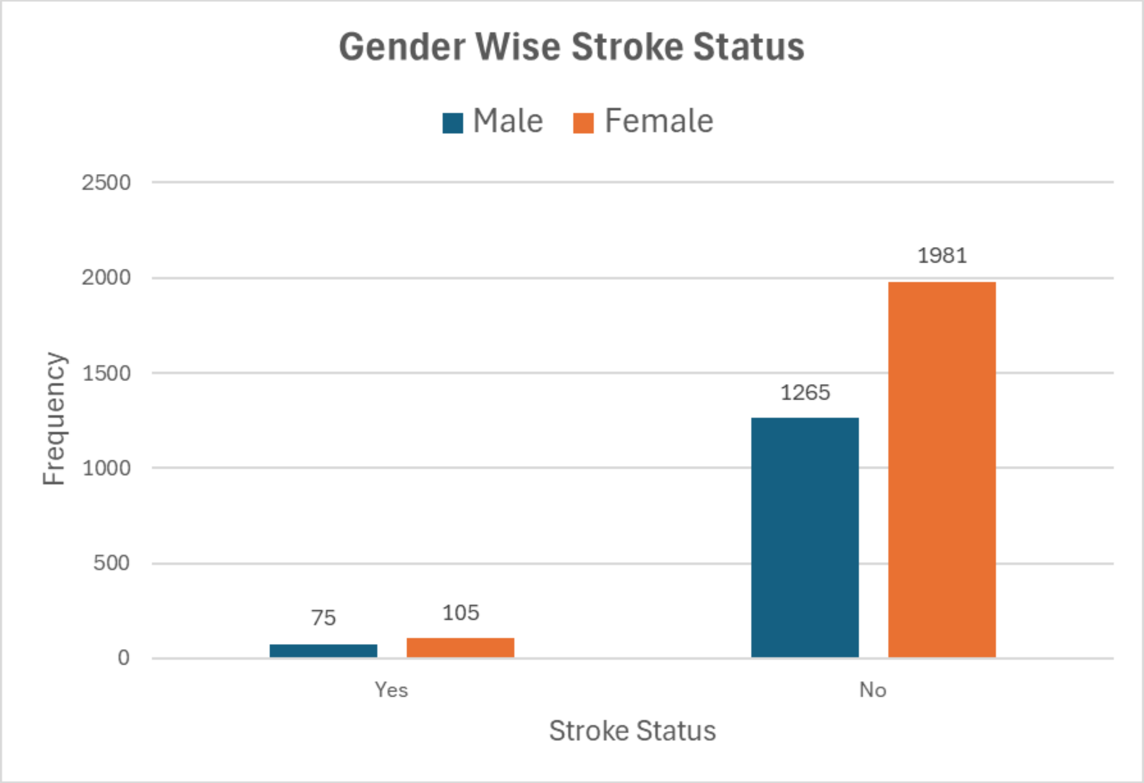


Fig 4: Gender wise Stroke Status

Fig 4 compares strokes patients according to gender. The bar plot shows that female patients are higher than male patients in stroke (yes) group. The female patients value in stroke (yes) group is 105. Moreover, 75 male patients have faced strokes issue. Moreover, male patient stroke (no) group number is less than females (no) group. The plot depicts 1265 male patients did not face stroke problems and the female patient's number was 1981.

Generalized Linear Models

From the processed Stroke dataset, I use stroke status as a response variable which is binary data (0 for no and 1 for yes). That's why I set and run several binomial generalized linear models according to different predictor variables and link functions. In Table 1, st1, st2, st3 model predictor variables are age, gender, work_type, smoking_status, Residency_status, ever_married, heart_diseases, hypertension, bmi, average_glucose_level and I used logit, probit, complementary log-log link functions, respectively. Moreover, st5, st6, st7 are forward regression, backward elimination, stepwise regression models respectively. Furthermore, model st4 based on interaction predictor variables. In st4 model, the explanatory variables are hypertension * heart_disease, work_type * avg_glucose_level, age, gender, residency_status, smoking_status, bmi and ever_married.

Models	AIC
st1	1164.3
st2	1161.5
st3	1165.5
st4	1154.7
st5	1154.6
st6	1154.6
st7	1154.6

Table 1: GLMs with AIC

We know lower AIC model performs better than higher AIC model. From Table 1, we can say that st5, st6, st7 (forward regression, backward elimination, stepwise regression) models AIC are similar and lower than other models.

```
Call:
glm(formula = stroke ~ age + avg_glucose_level + hypertension +
    heart_disease + work_type + smoking_status, family = binomial(link = "pro
bit"), data = st)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.7649643  0.2055815 -18.314  < 2e-16 ***
age                0.0338711  0.0031333  10.810  < 2e-16 ***
avg_glucose_level  0.0023580  0.0006995   3.371 0.000749 ***
hypertension       0.3123936  0.0956283   3.267 0.001088 **
heart_disease      0.2151825  0.1196874   1.798 0.072198 .
work_type1        -0.1489618  0.0844023  -1.765 0.077580 .
smoking_status1   -0.1705934  0.0991469  -1.721 0.085321 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1411.0  on 3425  degrees of freedom
Residual deviance: 1140.6  on 3419  degrees of freedom
AIC: 1154.6

Number of Fisher Scoring iterations: 7
```

```
Call:
glm(formula = stroke ~ age + hypertension + heart_disease + work_type +
    avg_glucose_level + smoking_status, family = binomial(link = "probit"),
    data = st)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.7649643  0.2055815 -18.314  < 2e-16 ***
age                0.0338711  0.0031333  10.810  < 2e-16 ***
hypertension       0.3123936  0.0956283   3.267 0.001088 **
heart_disease      0.2151825  0.1196874   1.798 0.072198 .
work_type1        -0.1489618  0.0844023  -1.765 0.077580 .
avg_glucose_level  0.0023580  0.0006995   3.371 0.000749 ***
smoking_status1   -0.1705934  0.0991469  -1.721 0.085321 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1411.0  on 3425  degrees of freedom
Residual deviance: 1140.6  on 3419  degrees of freedom
AIC: 1154.6

Number of Fisher Scoring iterations: 7
```

```
Call:
glm(formula = stroke ~ age + avg_glucose_level + hypertension +
    heart_disease + work_type + smoking_status, family = binomial(link = "pro
bit"), data = st)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.7649643  0.2055815 -18.314  < 2e-16 ***
age                0.0338711  0.0031333  10.810  < 2e-16 ***
avg_glucose_level  0.0023580  0.0006995   3.371 0.000749 ***
hypertension       0.3123936  0.0956283   3.267 0.001088 **
heart_disease      0.2151825  0.1196874   1.798 0.072198 .
work_type1        -0.1489618  0.0844023  -1.765 0.077580 .
```

```
smoking_status1   -0.1705934  0.0991469  -1.721 0.085321 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1411.0  on 3425  degrees of freedom
Residual deviance: 1140.6  on 3419  degrees of freedom
AIC: 1154.6

Number of Fisher Scoring iterations: 7
```

All three models’ coefficients are similar and significant by p-value (significant levels = **0.1**). Moreover, age (0.0338711) indicates that every unit increase of age, the response variable slightly increase the log-odds of 0.0338711 unit, when all other predictor variables are constant. Furthermore, avg_glucose_level (0.0023580) suggests that every unit increases of avg_glucose_level the log-odds of 0.0023580 units increase of response variable. In addition, hypertension (0.3123936) & heart_disease (0.2151825) illustrates that every unit increase of hypertension & heart_disease, the stroke risk increase log odds of 0.3123936 and 0.2151825 units when all other predictor variables are constant. Also, work_type1 (-0.1489618) suggests that every unit increase of work_type1 (children, govt_job, never_worked, self_employed) the log-odds of the stroke outcome decrease by 0.1489618 units compared to work_type0 (private). Smoking_status1 (-0.1705934) suggests every unit increase of smoking_status1 (formerly_smoked, never_smoked) the log-odds of the stroke outcome decrease by 0.0.1705934 units compared to smoking_status0 (smokes).

Then I checked multicollinearity by Variance Inflation Factor (VIF) and found all variables are free from high correlation issues in this model.

Variables	VIF
Average Glucose level	1.043697
Age	1.162920
Hypertension	1.040772
Heart Disease	1.069594
Work Type	1.044214
Smoking Status	1.041791

Table 2 : Multicollinearity check by VIF

From Table 2, we can say all the variables VIF value is less than 5 & 10. So, there is not any multicollinearity issue.

Our response variable data is highly imbalanced. In this case, Wald test is not trustworthy. Then I draw Score test for checking significancy of explanatory variables.

Score Test:

z-score	p-value
40.57449	0.00000

From this Score test, we get p-value is less than 0.05. That means the explanatory variables are significant.

Likelihood Ratio Tests to Compare Nested Models:

```
Model 1: stroke ~ age + avg_glucose_level + hypertension + heart_disease +
         work_type + smoking_status
Model 2: stroke ~ age + gender + hypertension + heart_disease + ever_married
+ work_type + Residence_type + avg_glucose_level + bmi + smoking_status

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3419     1140.7
2      3415     1139.5  4   1.1463   0.8869
```

The residual deviance is model 1 (st5, st6, st7) and model 2 (st2) is 1140.7, 1139.5, respectively. Here, p-value is 0.8869 which is much greater than 0.05. So, we can say the additional variable is not significant. That means the model 1 (st5, st6, & st7) is sufficient for the data.

After that, I check influential observation according to cook's distance, DFFITS and covariance ratio. I don't find any influential measures in my dataset. I illustrate first 50 observations influential measures here.

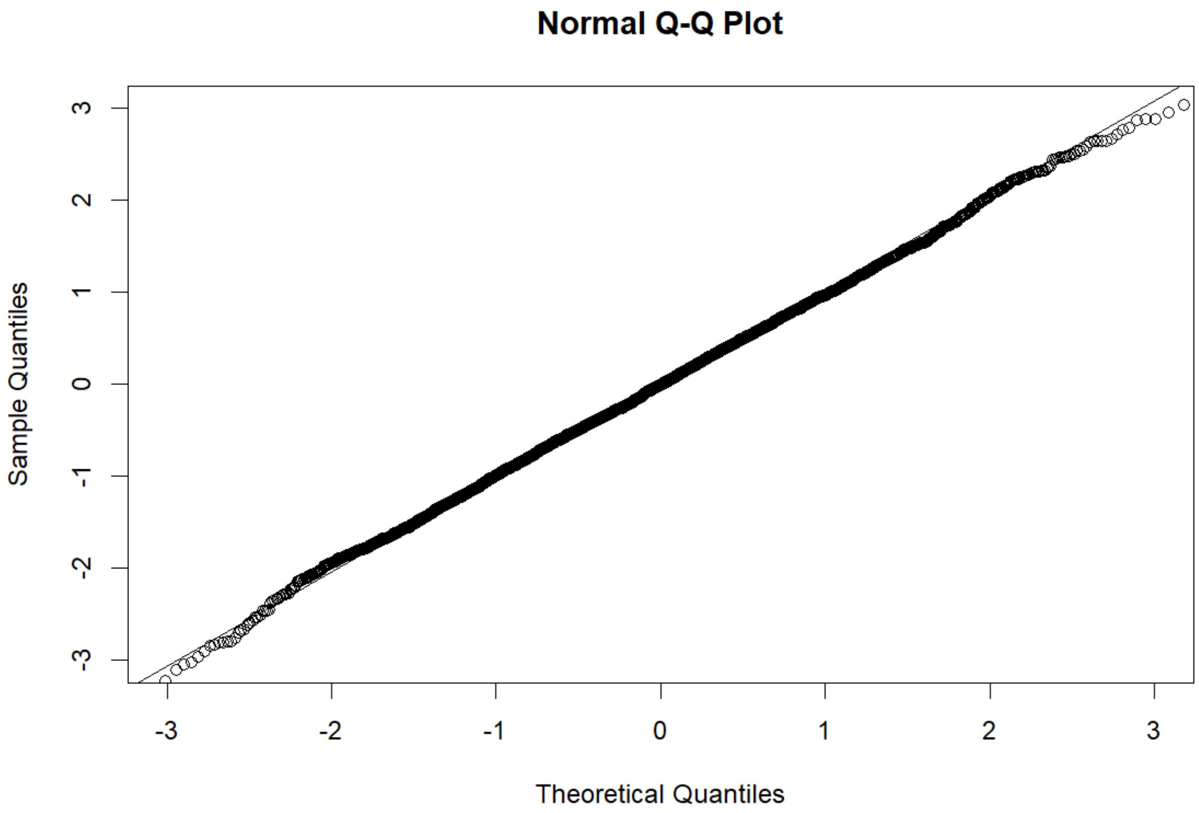
```
   dfb.1_    dfb.age dfb.hypr  dfb.hrt_ dfb.wr_1 dfb.av__ dfb.sm_1 dffit cov.r   cook.d      hat
1  -0.03573 -0.028327  -0.0794  0.209141  -0.0658  0.14060   0.0610 0.298 0.989 0.00560 0.008521
2  -0.04866  0.077482  -0.0675  0.203192  -0.0805 -0.05639   0.0504 0.274 0.992 0.00413 0.008188
3   0.08994 -0.044133  -0.0383 -0.035528  -0.0484  0.08423  -0.1397 0.208 0.966 0.00706 0.002286
4  -0.08769  0.055229   0.1322 -0.058496   0.0842  0.04568   0.0196 0.221 0.991 0.00254 0.005745
5  -0.12371  0.128384  -0.0745 -0.068729  -0.0976  0.08804   0.0225 0.212 0.989 0.00247 0.004985
6   0.01402  0.016180   0.1511  0.210708  -0.0745 -0.13063   0.0559 0.314 0.995 0.00532 0.010931
7  -0.01577  0.059499  -0.0485 -0.038551  -0.0807 -0.04317   0.0412 0.153 0.974 0.00243 0.001565
8  -0.06078  0.116467   0.1489 -0.059510  -0.0979 -0.09453   0.0241 0.246 0.992 0.00316 0.007085
9   0.09507 -0.051146  -0.0527  0.221319   0.1303 -0.01675  -0.1450 0.316 0.979 0.01002 0.006739
10  0.09383 -0.004818  -0.0317 -0.033930  -0.0540 -0.01088  -0.1427 0.187 0.966 0.00564 0.001878
11 -0.08941  0.053228  -0.0815  0.181008  -0.0766  0.10343   0.0441 0.274 0.997 0.00350 0.010040
12  0.06629 -0.128533   0.1465 -0.014872   0.1136  0.04964   0.0506 0.236 0.968 0.00886 0.002980
13  0.04085 -0.005102  -0.0669  0.200140  -0.0589  0.09248  -0.1529 0.311 0.990 0.00620 0.009111
14 -0.04859  0.087524   0.1231 -0.081762  -0.0822  0.11729  -0.1736 0.283 1.001 0.00338 0.012334
15  0.03251 -0.000427  -0.0370 -0.022850  -0.0634 -0.04184   0.0481 0.132 0.965 0.00280 0.000928
16 -0.02718  0.063956  -0.0679 -0.064793   0.1031  0.11478  -0.1873 0.270 0.985 0.00527 0.006245
17  0.02992 -0.132680   0.1435 -0.027601   0.1180  0.14961   0.0485 0.282 0.975 0.00923 0.004969
18 -0.12939  0.074933  -0.0836 -0.061831   0.0884  0.15486   0.0209 0.240 0.988 0.00349 0.005752
19 -0.03091  0.072770  -0.0523 -0.043164  -0.0845 -0.03205   0.0389 0.158 0.976 0.00232 0.001776
20 -0.07650  0.099779  -0.0629 -0.046224   0.0888 -0.03739   0.0261 0.192 0.980 0.00287 0.002973
21  0.00276  0.030858  -0.0452 -0.032837  -0.0737 -0.03005   0.0446 0.139 0.970 0.00234 0.001179
22  0.01672 -0.007127  -0.0737  0.203108   0.1212  0.09368  -0.1614 0.330 0.992 0.00679 0.010476
23  0.05040 -0.025688   0.1553 -0.063668  -0.0679  0.12882  -0.1724 0.297 0.987 0.00617 0.007713
24 -0.11016  0.069612   0.1213 -0.063520   0.0772  0.07511   0.0144 0.227 0.994 0.00239 0.007014
25 -0.10768  0.020978  -0.0914  0.186592   0.0974  0.15909   0.0393 0.309 0.999 0.00452 0.012399
26  0.07425 -0.079738  -0.0216  0.002761   0.0884 -0.03223   0.0445 0.140 0.949 0.00869 0.000732
27 -0.03341  0.063524   0.1315  0.182594  -0.0788 -0.10598   0.0422 0.290 1.001 0.00364 0.012394
28 -0.10325  0.077915  -0.0760 -0.064396  -0.0905  0.14577   0.0298 0.221 0.986 0.00307 0.004716
29 -0.01002  0.058914   0.1628 -0.047383  -0.0892 -0.10118   0.0363 0.236 0.984 0.00384 0.004941
30  0.08062 -0.052845  -0.0196 -0.003769  -0.0405 -0.06075   0.0499 0.132 0.953 0.00597 0.000700
31 -0.05848  0.125835  -0.0546 -0.050722  -0.0949 -0.07535   0.0322 0.198 0.981 0.00301 0.003232
32  0.06320 -0.031926  -0.0267 -0.011364  -0.0505 -0.05718   0.0500 0.132 0.959 0.00418 0.000788
33 -0.03366  0.114490  -0.0669  0.162921  -0.0748  0.00794  -0.1564 0.293 1.001 0.00372 0.012690
34  0.03870 -0.041310   0.1506 -0.057187   0.1156  0.12733  -0.1786 0.310 0.986 0.00705 0.008106
35 -0.10330  0.057622   0.1185 -0.070576  -0.0881  0.14472   0.0209 0.249 0.997 0.00276 0.008833
36  0.03447 -0.014965  -0.0380 -0.022252  -0.0606 -0.01558   0.0490 0.126 0.965 0.00266 0.000832
37 -0.07985  0.107470  -0.0630 -0.047080   0.0876 -0.04546   0.0251 0.197 0.981 0.00297 0.003200
38  0.17585 -0.101232   0.1550 -0.011785  -0.0330 -0.07764  -0.1115 0.248 0.959 0.01642 0.002696
39 -0.05302  0.067534  -0.0701  0.201088  -0.0792 -0.02326   0.0503 0.266 0.992 0.00386 0.007870
40 -0.05549  0.049975   0.1464 -0.050213   0.0897 -0.02854   0.0251 0.221 0.987 0.00296 0.004868
41  0.00174 -0.051721  -0.0706  0.220282  -0.0603  0.09394   0.0678 0.283 0.984 0.00622 0.006472
42 -0.07192  0.156404  -0.0548 -0.054121  -0.1002 -0.10696   0.0282 0.229 0.984 0.00367 0.004640
43 -0.06668  0.012569   0.1369 -0.054020   0.0941  0.08249   0.0264 0.229 0.988 0.00311 0.005393
44  0.04255 -0.110700   0.1485 -0.023061   0.1142  0.06961   0.0479 0.240 0.972 0.00719 0.003457
45  0.01189 -0.038426  -0.0507 -0.030484  -0.0612  0.09616   0.0496 0.163 0.968 0.00377 0.001518
46 -0.05331  0.111474   0.1520 -0.057819  -0.0974 -0.10309   0.0258 0.249 0.991 0.00335 0.006923
47  0.07150  0.028624   0.1707 -0.046084   0.1071 -0.11156  -0.1786 0.302 0.983 0.00759 0.006991
48  0.08241 -0.073624  -0.0206 -0.002695  -0.0355 -0.02126   0.0504 0.125 0.951 0.00606 0.000613
49 -0.05985  0.092250   0.1463 -0.059825  -0.0947 -0.04409   0.0261 0.224 0.990 0.00266 0.005733
50 -0.08242  0.050326  -0.0733 -0.059054  -0.0856  0.15197   0.0348 0.216 0.983 0.00333 0.004043
```

For checking random component, I use quantile residuals to draw Q-Q plot. For binomial families, quantile residual performance better than deviance residual. At first, I found quantile residuals for model st5, st6 & st7. I illustrate first 50 observations quantile residuals here.

```
   [1]  1.0496522673  2.3112645873  1.7149283528  1.2882447473  0.9660476777  0.8564715150  2.200
0489536
   [8]  1.4065778054  2.0168502436  2.0985311593  1.9077060190  1.8152745104  1.7068693302  1.865
2445422
  [15]  1.9043997864  1.3238466239  1.8332651527  0.9193288015  1.8937032569  1.1861843556  1.613
8907294
  [22]  1.7233481371  2.1907156190  0.5709549159  1.1175640073  3.2296286058  0.5903405126  1.104
1565365
  [29]  1.7190036887  2.3650431084  2.1286096291  2.0728578096  1.2881915734  1.1274697989  0.593
7157219
  [36]  1.8519359037  1.2023930985  2.2648808505  3.1085308623  0.9269916533  2.4783381686  1.125
5875161
  [43]  0.8770650989  1.6423496272  1.6795035718  1.0477911618  1.3700426952  2.3222888547  0.779
3545827
```

```
  [50]  2.0974115897  2.1088294671  2.0173544903  2.3266024448  1.5655408443  1.3614812869  3.293
8566211
```

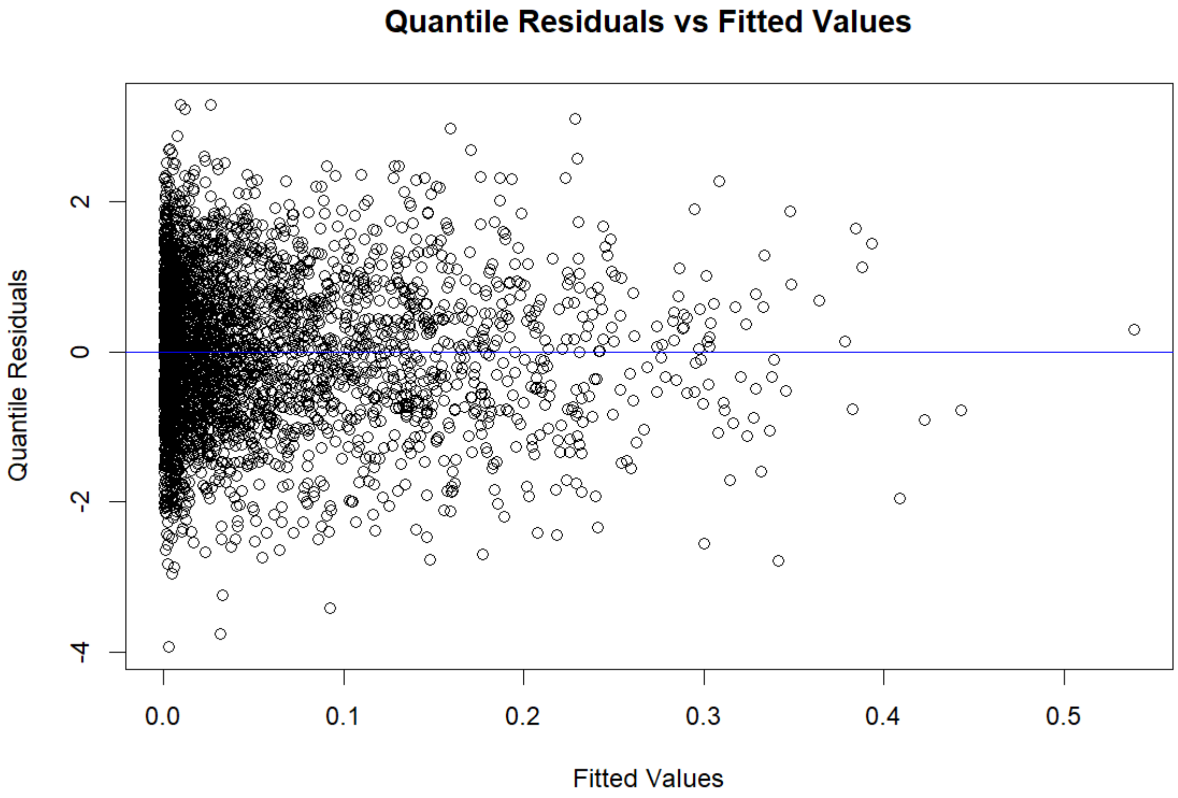
Then I drew the Q-Q plot based on quantile residuals of model st5, st6 & st7. I found all plots were similar. From the Normal Q-Q plot, we show quantile residuals have an exact normal distribution.



Then, I was depicted the plot of fitted values against quantile residuals to check constant variance. In this case, I calculate the fitted values of model st5, st6 & st7 (all are similar). I illustrate first 50 observations fitted values here.

```
   1            2            3            4            5            6            7            8         
0.1809480720 0.2233504129 0.0444228656 0.2464410048 0.2257509824 0.2307900987 0.0844295229 0.2451127299 
           9           10           11           12           13           14           15           16 
0.0887390817 0.0455628614 0.2946910208 0.0461085892 0.1761304227 0.3481078259 0.0452363025 0.1464536092 
          17           18           19           20           21           22           23           24 
0.0720450093 0.1923072272 0.0988484367 0.1293823965 0.0673497280 0.1837188712 0.1534304348 0.2979324847 
          25           26           27           28           29           30           31           32 
0.2864681086 0.0119055940 0.3327225039 0.1811492382 0.1566152516 0.0164932587 0.1335821916 0.0262762038 
          33           34           35           36           37           38           39           40 
0.3333520304 0.1429995910 0.3175374044 0.0428879444 0.1342033491 0.0230421161 0.2283151017 0.1917811660 
          41           42           43           44           45           46           47           48 
0.1307927543 0.1542794989 0.2005251072 0.0647026023 0.0545619764 0.2306033887 0.1177887018 0.0142488738 
          49           50 
0.2373849950 0.1486843738 
```

Later, I draw a plot of fitted values vs quantile residuals. We know any trends appearing in these plots indicate that the systematic component can be improved. On the other hand, if the random component is correct, the variance of the points is approximately constant. From these plots, we do not denote any trend. In this reason, the models (st5, st6 & st7) are adequate.



Goodness of Fit test

For binomial generalized linear model, I use Hosmer-Lemeshow test for checking goodness-of-fit of the model rather than R-square and adjusted R-square. It compares the observed number of events with the expected number of events based on the model's predictions. The p-value is used to test the null hypothesis that there is no significant difference between observed and expected outcomes. A low p-value (typically < 0.05) suggests rejection of the null hypothesis, indicating poor fit. A high p-value suggests that the model fits the data well.

p-value	0.9618

We get higher p-value (>0.05) from the Hosmer-Lemeshow tests. Now, we can say the model's predictions align with the observed outcomes. This indicates that the model fits the data well and that there's no significant deviation.

Odds Ratio

Odds ratios to convey the relationship between predictors and the binary outcome. An odds ratio (OR) quantifies the odds of an event occurring in one group relative to another group. Odds Ratio (>1) indicates an increased likelihood of the event occurring in one group compared to the reference group. Moreover, Odds Ratio (< 1) describes a decreased likelihood of the event occurring in one group compared to the reference group. Furthermore, Odds Ratio (= 1) denotes no difference in the odds between groups. For discreate and continuous variables, odds ratio are treated as predictors that are scaled by one-unit changes in their values.

Variables	OR	Reference Level
Age	1.034	
Average Glucose Level	1.002	
Hypertension1	1.37	Hypertension0
Heart Disease1	1.24	Country Disease0
Work Type1	0.86	Work Type0
Smoking Status1	0.84	Smoking Status0

Table 3: Odds Ratio

From Table 3, we described one year increase of age, the odds of stroke increase by 3.4%. Moreover, one unit increase of average glucose level, the odds of stroke slightly increase by 0.2%. The odds that hypertension patient stroke risk 37% higher than those do not have hypertension. Moreover, the odds that the stroke risk is 24% higher for heart disease patient rather than no heart disease patient. Subsequently, the odds that the stroke risk is 16% lower for Smoking Status1 (never_smoked, formerly_smoked) patient rather than smoker patients. In addition, the odds that stroke risk is 14% lower for Work Type1 (children, govt. job, self employed, never worked) rather than private job holder.

Limitations:

The Stroke predict dataset has 5110 observation and the bmi variable has a lot of missing values. After removing missing values, the total observation decreased (3426) and later in the model analysis, I found bmi variable was not significant. Moreover, the response variable is highly imbalanced.

Conclusion:

This project aims to leverage predictive analytics to forecast stroke and provide actionable insights for optimizing stroke risk. Specifically, constructing an optimum generalized linear model with the help of R language to predict stroke. According to the EDA analysis, the female patients stroke percentage is higher than male patients. Moreover, I find better generalized linear model by AIC when I use forward, backward and stepwise algorithm. According to score test, the explanatory variables (age, avg_glucose_level, hypertension, heart_disease, work_type, smoking_status) of the models are significant. In addition, the diagnostics (Q-Q plot, Quantile Residuals vs Fitted values plot, Goodness-of-fit test) indicates the model performance is well. Overall, these results offer a guideline for predicting stroke with influential predictor variables and mitigate stroke risk with enhance treatment strategies.

Reference:

[1] Ke, L., Zhang, H., Long, K., Peng, Z., Huang, Y., Ma, X.J., & Wu, W. (2024). Risk factors and prediction models for recurrent acute ischemic stroke: a retrospective analysis. *PeerJ, 12*.

[2] Lu, T., & Wang, Y. (2022). Prediction Model Construction for Ischemic Stroke Recurrence with BP Network and Multivariate Logistic Regression and Effect of Individualized Health Education. *Computational and mathematical methods in medicine*, *2022*, 4284566. https://doi.org/10.1155/2022/4284566.

[3] Nakashima, H., et al. (2024). Application of stroke prediction models to evaluation of worksite health status. *Environmental and Occupational Health Practice, 6*(1), n/a. Retrieved from https://doi.org/10.1539/eohp.2024-0002-FS.

[4] Wu, D., Zhang, X., & Zhu, X. (2024). A machine learning-based model for stroke prediction. *Proceedings of the 2nd International Conference on Software Engineering and Machine Learning*. https://doi.org/10.54254/2755-2721/67/20240645

[5] Qu, S., Zhou, M., Jiao, S., Zhang, Z., Xue, K., Long, J., Zha, F., Chen, Y., Li, J., Yang, Q., & Wang, Y. (2022). Optimizing acute stroke outcome prediction models: Comparison of generalized regression neural networks and logistic regressions. *PloS one, 17*(5), e0267747. https://doi.org/10.1371/journal.pone.0267747.

[6] Wang, K., Liu, J., Li, F., et al. (2024). Predicting in-hospital death of patients with acute stroke in the ICU using stacking model. *Preprint* (Version 1). Research Square. https://doi.org/10.21203/rs.3.rs-4908107/v1

[7] Dunn, P. K., & Smyth, G. K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244. https://doi.org/10.2307/1390802.

[8] https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset