MA678 Final Project

Yuanming LENG

11/30/2021

Abstract

Stroke is the 2nd leading cause of death globally in WHO report. And several effects influence the probability of stroke. Based on data from Kaggle, we find that hypertension, heart disease, and age have a high related to stroke. A model also includes individuals' average glucose level and BMI, giving good predictability of stroke. When setting the threshold as 0.1, which means when the probability of stroke is more significant than 0.1, we say the individual has high stroke risk, the model accuracy is 0.911.

Introduction

A stroke is a medical condition in which poor blood flow to the brain causes cell death. Signs and symptoms of a stroke may include an inability to move or feel on one side of the body, problems understanding or speaking, dizziness, or loss of vision to one side. According to the World Health Organization (WHO), stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. In this project, I would like to predict the probability a patient gets a stroke based on a data set I get from the Kaggle website. The data set contains each patient's information, including gender, age, marital condition, work type, residence type, average glucose level, BMI, smoking status, and whether suffer from hypertension, heart_disease. And the outcome is a binary column indicating whether each patient gets a stroke. I believe this project can give me a better understanding of the main risk factors leading to stroke while helping people to improve their health conditions and prevent getting a stroke.

Method

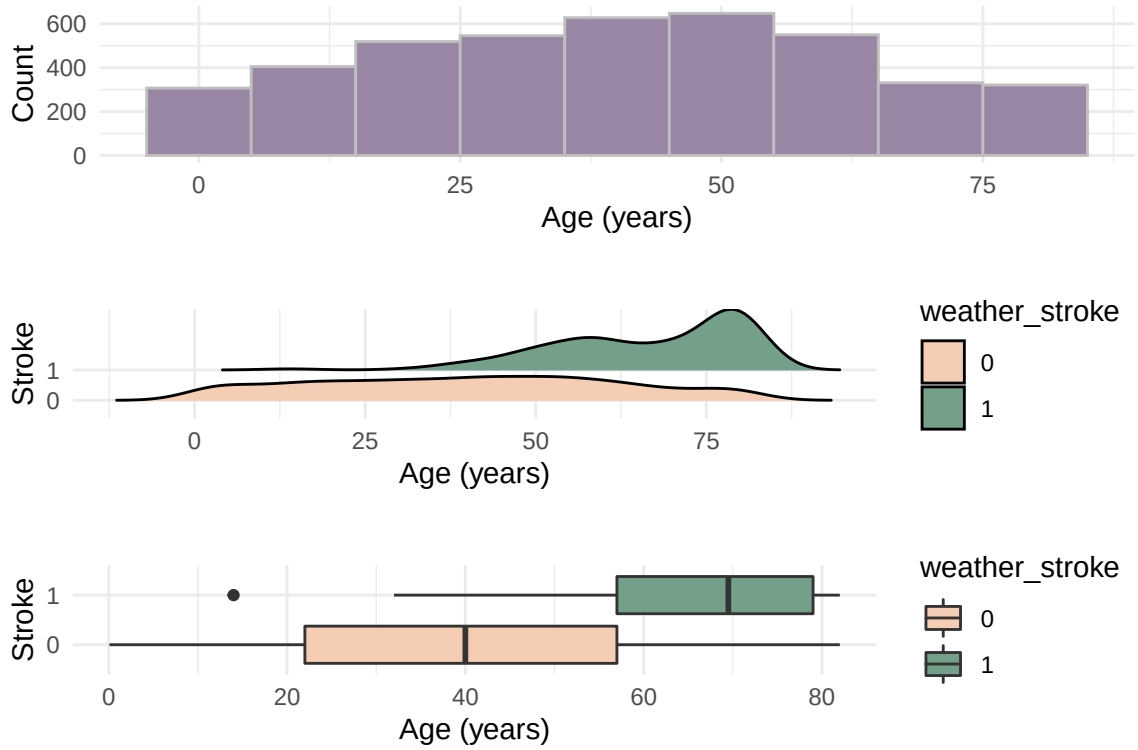
Data Cleaning and Processing

The data come from Kaggle: Stroke Prediction Dataset. This data set includes 12 variables and 5110 individuals, including their basic information and two kinds of stroke-related diseases: hypertension and heart disease.

To better fit the model, I dropped some outliers in our data set's average glucose level, BMI, age, and three continuous variables. Also, Considered strokes are highly influenced by age; I divided age into groups. Furthermore, Since I want to build a prediction model for stroke, I leave 20 percent of the data to act as a test set and 80 percent as a training dataset.

I get 4255 individuals after dropping outliers and 13 variables after adding one variable(*age_group*). And Training data set has 3390 individuals, test data set has 865 individuals.

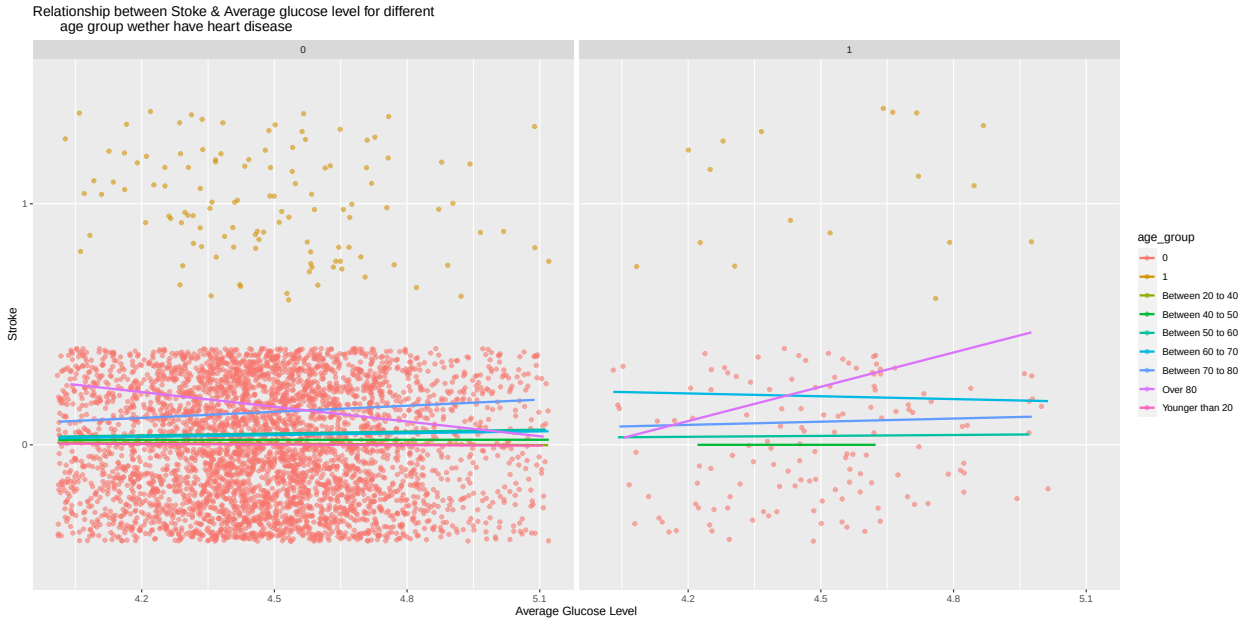
Exploratory Data Analysis



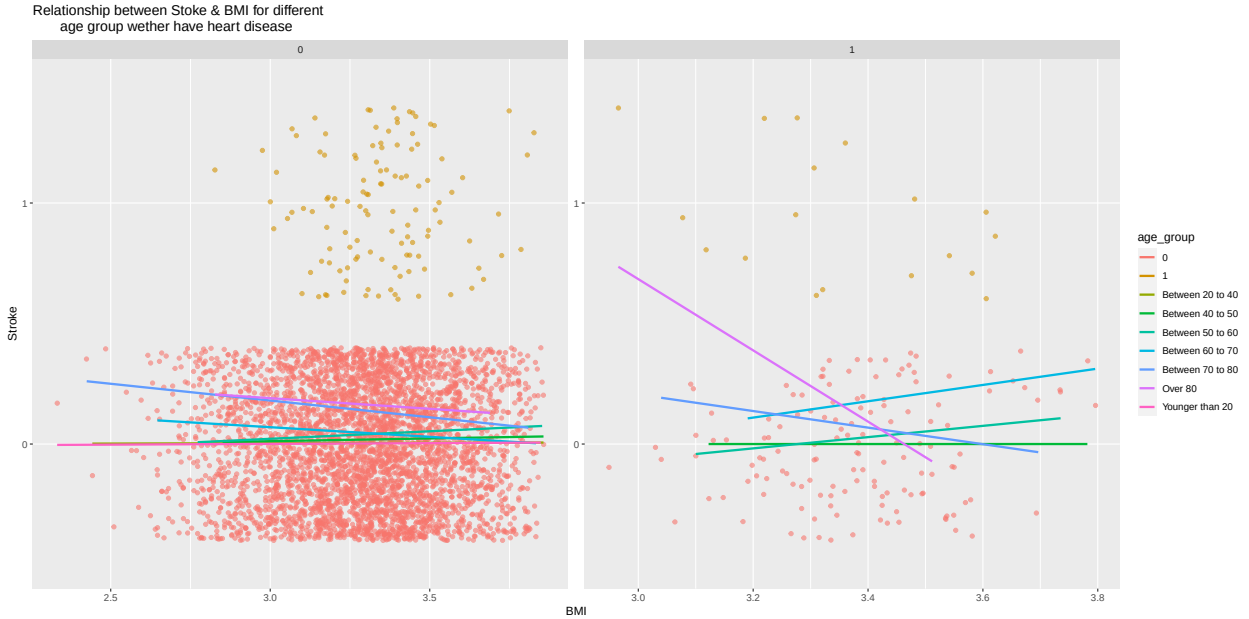
The upper figure shows most age of our individuals are 25 to 55 years old. The medium figure shows that the distribution of age in the stroke group and no stroke group are different. The lower figure shows that strokes are highly related to age in this data. The stroke group distributes in older individuals and concentrates on more than 55 years old. And people with no stroke focused on 20 to 55 years old. So I decided age group into 7 groups: younger than 20, 20 to 40, then 40 to 80 gap every 10 years old, older than 80.

Variable	weather_s	Stat.	P_value
gender	0 - 1	0.215	0.83
age	0 - 1	13.462	< 0.001
hypertension	0 - 1	7.837	< 0.001
heart_disease	0 - 1	6.068	< 0.001
ever_married	0 - 1	5.848	< 0.001
work_type	0 - 1	5.53	< 0.001
Residence_type	0 - 1	0.01	0.992
avg_glucose_level	0 - 1	0.157	0.875
bmi	0 - 1	2.301	0.021
smoking_status	0 - 1	4.387	< 0.001

The table uses a pairwise test for multiple comparisons of mean rank sums (Dunn's-Test) to check whether there is a significant difference between two groups (people with stroke & with no stroke) for each variable. The table supports that gender and Residence_type have no significant difference between the two groups. So I will drop these two variables in the model.



The figure shows the relationship between stroke and average glucose level for people with heart disease or not. We can see that the glucose level has a different trend for stroke for different age groups. Individuals who have an age over 80 have a distinct tendency compared with other age groups. Generally, there are different intercepts and slopes for every age group.



Similarly, the figure shows the relationship between stroke and BMI for different age groups. And the intercepts and slopes for each group are different.

Model Fitting

In the beginning, I transfer the average glucose level and BMI into a log scale. Then according to the result of EDA, I use age group as a category, add variables in the model. And for each class, the model has different intercepts and slopes. Except for gender and resident type, I also find that ever_maried and smoking_status are useless to make a difference to build an accurate model. After choosing the smallest AIC, certain final

model is :

$$logit(stroke = 1) = \beta_0 + \beta_1 hypertension + \beta_2 heart_disease + \alpha$$
$$\alpha = c_1 log(average\ glucose) + c_2 log(BMI)$$

Age_group	log(avg_glucose_level)	log(bmi)	(Intercept)	hypertension1	heart_disease1
Between 20 to 40	-1.3656109	1.2383967	-3.244589	0.7572017	0.5430504
Between 40 to 50	-0.4278746	0.3880158	-3.244589	0.7572017	0.5430504
Between 50 to 60	0.1225320	-0.1111173	-3.244589	0.7572017	0.5430504
Between 60 to 70	0.2161378	-0.1960034	-3.244589	0.7572017	0.5430504
Between 70 to 80	0.6475997	-0.5872722	-3.244589	0.7572017	0.5430504
Over 80	0.8432851	-0.7647284	-3.244589	0.7572017	0.5430504
Younger than 20	-1.7695603	1.6047162	-3.244589	0.7572017	0.5430504

Result

Model coefficient

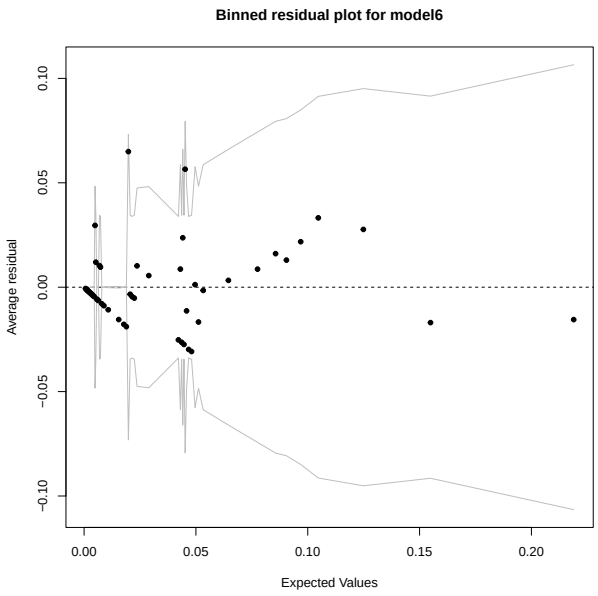
The model is varing slope between age groups and varing intercepts between disease groups (hypertension and heart disease). For example, if someone over 80 and who has hypertension and heart disease, the model predict wheather he or she stroke is :

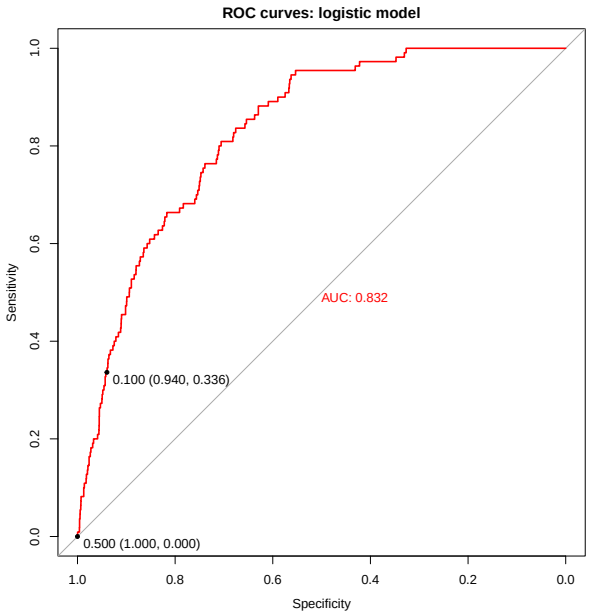
$$logit(stroke) = -1.9 + 0.84 log(glucose_level) - 0.76 log(BMI)$$

The -1.9 eaquals to $-3.2 + 0.76 + 0.54$.

Model validation

For check the model result, we plot the binned residual plot and ROC curve.





The most point located between the grey lines means most bins of residuals are significant. Also, for the right plot, the AUC of the model is 0.832. When we set the decision threshold to .1, the sensitivity was 0.33, and the specificity was 0.94. Likewise, when we set the decision threshold to 0.5, the sensitivity was 0, and the specificity was 1.

Also when I use the model to predict in the test data set:

Prediction	Reference	Freq
Yes	Yes	13
No	Yes	13
Yes	No	64
No	No	775

The Accuracy = 1 - (FP + FN) / Total, the accuracy of this model is 0.911.

Discussion

Generally speaking, the model fitted includes almost all information od individuals and has a good ability to predict individuals who have similar characteristics with the individuals in the dataset. And through the EDA, we also find people over 80 who have hypertension and heart disease will have a relatively high risk of stroke compared with other age groups in the same disease group or other disease groups who are also over 80.

However, since this data set is an imbalanced data set, we should balance the imbalance at first and then do the analysis. Also, the model same to be overfit and with a lower sensibility. What's more,an essay from CDC states that women exposure a higher risk of stroke with age increasing. Therefore, in the future, we should add more represent sample and more related variables to get a better fit.

Citation

[1] Course Notes for IS 6489, Statistics and Predictive Analytics, Jeff Webb https://bookdown.org/jefftemplewebb/IS-6489/

[2] https://en.wikipedia.org/wiki/Stroke.

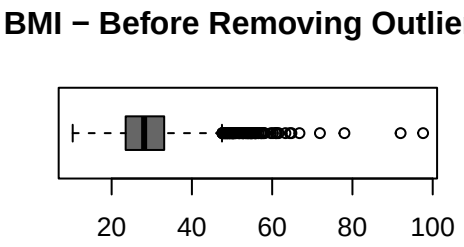
[3] https://www.kaggle.com/collinsakal/stroke-prediction-eda.

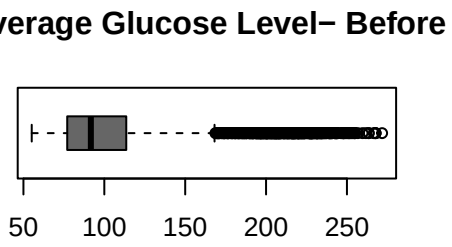
[4] Regression and othe stories, Adrew Gelman etc.

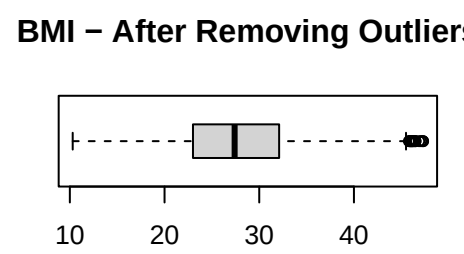
Appendix

data cleaning

removing outliers








```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction Yes  No
##        Yes  13  64
##        No   13 775
##
##                Accuracy : 0.911
##                  95% CI : (0.89, 0.9291)
##     No Information Rate : 0.9699
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2172
##
##  Mcnemar's Test P-Value : 1.212e-08
##
##             Sensitivity : 0.50000
##             Specificity : 0.92372
##          Pos Pred Value : 0.16883
##          Neg Pred Value : 0.98350
##              Prevalence : 0.03006
##          Detection Rate : 0.01503
##    Detection Prevalence : 0.08902
##       Balanced Accuracy : 0.71186
```

```
##
##        'Positive' Class : Yes
##
```

Model fit

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## weather_stroke ~ heart_disease + ever_married + work_type + avg_glucose_level +
##     bmi + smoking_status + (1 + hypertension | age_group)
##    Data: train
##
##      AIC      BIC   logLik deviance df.resid
##    863.7    955.6   -416.8    833.7     3375
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.7395 -0.2210 -0.0853 -0.0605 22.5643
##
## Random effects:
##  Groups    Name          Variance Std.Dev. Corr
##  age_group (Intercept)   2.1655   1.4716
##            hypertension1 0.3763   0.6134   1.00
## Number of obs: 3390, groups:  age_group, 7
##
## Fixed effects:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -4.310010   1.452928  -2.966  0.00301 **
## heart_disease1              0.522376   0.300252   1.740  0.08190 .
## ever_marriedYes            -0.345036   0.300859  -1.147  0.25145
## work_typeGovt_job           0.449147   1.362767   0.330  0.74171
## work_typeNever_worked      -8.990180  57.972794  -0.155  0.87676
## work_typePrivate            0.518335   1.331471   0.389  0.69706
## work_typeSelf-employed      0.562697   1.353650   0.416  0.67764
## avg_glucose_level           0.000340   0.004422   0.077  0.93871
## bmi                         0.011057   0.018074   0.612  0.54068
## smoking_statusnever smoked -0.200570   0.248404  -0.807  0.41942
## smoking_statussmokes       -0.068539   0.308506  -0.222  0.82419
## smoking_statusUnknown      -0.447980   0.320195  -1.399  0.16179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) hrt_d1 evr_mY wrk_G_ wrk_N_ wrk_tP wrk_S- avg_g_ bmi
## heart_diss1  -0.015
## ever_mrrdYs  -0.137  0.031
## wrk_typGvt_  -0.795  0.027  0.012
## wrk_typNvr_  -0.002 -0.001  0.000  0.002
## wrk_typPrvt  -0.808  0.025  0.005  0.976  0.002
## wrk_typSlf-  -0.804  0.028  0.010  0.970  0.002  0.984
## avg_glcs_lv  -0.265 -0.036 -0.029  0.011  0.000  0.007  0.010
## bmi          -0.246 -0.024 -0.023 -0.099  0.000 -0.092 -0.091 -0.054
```

```
## smkng_sttsns -0.096  0.031  0.045 -0.028  0.000 -0.024 -0.016  0.032 -0.003
## smkng_sttssm -0.075 -0.081  0.028 -0.021  0.000 -0.020 -0.016  0.001  0.010
## smkng_sttsU  -0.170  0.018  0.029  0.089  0.000  0.092  0.093 -0.006 -0.031
##              smkn_s smkng_
## heart_diss1
## ever_mrrdYs
## wrk_typGvt_
## wrk_typNvr_
## wrk_typPrvt
## wrk_typSlf-
## avg_glcs_lv
## bmi
## smkng_sttsns
## smkng_sttssm  0.475
## smkng_sttsU   0.460  0.365
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00237245 (tol = 0.002, component 1)
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: weather_stroke ~ hypertension + ever_married + heart_disease +
##     smoking_status + (1 + log(avg_glucose_level) + log(bmi) |      age_group)
##    Data: train
##
##      AIC      BIC   logLik deviance df.resid
##    855.6    935.3   -414.8    829.6     3377
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.8274 -0.2144 -0.1024 -0.0549 14.9706
##
## Random effects:
##  Groups    Name                   Variance Std.Dev. Corr
##  age_group (Intercept)            0.0000   0.0000
##            log(avg_glucose_level) 1.0708   1.0348     NaN
##            log(bmi)               0.8466   0.9201     NaN -1.00
## Number of obs: 3390, groups:  age_group, 7
##
## Fixed effects:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.77374    0.56102  -4.944 7.65e-07 ***
## hypertension1                0.71588    0.25227   2.838  0.00454 **
## ever_marriedYes             -0.34842    0.30096  -1.158  0.24699
## heart_disease1               0.50923    0.29944   1.701  0.08901 .
## smoking_statusnever smoked -0.19370    0.24843  -0.780  0.43557
## smoking_statussmokes       -0.06976    0.30932  -0.226  0.82158
## smoking_statusUnknown      -0.42185    0.31947  -1.320  0.18667
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
```

```
##              (Intr) hyprt1 evr_mY hrt_d1 smkn_s smkng_
## hypertensn1  -0.150
## ever_mrrdYs  -0.475  0.030
## heart_diss1  -0.075 -0.063  0.035
## smkng_sttsns -0.287 -0.038  0.038  0.028
## smkng_sttssm -0.221 -0.015  0.024 -0.084  0.477
## smkng_sttsU  -0.206  0.092  0.014  0.017  0.464  0.367
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: weather_stroke ~ hypertension + ever_married + smoking_status +
##     (-1 + log(avg_glucose_level) + log(bmi) | age_group) + (1 |
##     heart_disease)
##    Data: train
##
##      AIC      BIC   logLik deviance df.resid
##    855.8    917.1   -417.9    835.8     3380
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.6862 -0.2165 -0.0865 -0.0580 22.6020
##
## Random effects:
##  Groups        Name                    Variance Std.Dev. Corr
##  age_group     log(avg_glucose_level) 0.00000  0.0000
##                log(bmi)               0.22476  0.4741    NaN
##  heart_disease (Intercept)            0.06092  0.2468
## Number of obs: 3390, groups:  age_group, 7; heart_disease, 2
##
## Fixed effects:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -3.13607    0.68246  -4.595 4.32e-06 ***
## hypertension1                0.70188    0.25183   2.787  0.00532 **
## ever_marriedYes             -0.30707    0.30379  -1.011  0.31211
## smoking_statusnever smoked -0.22946    0.24785  -0.926  0.35454
## smoking_statussmokes        -0.06753    0.30825  -0.219  0.82660
## smoking_statusUnknown       -0.48611    0.32009  -1.519  0.12884
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) hyprt1 evr_mY smkn_s smkng_
## hypertensn1  -0.062
## ever_mrrdYs  -0.299  0.032
## smkng_sttsns -0.222 -0.033  0.043
## smkng_sttssm -0.207 -0.012  0.028  0.473
## smkng_sttsU  -0.207  0.089  0.026  0.463  0.365
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## boundary (singular) fit: see ?isSingular
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
```

```
##  Family: binomial  ( logit )
## Formula: weather_stroke ~ hypertension + ever_married + heart_disease +
##     smoking_status + (-1 + log(avg_glucose_level) + log(bmi) |      age_group)
##    Data: train
##
##      AIC      BIC   logLik deviance df.resid
##    849.6    910.9   -414.8    829.6     3380
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -0.8274 -0.2144 -0.1024 -0.0549 14.9716
##
## Random effects:
##  Groups    Name                    Variance Std.Dev. Corr
##  age_group log(avg_glucose_level) 1.0759   1.037
##            log(bmi)               0.8519   0.923    -1.00
## Number of obs: 3390, groups:  age_group, 7
##
## Fixed effects:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.77309    0.56062  -4.947 7.56e-07 ***
## hypertension1               0.71594    0.25227   2.838  0.00454 **
## ever_marriedYes            -0.34829    0.30097  -1.157  0.24717
## heart_disease1              0.50902    0.29945   1.700  0.08916 .
## smoking_statusnever smoked -0.19320    0.24844  -0.778  0.43678
## smoking_statussmokes       -0.06939    0.30934  -0.224  0.82250
## smoking_statusUnknown      -0.42139    0.31947  -1.319  0.18716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) hyprt1 evr_mY hrt_d1 smkn_s smkng_
## hypertensn1  -0.150
## ever_mrrdYs  -0.476  0.030
## heart_diss1  -0.075 -0.063  0.035
## smkng_sttsns -0.287 -0.038  0.038  0.028
## smkng_sttssm -0.221 -0.015  0.024 -0.084  0.477
## smkng_sttsU  -0.207  0.092  0.014  0.017  0.464  0.367
## optimizer (Nelder_Mead) convergence code: 0 (OK)
## Model failed to converge with max|grad| = 0.00488297 (tol = 0.002, component 1)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## weather_stroke ~ hypertension + heart_disease + log(avg_glucose_level) +
##     (-1 + log(bmi) | age_group)
##    Data: train
##
##      AIC      BIC   logLik deviance df.resid
##    847.0    877.6   -418.5    837.0     3385
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -0.7585 -0.2191 -0.0773 -0.0603 21.3936
##
## Random effects:
##  Groups    Name     Variance Std.Dev.
##  age_group log(bmi) 0.2023   0.4498
## Number of obs: 3390, groups:  age_group, 7
##
## Fixed effects:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -3.87325    1.93187  -2.005  0.04497 *
## hypertension1           0.73461    0.24946   2.945  0.00323 **
## heart_disease1          0.54530    0.29772   1.832  0.06701 .
## log(avg_glucose_level)  0.03361    0.41426   0.081  0.93535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) hyprt1 hrt_d1
## hypertensn1 -0.047
## heart_diss1  0.024 -0.070
## lg(vg_glc_) -0.957  0.039 -0.036
```