ANALYSIS OF
STROKE RISK FACTORS

Ha HeeJu
Sandy Seah Jin San
Tang Wei Feng
Yeo Shen Kai
Yeo Wei Jern

ANALYSIS OF

STROKE RISK FACTORS

TABLE OF CONTENTS

01
INTRODUCTION

Background, Problem statement, Assumptions

02
DATASET

Description and Investigation of dataset and the variables

03
STATISTICAL METHOD

04
HYPOTHESIS

3 Hypothesis

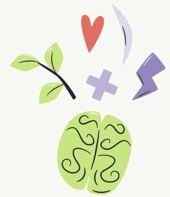
05
CONCLUSION

06
ADDITIONAL DISCUSSION

QUESTION

What attributes of a patient are linked to increase risk of having a stroke?

INTRODUCTION



BACKGROUND INFORMATION

World Health Organization (WHO):

“ one out of four people are likely to get a stroke in their lifetime and it is considered the most frequent disease that causes death globally”

“15 million people worldwide suffer a stroke annually, and of these 5 million die and another 5 million are left permanently disabled”

MOTIVATION

- Stoke is one of the most common causes of death and disabilities
- If stroke is detected early, 85% of the cases can be prevented

DATASET

Description of dataset and variables

DATASET

- **5510 Observations**
- **12 Attributes**
    - 10 Predictors

Binary Response: Stroke

- **1: had a stroke**
- **0: otherwise**

Type	Predictor Name	Data Type	Data Description
Biological	gender	Categorical	Male, Female, or Other
	age	Numerical	Age of the patient
Health	hypertension	Categorical	0: no hypertension, 1: has hypertension
	heart_disease	Categorical	0: no heart disease, 1: has a heart disease
	avg_glucose_level	Numerical	Average glucose level in blood
	bmi	Numerical	Body mass index
Lifestyle	ever_married	Categorical	No or Yes
	work_type	Categorical	Children, Govt_jov, Never_worked, Private, or Self-employed
	Residence_type	Categorical	Rural or Urban
	smoking_status	Categorical	Formerly_smoked, Never_smoked, Smokes, or Unknown

INVESTIGATING DATA

Patient ID — No repeated IDs

“smoking_status” — “Unknown”: 1544 observations (30.2%)

“stroke” — 1: stroke (4.87%) 0: otherwise (95.12%)

CLEANING DATA

Not a predictor variable

Remove ID

Filter bmi

"N/A": 201 observations

"Never_worked": 22 observations

Filter work_type

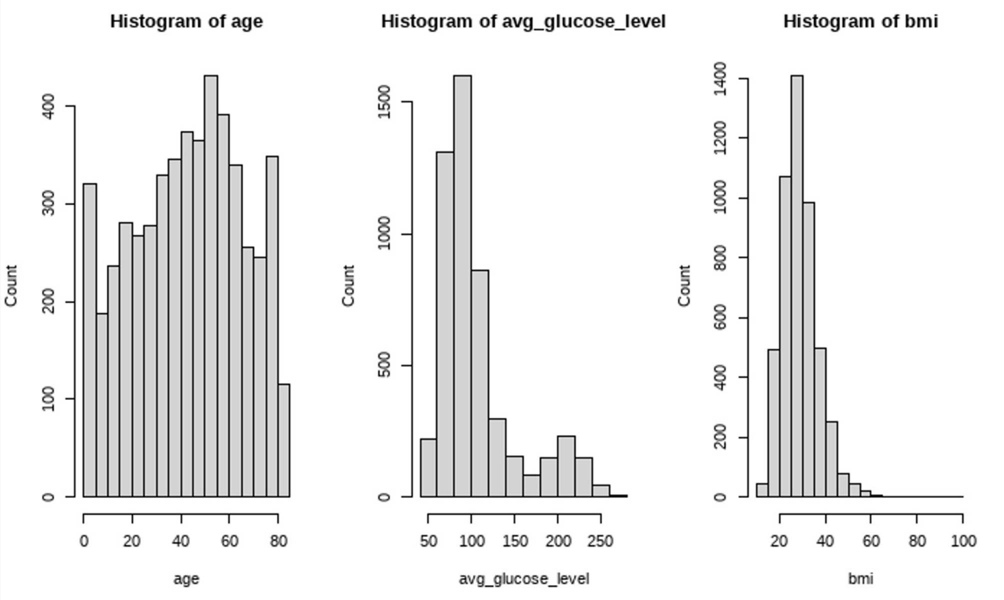
Filter gender

"Other": 1 observation

4886 Observations

FEATURE ENGINEERING

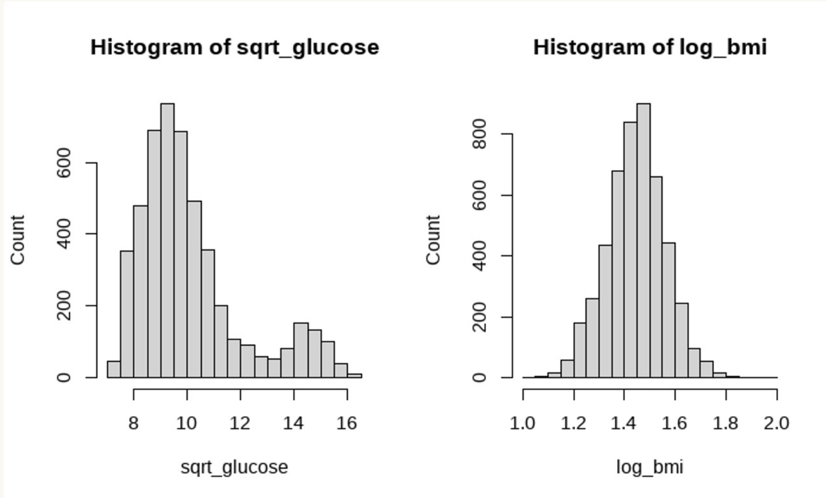
described_variables <chr>	n <int>	na <int>	mean <dbl>	sd <dbl>	se_mean <dbl>	IQR <dbl>	skewness <dbl>	kurtosis <dbl>
age	5110	0	43.22661	22.612647	0.3163304	36.000	-0.1370593	-0.9910102
avg_glucose_level	5110	0	106.14768	45.283560	0.6334759	36.845	1.5722839	1.6804785
bmi	4909	201	28.89324	7.854067	0.1120981	9.600	1.0553402	3.3626592



FEATURE ENGINEERING

avg_glucose_level → **sqrt_glucose**

bmi → **log_bmi**



described_variables <chr>	skewness <dbl>	kurtosis <dbl>
sqrt_glucose	1.267398945	0.9686552
log_bmi	-0.000482973	0.2446951

STATISTICAL METHOD

Categorical Response Variable: stroke

Binary:
1 – had a stroke
0 – otherwise

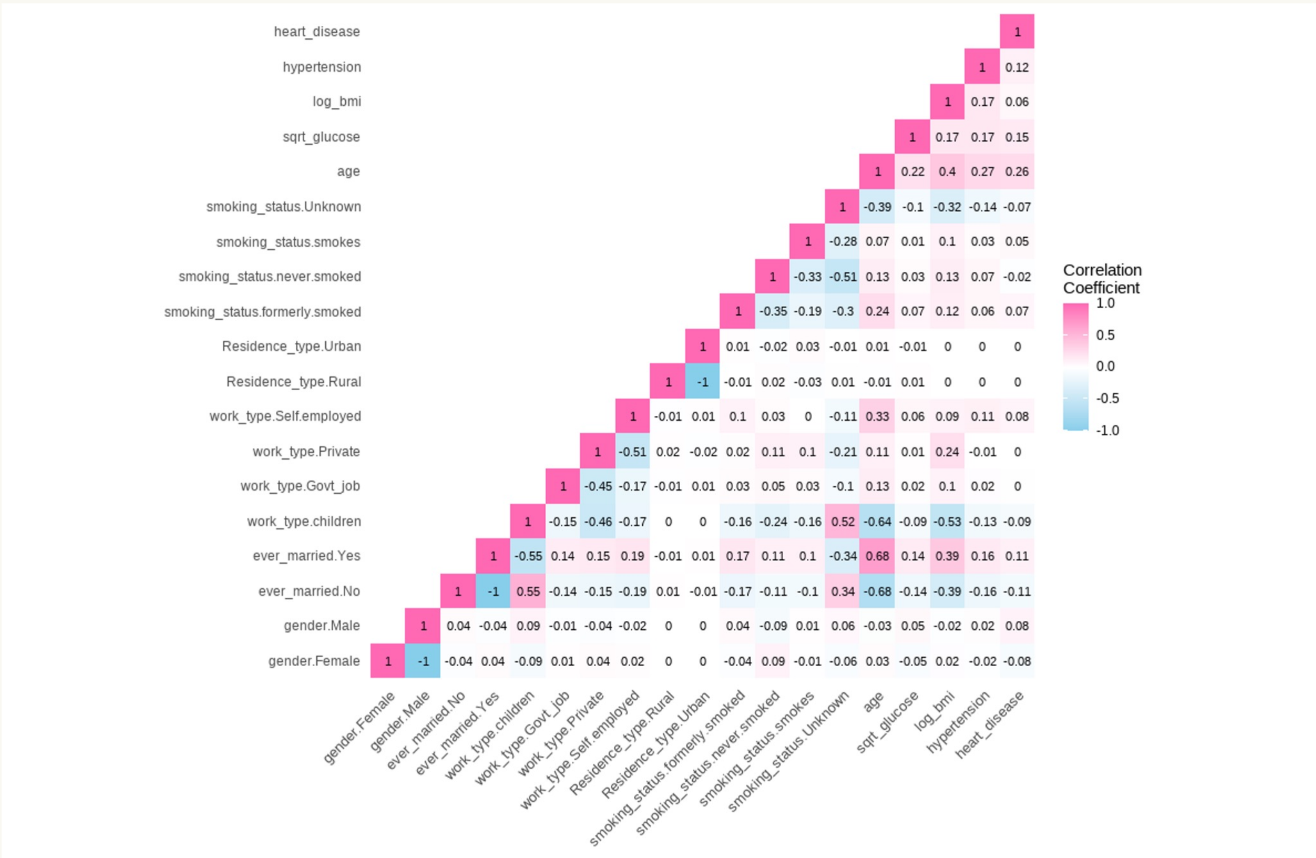
LOGISTIC REGRESSION

LOGISTIC REGRESSION
ASSUMPTIONS

1. Binary Response Variable
2. Independent Observations
3. No Multicollinearity
4. Linearity of Numerical Variables and Log odds
5. Large sample size

ASSUMPTIONS

NO MULTICOLLINEARITY



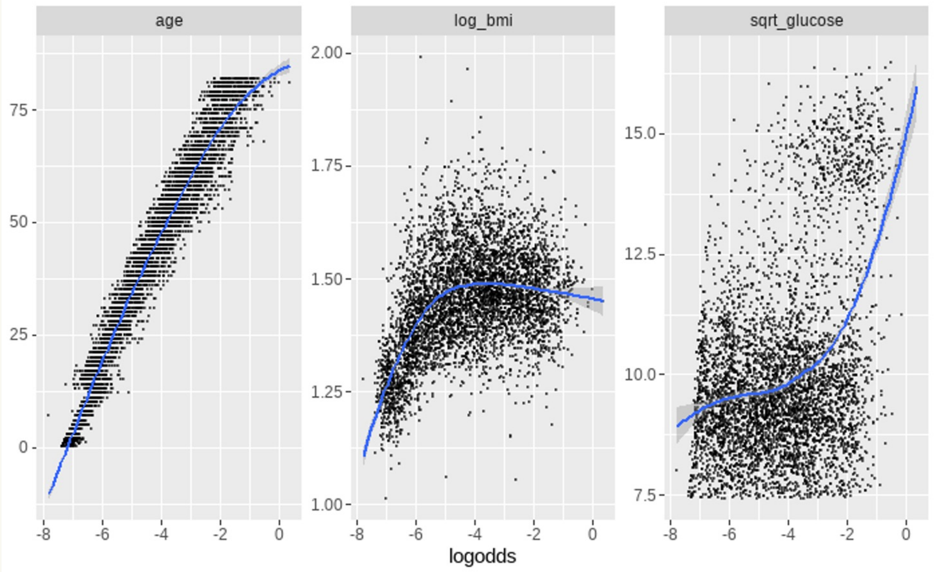
0.68

age & ever_married

ASSUMPTIONS

LINEARITY OF VARIABLES & LOG ODDS

	MLE of lambda	Score Statistic (z)	Pr(>|z|)	
age	1.1659	16.8620	< 2.2e-16	***
sqrt_glucose	1.5638	3.2509	0.001150	**
log_bmi	4.5680	2.6798	0.007368	**



ASSUMPTIONS

LARGE SAMPLE SIZE

FORMULA = (10 * # of independent variables) / expected probability of outcome

(10 * 17)/0.05 = 3400

4886

Our sample size

HYPOTHESIS

3 Hypothesis + Limitations and Possible Endogeneity

HYPOTHESIS 1

Higher BMI → Higher risk of stroke?

Coefficient: 2.3123
Odds increases by: 205.258
Statistically significant: Yes
R-Squared: 0.008193

```
Call:
glm(formula = stroke ~ log_bmi, family = "binomial", data = stroke_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5321  -0.3150  -0.2895  -0.2635   2.7092

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.4834     0.9092  -7.131 9.98e-13 ***
log_bmi       2.3123     0.6151   3.759 0.000171 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1726.4  on 4885  degrees of freedom
Residual deviance: 1712.2  on 4884  degrees of freedom
AIC: 1716.2

Number of Fisher Scoring iterations: 6
```

HYPOTHESIS 1

Higher BMI → Higher risk of stroke?

Control:

1. Gender
2. Age

Coefficient: 1.276
Odds increases by: 18.881
Statistically significant: No
R-Squared: 0.1863

```
Call:
glm(formula = stroke ~ log_bmi + gender + age, family = "binomial",
    data = stroke_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7949  -0.3119  -0.1667  -0.0741   3.5772

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.361634   1.329823  -7.040 1.93e-12 ***
log_bmi      1.276020   0.828489   1.540    0.124
genderMale   0.092099   0.149095   0.618    0.537
age          0.076000   0.005479  13.872  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1726.4  on 4885  degrees of freedom
Residual deviance: 1404.8  on 4882  degrees of freedom
AIC: 1412.8

Number of Fisher Scoring iterations: 7
```

HYPOTHESIS 1

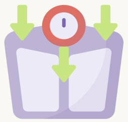
LIMITATIONS / ENDOGENEITY

Omitted Variable Bias

- Low R-squared value
- Other confounding variables

Reverse Causality

- BMI → Stroke?
- Stroke → BMI?



Errors In Variable (EIV) Bias

- NA values in BMI

HYPOTHESIS 1

CONCLUSION

Not enough evidence to imply causality between BMI and stroke

- Low correlation
- High presence of endogeneity

Improvements

- More control variables
- Better data collection

HYPOTHESIS 2

Different residence type (rural instead. of urban) → Higher risk of stroke?

Coefficient: 0.06266
Odds increases by: 1.155
Statistically significant: No
R-Squared: 0.0001

```
Call:
glm(formula = stroke ~ Residence_type, family = "binomial", data = stroke_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.3001  -0.3001  -0.2911  -0.2911   2.5229

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.14027    0.10214 -30.745   <2e-16 ***
Residence_typeUrban  0.06266    0.14153   0.443    0.658
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1726.4  on 4885  degrees of freedom
Residual deviance: 1726.2  on 4884  degrees of freedom
AIC: 1730.2

Number of Fisher Scoring iterations: 6
```

HYPOTHESIS 2

Different residence type (rural instead. of urban) → Higher risk of stroke?

Control:

1. Gender
2. Ever_married
3. Age

Coefficient: 0.011170
Odds increases by: 1.026
Statistically significant: No
R-Squared: 0.185

```
Call:
glm(formula = stroke ~ Residence_type + gender + ever_married +
    age, family = "binomial", data = stroke_df2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7621  -0.3100  -0.1666  -0.0800   3.5539

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -7.366188   0.405091 -18.184   <2e-16 ***
Residence_typeUrban  0.011170   0.147700   0.076    0.940
genderMale           0.099289   0.149284   0.665    0.506
ever_marriedYes     -0.083318   0.239177  -0.348    0.728
age                  0.075218   0.005374  13.996   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1726.4  on 4885  degrees of freedom
Residual deviance: 1407.0  on 4881  degrees of freedom
AIC: 1417

Number of Fisher Scoring iterations: 7
```

HYPOTHESIS 2

LIMITATIONS / ENDOGENEITY

Omitted Variable Bias

- Other confounding variables - income, ethnicity, etc.

Errors In Variable (EIV) Bias

- No error-in-variable bias

HYPOTHESIS 2

CONCLUSION

Not enough evidence to imply causality between Residence_Type and stroke

Improvements

- Supporting research on other possible confounders
- Increased data collection on these possible confounders (while considering privacy)

HYPOTHESIS 3

Average glucose level → Higher risk of stroke?

Coefficient: 0.26021
Risk increases by: 0.06771
Statistically significant
R-Squared: 0.0411280

```
Call:
glm(formula = stroke ~ sqrt_glucose, family = "binomial", data = stroke_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6038  -0.2939  -0.2562  -0.2299   2.8065

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.86767    0.33665 -17.429   <2e-16 ***
sqrt_glucose  0.26021    0.02938   8.857   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
```

HYPOTHESIS 3

Average glucose level → Higher risk of stroke?

Control:

- Smoking status
- Gender
- age

Coefficient: 0.128070
Risk increases by: 0.01640
Statistically significant
R-Squared: 0.1996651

```
Call:
glm(formula = stroke ~ sqrt_glucose + smoking_status + gender +
    age, family = "binomial", data = stroke_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9186  -0.3013  -0.1604  -0.0731   3.7255

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -8.632113   0.504134 -17.123  < 2e-16 ***
sqrt_glucose                0.128070   0.029365   4.361 1.29e-05 ***
smoking_statusnever smoked -0.050068   0.187048  -0.268    0.789
smoking_statussmokes        0.342679   0.227036   1.509    0.131
smoking_statusUnknown      -0.303775   0.243737  -1.246    0.213
genderMale                  0.019290   0.152429   0.127    0.899
age                         0.073025   0.005581  13.084  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

HYPOTHESIS 3

LIMITATIONS / ENDOGENEITY

Omitted Variable Bias

- Diabetes affects both the risk of strokes and average glucose levels.

Reverse Causality

- Stroke victims likely to have high glucose level, leading to a higher risk of stroke.

Errors In Variable (EIV) Bias

- No error-in-variable bias

HYPOTHESIS 3

CONCLUSION

Statistically significant positive causal effect.

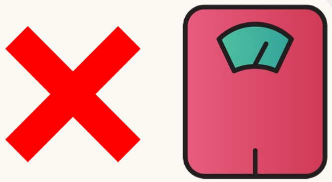
Improvements

- R-Squared is low, more research tobe done to control for possible confounders.
- Increased data collection on these possible confounders (while considering privacy)

KEY FINDINGS

After controls incorporated in regression model:

BMI: Not statistically significant



Residence_type: Not statistically significant

Glucose level: Statistically significant

POLICY IMPLICATIONS

BMI + Residence type

- Implementing policies
- May not be as relevant for proactive health interventions (promotional materials) specifically to target stroke-vulnerable populations

SUGAR

- Sugar tax
- Promotional materials, etc.

Thank you!