Introduction

Background

A stroke is a medical emergency that happens when the blood flowing to the brain is blocked. According to the World Health Organization[1], one out of four people are likely to get a stroke in their lifetime, and it is considered the second most frequent disease that causes death globally. Furthermore, 15 million people worldwide suffer a stroke annually, and of those 5 million die and another 5 million are permanently disabled.[2]

Problem Statement

Research has shown that if stroke is detected or diagnosed early, death and severe damage to the brain can be prevented in 85% of cases[3]. Therefore, our group was motivated to utilise a dataset that collected different features from patients to predict which parameters may increase the risk of getting a stroke. The causal relationship found can be useful in suggesting solutions to decrease the likelihood of suffering from a stroke. Our group used a logistic regression model to investigate the possible causality between getting a stroke with 3 factors: Body Mass Index (BMI), residence type (urban or rural), and average glucose level.

Dataset

Description of Dataset

The dataset is obtained from "Stroke Prediction Dataset" by Frederico Soriano on kaggle[4]. It contains 5510 observations and 12 attributes regarding a patient. The source and collection methodology of the dataset is confidential to protect the personal information of patients, and the data was made available for educational purposes. Though the credibility of the dataset and its data is unclear, the dataset has claimed a gold medal in the kaggle community.

Description of Variables

12 attributes of a patient have been provided in the dataset:

1. **id**: unique identifier of the patient
2. **gender**: "Male", "Female" or "Other"
3. **age**: age of the patient
4. **hypertension**: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5. **heart_disease**: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

---

[1] World Health Organization. (n.d.). *World stroke day*. World Health Organization. Retrieved April 9, 2023, from https://www.who.int/southeastasia/news/detail/28-10-2021-world-stroke-day

[2] WHO EMRO. (n.d.). *Stroke, Cerebrovascular accident.* World Health Organization - Regional Office for the Eastern Mediterranean. Retrieved April 9, 2023, from https://www.emro.who.int/health-topics/stroke-cerebrovascular-accident/index.html

[3] Kaur, M., Sakhare, S. R., Wanjale, K., & Akter, F. (2022, April 11). *Early stroke prediction methods for prevention of Strokes*. Behavioural neurology. Retrieved April 9, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9017592/

[4] Fedesoriano. (2021, January 26). *Stroke prediction dataset*. Kaggle. Retrieved April 9, 2023, from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

6. **ever_married**: “No” or “Yes”
7. **work_type**: “children”, “Govt_jov”, “Never_worked”, “Private” or “Self-employed”
8. **Residence_type**: “Rural” or “Urban”
9. **avg_glucose_level**: average glucose level in blood of the patient
10. **bmi**: body mass index of the patient
11. **smoking_status**: “formerly smoked”, “never smoked”, “smokes” or “Unknown”
12. **stroke**: 1 if the patient had a stroke , or 0 if not

The predictive variables can be grouped into three distinct types:

1. Biological: gender, age
2. Health: hypertension, heart_disease, avg_glucose_level, bmi
3. Lifestyle: ever_married, work_type, Residence_type, smoking_status

The other two remaining variables are the unique identifier (id) and the target variable (stroke). It is important to note that variables within the same group could be correlated with each other, which can create confounding effects. For example, a person with hypertension is possibly more likely to have heart disease as well. To ensure accurate analysis of causality, it is essential to control for potential confounding variables when analysing the data in our hypothesis.

Investigating the Variables

Upon further examination of each variable in the dataset, we determined that there are no instances of duplicate patient IDs. However, smoking status includes a category labelled “Unknown” which comprises a substantial portion of the data with 1544 observations, accounting for 30.2% of the total dataset. Due to its significant proportion, we chose to retain this category as a separate group in our analysis.

A few variables with a limited number of observations in each category were also discovered. For instance, the variable gender has 1 observation with category “Other” (0.02%), and work type has 22 observations with category “Never_worked” (0.43%). BMI also has 201 observations with value “N/A” (3.93%). As these records make up a relatively small portion of the overall dataset, we chose to exclude them to improve the causal inferences in our analysis.

Preliminary analysis revealed that the numerical variables avg_glucose_level and bmi were significantly skewed, potentially impacting the accuracy of our models (Figure 1). Therefore, we performed a square root transformation on avg_glucose_level to derive sqrt_glucose, and a log base 10 transformation on bmi to derive log_bmi. These transformations were done to enhance the suitability of the data for analysis, by producing more normally distributed variables, which we believe could help improve the overall fit of the models.

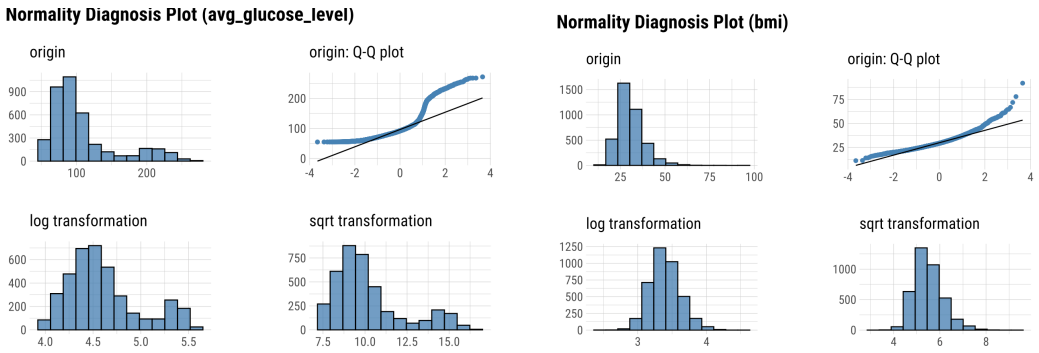

Figure 1. Normality plot of avg_glucose_level (left) and bmi (right)

Assumptions

In order to conduct a binary classification, the logistic regression follows a different set of assumptions[5] compared to linear regression.

First, the binary logistic regression needs the dependent variable to be binary. Since the dependent variable is discrete with 2 possible values, 1 being a patient with stroke and 0 otherwise, we meet the first criteria.

Second, the observations in the data should be independent of each other. Since there was no repeated patient ID in the dataset, the second assumption is met.

Third, there should be little or no multicollinearity in the independent variables. According to the correlation matrix plot (Figure 2), highest correlation observed is 0.68, between the variables age and ever_married, which implies that all the correlation coefficients are below 0.7[6]. Therefore, there is no multicollinearity between the independent variables.

[5] *Assumptions of logistic regression*. Statistics Solutions. (2021, August 11). Retrieved April 9, 2023, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

[6] M, R. (2020, March 23). *Correlation and collinearity-how they can make or break a model*. Medium. Retrieved April 9, 2023, from https://blog.clairvoyantsoft.com/correlation-and-collinearity-how-they-can-make-or-break-a-model-9135fbe6936a

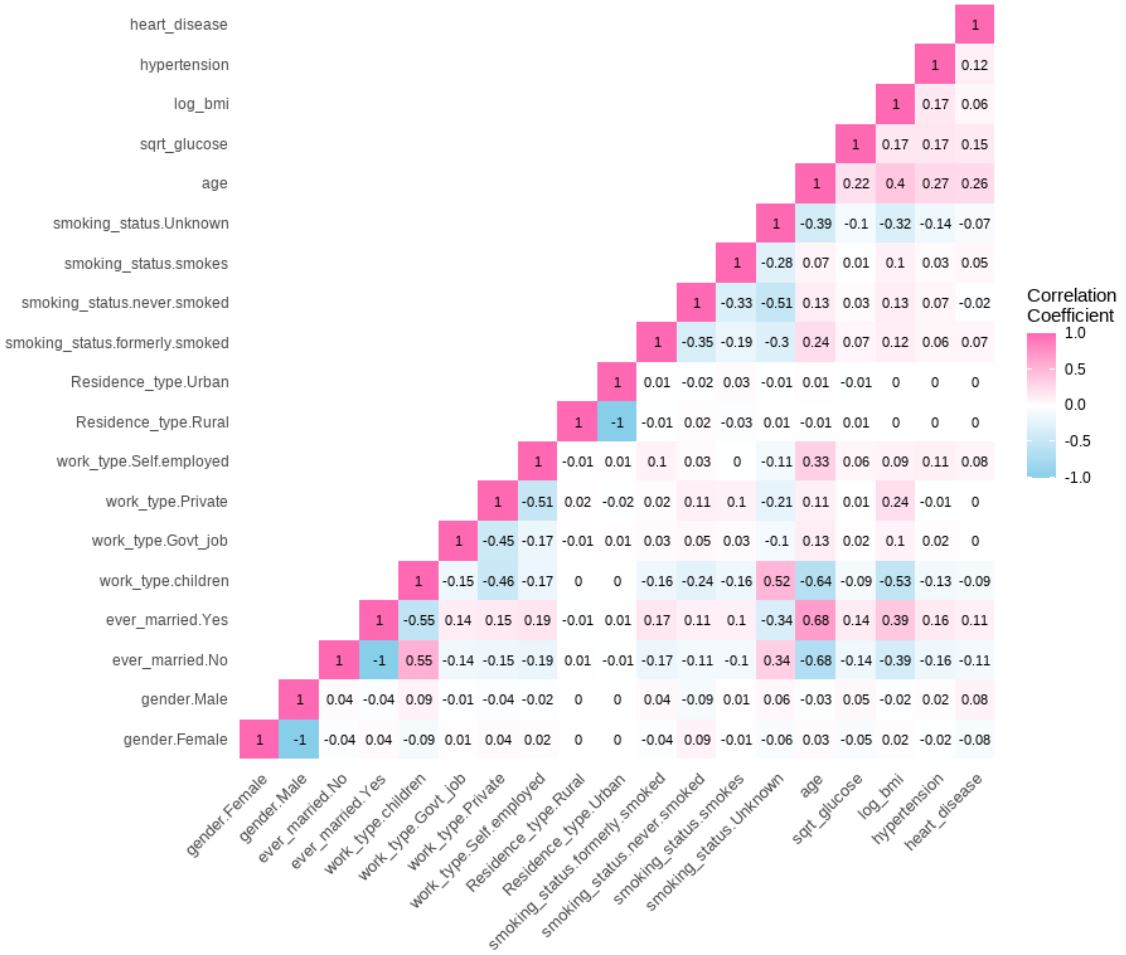

Figure 2. Correlation matrix of stroke dataset

Fourth, logistic regression requires the relationship between continuous independent variables and log odds to be linear. After conducting the Box-Tidwell test, p-values of some of the independent variables were statistically significant (i,e, $p \leq 0.5$) implying that the independent variables and natural log of the variable are not linear. Therefore, the fourth hypothesis is not satisfied.

	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	**

Figure 3. Results of Box-Tidwell test

Finally, a large sample size is required for accurate analysis. Our data has a total of 17 independent variables and the probability of getting a stroke is .05. When following the general guideline to calculate a minimum sample size (10 * 17 / 0.05), performing a logistic regression on our dataset requires at least 3400 observations. After performing our data cleaning, we were left with 4886 data points, which is a sufficiently large sample size.

Hypotheses

BMI

The first hypothesis tested whether there is a direct causal relationship between a higher BMI and having an increased risk of having a stroke.

The most common type of stroke, ischemic stroke, happens when the brain's blood vessels become narrowed or blocked, which greatly reduces blood flow. Similar to that of a heart attack, this occurs when the blood vessels are blocked or narrowed by fatty deposits. Therefore, being overweight could be a major risk factor contributing to stroke occurrence, which leads us to examine the effect of BMI, a measurable indicator of healthy weight levels.

To investigate the relationship between BMI and stroke risk, a single-variable logistic regression model was used to fit the log odds of stroke to the log_bmi variable. The following results were obtained from the analysis:

```
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
```

Figure 4. log_bmi vs stroke naive regression model

The regression output shows that a unit increase in log BMI increases the log odds of a stroke by 2.3123 on average. Consequently, this means that when BMI increases by 10 times, the odds of a stroke increases by 205.258, and the probability of a stroke increases by 0.995 on average. The p-value of the coefficient for log BMI is below 0.05, which means that it is statistically significant at a 95% confidence level. Therefore we can conclude that BMI is indeed correlated with the probability of having a stroke. The R-squared value, calculated as McFadden's pseudo R-squared value for logistic regression is 0.008193. This means that around 0.8% of the variance in log odds of stroke is explained by log BMI, which is extremely low.

Attempting to improve on the model, a more extensive logistic regression was fitted, which included gender and age as controls. Both of those are added to the model as they are uncontrollable biological

factors that affect each and every patient, and gender[7] and age[8] are also known risk factors of stroke. The updated model is as follows:

```
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
```

Figure 5. log_bmi vs stroke logistic model with controls

The output now shows that a unit increase in log BMI increases the log odds of a stroke by 1.276 on average, holding other variables constant. This means that when BMI increases by 10 times, the odds of a stroke increases by 18.881, and the probability of a stroke increases by 0.950 on average, holding other variables constant. However, the p-value of this coefficient is above 0.05, which means that it is no longer significant, therefore is not different from 0. With the control variables included, BMI is not correlated with the odds of a stroke. The R-squared value has increased to 0.1863, which means that the model explains around 18.6% of the variation in the log odds of having a stroke. The increase in R-squared value is likely to be due to the age control variable.

Overall, we do not find enough evidence to imply causality between BMI and stroke, due to the low correlation and high presence of endogeneity. Endogeneity concerns are addressed under the limitations section below. To improve the model, more controlling variables need to be added, and better data collection to be done.

[7] Wyller T. B. (1999). Stroke and gender. The journal of gender-specific medicine : JGSM : the official journal of the Partnership for Women's Health at Columbia, 2(3), 41–45.

[8] 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

Residence

The second hypothesis tested whether there is a direct causal relationship between the region a person lives in (rural or urban) and if the person has a stroke.

In some countries, it has been found that rural stroke patients tend to have higher mortality rates than urban stroke patients. One potential explanation is that people living in rural areas have a higher incidence of stroke than people living in urban areas. Further investigation of whether living in a rural area has a direct causal effect on the probability of stroke can help identify possible causal mechanisms and hence, guide regulation to reduce stroke incidence and mortality.

A single-variable logistic regression model was used to fit the log odds of stroke to the Residence_type variable. The following results were obtained from the analysis:

```
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
```

Figure 6. Residence_type vs stroke naive regression model

The regression output shows that the p-value of Residence_type being urban (instead of rural) is at 0.658, which is greater than 0.05, and hence we find that the impact of residence_type on the probability of having had a stroke is statistically insignificant at a 95% confidence level. The naive regression hence suggests that there is likely no causal relationship between a person’s residence_type (rural or urban) and his likelihood of having had a stroke. McFadden's pseudo R-squared was very low at 0.0001, which means that only around 0.01% of the variance in log odds of stroke is explained by Residence_type, suggesting that it has very poor explainability for the likelihood of a stroke.

Moving on to a more comprehensive model. Instead of solely regressing on residence_type, confounders like age, along with other variables that may have a causal relationship on stroke such as ever_married, gender were added. Age was identified as a possible confounder as it has a strong

causal effect on stroke[9], while also having a likely causal effect on residence_type through 2 direct effects: a rural to urban migration trend of the young, and a urban to rural migration of the old[10].

We choose not to regress on behaviour and lifestyle-related factors, such as glucose level, bmi, smoking status, and hypertension as such factors may act as mediators between a person's residence type and his susceptibility to having a stroke. The updated results as follows:

```
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
```

Figure 7. Residence_type vs stroke regression model with controls

From the results above, we find that even after controlling for confounding variables and other relevant variables, residence_type p-value of 0.940, which is greater than 0.05. Hence, it is statistically insignificant, implying that it is unlikely that there is causality between the region a person lives in and their likelihood of having a stroke. McFadden's pseudo R-squared for this more comprehensive logistic regression model 0.185, suggesting that incorporating gender, past marriage status and age significantly improved the model's explainability for the likelihood of a stroke.

From the studies above, we may reject the suggested causal explanation of higher stroke mortality in rural areas as one of higher stroke incidence among people living in the United States' rural areas

[9] Yousufuddin, M., & Young, N. (2019). Aging and ischemic stroke. Aging, 11(9), 2542–2544. https://doi.org/10.18632/aging.101931

[10] Cromartie, P. (2018, December 20). Rural aging occurs in different places for very different reasons. Retrieved April 6, 2023, from https://www.usda.gov/media/blog/2018/12/20/rural-aging-occurs-different-places-very-different-reasons

ceteris paribus. Therefore, further studies may instead be conducted to investigate other potential causal mechanisms to higher stroke mortality.

Glucose Level

The third hypothesis tested whether there is a direct causal relationship between a person's blood glucose level and their likelihood of experiencing a stroke.

High glucose levels have been shown to be correlated with an increased risk of experiencing a stroke. Hyperglycemia is a condition characterised by high blood sugar level or glucose levels, and studies have shown that a large proportion of stroke victims have a history of hyperglycemia[11]. Over time, hyperglycemia can causes damage to blood vessels[12], leading to the prevalence of blood clots and an increased risk of having a stroke.

To investigate this potential causal relationship, a single-variable logistic regression model was used to fit the log odds of stroke to the sqrt_glucose variable. The following results were obtained from the analysis:

```
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)
```

Figure 8. sqrt_glucose vs stroke naive regression model

The regression output shows that the coefficient of sqrt_glucose is 0.26021, thus for every 1 unit increase in the glucose level, the odds of having a stroke increases by 0.06771. Furthermore, this coefficient is statistically significant at a 95% confidence level, as the p-value is below 0.05, suggesting that the coefficient of sqrt_glucose differs from 0. To estimate the explanatory power of the model, McFadden's pseudo R-squared was used to get an approximate R-Squared value of 0.041128 which is rather low. Thus, the changes in the glucose level explains 4.11% of the changes in stroke.

Further edits were made to the model with the aim to improve the explanatory power. To control for possible confounders, more variables were added to act as controls such as smoking status, gender, and age, which could possibly be confounders. Running the regression model again, the results are as follows:

---

[11] *Hyperglycemia in acute stroke | stroke*. (n.d.). Retrieved April 9, 2023, from https://www.ahajournals.org/doi/10.1161/01.STR.0000115297.92132.84

[12] *Diabetes & stroke: Causes, symptoms, treatment & prevention. Cleveland Clinic. (n.d.). Retrieved April 9, 2023, from https://my.clevelandclinic.org/health/diseases/9812-diabetes-and-stroke#:~:text=Over%20time%2C%20high%20 glucose%20 levels,Heart%20disease*

```
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)
```

Figure 9. sqrt_glucose vs stroke regression model with controls

As shown above, the new coefficient for the variable sqrt_glucose is 0.12807, thus for every 1 unit increase in glucose level, the odds of having a stroke increases by 0.016402, keeping all other variables constant. Furthermore, this is statistically significant at a 95% confidence level, as the p-value is lesser than 0.05. McFadden's pseudo R-squared returns a new R-Squared value of 0.19966, which implies that after including the new controls, glucose level explains 19.97% of the changes in stroke, implying that the explanatory power has improved .

To conclude for the third hypothesis, we find that there is a statistically significant causal effect of average glucose level of a person on the incidence of stroke of an individual, despite the low R-squared value. More research could be done to control for possible confounders, such as historical health conditions and medication history, which should help to improve the model explainability.

Limitations

General

General limitation that we faced when conducting our causal analysis via logistic regression was about our assumptions. While most assumptions were satisfied, such as no multicollinearity, the linearity assumption of the log-odds data to BMI and glucose level was not satisfied. While we applied a transformation to both of these variables, linearity was still not at the desired level required to fulfil the assumption. Further improvements to the model can be made to reduce bias.

BMI

To be able to imply causality between bmi and stroke, various assumptions need to be fulfilled. Gauss-Markov assumptions are less applicable in logistic regression, but some parts are still similar. There is no need for the homoscedasticity and autocorrelation assumptions, but others remain. Firstly, the linearity assumption is satisfied, as the non-linear variables, such as bmi, have been transformed to make it more linear.

Secondly, there are no significant outlier values, as seen from the Cook's distance plot.

Third, no multicollinearity has been satisfied, as determined earlier.

Exogeneity condition is not fulfilled, as there is likely to be an omitted variable bias, as seen from the low R-squared value, especially in the single variable model used. Reverse causality is likely to be present as well, as stroke may affect BMI instead of the other way round. Experiencing a stroke usually has a large impact on one's physical health and lifestyle, which could lead to extreme changes in their BMI. Error in variable bias may also exist, as seen from the "N/A: values in BMI, which may contain important data.

Residence

The residence_type model has certain limitations, an example being due to patient confidentiality, the data source is not available to the public. Hence, despite its credible provenance, certain characteristics, such as the country in which the data is sourced from, were not identified. This has implications for endogeneity, as country-specific causal mechanisms for confounders cannot be identified and investigated, and may introduce hidden bias in the model via omitted variable bias. Hence, this issue cannot be easily resolved, as greater access to data is not a modelling issue, but rather an ethical issue.

Glucose

The model has a few limitations, as there is evidence of possible endogeneity in the model. There is evidence of omitted variable bias being present, as possible confounders such as pre-existing health conditions are not accounted for. Specifically type 1 diabetes and hyperglycemia, as people with diabetes are twice as likely to have a stroke compared to those without diabetes[13], and people with hyperglycemia[14] are unable to effectively convert glucose into glycogen which results in higher blood glucose level.

Additionally, reverse causality may be present, where there is a two-way causal relationship between the average glucose level and risk of stroke. This arises from the fact that research has shown that stroke victims are more likely to develop hyperglycemia[15], resulting in a high glucose level which in turn increases the risk of experiencing a stroke. To overcome these biases, the model should include historical health conditions and medication history after the stroke, as it is possible that the medication will counteract the effect of the stroke on glucose levels.

---

[13] *Stroke*. Stroke | ADA. (n.d.). Retrieved April 9, 2023, from https://diabetes.org/diabetes/stroke#:~:text=If%20you%20have%20a diabetes%2C%20 your,risk%20of%20 getting%20a%20 stroke

[14] *Hyperglycemia in acute stroke | stroke*. (n.d.). Retrieved April 9, 2023, from https://www.ahajournals.org/doi/10.1161/01.STR.0000115297.92132.84

[15] U.S. Department of Health and Human Services. (2019, July 23). *Researchers get a handle on how to control blood sugar after a stroke*. National Institutes of Health. Retrieved April 9, 2023, from https://www.nih.gov/news-events/news-releases/researchers-get-handle-how-control-blood-sugar-after-stroke#:~:text=Hyperglycemia%2C%20or%20high%20levels%20of,to%20normal%20blood%20sugar%20levels

Conclusion

Key findings

In conclusion, we found that there is a statistically significant causal effect of average glucose level on the incidence of stroke of an individual. While our findings are statistically significant, it is important to acknowledge the limitations to our study, such as the low R-squared value and possible confounders not being included.

However, we found that there is no statistically significant causal effect by either bmi or residence_type on a person's incidence of stroke. Further studies can be conducted on the causal effect of these independent factors on the incidence of stroke, while controlling for more confounders and using a more comprehensive set of data. Nevertheless, the above studies still have its merits and can be considered in guiding both further research and health policy decisions taken to reduce stroke incidence at regional and national levels.

Policy implications

Given the challenges faced by policymakers in allocating resources to promote favourable health outcomes, including stroke prevention, we wish to acknowledge the difficult decisions they must make. Therefore, the below recommendations are to be considered and balanced with other policy objectives, to help policymakers make informed decisions that will best serve the needs of their communities.

For government bodies looking to significantly reduce stroke incidence within their governed population, we believe that there is little benefit in conducting policy-targeting on populations with high bmi, or specifically targeting either rural or urban populations. While such strategies may have positive health benefits on other indicators of overall health of the population, it is not so for stroke.

Instead, governments could look into allocating resources to lowering the average glucose level of their population, particularly among those with other high risk factors. This can be done through a few different strategies, such as implementing new or raising existing sugar taxes, providing subsidies for healthier food options, and conducting promotional campaigns to raise awareness of alternatives to high-sugar foods. In this manner, the focusing of policy efforts on reducing sugar intake should results in lowering the average glucose levels of the population, resulting in an overall positive effect in reducing stroke incidence.