A REPORT ON DIABETES RISK FACTORS IN AMERICA

Getrude Moraa Nyabuto

April 23, 2024

1 Introduction

Diabetes is a chronic disease that occurs when the body is unable to produce or use insulin properly. The disease has been associated with several risk factors such as high blood pressure, high cholesterol, obesity, smoking, and lack of physical activity. In this report, we analyze the factors responsible for diabetes using logistic regression analysis.

1.1 Dataset

The data used in this analysis was obtained from Kaggle an open-access platform from a Diabetes Health indicator done by CDC BRFSS2015. After data cleaning and manipulation we end up with 3600 survey responses.

```
##     Diabetes       HighBP          HighChol           BMI       
##  Min.   :0.0   Min.   :0.0000   Min.   :0.0000   Min.   :12.00  
##  1st Qu.:0.0   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:25.00  
##  Median :0.5   Median :1.0000   Median :1.0000   Median :29.00  
##  Mean   :0.5   Mean   :0.5635   Mean   :0.5257   Mean   :29.86  
##  3rd Qu.:1.0   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:33.00  
##  Max.   :1.0   Max.   :1.0000   Max.   :1.0000   Max.   :98.00  
##      Smoker           Stroke        HeartDiseaseorAttack  PhysActivity   
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000       Min.   :0.000  
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000       1st Qu.:0.000  
##  Median :0.0000   Median :0.00000   Median :0.0000       Median :1.000  
##  Mean   :0.4753   Mean   :0.06217   Mean   :0.1478       Mean   :0.703  
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.0000       3rd Qu.:1.000  
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000       Max.   :1.000
```

1.2 Research Question

1. What are the risk factors for diabetes in America?

2 Methodology

The aim of this study is to investigate the factors responsible for diabetes by fitting a multiple logistic regression model using R. The next step involved data manipulation, which included removing missing values and removing duplicates. The third step involved converting factor variables to numeric variables.

A multiple logistic regression model is fitted to the data using the *glm*() function. The family is set to binomial to specify the distribution to be used for the response variable. The summary() function is used to print the model's results, which include the deviance residuals, coefficients, standard errors, z-values, and p-values for each variable.

```
## Warning in system("timedatectl", intern = TRUE): running command
'timedatectl' had status 1
```

```
## [1] 3600
##
## Call:
## glm(formula = Diabetes ~ HighBP + HighChol + BMI + Smoker + Stroke +
##     HeartDiseaseorAttack + PhysActivity, family = "binomial",
##     data = newdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8432  -1.2390   0.8609   1.0628   1.4487
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.755965   0.129659  -5.830 5.53e-09 ***
## HighBP                0.109712   0.069809   1.572   0.1160
## HighChol              0.076816   0.068722   1.118   0.2637
## BMI                   0.021138   0.002589   8.163 3.26e-16 ***
## Smoker               -0.042529   0.068663  -0.619   0.5357
## Stroke                0.373421   0.074628   5.004 5.62e-07 ***
## HeartDiseaseorAttack  0.384631   0.071248   5.398 6.72e-08 ***
## PhysActivity         -0.136956   0.068515  -1.999   0.0456 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4919.3  on 3599  degrees of freedom
## Residual deviance: 4800.4  on 3592  degrees of freedom
```

```
## AIC: 4816.4
##
## Number of Fisher Scoring iterations: 4
```

Calculating P- Values for the multiple logistic regression model

```
## [1] 1.706166e-38
```

A reduced model is fitted and includes only the significant variables (BMI, Stroke, HeartDiseaseorAttack, and PhysActivity). The summary() function was used to print the model's results.

```
##
## Call:
## glm(formula = Diabetes ~ BMI + Stroke + HeartDiseaseorAttack +
##     PhysActivity, family = "binomial", data = newdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8448  -1.2415   0.8655   1.0606   1.4125
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.701154   0.117219  -5.982 2.21e-09 ***
## BMI                   0.021650   0.002576   8.403  < 2e-16 ***
## Stroke                0.386011   0.074197   5.203 1.97e-07 ***
## HeartDiseaseorAttack  0.395951   0.070850   5.589 2.29e-08 ***
## PhysActivity         -0.139369   0.068448  -2.036   0.0417 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4919.3  on 3599  degrees of freedom
## Residual deviance: 4804.7  on 3595  degrees of freedom
## AIC: 4814.7
##
## Number of Fisher Scoring iterations: 4
```

Calculating P- Values

```
## [1] 1.536209e-38
```

3 Confidence Interval

Calculate odds ratios and confidence intervals

```
## Waiting for profiling to be done...

##                              OR     2.5 %    97.5 %
## (Intercept)          0.4960127 0.3938053 0.6235700
## BMI                  1.0218862 1.0167780 1.0271024
## Stroke               1.4711014 1.2724578 1.7020626
## HeartDiseaseorAttack 1.4857962 1.2935154 1.7076657
## PhysActivity         0.8699074 0.7606461 0.9947678
```

4 Results

The multiple logistic regression analysis aimed to identify factors responsible for diabetes. The initial model included all predictors, but only BMI, Stroke, Physical activity and HeartDiseaseorAttack were found to be significant. The final model, including only these significant predictors, was found to have a better fit than the initial model. The p-values for both models were very small, indicating that the models were statistically significant. Therefore, we can conclude that BMI, Stroke, and HeartDiseaseorAttack are significant predictors of diabetes.

The output of the logistic regression analysis shows the odds ratios (OR) and corresponding 95% confidence intervals for each predictor in the model. The Intercept has an OR of 0.496, indicating that the odds of Diabetes are about half as high when all other predictors are held constant. The predictor BMI has an OR of 1.022, indicating that for every one-unit increase in BMI, the odds of Diabetes increase by a factor of 1.022. In addition, people with a history of stroke or heart disease/attack have higher odds of Diabetes, with odds ratios of 1.471 and 1.486, respectively. Finally, the predictor for physical activity has an OR of 0.870, indicating that people who engage in physical activity have lower odds of having Diabetes compared to those who do not engage in physical activity.

The confidence intervals around these estimates provide a range of plausible values for the true odds ratios with 95% confidence. For example, the 95% confidence interval for the OR of BMI ranges from 1.017 to 1.027, indicating that if we were to repeat the study many times, we would expect the true odds ratio to lie between these values in 95% of the studies

5 Conclusion

These findings suggest that maintaining a healthy weight, managing stroke and heart disease, and living an active lifestyle can help prevent diabetes.