PRACTICE DATA PROJECT REPORT

CONTENTS

1. Introduction/Abstract ………………………………………………. 2

2. Exploratory Data Analysis …………………………………………. 3

3. Logistic Regression for Each 'X' ………………………………... 4

4. Logistic Regression for all 'X' …………………………………... 6

5. Bootstrapped Logistic Regression ……………………….………. 7

6. Likelihood Ratio Testing(LRT) ……………………...……….. 8

7. Forward/Backward Selection Models ……………………..……… 9

8. Interactions of Variables ...…………………………………..…… 11

9. Predicted Probability for Specific Combination of Variables ……… 12

10. Adjusted Probability of Outcomes …………………………………. 12

11. Goodness of Fit Tests ……………………………………………… 13

INTRODUCTION

Hypertension, also known as high blood pressure, is a prevalent health condition that affects millions of individuals worldwide. Understanding the underlying factors contributing to hypertension is of utmost importance for healthcare professionals, researchers, and policymakers. In this data science project, we aim to explore the relationships between hypertension and three variables: Smoking, Obesity, and Snoring. By applying logistic regression, a powerful statistical technique for binary classification, we will uncover insights into how these variables may influence the likelihood of developing hypertension.

Smoking has long been identified as a major risk factor for various cardiovascular diseases, including hypertension. The harmful chemicals present in tobacco smoke can lead to narrowing of blood vessels, increased heart rate, and elevated blood pressure levels. By analyzing a comprehensive dataset, we will investigate the extent to which smoking influences the probability of developing hypertension.

Obesity, characterized by excessive body fat accumulation, is another well-established risk factor for hypertension. Excess weight can lead to hormonal imbalances, insulin resistance, and changes in the structure and function of blood vessels, all of which contribute to elevated blood pressure. Through our logistic regression analysis, we will examine the relationship between obesity and hypertension, shedding light on the magnitude of this association.

Snoring, often an indicator of sleep-related breathing disorders, has recently gained attention as a potential risk factor for hypertension. Frequent and loud snoring can disrupt the quality of sleep and contribute to the development of hypertension through various mechanisms such as sleep apnea, increased sympathetic activity, and oxidative stress. By investigating the impact of snoring on hypertension using logistic regression, we aim to assess its significance as an independent predictor.

EXPLORATORY DATA ANALYSIS

1. CONTINGENCY TABLES :

a. Hypertension and Smoking :

Hypertension\Smoking	Yes	No
Yes	23	56
No	104	250

- Odds of developing Hypertension in people who Smoke = 23/104 = 0.2211
- Odds of developing Hypertension in people who do not Smoke = 56/250 = 0.224
- Odds Ratio = 0.2211/0.224 = **0.9871 ~ 1**
- Hence, we can conclude that the odds of developing Hypertension in people who Smoke is approximately equal to the odds of developing Hypertension in people who don't.

b. Hypertension and Obesity :

Hypertension\Obesity	Yes	No
Yes	24	55
No	60	294

- Odds of developing Hypertension in people who are Obese = 23/60 = 0.4
- Odds of developing Hypertension in people who are not Obese = 55/294 = 0.1871
- Odds Ratio = 0.4/0.1871 = **2.1379**
- Hence, we can conclude that the odds of developing Hypertension in people who are Obese is approximately 2 times the odds of developing Hypertension in people who aren't Obese.

c. Hypertension and Snoring :

Hypertension\Snoring	Yes	No
Yes	71	8
No	275	79

- Odds of developing Hypertension in people who Snore
= 71/275 = 0.2582
- Odds of developing Hypertension in people who do not Snore
= 8/79 = 0.1013
- Odds Ratio = 0.2582/0.1013 = **2.5489**
- Hence, we can conclude that the odds of developing Hypertension in people who Snore is 2.5 times the odds of developing Hypertension in people who don’t.

2. SIMPLE LOGISTIC REGRESSION FOR EACH ‘X’ :

❖ SMOKING :

VARIABLE	BETA ESTIMATE	P-VALUE
Smoking	-0.009042	0.963

Inference :

- As the beta estimate is approximately 0 we can say that smoking plays a very small or a negligible part in determining Hypertension.
- Also, a p-value > 0.05 confirms that the null hypothesis(H0 : $\boldsymbol{\beta}$ = 0) is accepted.

- **OBESITY :**

VARIABLE	BETA ESTIMATE	P-VALUE
Obesity	0.5374	0.00718

Inference :

- As the beta estimate is approximately 0.5 we can conclude that obesity of an individual plays a moderate but significant part in determining Hypertension.
- Also, the p-value < 0.05 confirms that the null hypothesis(H0 : $\boldsymbol{\beta} = 0$) is rejected and a statistically significant, non-zero relationship is present.

- **SNORING :**

VARIABLE	BETA ESTIMATE	P-VALUE
Snoring	0.6618	0.0176

Inference :

- As the beta estimate is 0.66 we can say that snoring plays a moderate but significant part in determining Hypertension.
- Also, the p-value < 0.05 confirms that the null hypothesis(H0 : $\boldsymbol{\beta} = 0$) is rejected and a statistically significant, non-zero relationship is present.

LOGISTIC REGRESSION MODEL USING ALL VARIABLES

VARIABLE	BETA ESTIMATE	P-VALUE
Intercept	-1.62792	2.66e-13
Smoking	-0.04792	0.8075
Obesity	0.49166	0.0147
Snoring	0.61655	0.0283

INFERENCE :

- Smoking: The beta estimate for Smoking (-0.04792) suggests that there is a small negative association between Smoking and the log-odds of Hypertension. However, the p-value (0.8075) is greater than the conventional significance level of 0.05, indicating that the relationship between smoking and Hypertension is not statistically significant in this analysis.

- Obesity: The beta estimate for Obesity (0.49166) suggests a positive association between Obesity and the log-odds of Hypertension. The p-value (0.0147) is less than 0.05, indicates that the relationship between Obesity and Hypertension is statistically significant at the 0.05 significance level.

- Snoring: The beta estimate for Snoring (0.61655) indicates a positive association between snoring and the log-odds of Hypertension. The p-value (0.0283) is less than 0.05, suggesting that the relationship between Snoring and Hypertension is statistically significant at the 0.05 significance level.

- Intercept: The intercept term depicts the estimated log odds of Hypertension when all variables are set to zero. We can apply inverse logit function to find the *baseline probability* (probability of Hypertension when all variables are zero).
  Hence, the baseline probability is 1/(1+e^-(-1.62792)) = 0.1641

BOOTSTRAPPED LOGISTIC REGRESSION

- Bootstrapping involves creating multiple datasets, with replacement, from the original dataset, fitting the regression model on each dataset, and, collecting all such beta estimates to provide a robust value of each beta coefficient.
- It is then used to assess the quality and reliability of the original model.
- The beta coefficients of the original model are then compared to the beta coefficients from the bootstrapped models.

VARIABLE	ORIGINAL	BIAS
Intercept	-1.62792	-0.06207
Smoking	-0.04792	-0.01242
Obesity	0.49166	0.00200
Snoring	0.61655	0.05714

The “Bias” column shows the difference between the original value and the robust value(from bootstrapped regression) of beta coefficients.

INFERENCE :

- ❖ We observe a small adjustment is needed to align the original estimate of intercept, Smoking, and Snoring with the average bootstrapped estimate of the respective variables.
- ❖ For Obesity, the bias is 0.002, indicating a negligible difference between the original and the bootstrapped estimate.

LIKELIHOOD RATIO TESTS(LRT)

Null Hypothesis : H0 = reduced model provides an equally good fit as the full model

Test Statistic :

Deviance = -2 * (log-likelihood of reduced model - log-likelihood of full model)

that follows χ^2 distribution with 1 df

Likelihood Ratio Tests are carried out the check the significance of each individual variable in the full model.

If the p-value > 0.05, we accept H0 i.e. the reduced model fits equally good as the full model and the regressors that were removed in the reduced model are **not** significant in determining the outcome variable.

If the p-value < 0.05, we reject H0 i.e. the reduced model fits poorly as compared the full model and the regressors that were removed in the reduced model are significant in determining the outcome variable.

- LRT for Smoking :

  We obtain a p-value of **0.8069** > 0.05

  Therefore, Smoking does not make any significant difference in the goodness of fit of the model and hence, it is not significant in explaining Hypertension.

- LRT for Obesity :

  We obtain a p-value of **0.01739** < 0.05

  Therefore, the presence of the variable Snoring significantly improves the fit of the model and hence, we can say that it is crucial in explaining Hypertension.

- LRT for Snoring :

We obtain a p-value of **0.01718** < 0.05

Therefore, the presence of the variable Snoring significantly improves the fit of the model and hence, we can say that it is crucial in explaining Hypertension.

FORWARD SELECTION / BACKWARD ELIMINATION MODELS

- FORWARD SELECTION :
  - In forward selection, we start with an empty model and continue to iteratively add variables based on their statistical significance in explaining the outcome variable.
  - AIC(Akaike Information Criterion) values are used to compare models before adding another variable
  - AIC values consider goodness of fit of model along with model complexity(number of parameters).

- **RESULTING MODEL :**

VARIABLE	COEFFICIENTS	VIF
Intercept	-1.6116	-
Obesity	0.4917	1.005232
Snoring	0.6120	1.005232

- AIC = 404.98
- VIF(Variance Inflation Factor) of both variables are less than 5 which suggests that there is no multicollinearity present in the model.

- BACKWARD ELIMINATION :
  - In backward elimination, we start with the full model that includes all predictors, and continue to iteratively remove variables based on their statistical significance in explaining the outcome variable.

- **RESULTING MODEL :**

VARIABLE	COEFFICIENTS	VIF
Intercept	-1.6116	-
Obesity	0.4917	1.005232
Snoring	0.6120	1.005232

- AIC = 404.98
- VIF(Variance Inflation Factor) of both variables are less than 5 which suggests that there is no multicollinearity present in the model.

Hence, through both forward selection and backward elimination we obtain the same model with only Snoring and Obesity as regressors.

INTERACTION OF VARIABLES

Interaction of variables is a statistical phenomenon where, the effect of one regressor(X1) on the outcome variable(Y), depends on the value of another regressor(X2). Interaction could exist between multiple regressors in a regression model.

To determine if the interactions present in the model are significant or not we perform LRT for the model with and without interactions.

RESULT :

- We get the p-value for LRT as 0.5623
- As p-value > 0.05, we accept the null hypothesis.
- Hence, we can conclude that, there is no significant difference between the model with interactions and the model without interactions.
- This suggests that the interaction term is not required in the model.

PREDICTED PROBABILITY FOR SPECIFIC COMBINATION OF VARIABLES(DOUBT IN THIS TABLE)

Following is a table of predicted probabilities of having Hypertension for each combination of variables :

SMOKING	OBESITY	SNORING	ODDS	PROB	CI95_LOW	CI95_HIGH
NO	NO	YES	0.08668	0.07977	0.03560	0.16346
NO	NO	NO	0.09276	0.08489	0.04218	0.16912
NO	YES	YES	0.17375	0.14803	0.06219	0.23914
NO	YES	NO	0.18593	0.15678	0.07238	0.25244
YES	NO	YES	0.20731	0.17172	0.11290	0.30702
YES	NO	NO	0.22185	0.18157	0.13539	0.31280
YES	YES	YES	0.41553	0.29355	0.18391	0.42488
YES	YES	NO	0.44467	0.30780	0.21113	0.43380

ADJUSTED PROBABILITY BY SNORING (DOUBT ABOUT THE OUTPUT)

SNORING	PROBABILITY
0	0.04719
1	0.08489

GOODNESS OF FIT TESTS

1. LIKELIHOOD RATIO TEST :
    - Through this test, we want to see if the model under consideration is performing better than the null model.
    - Null Hypothesis (H0) : Null model fits equally as good as the given model.
    - RESULT :
      The p-value = 0.00532 < 0.05; hence we reject H0.
    - This implies that the given model performs significantly better than the null model.

2. CHI – SQUARE TEST :
    - We compare the observed frequencies with the expected frequencies using the Pearson's chi-square test.
    - Null Hypothesis(H0) : No significant difference between observed and expected frequencies.
    - RESULT :
      The p-value = 0.2289 > 0.05; hence we accept H0. This implies that there is no significant difference between the observed and expected frequencies.
    - Hence, the observed model does not significantly deviates from the expected.

3. HOSMER-LEMESHOW GOODNESS-OF-FIT TEST :
    - This test evaluates whether there is a significant difference between the observed and expected frequencies which indicates a lack of fit.
    - H0 : There is no significant difference between the observed and predicted values in the model. The model accurately predicts the probabilities of the outcome(Hypertension).
    - RESULT :
      The p-value = 0.9998 > 0.05; hence we accept H0. This implies that there is no significant difference between the observed and expected frequencies.