**Smoke Signals: Unraveling the Complex Nexus of Smoking, Diet, and Physical Health Impact on Human Weight**

Paul Bedu-Osei[1] and Olakunle Ogunyemi[2]

1. Syracuse University, New York, United States.
2. State University of New York College of Environmental Science and Forestry, Syracuse, New York, United States.

---

**Abstract:**

Research has shown that most smokers generally add more weight the moment they quit smoking; this momentarily effects on weight can lessen some of the health benefits of quitting smoking. In this study, we present a comprehensive review of some useful literature to understand how smoking habits, dietary choices, and physical health exercises collectively influence on weight in the human body.

## 1.0 Introduction

This research primarily centers on exploring the intricate relationship between smoking, diet type, and physical health, with a specific emphasis on the impact of smoking on human weight. Additionally, the objective is to substantiate the notion that nicotine influences hunger, elucidating why many smokers experience weight loss. The secondary focus of this study is to investigate the potential influence of smoking on weight, a significant health factor, and how other factors such as your diet and the frequency of your exercise contribute to your overall well-being. Despite a limited understanding of the mechanisms linking smoking, diet, and physical health to blood

pressure and weight, delving into relevant literature is essential to comprehend the nuances of these relationships. Understanding the interplay among these factors is crucial for promoting overall well-being. The main problem we aim to solve in this study is to understand, using statistical knowledge, the effect of smoking on an individual's weight. Weight gain is a likely outcome of smoking cessation [1]. There are several explanations for weight gain which may be associated with age, body consumption, the number of cigarettes smoked per day, etc. [2], yet the general public tend to believe that weight gain could largely be as a result of cessation in smoking for people who smoke and this is one of the primary cause of smokers returning to the habit even after successfully quitting [3]. An additional explanation is that the appetite-suppressant effects of nicotine on the brain are reversed and the presence of nicotine, a stimulant that prompts the release of neurotransmitters such as dopamine and serotonin in the brain—acting akin to the brain's reward system can exert a suppressing effect on appetite [4]. This suppression, influencing the hypothalamus, the brain region responsible for regulating hunger and satiety, coupled with a diminished sense of taste, may render food less appealing and potentially lead to weight gain. Several studies have shown that smokers weigh less than non-smokers [5,6] this is because of the effect of smoking to inhibiting the hinger drive arising from gastric contractions [7], including research conducted by Audrain-McGovern and Benowitz [8] underscores the essential effects of smoking and nicotine, emphasizing their potential influence on the intricate relationship between smoking and weight. “And healthy eating for healthy weight, 2023” accentuates the impact of dietary modifications and regular exercise as effective strategies to healthy weight, which could help if individuals stop smoking.

Regression modeling is a common and interesting method in the field of statistics that is often used to explore and provide a succinct description of the relationship between an outcome or dependent/response variable and a set of independent predictors. Logistic regression is concerned with the special situation in regression modeling, where the outcome is of a binary or dichotomous

(yes/no, smoker/non-smoker) nature [9]. In this study, the binary variable which we want to use the statistical power of linear regression analysis to classify based on other predictor variables such as diet type, weight, physical health are the levels of smokers. i.e., smokers or non-smokers.

### 1.2 Secondary Problem

Beyond the primary emphasis, the secondary focus of this study involves comprehending the association of the likelihood of a smoker and its relationship to one's exercise habits, diet, blood pressure, height and as well as the weight. It delves into understanding the moderating role of lifestyle choices, particularly diet and physical activity, and the likelihood of the person being a smoker within the context of the logistic regression model.

### 1.3 Background Knowledge

Approximately 19% of women and 23% of men in the United States are current smokers. Each year, cigarette smoking causes more than 400,000 premature deaths in the United States from cardiovascular and respiratory diseases and cancer [10]. The second leading cause of premature morbidity and mortality is excess body weight due to poor diet and insufficient physical activity [10]. In the United States, ~30% of women and ~43% of men are overweight, and ~26% of women and ~29% of men are obese [11]. The idea that cigarette smoking is helpful in controlling body weight has been part of popular culture for many years. Cigarette advertisements from the 1930s suggested that women should “reach for a cigarette instead of a sweet.” For many smokers, the anticipation of weight gain can hinder smoking cessation success [12]. Gaining insight into nicotine's effects on body weight can offer valuable information for interventions addressing nicotine dependence. Quitting smoking and lifestyle adjustments has the potential to improve life quality and control the weight of individuals. Most health-care providers would agree that the decrease in morbidity and mortality associated with smoking cessation far outweighs the health

risks associated with post-cessation weight gain. Nevertheless, weight gain can reduce some of the health benefits of quitting smoking. Optimizing the health benefits of smoking cessation requires greater understanding of the behavioral and biological relationships between smoking and dietary habits to prevent weight gain after quitting smoking. The mechanisms through which smoking decreases body weight are complex and incompletely understood. Most of the effects of cigarette smoking on body weight are mediated by nicotine, although smoking a cigarette may also serve as a behavioral alternative to eating, resulting in decreased food intake. Body weight is determined by the balance of caloric intake and daily energy expenditure. Daily energy expenditure is determined by resting metabolic rate, physical activity, and the thermic effects of food. Nicotine reduces body weight by raising the resting metabolic rate while blunting the expected increase in food intake in response to the increase in metabolic rate. When smokers quit smoking, the loss of the metabolic boost and appetite suppression conferred by nicotine is often accompanied by increased caloric intake but no increase in physical activity. This positive energy balance leads to weight gain. Behavioral interventions to manage post-cessation weight gain have therefore focused on managing caloric intake, increasing physical activity, or both.

## 2.0 Data Description

**Title of the Dataset:** Health Profile Dataset

**Curated by:** Wellness Institute.

The Health Profile Dataset is a compilation of health-related information collected by the Health Metrics Institute. The dataset includes various parameters to assess and monitor individuals' health.

The dataset's underlying population comprises individuals whose health profiles were of interest to the "Wellness Institute." The ideal population is envisioned to be diverse, including individuals of varying ages, heights, weights, blood pressure, genders, and lifestyles (smoking status, exercise

habit and diet type), aligning with the institute's overarching goals. This diversity implies that the sample is likely to be a random one, as there is no discernible pattern or specific criteria for inclusion or exclusion. Everyone within the broader demographic had an equal opportunity to be selected, reinforcing the notion of a random sample in this context.

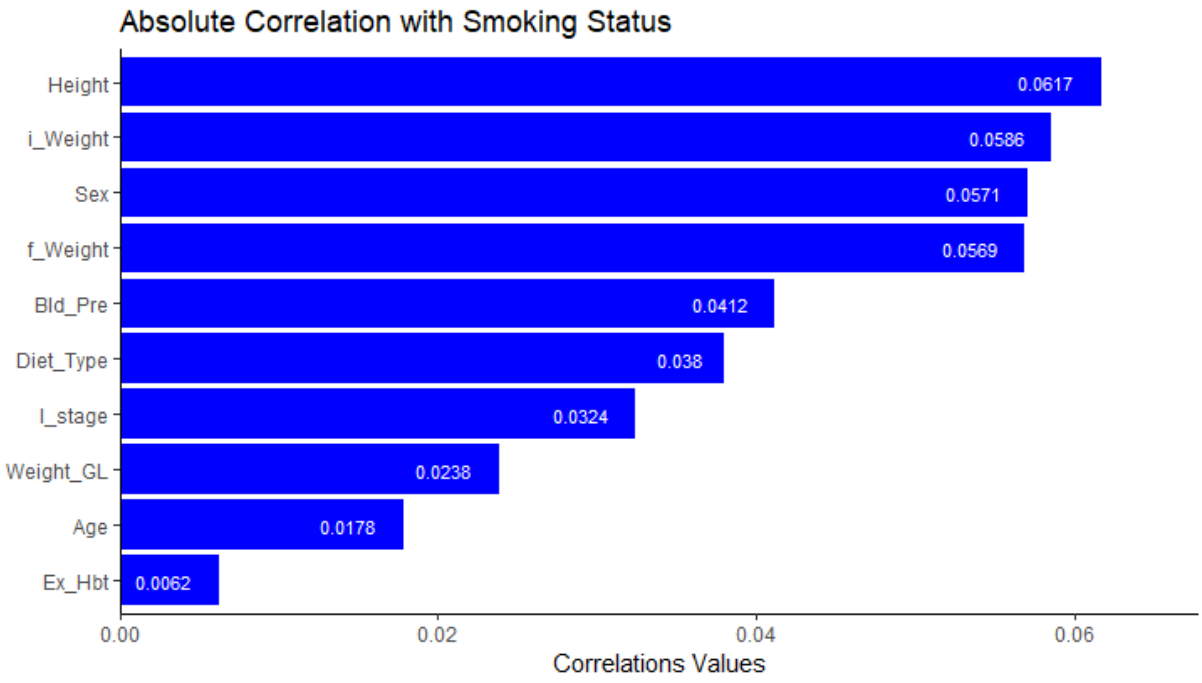



**Figure 1**: Absolute correlation with smoking status.

The datasets used for this study consist of a pool of one thousand patients who were either classified as a smoker with the code of 1 or non-smoker with the code of 0. **Figure** 1 shows the Pearson correlation between other numeric variables who correlations values are positive in the dataset with smoking status. The distribution of the smoking status shows that 50.3% are smokers while 49.7% of the patients are non-smokers; and the distribution by sex shows that 34.1% of these patients are female and 65.9% are male as shown in **Figure 2** below:

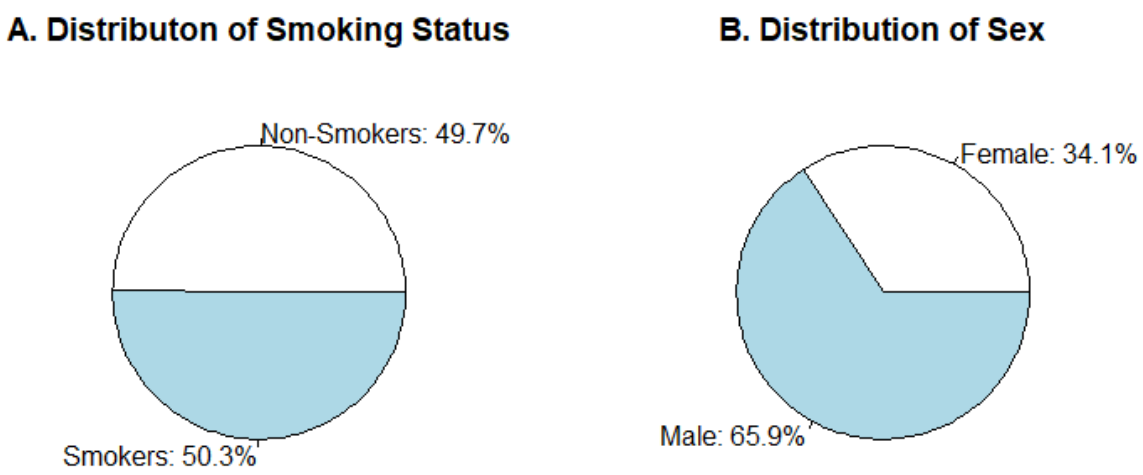



**Figure 2**: Distribution of sex and smoking status across patients.

Weight gain is a derived variable in this dataset, and it is obtained from the difference between the final weight and the initial weight of the patient. In the end, positive values are referred to as gain while a negative value is referred to as loss as shown in the table below.

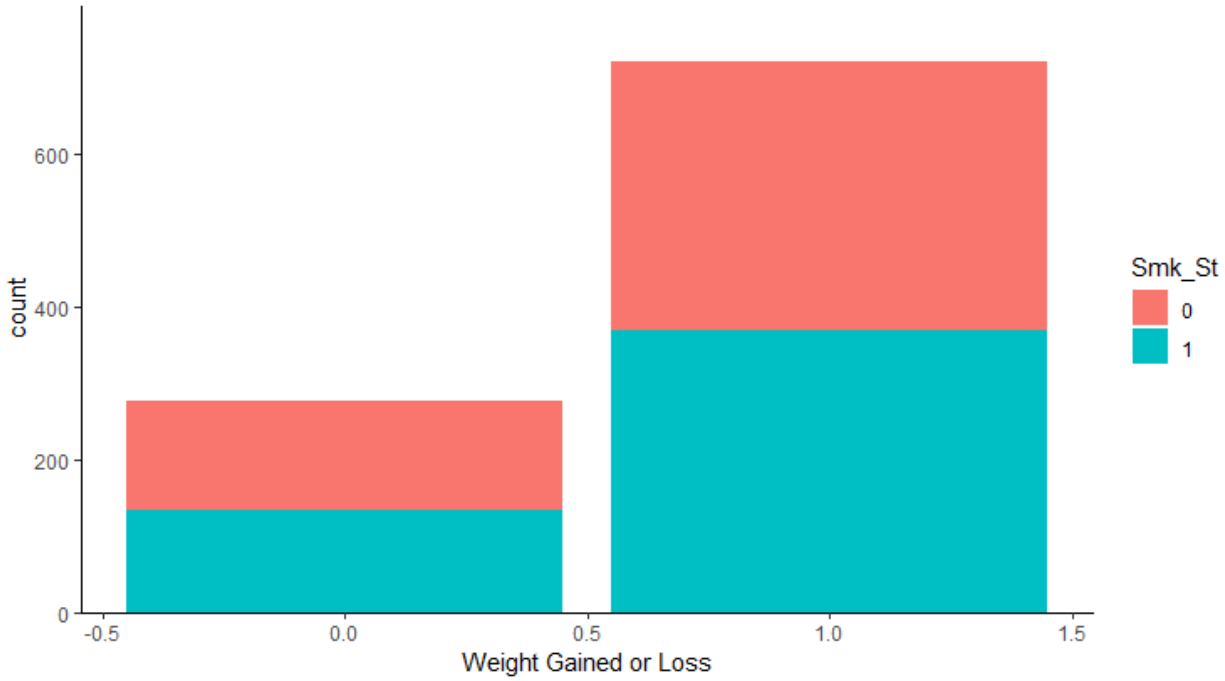



**Figure 3:** Distribution of weight gain or loss across smoking status.

In **Figure 3**, information displayed through a bar chart shows that 723 patients gained weight at the end of the study and 51% of them are smokers while the remaining 49% are non-smokers. Additionally, out of the 277 patients that experienced weight loss, about 48% of them are smokers while the rest 52% are non-smokers.

## 3.0 Data Analysis and Discussion

### 3.1.1 Multiple Linear Regression

In this discussion, we will primarily focus on investigating the potential relationship between nicotine exposure and an individual's weight. The underlying hypothesis is that the presence of nicotine may be associated with lower body weight, suggesting that regular smokers might generally have a lower weight compared to non-smokers. Conversely, the absence of such a correlation would imply that nicotine does not significantly influence an individual's weight.

To delve into this hypothesis, we will explore the impact of smoking habits on weight. Specifically, we anticipate that individuals who smoke regularly may exhibit a tendency to have a reduced weight, and conversely, non-smokers may demonstrate higher weights. This exploration aims to provide insights into the potential weight-altering effects of nicotine.

Should the analysis reveal that smoking does not have a substantial impact on weight, our attention will shift towards investigating other potential predictors. Important and obvious predictors such as exercise habits ( Ex_Hbt ) and diet type (Diet_Type ) and other possible predictors like height (Height), and blood pressure (Bld_Pre ) will be scrutinized to determine whether they play a significant role in influencing an individual's weight. By examining these additional factors, we aim to comprehensively assess the various elements that might contribute to variations in weight among individuals. This holistic approach will contribute to a nuanced understanding of the multifaceted factors influencing body weight beyond just smoking habits.

- **Outcome Variable (Dependent Variable):** Weight
- **Potential Predictors (Independent Variables):** Smoking status (Smk_St)(main), Exercise habit ($\text{Ex\_Hbt}_i$), Diet type (Diet_Type ), Height (Height) and Blood pressure (Bld_Pre ).

The population regression model is stated below:

$$Weight_i = \beta_0 + \beta_1 \text{Smk\_St}_i + \beta_2 \text{Ex\_Hbt}_i + \beta_3 \text{Diet\_Type}_i + \beta_4 \text{Height}_i + \beta_5 \text{Bld\_Pre}_i + \varepsilon_i$$

The standard assumptions are:

i. expectation. $E(\varepsilon_i) = 0$

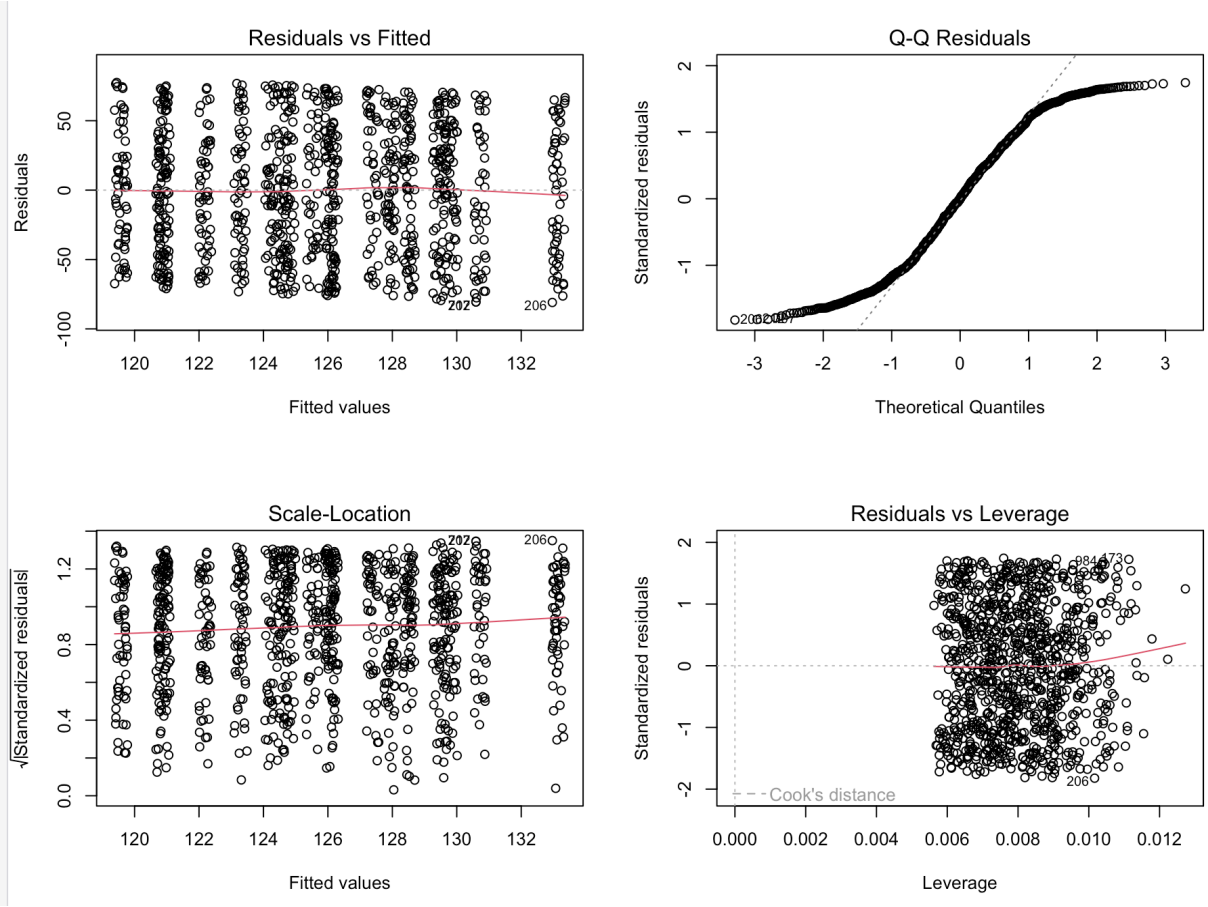
ii. $\varepsilon_i$ and $X_i$ are independent.

iii. variance $Var(\varepsilon_i|X_i) = \sigma^2$

iv. $\varepsilon_i$ and $\varepsilon_j$ are independent.

In this work, we will conduct some hypothesis tests to ensure that all the listed assumptions above are not violated.

**Checking assumptions.**



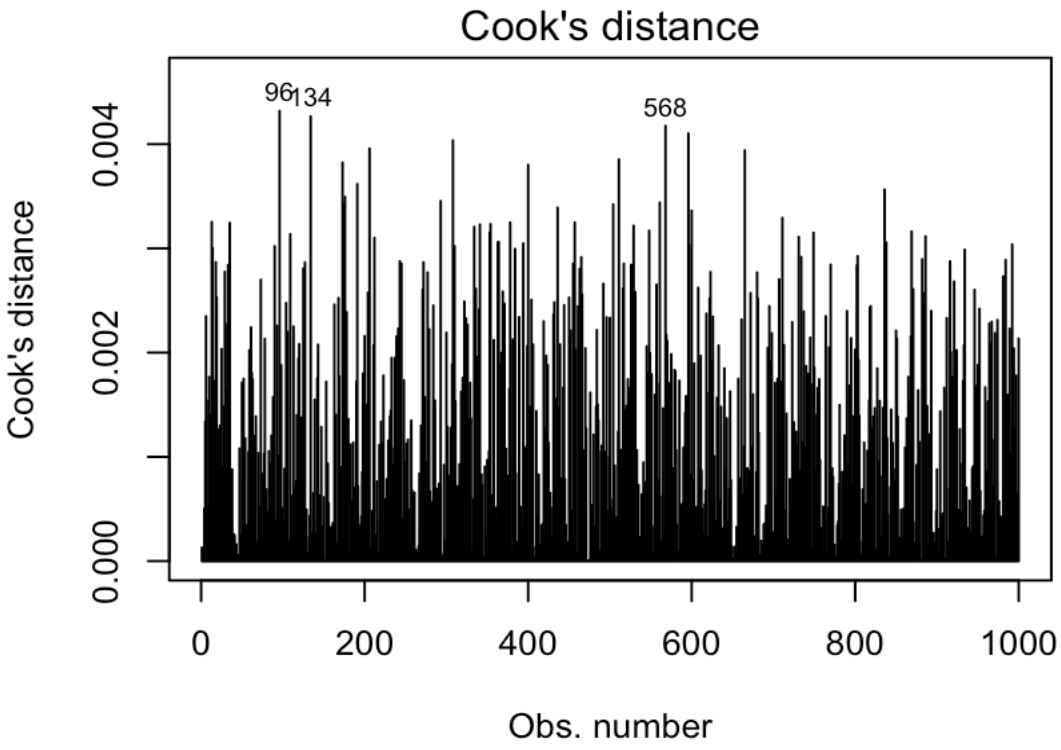

Before delving into the findings of this study, an examination of the assumptions underlying the regression model was conducted by examining the residual plots above. The Q-Q plot of residuals indicates a slight deviation at the tails of the line, but overall, most of the data adheres well to the line, suggesting that the assumption of normal distribution is not significantly violated.

Turning attention to the Residuals versus Fitted Values plot, the points exhibit an even scattering around the red line, and the line itself appears fairly straight and close to zero. This suggests that the data exhibits linearity, and there is no apparent issue of heteroskedasticity. Further assessment with the Scale-Location plot reveals a consistently straight line, and the data points appear to be evenly distributed. This observation suggests that the assumption of constant variance is not violated.

In summary, the diagnostic plots indicate that the normality assumption is reasonably satisfied, the data exhibits linearity, and there are no evident issues of heteroskedasticity or violation of constant variance assumptions. It's important to consider these findings when interpreting the results of the regression model.

**Cooks Distance**

From the cook distance in the diagram below we realize that 96, 134 and 568 are influential point. But these observations are not substantially different from the couple of the other observations.




**Checking multicollinearity**

```
> print(vif_values)
   Smk_St    Ex_Hbt Diet_Type    Height   Bld_Pre
 1.006737  1.003233  1.001849  1.005623  1.006198
```

As the VIF values for all variables are very close to 1 in the image above, which is generally considered an indication of low multicollinearity we can be more confident in the independence of the predictors in the multiple linear regression model and move further with the analysis.

**Correlation matrix**

| | Height | Weight | Bld_Pre | Smk_St | Ex_Hbt | Diet_Type |
|---|---|---|---|---|---|---|
| Height | 1.000000000 | 0.003604667 | 0.039511339 | 0.061703701 | 0.006416732 | 0.020621541 |
| Weight | 0.003604667 | 1.000000000 | 0.001278849 | 0.060480225 | -0.012217074 | 0.026622630 |
| Bld_Pre | 0.039511339 | 0.001278849 | 1.000000000 | 0.041182385 | -0.055493641 | -0.001677819 |
| Smk_St | 0.061703701 | 0.060480225 | 0.041182385 | 1.000000000 | -0.006237869 | 0.038034539 |
| Ex_Hbt | 0.006416732 | -0.012217074 | -0.055493641 | -0.006237869 | 1.000000000 | 0.007062218 |
| Diet_Type | 0.020621541 | 0.026622630 | -0.001677819 | 0.038034539 | 0.007062218 | 1.000000000 |

Examining the correlation matrix in the diagram above, our primary focus in this analysis is to assess the impact of smoking on an individual's weight. Given that cigarettes contain nicotine, which is known for its appetite-suppressant effects on the brain, we anticipate observing a lower weight among smokers compared to non-smokers. However, the correlation matrix suggests a weak positive correlation (0.0605) between smoking status and weight, indicating a slight tendency for individuals who smoke to have slightly higher weights. This relationship, however, is not robust.

Additionally, there is a weak negative correlation (-0.0122) between exercise habits and weight, suggesting a slight tendency for individuals with higher exercise habits to have slightly lower weights. This finding aligns with expectations, as regular exercise is often associated with weight management.

Moreover, the weak positive correlation (0.0266) between diet type and weight indicates a minor tendency for individuals with a specific diet type to have slightly higher weights. This observation underscores the expected connection between dietary habits and weight.

## ANOVA Interpretation

```
Call:
lm(formula = Weight ~ Smk_St + Ex_Hbt + Diet_Type + Height +
    Bld_Pre, data = team_5clean)

Residuals:
    Min      1Q  Median      3Q     Max
-78.983 -40.150  -0.081  37.453  77.352

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 122.590001  23.102882   5.306 1.38e-07 ***
Smk_St        5.317420   2.834302   1.876   0.0609 .
Ex_Hbt       -0.656365   1.717455  -0.382   0.7024
Diet_Type     1.353543   1.753791   0.772   0.4404
Height       -0.001599   0.118830  -0.013   0.9893
Bld_Pre      -0.005380   0.095474  -0.056   0.9551
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.66 on 994 degrees of freedom
Multiple R-squared:  0.004398,  Adjusted R-squared:  -0.0006099
F-statistic: 0.8782 on 5 and 994 DF,  p-value: 0.495
```

From the ANOVA table above with the hypothesis:

Null Hypothesis(H0): $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ (there is no linear relationship between $Y$ and X1 or $X2$).

Alternative Hypothesis (H1): At least one of $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4 \; or \; \beta_5$ is not equal to 0 (at least one of X1or X2 has a non-zero effect on Y).

We realized that since the p-values for the model, 0.495 is greater than the alpha, 0.05, we fail to reject the null hypothesis and conclude that there is not enough evidence of a significant linear relationship between Weight and at least one of the predictors: exercise habits ( Ex_Hbt ) , diet

type (Diet_Type ) , height , and blood pressure (Bld_Pre ) and their coefficients $\beta_1, \beta_2, \beta_3, \beta_4 \; or \; \beta_5$ are not significantly different from zero.

Upon closer examination of the output, it is evident that the p-value for smoking (0.0609) is in proximity to the significance level of 0.05. Although slightly above this threshold, the closeness suggests that the effect of smoking on weight may be on the border of statistical significance.

Comparing p-values across predictors, particularly in relation to diet and exercise habits, reinforces the notion that the smoking variable has a relatively stronger effect, at least based on this analysis. The proximity of its p-value to 0.05 implies a potential significance that distinguishes it from the other predictors.

But from the correlation matrix in the correlation part of this discussion we found out that diet types and exercise habits had a decent linear correlation with weight, since we are focusing on the effects of smoking on weight, let's work with smoking and either diet types and exercise habits to assess whether the effect of smoking status on weight depends on the diet type or exercise habits, and vice versa.

**ANOVA of smoking and diet types along with the interaction**

Looking at the diagram below, none of the individual predictors or the interaction term is statistically significant as 0.358, 0.470, and 0.803 are all greater than the significance level of 0.05 for Smoking status, Diet type and their interaction respectively. And also, the overall model fit, as indicated by R-squared and the F-statistic, is not strong, The F-statistic and its associated p-value (0.2302) suggest that the overall model fit is not statistically significant. Meaning the effect of smoking status on weight does not depend on the diet type and vice versa.

```
Call:
lm(formula = Weight ~ Smk_St * Diet_Type, data = team_5clean)

Residuals:
    Min      1Q  Median      3Q     Max
-78.434 -39.851  -0.176  37.659  77.715

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      119.5016     5.3877  22.180   <2e-16 ***
Smk_St             7.1164     7.7406   0.919    0.358
Diet_Type          1.7832     2.4688   0.722    0.470
Smk_St:Diet_Type  -0.8755     3.5036  -0.250    0.803
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.62 on 996 degrees of freedom
Multiple R-squared:  0.004313,  Adjusted R-squared:  0.001314
F-statistic: 1.438 on 3 and 996 DF,  p-value: 0.2302
```

---

The hypotheses for this F-test are:

H0: $\beta_1 = \beta_3 = 0$ (There is no difference in Diet Type or the effect of Diet Type between smokers and non-smokers)

H1: $\beta_1 \neq 0$ or $\beta_3 \neq 0$(There is a difference in Diet Type or the effect of Diet Type between smokers and non-smokers)

```
Analysis of Variance Table

Model 1: Weight ~ Smk_St * Diet_Type
Model 2: Weight ~ Diet_Type
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    996 1983006
2    998 1990184 -2   -7177.6 1.8025 0.1654
```

In this case, the F-test for group differences was performed, the p-value is 0.7208, which is greater than 0.05. The result suggests that the additional predictor Diet Type in Model 1 does not significantly improve the model compared to the simpler Model 2, which includes only Smk_St. The null hypothesis, that the additional predictor Diet Type does not contribute significantly, is not rejected.

**ANOVA of smoking and Exercise Habits along with the interaction**

Looking at the diagram below we also realize that 0.131, 0.742, 0.394 are all greater than the significance level of 0.05 for Smoking status, Exercise habits and their interaction respectively. And the overall model fit, as indicated by R-squared and the F-statistic, is not strong, The F-statistic and its associated p-value (0.2107) suggest that the overall model fit is not statistically significant. Meaning the effect of smoking status on weight does not depend on the Exercise Habits and vice versa.

```
Call:
lm(formula = Weight ~ Smk_St * Ex_Hbt, data = team_5clean)

Residuals:
    Min      1Q  Median      3Q     Max
-80.680 -39.551   0.114  38.449  76.908

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   121.5103     5.2628  23.089   <2e-16 ***
Smk_St         11.2987     7.4837   1.510    0.131
Ex_Hbt          0.7911     2.3999   0.330    0.742
Smk_St:Ex_Hbt  -2.9201     3.4264  -0.852    0.394
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 44.62 on 996 degrees of freedom
Multiple R-squared:  0.004524,  Adjusted R-squared:  0.001526
F-statistic: 1.509 on 3 and 996 DF,  p-value: 0.2107
```

**Hypothesis:**

H0: $\beta_1 = \beta_3 = 0$ (There is no difference in Exercise Habits or the effect of Exercise Habits between smokers and non-smokers)

H1: $\beta_1 \neq 0$ or $\beta_3 \neq 0$(There is a difference in Exercise Habits or the effect of Exercise Habits between smokers and non-smokers)

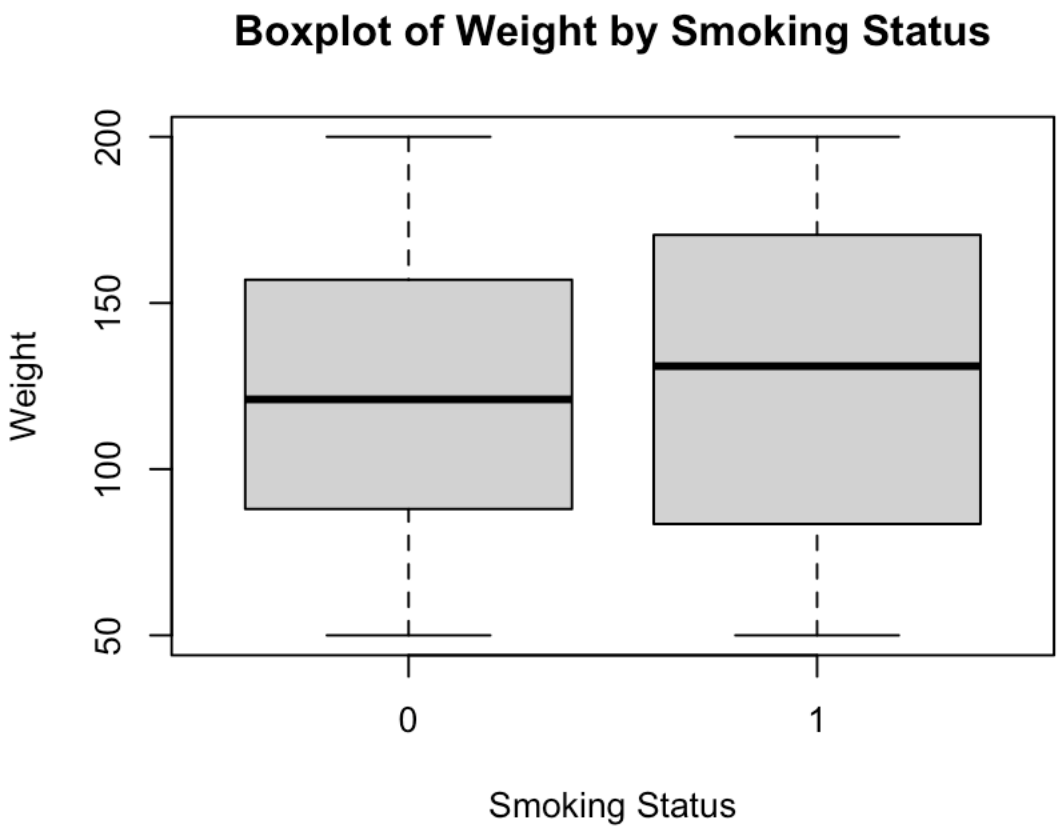
```
Analysis of Variance Table

Model 1: Weight ~ Smk_St * Ex_Hbt
Model 2: Weight ~ Smk_St
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    996 1982585
2    998 1984310 -2   -1724.9 0.4333 0.6485
```

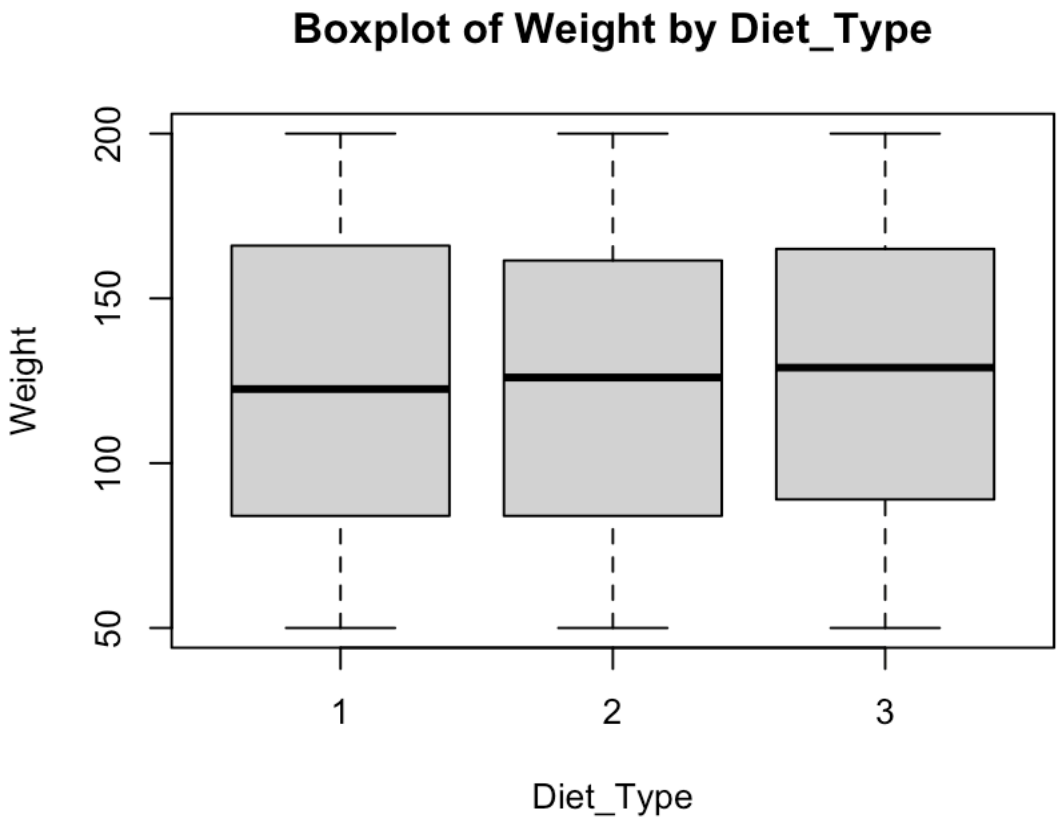
In this case, looking at the diagram above an F-test for group differences was performed, the p-value is 0.6485, which is greater than 0.05.The result suggests that the additional predictor Exercise Habit in Model 1 does not significantly improve the model compared to the simpler Model 2, which includes only Smoking Status .The null hypothesis, that the additional predictor Exercise Habit does not contribute significantly, is not rejected.

**Boxplots to compare.**

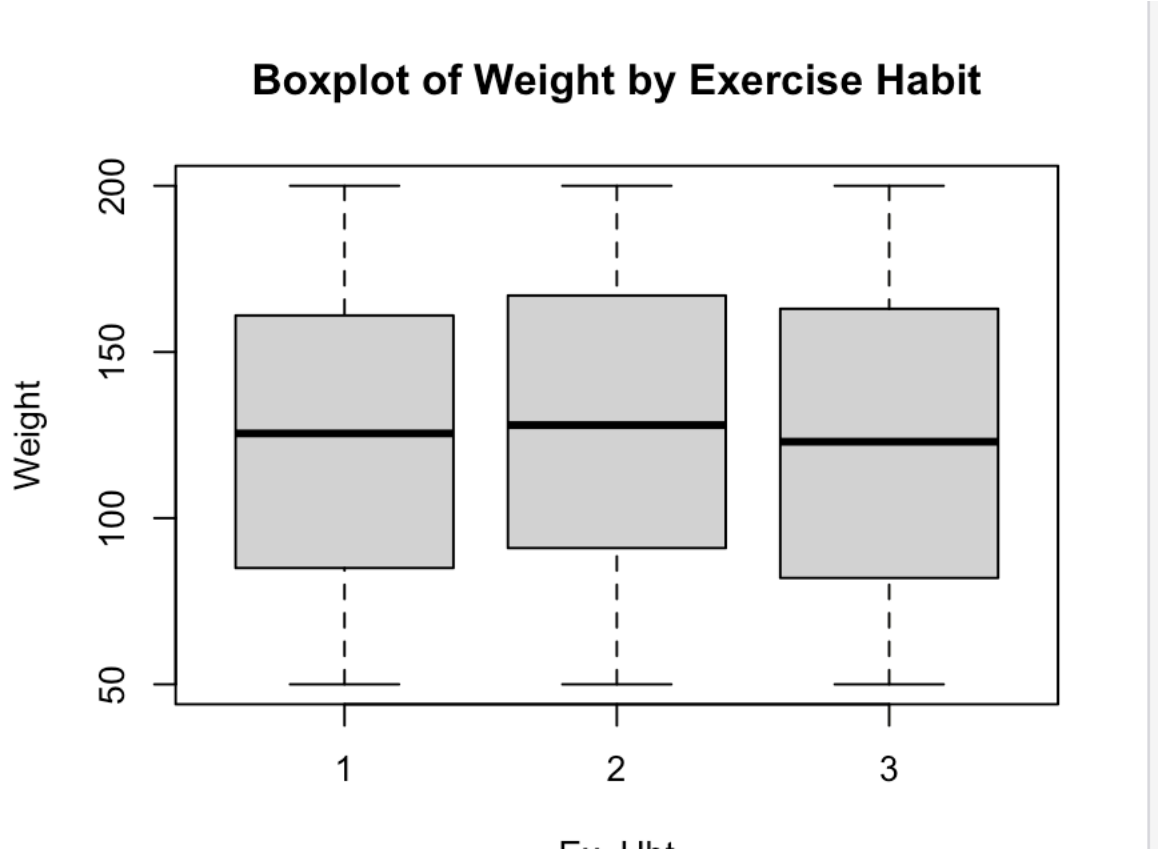
Examining the box plot below comparing weight between non-smokers and smokers, it's noteworthy that individuals who smoke regularly surprisingly exhibit a higher average weight than non-smokers. This observation suggests that, contrary to expectations, nicotine does not appear to have a significant influence on the weight of individuals based on the collected data. In fact, the data suggests that nicotine might be associated with an increase in weight rather than suppressing appetite, as previously hypothesized and explored in the correlation matrix discussed in an earlier paragraph.




Examining additional potential predictors such as exercise habits and diet type that may exert an impact on weight.




In the diagram above. It was checked whether the average for weight in different diet types and realized the average was pretty much close to each other on different diet types.




And the same with the Exercise habits. It was checked whether the average for weight in different Exercise habits. and realized the average was pretty much close to each other on different diet types.

In conclusion this makes sense since a lot of people I know are fat and still get fat even though they smoke a lot, as weight could be hereditary. And we also realized from the data that none of these other predictors could predict the weight of an individual and smoking was the only one near to predict the weight of an individual but was still insignificant.

**Slope Estimates**

```
> coef
  (Intercept)        Smk_St        Ex_Hbt     Diet_Type        Height       Bld_Pre
122.590001195   5.317420336  -0.656365439   1.353542791  -0.001598986  -0.005380195
```

From the estimates we realize that the coefficient associated with the variable "Smk_St" (smoking status). The positive coefficient suggests a positive relationship between smoking

status and the Weight, and the value (5.317420336) indicates the estimated change in the dependent variable for a one-unit change in smoking status, holding the other variables constant.

**Confidence Intervals**

```
> conf_int
                  2.5 %      97.5 %
(Intercept) 77.2539818 167.9260206
Smk_St      -0.2444816  10.8793223
Ex_Hbt      -4.0266193   2.7138885
Diet_Type   -2.0880154   4.7951010
Height      -0.2347862   0.2315882
Bld_Pre     -0.1927337   0.1819734
```

Created a confidence interval to see if any of the predictors are have a significant effect on weight, but none of the individual predictors (Smk_St, Ex_Hbt, Diet_Type, Height, Bld_Pre) show a statistically significant effect on weight, as their confidence intervals include zero. This implies that, based on the given data and the chosen significance level, there is insufficient evidence to conclude that these predictors have a significant impact on weight.

### 3.1.2 Logistic Regression

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability of an instance of belonging to a given class. It is used for classification algorithms; its name is logistic regression. It's referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. The difference between linear regression and logistic regression

is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not. In this study, logistic regression is applied to understand the effect of smoking on weight gain. Weight gain or loss is a derived variable that has two levels. Gain for scenarios where the final weight is greater than the initial weight and loss for scenarios where the initial weight is greater than the final weight. The question we aimed to address using logistic regression is to better understand how to classify an observation into whether such a person gained or loss weight by quitting smoking or continuing in the habit of smoking.

The logistic regression model used is stated below:

$$p_1(x;\ \beta_0, \beta_1) := P(Y = 1 \mid X = x) = \frac{\exp(\beta_0 + \ x'\ \beta_1)}{1 + \exp(\beta_0 + \ x'\ \beta_1)}$$

Where, Y is the response variable with 0 or 1 values and X represents the predictors.

### 3.1.3 Descriptive statistics

**Table 1:** Summary statistics of variables used in the study.

| Variable | Minimum | 1st Quart. | Median | Mean | 3rd Quart. | Maximum | SD |
|---|---|---|---|---|---|---|---|
| Height | 150.00 | 159.00 | 169.00 | 169.30 | 179.00 | 190.00 | 11.92 |
| i_weight | 110.00 | 127.00 | 142.00 | 140.40 | 153.00 | 170.00 | 15.93 |
| f_weight | 107.00 | 129.00 | 143.00 | 142.50 | 156.00 | 175.00 | 16.18 |
| weight_GL | 0.00 | 0.00 | 1.00 | 0.72 | 1.00 | 1.00 | 0.45 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Blood pressure | 90.00 | 103.00 | 115.00 | 115.30 | 128.00 | 140.00 | 14.85 |
| Smoking status | 0.00 | 0.00 | 1.00 | 0.50 | 1.00 | 1.00 | 0.50 |
| Diet type | 1.00 | 1.00 | 2.00 | 2.06 | 3.00 | 3.00 | 0.81 |
| Exercise Habits | 1.00 | 1.00 | 2.00 | 2.02 | 3.00 | 3.00 | 0.82 |
| Age | 20.00 | 31.00 | 44.50 | 44.38 | 57.00 | 70.00 | 14.99 |
| Sex | 0.00 | 0.00 | 1.00 | 0.66 | 1.00 | 1.00 | 0.47 |

### 3.1.4 Model Assumptions

#### 3.1.4.1 Multicollinearity Assumptions

The test on multicollinearity is carried out to check for any form of linear association in the linear part of the logistic regression model. The result as shown in the ***Table 2*** below reveals there is not any form of deviation away from the assumptions of multicollinearity. This test is also in line with the correlation results within the variables that suggest the independence of each variable.

**Table 2:** Test for multicollinearity between variables.

```
> library(car)
> vif(lrm)
               GVIF Df GVIF^(1/(2*Df))
Sex        1.012142  1        1.006053
Age        1.009058  1        1.004519
Bld_Pre    1.009140  1        1.004560
Ex_Hbt     1.013843  1        1.006898
Diet_Type  1.011700  2        1.002912
Smk_St     1.006725  1        1.003357
```

### 3.1.4.2 Comparing weight potential by sex and smoking status.

First, grouping the probability for whether a patient gained or lost weight by their sex, the male has a probability of about 73% and female has 72% of chance suggesting that the potential to either gain or lost weight is not totally dependent on the sex of the patient as they both have roughly equal chances. This shows that if we increase sex by just one unit, while keeping the other predictors fixed, is associated with a multiplication of the odds by the 1.03, the odds ratio between male and female. Giving a similar explanation to the smoking status of the patient, the probability of gaining or losing weight for the smokers is 73% and 71% chances for non-smokers.

### 3.1.4.2 Logistic Regression Model

**Table 3**: Output of the logistic regression model.

```
Call:
glm(formula = Weight_GL ~ Sex + Age + Ex_Hbt + Diet_Type + Smk_St,
    family = binomial(link = "logit"), data = dat)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.084666   0.263726   4.113 3.91e-05 ***
Sex0        -0.041156   0.149959  -0.274    0.784
Age         -0.003002   0.004741  -0.633    0.527
Ex_Hbt1      0.114346   0.170221   0.672    0.502
Ex_Hbt2      0.280567   0.174908   1.604    0.109
Diet_Type1  -0.101052   0.176303  -0.573    0.567
Diet_Type2  -0.072509   0.170529  -0.425    0.671
Smk_St0     -0.092003   0.142072  -0.648    0.517
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1180.2  on 999  degrees of freedom
Residual deviance: 1176.2  on 992  degrees of freedom
AIC: 1192.2
```

In the result shown in ***Table 3***, variables including sex, age, exercise habits, diet type and smoking status was used in fitting the logistic regression model with the response variable being the binary variable for weight gain (1) or weight loss (0). The result shows the intercept of the model is significant which implies that every interpretation given for the coefficients of the predictor variables as the possible impact on the response is valid.

**Table 4**: Odds ratios and confidence intervals for odds ratio.

```
> print(odds_ratios)
                    OR     2.5 %   97.5 %
(Intercept) 2.9584506 1.7716459 4.985895
Sex0        0.9596792 0.7164378 1.290272
Age         0.9970024 0.9877684 1.006312
Ex_Hbt1     1.1211398 0.8033947 1.566646
Ex_Hbt2     1.3238801 0.9407322 1.868667
Diet_Type1  0.9038863 0.6398137 1.277880
Diet_Type2  0.9300574 0.6654926 1.299350
Smk_St0     0.9121026 0.6901664 1.204964
```

Thus, in this study, the coefficient for sex when its female is -0.04 which means that the female group has 0.96 times the odds of the male base group of having gain weight. The confidence interval is not so wide and indicates a not so strong and insignificant effect on weight gain as the interval does includes 1. For smoking status, the bases group used are the non-smokers and they have a coefficient of -0.09 indicating that the group has 0.91 times the odds of the smoker's group of having gain weight. The interval is also not so wide and indicates a not so strong and insignificant effect as well on weight gain as it contains 1. Similar interpretations can be provided to understand the impact of each coefficient for categorical predictors on the response variable of the logistic model. For age, the coefficient is -0.003, which implies that an increase in 1 standard deviation in the age is associated with just a 1% increase in weight gain. The interval of the odds ratios is very small and still does not imply any significance or strong relationship on weight gain as the interval though small contains 1. The implication of this is that the possibility of any patient to either gain or lose weight is not a function of age, but much tied to other factors as explained in most of the literature cited in this study.

### 3.1.4.3 Diagnostic check

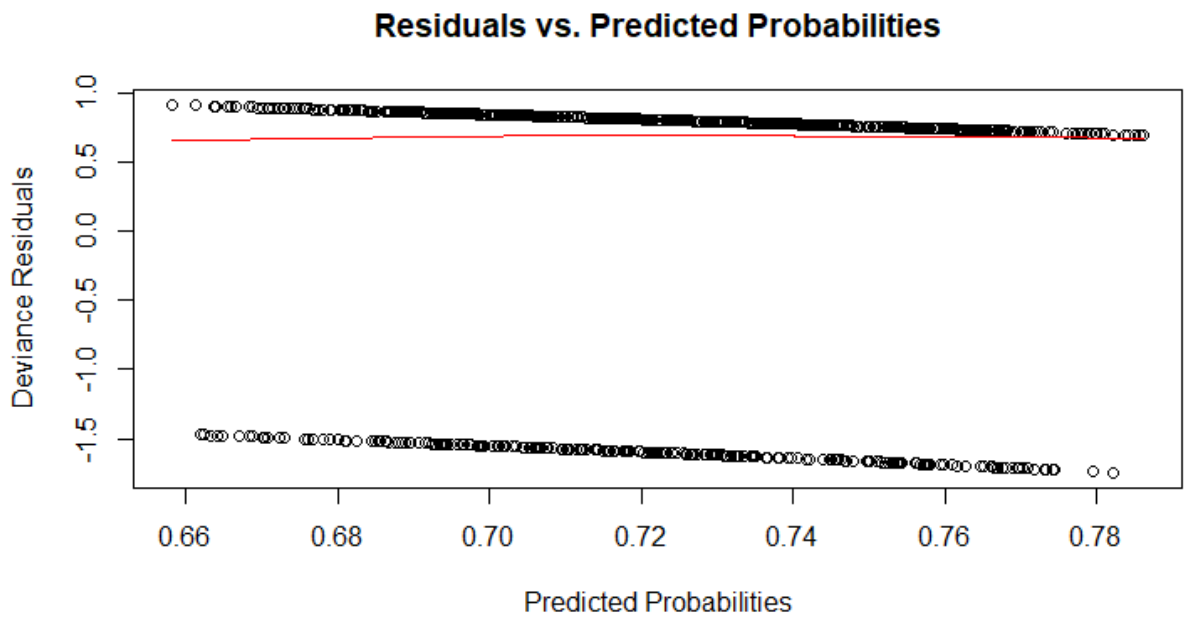



**Figure 4:** Residuals against the predicted probabilities.

This plot in ***Figure 4*** is used to assess the goodness of fit of the model and for this, the ideal situation would be to have the red line fall horizontally around the zero value of the residuals but in this case, the red line is slightly horizontal which shows that the model is not correctly specified with the variable used. This is further confirmed by carrying out the Hosmer-Lemeshow test for accessing the goodness of fit or a logistic regression model. The result of the test is shown below:

```
> print(hosmer_lemeshow_test)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  y_train, fitted(lrm)
X-squared = 0.73228, df = 3, p-value = 0.8656
```

The p-value is much bigger than the significance level of 0.05 which is commonly used, suggesting that our model does not adequately describe the data. Again, this result can be further corroborated with the large deviance 1176 obtained. This may also as a result of some outlier's presence as shown in the ***Figure 5*** and ***Figure 6*** below.

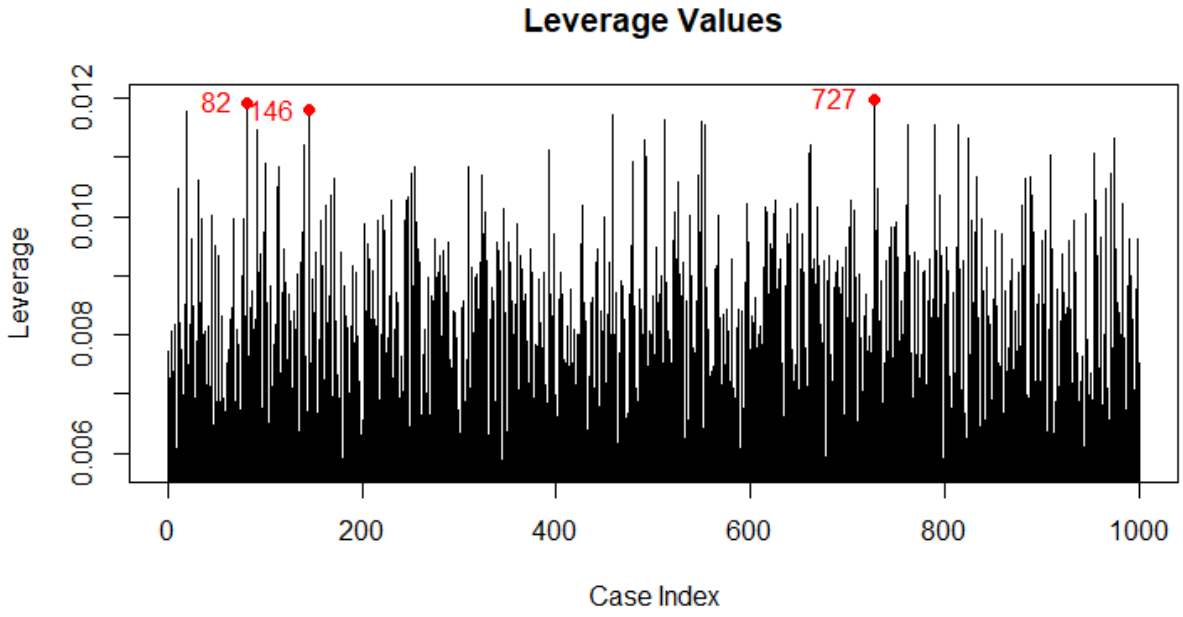



**Figure 5:** Leverage values vs index.

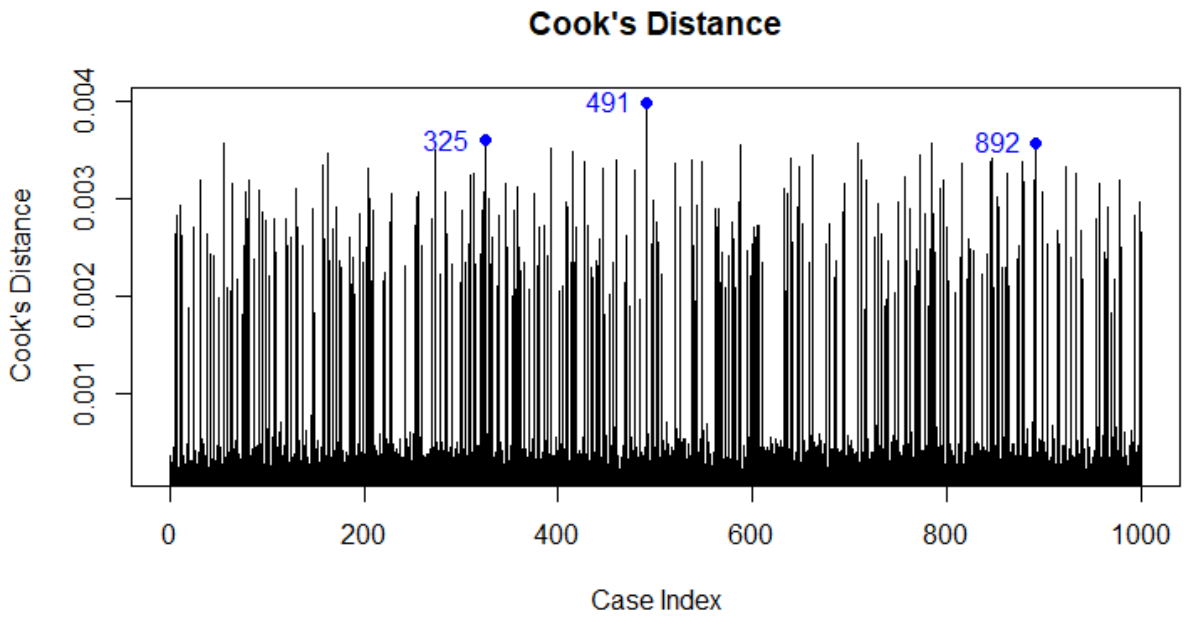



**Figure 6:** Cook's distance vs index.

In this study, we compared the predicted classes with the actual classes by taking the mean and we obtained a 72% in-sample accuracy.

## 4.0 Conclusion

The Outcome of the Multiple regression goes to prove that there are most likely other predictors to weight other than the nicotine in cigarettes, element that suppresses appetite, other predictors that could influence weight can be genetics from family among others. While also the outcome of this study from the application of the logistic regression model suggests that the potential to gain weight or lose weight by patients is not necessarily an effective way to prevent obesity, and in fact

increases it, especially among ex-smokers and smokers who continue their habit. This findings is in line with the work of [13].

## 5.0 References:


1. Brožek, J.; Keys, A. Changes of Body Weight in Normal Men Who Stop Smoking Cigarettes. *Science* **1957**, *125*, 1203–1203, doi:10.1126/science.125.3259.1203.
2. Bossé, R.; Garvey, A.J.; Costa, P.T. Predictors of Weight Change Following Smoking Cessation. *Int. J. Addict.* **1980**, *15*, 969–991, doi:10.3109/10826088009040072.
3. Cornelius, M.E.; Loretan, C.G.; Jamal, A.; Davis Lynn, B.C.; Mayer, M.; Alcantara, I.C.; Neff, L. Tobacco Product Use Among Adults – United States, 2021. *Morb. Mortal. Wkly. Rep.* **2023**, *72*, 475–483, doi:10.15585/mmwr.mm7218a1.
4. Benowitz, N.L. Pharmacology of Nicotine: Addiction, Smoking-Induced Disease, and Therapeutics. *Annu. Rev. Pharmacol. Toxicol.* **2009**, *49*, 57–71, doi:10.1146/annurev.pharmtox.48.113006.094742.
5. Keys, A.; Aravanis, C.; Blackburn, H.W.; Van Buchem, F.S.; Buzina, R.; Djordjević, B.D.; Dontas, A.S.; Fidanza, F.; Karvonen, M.J.; Kimura, N.; et al. Epidemiological Studies Related to Coronary Heart Disease: Characteristics of Men Aged 40-59 in Seven Countries. *Acta Med. Scand. Suppl.* **1966**, *460*, 1–392.
6. Karvonen, M.; Orma, E.; Keys, A.; Fidanza, F.; Brozek, J. Cigarette Smoking, Serum-Cholesterol, Blood-Pressure, and Body Fatness. Observations in Finland. *Lancet* **1959**, 492–494.
7. Carlson, A.J.; Lewis, J.H. Contributions to the Physiology of the Stomach. *Am. J. Physiol.-Leg. Content* **1914**, *34*, 149–154, doi:10.1152/ajplegacy.1914.34.2.149.
8. Cigarette Smoking, Nicotine, and Body Weight - Audrain-McGovern - 2011 - Clinical Pharmacology & Therapeutics - Wiley Online Library Available online: https://ascpt.onlinelibrary.wiley.com/doi/abs/10.1038/clpt.2011.105 (accessed on 18 December 2023).

9. Cole, T.J. Applied Logistic Regression. D. W. Hosmer and S. Lemeshow, Wiley, New York, 1989. No. of Pages: Xiii + 307. Price: £36.00. *Stat. Med.* **1991**, *10*, 1162–1163, doi:10.1002/sim.4780100718.
10. Ah, M. Actual Causes of Death in the United States, 2000. *Jama* **2004**, *291*, 1238–1245.
11. Actual Causes of Death in the United States, 2000 | Lifestyle Behaviors | JAMA | JAMA Network Available online: https://jamanetwork.com/journals/jama/article-abstract/198357 (accessed on 18 December 2023).
12. Meyers, A.W.; Klesges, R.C.; Winders, S.E.; Ward, K.D.; Peterson, B.A.; Eck, L.H. Are Weight Concerns Predictive of Smoking Cessation? A Prospective Analysis. *J. Consult. Clin. Psychol.* **1997**, *65*, 448–452, doi:10.1037/0022-006X.65.3.448.
13. Javier Basterra-Gortari, F.; Forga, L.; Bes-Rastrollo, M.; Toledo, E.; Alfredo Martínez, J.; Martínez-González, M.A. Effect of Smoking on Body Weight: Longitudinal Analysis of the SUN Cohort. *Rev. Esp. Cardiol. Engl. Ed.* **2010**, *63*, 20–27, doi:10.1016/S1885-5857(10)70005-0.

## 6.0 Appendices or supplementary results.