Understanding Obesity: A Statistical Approach to Lifestyle and Health

1. Introduction and Hypotheses

Obesity is a growing global health concern, affecting millions of individuals and placing a significant burden on healthcare systems worldwide. As stated by WHO (World Health Organization, 2024), as of 2022, 1 in 8 people suffer from obesity, which causes other diseases such as high blood pressure, cholesterol, breathing problems and can even cause cancer. Obesity is commonly caused by an imbalanced diet and a lack of activity, whilst other causes, such as social causes, are the secondary reason why people are suffering from obesity. Understanding the key contributors to obesity is crucial for designing effective interventions and promoting healthier lifestyles. Using generalized linear models (GLM), binary logistic regression, and generalized additive models (GAM), we analyze patterns in eating habits and physical activity to identify significant risk factors for obesity. We further refine our models through backward selection to enhance predictive accuracy and support early intervention strategies.

The primary objective of this study is to investigate the relationship between lifestyle factors and obesity levels. By analyzing various health indicators, including dietary habits, physical activity, and family history, we aim to develop a predictive model that estimates obesity risk. This study seeks to provide insights that can inform public health initiatives and personalized weight management strategies.

1.1 Scientific Questions

This study aims to explore key health and lifestyle factors contributing to obesity by addressing the following questions:

- Which factor has the largest impact (most strongly associated) on obesity?
- Are there significant differences in obesity levels based on gender and family history?
- How does physical activity influence obesity level?
- Can we estimate obesity levels based on lifestyle factors?

2. Data Description

The dataset selected for analysis is the **"Estimation of Obesity Levels Based On Eating Habits and Physical Condition", 2019** dataset from the UCI Machine Learning Repository (Palechor and Manotas, 2019). It contains information on the eating habits and physical conditions of individuals from Mexico, Peru and Colombia. The data comprises 17 attributes and 2111 records of categorical and numerical variables. 77% of the data was generated synthetically using the Weka tool and the SMOTE filter, and 23% of the data was collected directly from users through a web platform. The Weka tool is an external tool functioning as a data center which enables user to do data processing, storage and management (Tatvasoft, 2023). The SMOTE filter stands for Synthetic Minority Over-sampling Technique, which uses the k-nearest neighbors concept to generate synthetic data points for low-volume classes. This is done to help balance imbalanced datasets without duplicating existing data (Swastik, 2025).

The data was collected by Fabio Mendoza Palechor and Alexis De la Hoz Manotas, studying obesity and health levels aiming to develop systems that monitor obesity levels. The data was collected and compiled in 2019 using a web platform with a survey where anonymous users answered each question.

2.1 Data Types and Structures

Table 1: Description of variables collected

Variable Name	Data Type	Description
Gender	Categorical	Gender
Age	Continuous	Age
Height	Continuous	Height
Weight	Continuous	Weight
family_history_with_overweight	Binary	Has a family member suffered or suffers from overweight?
HICAL	Binary	Do you eat high caloric food frequently?
VEGE	Ordinal	Do you usually eat vegetables in your meals?
NMEAL	Ordinal	How many main meals do you have daily?
SNACK	Categorical	Do you eat any food between meals?
SMOKE	Binary	Do you smoke?
WATER	Ordinal	How much water do you drink daily?
CAL	Binary	Do you monitor the calories you eat daily?
PHY	Ordinal	How often do you have physical activity?
TECH	Integer	How much time do you use technological devices such as cell phone, videogames, television, computer and others?
ALC	Categorical	How often do you drink alcohol?
MTRANS	Categorical	Which transportation do you usually use?
OBSLVL	Categorical	Obesity level

Table 2: Numerical Variables

Variable Name	Variable Values
Age	min value = 14, max value = 61
Weight	min value = 39, max value = 173
Height	min value = 1.45, max value = 1.98

We categorized certain numerical variables to match predefined survey response options. The obesity level variable (`OBSLVL`) was treated as both a binary variable (1 = obese, 0 = non-obese) for primary analysis and as a multi-category variable for secondary analyses to explore more detailed relationships. The example values for the variables can be found in Table 3 below.

Table 3: Categorical Variables

Variable Name	Variable Values
Gender	Male or Female
family_history_with_overweight	Yes or No
HICAL	Yes or No
VEGE	1, 2 or 3
NMEAL	1, 2, 3 or 4
SNACK	No, Sometimes, Frequently or Always
SMOKE	Yes or No
WATER	1 = Less than a litre, 2 = Between 1 and 2 litres or 3 = More than 2 litres
CAL	Yes or No
PHY	0, 1, 2 or 3
TECH	0, 1 or 2
ALC	No, Sometimes, Frequently or Always
MTRANS	Automobile, Bike, Motorbike, Public Transport or Walking
OBSLVL	Insufficient Weight, Normal Weight, Obesity Type I, Obesity Type II, Obesity Type III, Overweight Level I, Overweight Level II

3. Regression Analysis

3.1 Logistic Regression Analysis for Obesity Prediction

In this section, a logistic regression model is applied to predict obesity (Obese = 1, Non-Obese = 0) based on various lifestyle and demographic factors. The dataset is preprocessed by converting categorical variables into factors and creating a binary obesity indicator. The model includes predictors - gender, family history of overweight, eating habits, physical activity, and mode of transportation. The results show which factors significantly influence obesity, with variables like family history, eating frequency, and transportation mode having notable effects.

```
##
## Call:
## glm(formula = Obese ~ Gender + family_history_with_overweight +
##     HICAL + VEGE_rounded + NMEAL_rounded + SNACK + SMOKE + WATER_rounded +
##     CAL + PHY_rounded + TECH_rounded + ALC + MTRANS, family = binomial,
##     data = data)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -13.06417  324.74465  -0.040 0.967911
## GenderMale                        0.12519    0.15548   0.805 0.420704
## family_history_with_overweightyes   2.74932    0.17623  15.600  < 2e-16 ***
## HICALyes                          0.55126    0.21270   2.592 0.009549 **
## VEGE_rounded2                     0.40229    0.31336   1.284 0.199213
## VEGE_rounded3                     0.33387    0.31829   1.049 0.294206
## NMEAL_rounded2                    1.19706    0.38853   3.081 0.002063 **
```

```
## NMEAL_rounded3                      -0.20410    0.19955  -1.023 0.306408
## NMEAL_rounded4                      -1.84263    0.30569  -6.028 1.66e-09 ***
## SNACKFrequently                     -0.50028    0.42337  -1.182 0.237338
## SNACKno                              2.77595    0.55239   5.025 5.02e-07 ***
## SNACKSometimes                       2.46945    0.38400   6.431 1.27e-10 ***
## SMOKEyes                            -0.20201    0.48930  -0.413 0.679718
## WATER_rounded2                      -0.17888    0.17672  -1.012 0.311453
## WATER_rounded3                       0.51827    0.22470   2.306 0.021083 *
## CALyes                              -0.03633    0.28861  -0.126 0.899838
## PHY_rounded1                        -0.24732    0.18247  -1.355 0.175285
## PHY_rounded2                        -0.88290    0.19664  -4.490 7.12e-06 ***
## PHY_rounded3                        -0.86004    0.30525  -2.817 0.004840 **
## TECH_rounded1                       -0.41978    0.15923  -2.636 0.008379 **
## TECH_rounded2                       -0.89846    0.21739  -4.133 3.58e-05 ***
## ALCFrequently                       11.31467  324.74430   0.035 0.972206
## ALCno                                9.76841  324.74410   0.030 0.976003
## ALCSometimes                        10.28183  324.74411   0.032 0.974742
## MTRANSBike                          -1.11466    0.96218  -1.158 0.246671
## MTRANSMotorbike                     -0.32612    0.82191  -0.397 0.691526
## MTRANSPublic_Transportation         -0.08588    0.18456  -0.465 0.641707
## MTRANSWalking                       -1.50503    0.42254  -3.562 0.000368 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2440.4  on 2110  degrees of freedom
## Residual deviance: 1416.1  on 2083  degrees of freedom
## AIC: 1472.1
##
## Number of Fisher Scoring iterations: 11
##                                    GVIF Df GVIF^(1/(2*Df))
## Gender                         1.286645  1        1.134304
## family_history_with_overweight 1.237339  1        1.112357
## HICAL                          1.110334  1        1.053724
## VEGE_rounded                   1.222064  2        1.051413
## NMEAL_rounded                  1.425671  3        1.060889
## SNACK                          1.515748  3        1.071777
## SMOKE                          1.040729  1        1.020161
## WATER_rounded                  1.217447  2        1.050419
## CAL                            1.107548  1        1.052401
## PHY_rounded                    1.441443  3        1.062836
## TECH_rounded                   1.263035  2        1.060117
## ALC                            1.273358  3        1.041098
## MTRANS                         1.349583  4        1.038185
```

After assessing collinearity using the Variance Inflation Factor (VIF), all GVIF values were below 1.5, indicating that collinearity is not an issue in our explanatory variables. Thus, our logistic regression model

is stable and interpretable.

Based on the results from fitting the GLM model, it can be seen that there are many variables that are significant and strongly associated with obesity from their p-value. The results show significance for a few variables:

- `family_history_with_overweightyes` has the strongest association with obesity (coefficient = 2.749, p-value $<$ 2e-16). This indicates that individuals with a family history of overweight/obesity are significantly more likely to be obese themselves.
- `SNACKSometimes` which represents eating food between meals sometimes (p-value = 1.27e-10). This indicates that irregular snacking is linked to higher obesity risk compared to consistent habits (eg, "always" or "never"), which may reflect disordered eating patterns or lack of routine.
- `PHY_rounded2` which represents having physical activity for 2-4 days (p-value = 7.12e-06). This indicates that moderate physical activity (2-4 days/week) is protective but less effective than daily activity, so limited exercise still reduces obesity risk significantly.
- `TECH_rounded2` which represents using technological devices more than 5 hours (p-value = 3.58e-05). Suggesting that the use of technology (sedentary behavior) indicates significance towards the risk of obesity.

Notably, smoking status and alcohol consumption did not demonstrate statistically significant associations with obesity (p-value $>$ 0.05). This could be due to many factors, such as the population studied having a limited representation of heavy smokers or drinkers, or these factors may influence obesity indirectly through other metabolic pathways not captured in the model.

3.2 Physical Activity Influence on Obesity

Now we will look into how physical activity influences the obesity level. In particular, the logistic regression model examines how physical activity influences obesity levels using three predictors, `PHY_rounded` (Frequency of Physical Activity), `TECH_rounded` (Time Using Electronic Devices), `MTRANS` (Mode of Transportation).

```
##
## Call:
## glm(formula = Obese ~ PHY_rounded + TECH_rounded + MTRANS, family = binomial,
##     data = data)
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   1.79623    0.14856  12.091  < 2e-16 ***
## PHY_rounded1                 -0.04498    0.12837  -0.350  0.72603
## PHY_rounded2                 -0.82176    0.13323  -6.168 6.91e-10 ***
## PHY_rounded3                 -0.87813    0.21725  -4.042 5.30e-05 ***
## TECH_rounded1                -0.12067    0.11173  -1.080  0.28015
## TECH_rounded2                -0.47770    0.16227  -2.944  0.00324 **
## MTRANSBike                   -1.38418    0.78557  -1.762  0.07807 .
## MTRANSMotorbike              -1.73970    0.62959  -2.763  0.00572 **
## MTRANSPublic_Transportation -0.38660    0.13444  -2.876  0.00403 **
## MTRANSWalking                -1.94663    0.31802  -6.121 9.29e-10 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2440.4  on 2110  degrees of freedom
## Residual deviance: 2314.6  on 2101  degrees of freedom
## AIC: 2334.6
##
## Number of Fisher Scoring iterations: 4
##                              Odds_Ratio   Lower_CI  Upper_CI
## (Intercept)                   6.0268755 4.52823597 8.1098713
## PHY_rounded1                  0.9560139 0.74296389 1.2292768
## PHY_rounded2                  0.4396570 0.33826998 0.5704165
## PHY_rounded3                  0.4155609 0.27204285 0.6385582
## TECH_rounded1                 0.8863289 0.71183691 1.1032484
## TECH_rounded2                 0.6202096 0.45219109 0.8547441
## MTRANSBike                    0.2505300 0.04759326 1.1821155
## MTRANSMotorbike               0.1755738 0.04855017 0.6104815
## MTRANSPublic_Transportation   0.6793645 0.51981175 0.8809336
## MTRANSWalking                 0.1427543 0.07512850 0.2629268
```

Table 4: Summary Statistics for Binomial Logistic Regression for Physical Activity

	Estimate	Std. Error	z value	Pr(>|z|)	Odds Ratio
PHY_rounded1	-0.04498	0.12837	-0.350	0.72603	0.9560139
PHY_rounded2	-0.82176	0.13323	-6.168	6.91e-10	0.4396570
PHY_rounded3	-0.87813	0.21725	-4.042	5.30e-05	0.4155609
TECH_rounded1	-0.12067	0.11173	-1.080	0.28015	0.8863289
TECH_rounded2	-0.47770	0.16227	-2.944	0.00324	0.6202096
MTRANSBike	-1.38418	0.78557	-1.762	0.07807	0.2505300
MTRANSMotorbike	-1.73970	0.62959	-2.763	0.00572	0.1755738
MTRANSPublic_Transportation	-0.38660	0.13444	-2.876	0.00403	0.6793645
MTRANSWalking	-1.94663	0.31802	-6.121	9.29e-10	0.1427543

Table 4 shows the results of binomial logistic regression examining the relationship between lifestyle factors and obesity risk. The analysis reveals a strong dose-response relationship for physical activity, where both moderate (Odds Ratio=0.44) and high activity levels (OR=0.42) significantly reduce obesity risk compared to inactivity, with moderate activity providing nearly maximal protective benefits. Interestingly, only high technology use demonstrated a significant protective effect (OR=0.62), contrary to conventional expectations about screen time, potentially indicating confounding factors or measurement issues. Regarding transportation methods, all active modes showed protective associations, particularly walking (OR=0.14), suggesting that incorporating physical activity into daily routines through commuting choices may be as impactful as structured exercise for obesity prevention.

To assess the model's predictive performance, a Receiver Operating Characteristic (ROC) curve was used. The ROC curve plots sensitivity (true positive rate) against 1-specificity (false positive rate) at various classification thresholds, providing a comprehensive measure of the model's discriminative ability.

(Hosmer et al., 2013) The Area Under the Curve (AUC) value, which ranges from 0.5 (fail) to 1.0 (excellent performance), serves as a key metric for evaluating classification performance. (Corbacioglu and Aksel, 2023)

In addition, a boxplot is employed to visualize obesity probability by physical activity levels.

```
## [1] "AUC: 0.655"
## [1] "Model Accuracy: 74.47 %"
```

```
plot(roc_curve, col="blue", main="ROC Curve for Obesity Prediction")
```

Figure 1: ROC Curve for Obesity Prediction

The AUC of 0.655 falls into the "poor discrimination" category (Hosmer et al., 2013), indicating that the model's ability to distinguish between obese and non-obese individuals is only slightly better than random guessing; its predictive power is limited. This implies the model correctly ranks 65.5% of obese cases higher than non-obese cases. The model correctly classifies 74.47% of individuals as either obese or non-obese when using a 0.5 probability threshold. From Figure 1, the curve's position closer to the diagonal than the top-left corner suggests room for improvement. To improve our model, we can add more predictive variables and address potential measurement error in the predictors.

As shown in Figure 2, the predicted probabilities demonstrate three coherent patterns: (1) Higher physical activity levels exhibit systematically lower obesity risk compared to sedentary individuals, consistent with their significant protective odds ratios (OR=0.44 and 0.42, respectively); (2) Automobile users show the highest median obesity probability, followed by public transportation (OR=0.68), confirming motorized transport as a risk factor; and (3) Active transportation modes—particularly walking (OR=0.14) and motorbike use (OR=0.18) show the most pronounced protective effects, likely due to incidental energy expenditure. The boxplots' interquartile ranges reveal overlapping distributions between groups, explaining the model's moderate discriminative capacity (AUC=0.655), while underscoring that lifestyle factors operate probabilistically rather than deterministically.

```
combined_plot
```

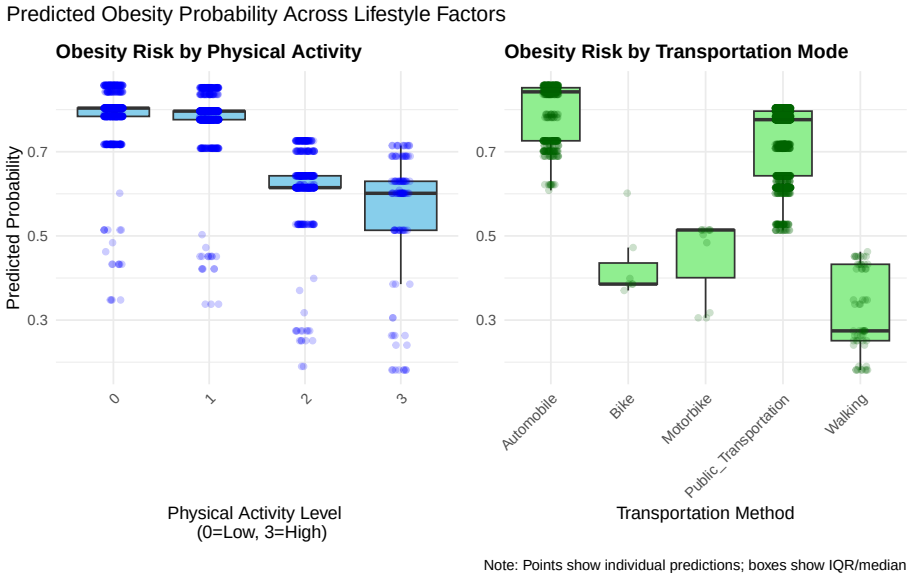

Figure 2: Predicted Obesity Probability Across Lifestyle Factors

3.3 Backward Selection

Now, to estimate obesity levels based on lifestyle factors, we will be doing Backward Selection to remove predictors that are deemed insignificant.

```
##                     (Intercept) GenderMale    Height     Weight
## Normal_Weight        343.352256   14.01462 -523.7059  8.1139174
## Obesity_Type_I        52.266924   34.61969 -136.2769  2.6926999
## Obesity_Type_II      184.121716   24.01398 -435.8857  7.3050720
## Obesity_Type_III       9.345408  -99.05184 -300.5829  8.5699254
## Overweight_Level_I  -353.354128   35.01596  290.3235 -1.7811089
## Overweight_Level_II -180.085717   37.07309  112.5178  0.1792264
##                     family_history_with_overweightyes   HICALyes      WATER
## Normal_Weight                           -8.1695849  -3.781915  -3.856454
## Obesity_Type_I                           3.6890112 -17.847504  -7.211233
## Obesity_Type_II                         -5.1591155 -15.068455 -19.784886
## Obesity_Type_III                        24.3265143  13.883885 -11.056996
## Overweight_Level_I                      -3.2473061 -22.017360  -8.339534
## Overweight_Level_II                      0.6291883 -24.475320  -7.322417
##                          PHY ALCFrequently       ALCno ALCSometimes
## Normal_Weight        0.1083526     112.54015  116.355404   114.456703
## Obesity_Type_I       2.4619842      25.78882    6.170226    20.307873
## Obesity_Type_II     -7.8135577      54.09271   66.888069    63.140934
## Obesity_Type_III     5.5685687      10.49263    3.753782    -4.901003
```

```
## Overweight_Level_I   2.2554383    -113.86669 -128.738145  -110.749297
## Overweight_Level_II  1.6578868     -54.83871  -70.094578   -55.152429
##                     TECH_rounded1 TECH_rounded2       Obese
## Normal_Weight            3.867923     3.2731148 -340.098193
## Obesity_Type_I          -7.810258    -1.2720041   -6.467639
## Obesity_Type_II        -11.546775   -16.7074024 -103.762939
## Obesity_Type_III        -7.569771   -40.2385890 -263.003606
## Overweight_Level_I      -7.421024    -0.8865811  190.083904
## Overweight_Level_II     -5.200161    -0.4488914  104.214206
```

The output of the code displays the effect of each predictor with regard to the odds of each weight category. Positive coefficients indicate a more likely odds of being in that weight category, and negative coefficients indicate a less likely odds.

Based on the output, the following variables have been removed from the Backward Selection process:

- `VEGE`: Whether or not the subject eats vegetables
- `TECH`: How much time the subject uses technological devices
- `MTRANS`: What mode of transportation the subject usually uses

A few significant findings through backward selection include, `Males` are more likely to have Obesity Type II but have dramatically lower odds of having Obesity Type III in comparison to `Females`. Whether or not the subject `SMOKE` also surprisingly increases the odds across all weight categories.

Other findings are as expected, such as `Age` and `SNACK` (Having food in between meals) are expected to increase the odds of being obese across all weight categories. There are some predictors which specifically affects the odds of being in a certain weight class, such as: `family_history_with_overweight` and `Obesity_Type_I` having positive relation, `HICAL` (especially when "Sometimes") and `Normal_Weight` having negative relation, and many more.

Based on the analysis, these findings emphasize the nature of obesity and suggest that public health interventions should focus on promoting structured eating habits, increasing access to physical activity opportunities, and addressing genetic predispositions. Future research should explore a broader range of variables and employ more precise measurement techniques to enhance predictive accuracy and develop more targeted obesity prevention strategies.

3.4 GAM

As Backward Selection only cover the possible linear relationship between the predictors and response variables. GAM would be able to cover non-linear relationships.

From the plot above, it can be shown that the predictors: "Age", "Height, and "TECH" might have a linear relationship with the output, whilst "Weight" might have a non-linear relationship

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## OBSLVL ~ Gender + s(Age, bs = "cr") + s(Weight, bs = "cr") +
##     family_history_with_overweight + HICAL + SMOKE + MTRANS
```

```
gam.check(gam_model)
```

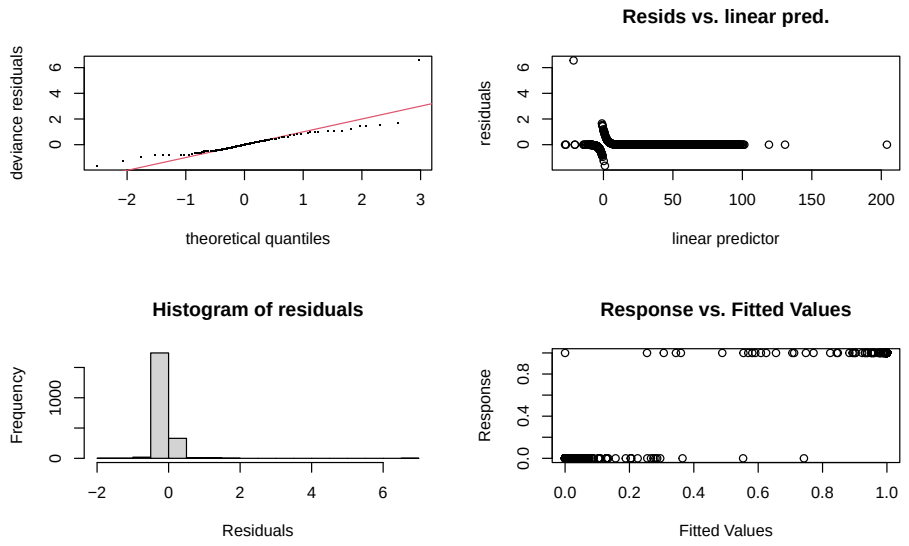

Figure 3: GAM Model Diagnostic

```
plot(gam_model, pages = 1, shade = TRUE, scheme = 1)
```

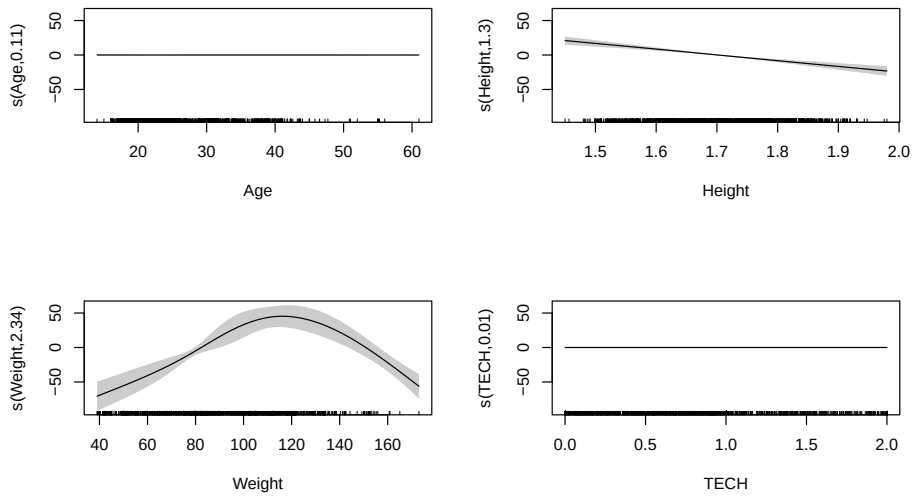

Figure 4: Visualization of Smoothing Terms

```
##
## Parametric coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       1.586e+01  1.377e+00  11.512  < 2e-16 ***
## GenderMale                       -2.092e+00  3.551e-01  -5.893 3.79e-09 ***
## family_history_with_overweightyes -9.056e-01  2.973e-01  -3.046 0.002320 **
## HICALyes                         -3.660e-01  3.558e-01  -1.029 0.303686
## SMOKEyes                          2.480e+00  1.386e+00   1.790 0.073490 .
## MTRANSBike                        2.717e+01  3.100e+05   0.000 0.999930
## MTRANSMotorbike                   2.727e+01  2.356e+05   0.000 0.999908
## MTRANSPublic_Transportation       1.347e+00  3.771e-01   3.573 0.000352 ***
## MTRANSWalking                     1.821e+00  7.323e-01   2.487 0.012875 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df  Chi.sq p-value
## s(Age)    1.147  1.279   4.985   0.042 *
## s(Weight) 1.000  1.000 149.624  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.736   Deviance explained = 75.6%
## fREML = 2108.2  Scale est. = 1         n = 2111
```

Now that we have have done both Backward Selection as well as fitting GAM to the data, we will be comparing the AIC and BIC values of both models to see which model explains the predictors better.

```
##                Model      AIC      BIC
## 1 Backward Selection 279.0472 702.8271
## 2                GAM 417.4024 476.6474
```

The backward selection model has a lower AIC (279.05 vs 417.40), indicating it provides a better balance of fit and simplicity when only considering fit quality. However, the large gap in BIC suggests this advantage may come from overfitting. The GAM has a significantly lower BIC (476.65 vs 702.83), meaning it's the strongly preferred model when penalizing complexity more heavily. BIC's stronger penalty on parameters indicates the backward model is likely overparameterized.

4. Discussion and Conclusion

The analysis using the GLM model identified several key factors strongly associated with obesity. The most significant predictor was a family history of overweight, which had the highest positive association with obesity (coefficient = 2.749, p-value < 2e-16). Additionally, irregular snacking habits (p-value = 1.27e-10) and sedentary behavior from excessive technology use (p-value = 3.58e-05) were also strongly linked to increased obesity risk. Conversely, moderate physical activity (2-4 days per week) was found to be protective, though less effective than daily exercise (p-value = 7.12e-06). Interestingly, smoking and alcohol consumption did not show statistically significant associations with obesity in this dataset. Importantly, the variance inflation factor (VIF) analysis confirmed that collinearity was not a concern, ensuring the reliability

of our findings. These results highlight the importance of genetic predisposition, dietary patterns, and sedentary lifestyle choices in obesity risk, emphasizing the need for targeted interventions focusing on both lifestyle modifications and public health policies.

Through binary logistic regression, the role of physical activity in reducing obesity risk was examined, with moderate (OR = 0.44) and vigorous (OR = 0.42) activity both significantly lowering the odds of obesity compared to inactivity. Walking showed the strongest protective effect among transportation modes (OR = 0.14), highlighting the importance of incorporating movement into daily routines. Surprisingly, high screen time was also associated with reduced obesity risk (OR = 0.62), which contradicts typical expectations and suggests potential confounding factors or measurement issues. With a moderate predictive ability (AUC = 0.655), the model suggests the need for additional factors like diet and smoking to improve accuracy. Future research should refine measurement tools and examine demographic subgroups to strengthen public health strategies against obesity.

The backward selection process refined our model by removing insignificant predictors, ensuring a more accurate estimation of obesity levels based on lifestyle factors. Key findings indicate that males are more likely to have Obesity Type II but have significantly lower odds of Obesity Type III compared to females. Additionally, smoking status unexpectedly increased the odds across all weight categories. Expected results, such as age and snacking behavior contributing to obesity risk, were reaffirmed. Specific associations were also observed, including a strong positive relationship between family history of overweight and Obesity Type I, as well as a negative correlation between high-caloric food consumption and normal weight.

The results show that while the simpler stepwise model fits the current data well due to having a smaller AIC value of 279.0472 (in comparison to 417.4024), the flexible GAM approach likely works better for real-world predictions by capturing subtle patterns in how factors like age and weight relate to obesity. If you need clear, straightforward explanations of risk factors, the stepwise model may be preferable due to it having a lower BIC value of 476.6474 (in comparison to 702.8271). But if accurate predictions matter most—especially when body measurements have complex effects—the GAM is probably the better choice. The best model depends on whether you prioritize easy interpretation or stronger predictive power for your specific needs.

These results reinforce the complexity of obesity and suggest that public health interventions should target structured eating habits, increased physical activity, and awareness of genetic predispositions. The removal of variables like vegetable consumption, technology use, and transportation mode suggests that other lifestyle factors may play a more dominant role in obesity prediction. Future research should integrate additional variables and employ more precise measurement methods to enhance model accuracy and develop tailored obesity prevention strategies.

5. References

Corbacioglu, S. K., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish journal of emergency medicine*, *23(4)*, 195–198.

Estimation of obesity levels based on eating habits and physical condition [Dataset]. (2019). https://doi.org/10.24432/C5H31Z (accessed on 2025-03-10).

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd) [Ebooks Corporation]. Wiley.

Palechor, F. M., & Manotas, A. D. (2019). Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from colombia, peru and mexico. *Data in Brief*, *25*.

Swastik. (2025, March). Smote for imbalanced classification with python. https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/ (accessed on 2025-04-01).

Tatvasoft. (2023, August). Data mining with weka. https://www.tatvasoft.com/blog/data-mining-with-weka/ (accessed on 2025-04-01).

World Health Organization. (2024). *Obesity and overweight*. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight (accessed on 2025-03-10).