Factors Associated with Obesity

Zhiyi Lu[1, a, †], Qianyu Sun[2, b, †], Yao Zhang[3, c, †], Yongyi Zhou[4, d, *, †]

[1]Applied Statistics, Mathematical and Statistical Institute, Nanjing University of Information Science and Technology, Nanjing, China

[2]Biopharmacy, College of Pharmaceutical Sciences, Soochow University, Suzhou, China

[3]Grade 12, Rosedale Academy, Zhengzhou, China

[4]Bioinformatics, Faculty of Health Science, University of Macao, Macao, China

*Corresponding author:
cc11265@connect.um.edu.mo, [b]201913870061@nuist.edu.cn, [c]1522246079@qq.com, [d]2507625562@qq.com

[†]These authors contributed equally to this work.

Keywords: obesity, mass body index, NObesity.

Abstract: Background Lots of public health studies have investigated the factors associated with obesity or the body mass index on different groups of individuals and have reached conclusive results. No study has been conducted among three developing countries in Americas: Peru, Colombia, and Mexico. We investigated the hypothesis that eating habits, daily habits, sports, and family history are all related to obesity. **Method**: The data was donated on Aug.27th of 2019. Individuals from the countries of Mexico, Peru, and Colombia, with ages between 14 and 61 and diverse eating habits and physical conditions are investigated. The data was collected by web platform survey and generated synthetically, then was processed obtaining 2111 records. The records are then labeled with the class variable NObesity (Obesity Level). Mixed-effects modeling was used to examine the associations between different factors and obesity. **Results:** All x variables including FAVC, NCP, SMOKE, SCC, TUE, family history with overweight, FCVC, CAEC, CH2O, FAF, CALC, MTRANS are correlated with the body mass index. SCC, CAEC and MTRANS have a negative association with mass_body_index. FAVC, CH20 have a positive associtation with mass_body_index.

1. Introduction

Nowadays, with the improvement of people's living standards, more and more people become overweight and even obese due to unscientific diets, lack of exercise or other reasons. In September 2021, UN Secretary-General Antonio Guterres pointed out at the first UN Food System Summit that more than 2 billion people around the world are overweight or obese. However, being obese brought a lot of inconvenience to people’s life, such as enabling them to move flexibly. Besides that, Obesity does great harm to people’s health. For example, it will increase the risk of developing diabetes and coronary heart disease. Our purpose of research is finding how the related factors influence the obesity level by analyzing the data.

Mexico, Peru and Colombia are all developing countries in Americas. The lifestyles and living habits of residents in these three countries are different from those developed countries in Americas. Therefore, we want to explore the factors that may associated with obesity in people in developing countries to help people prevent them in their daily lives.

Reading the data through the R language to get the data content and a simple data analysis doned with data and images which includes exploratory data analysis and regression analysis. More details will be showed later.

2. Method

2.1 Result obtaining

The body mass index is calculated from the equation $Mass\ body\ index = \frac{Weight}{height*height}$. Then the index is compared with data provided by WHO and Mexican Normativity and obtain the estimation of obesity levels for each individual based on eating habits and physical condition. [1]

Table 1. Definition of NObesity (obesity levels)

Mass body index	**Obesity level**
Less than 18.5	Underweight
18.5-24.9	Normal
25.0-29.9	Overweight
30.0-34.9	Obesity I
35.0-39.9	Obesity II
Higher than 40.0	Obesity III

2.2 Study population

Individuals from the cities: Barranquilla (Colombia), Lima (Peru), City of Mexico (Mexico), with ages between 14 and 61 and diverse eating habits and physical condition are investigated. The data was collected: 23% by using a web platform with a survey where anonymous users answered each question, the other 77% of the data was generated synthetically using the Weka tool and the SMOTE filter. Then the data was processed obtaining 17 attributes and 2111 records.

Table 2. Distribution of selected characteristics of the selected distribution (*N*=2111)

Characteristics	***N***	**Percentage (%)**
Frequent consumption of high caloric food (FAVC)		
Yes	1866	88.4
No	245	11.6
Frequency of consumption of vegetables (FCVC)		
Never	102	4.8
Sometimes	1013	48.0
Always	996	47.2
Number of main meals (NCP)		
One	316	15.0
Two	176	8.3
Three	1470	69.6
More than three	149	7.1
Consumption of food between meals (CAEC)		
No	51	2.4
Sometimes	1765	83.6

Frequently	242	11.5
Always	53	2.5
Consumption of water daily (CH20)		
Less than a liter	485	23.0
Between 1 and 2 L	1110	52.6
More than 2 L	516	24.4
Consumption of alcohol (CALC)		
Never drink	639	30.3
Sometimes	1401	66.4
Frequently	70	30.3
Always	1	0.05
SMOKE		
Yes	44	2.1
No	2067	97.9
Calories consumption monitoring (SCC)		
Yes	96	4.5
No	2015	95.6
Physical activity frequency (FAF)		
do not have	720	34.1
1 or 2 days	776	36.8
2 or 4 days	496	23.5
4 or 5 days	119	5.6
Time using technology devices (TUE)		
0–2 hours	952	45.1
3–5 hours	915	43.3
More than 5 hours	244	11.6
Transportation used (MTRANS)		
Automobile	457	21.6
Motorbike	11	0.5
Bike	7	0.3
Public Transportation	1580	74.8
Walking	56	2.7
Family member who suffered or suffers from overweight		
Yes	1726	81.8
No	385	18.2

2.3 Data collection

The initial data was collected using the form of a web page questionnaire, and the survey was accessible online for 30 days so the users could evaluate their eating habits and some aspects that helped to identify their physical condition. [1] Then the data was processed by the Weka tool and SMOTE filter. Weka is a data science tool, and SMOTE is a Weka oversampling technique that uses to increase the minority group by generating synthetic samples. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together. The filter required to indicate the class for the generation of synthetic data, the number of nearest neighbors used, the percentage that you need to increase the selected class, and the random seed used for random sampling. Other aspects analyzed were the identification of atypical and missing data. After the filter was applied to each category, the final result was 2111 records. [1-3]

2.4 Statistical analysis

All analysis was performed separately for each individual. For all analysis, the software R studio was used. [4-7] The independent variable is mass body index which is continuous data and the dependent variables are all categorical variables. EDA (exploratory data analysis) was processed first, histogram, density plot, and box plots were produced in this part, and the p-values of t-test for each y were calculated. Histogram and density plot could show the distribution of the body mass index of people. Various box plot for each x to y was used to explore the relationship between them and the t-test can present the linear correlation by comparing the calculated p-value with the critical value. When the box plots were synthesized, the category variables were recoded by number, which can be seen in table 3. Then regression analysis was processed, in this dataset, both linear regression and polynomial regression were applied and calculated the R square. After that, the interaction term was added in the regression and found. Lastly, variable selection and performance evaluation was conducted. Different variable selection methods were chosen and conducted for this data-set, which is ridge regression, forward and backward stepwise selection.

Table 3. Recode of selected characteristics

Characteristics	Recode
Frequent consumption of high caloric food (FAVC)	
Yes	0
No	1
Frequency of consumption of vegetables (FCVC)	
Never	1
Sometimes	2
Always	3
Number of main meals (NCP)	
One	1
Two	2
Three	3
More than three	4
Consumption of food between meals (CAEC)	
No	1
Sometimes	2

Frequently	3
Always	4
Consumption of water daily (CH20)	
Less than a liter	1
Between 1 and 2 L	2
More than 2 L	3
Consumption of alcohol (CALC)	
Never drink	1
Sometimes	2
Frequently	3
Always	4
SMOKE	
Yes	0
No	1
Calories consumption monitoring (SCC)	
Yes	0
No	1
Physical activity frequency (FAF)	
do not have	0
1 or 2 days	1
2 or 4 days	2
4 or 5 days	3
Time using technology devices (TUE)	
0–2 hours	0
3–5 hours	1
More than 5 hours	2
Transportation used (MTRANS)	
Automobile	1
Motorbike	2
Bike	3
Public Transportation	4
Walking	5
Family member who suffered or suffers from overweight	
Yes	1
No	0

3. Results

Table 1 (shown above) shows the distributions of selected characteristics of the study population. More than three quarters of people are non-smokers (97.9%), consume vegetables (95.2%) and high caloric food (88.4%), consume food between meals sometimes (83.6%) and have family members who suffered or suffers from overweight (81.8%).

More than half of the people have three main meals daily (69.6%), take public transportation (74.8%), consume water between 1-2L daily (52.6%) and drink alcohol sometimes (66.4%). Few people never comsume food between meals (2.4%), always drink alcohol (0.05%), smoke (2.1%), do calories consumption monitoring (4.5%), take part in physical activity 4 or 5 days a week (5.6%), use technology devices more than 5 hours (11.6%) and walk (2.7%).

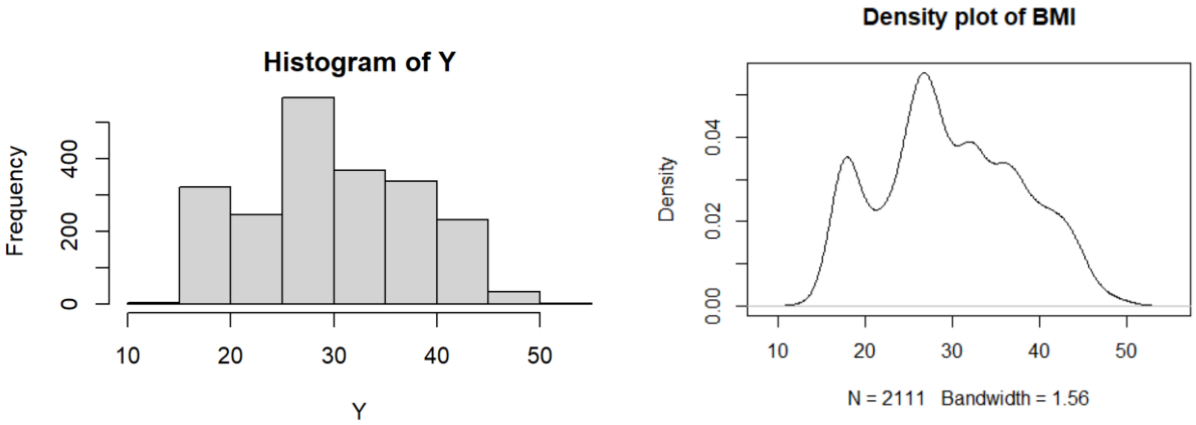

Figure 1. The histogram and density plot of BMI index.

Figure 1 shows the distribution of the mass_body_index of people. People who have a mass_body_index between 25-30 which means they are overweight are the most, accounting for more than 25% of people. Besides that, nearly half of the people have a mass_body_index more than 30 which means they are obese. Among them about 350 people have a mass_body_index between 30-35 which means they are on obesity level 1, about 330 people have a mass_body_index between 35-40 which means they are on obesity level 2, about 250 people have a mass_body_index more than 40 which means they are on obesity level 3. There are approximately 250 people have a mass_body_index less than 25 which means they are normal. In addition to that, people who have a mass_body_index less than 18.5 which means they are underweight makes up about 10% of people as shown in the figure, there are only about 500 people are non-obese. More than 75% people have a mass_body_index more than 25 which means they are overweight or obese.

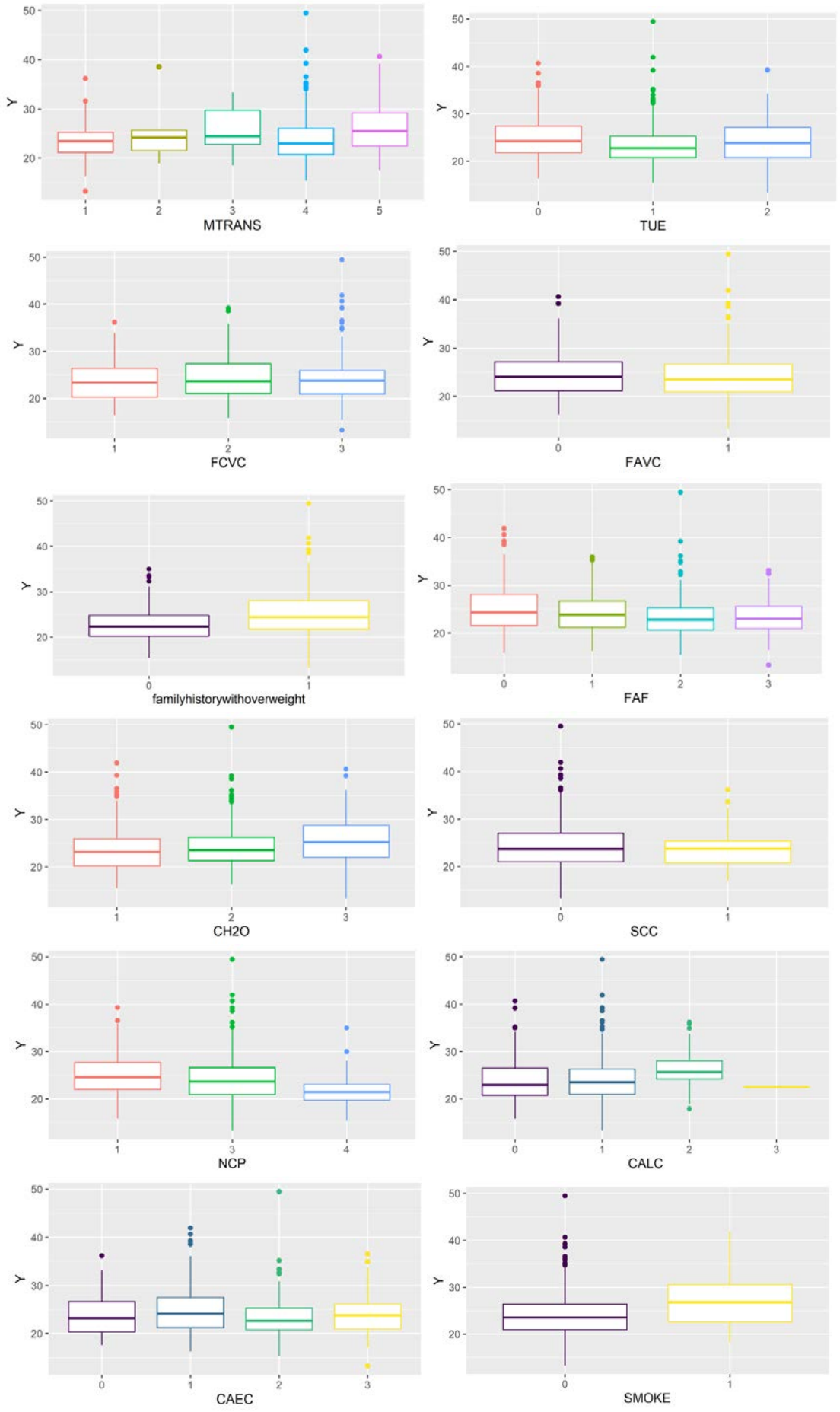

Figure 2. The box plots between different x variables and the BMI index.

Figure 2 presents the associations between 12 x variables and mass_body_index. When X is transportation, the box plot shows that people who walk have a lower mass_body_index than people who take public transportation, people who take public transportation have a lower mass_body_index than people who drive cars. Therefore, the variable "MTRANS" is related to BMI. When X is TUE, people using technology devices more than 5 hours have a lower mass_body_index than people who use technology devices less than 5 hours. People who use technology devices between 3- 5 hours have a higher mass_body_index than people who use technology devices between 0- 2 hours. Therefore, the variable "TUE" is related to BMI.

All X variables are correlated with the body mass index, because all the calculated p-value of t test are smaller than 0.05, which is "p-value<2.2e-16". The R square of the linear regression model is 0.4361. When we add poly (FCVC,2) and poly (MTRANS,2) into the model, the R square is 0.4779, which is higher than that in lm1 (0.4361). Including the interactions between FAF and MTRANS and the interaction between CAEC and CALC will increase the R square from 0.4779 to 0.4818. The interaction term is significant under 0.01 significant level.

Table 4. Association between all X variables and mass_body_index.

variables	linear regression		Polynomial regression		Interactions	
	β^{α}	p-value	β^{α}	p-value	β^{α}	p-value
Age	0.16200	4.21e-11	0.32090	< 2e-16	0.31398	< 2e-16
FAVC	1.97872	2.12e-10	1.71699	1.08e-08	1.72711	8.02e-09
NCP	0.38697	0.025651	0.41828	0.0128	0.44848	0.007659
SMOKE	-0.34642	0.599901	-0.60801	0.3395	-0.57421	0.365470
SCC	-1.88423	5.12e-05	-1.82480	4.73e-05	-1.91467	1.94e-05
TUE	-0.54638	0.017534	-0.41987	0.0583	-0.46992	0.033839
Family history with overweight	5.42923	< 2e-16	5.37339	< 2e-16	5.33329	< 2e-16
FCVC	3.74477	< 2e-16	NA	NA	NA	NA
poly (FCVC, 2)1		NA	88.19459	< 2e-16	89.88746	< 2e-16
poly (FCVC, 2)2		NA	42.23914	2.83e-12	42.13416	2.83e-12
CAEC	-3.76252	< 2e-16	-3.90012	< 2e-16	-6.44600	4.16e-14
CH2O	0.66536	0.003116	0.67399	0.0019	0.71702	0.000943
FAF	-1.18035	3.78e-13	-0.90937	8.14e-09	0.84215	0.277967
CALC	1.92118	3.35e-13	1.64932	1.07e-10	-1.59637	0.130734
MTRANS	-0.84414	0.000105	NA	NA	NA	NA
poly (MTRANS, 2)1	NA	NA	-39.25405	4.22e-09	-20.72804	0.048103
poly (MTRANS, 2)2	NA	NA	-71.43725	< 2e-16	-74.80910	< 2e-16
I (FAF * MTRANS)	NA	NA	NA	NA	-0.43326	0.020269
I (CAEC * CALC)	NA	NA	NA	NA	1.44303	0.001569

Table 4 shows the associations between all X variables and mass_body_index. Multiple factors were found to be associated with mass_body_index. For example, SCC(β^{α} = -1.88423, p-value = 5.12e-05), CAEC(β^{α} = -3.76252, p-value < 2e-16) and MTRANS(β^{α} = -0.84414, p-value =

0.000105) have a negative association with mass_body_index.FAVC(β^{α} = 1.97872, p-value = 2.12e-10), CH20(β^{α} = 0.66536, p-value = 0.003116) have a positive association with mass_body_index

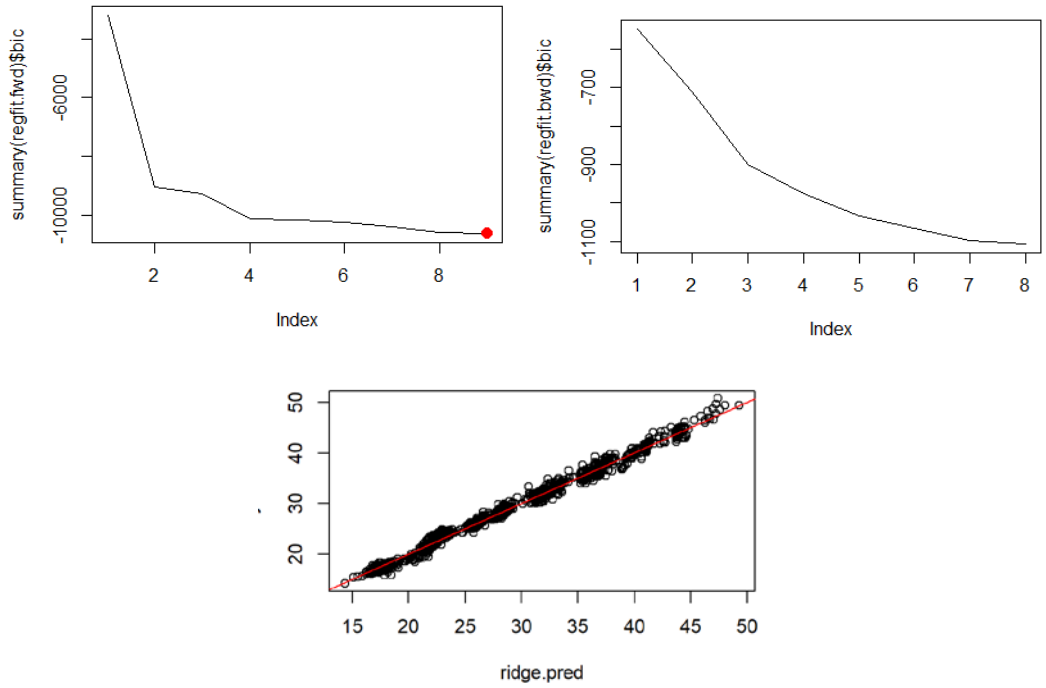

Figure 3. The result of forward stepwise selection, backward stepwise selection, and ridge regression.

As can be seen in Figure 3, the best model selected by forward method select 9 variables. The variable includes NCP, familyhistoryoverweight. L. FAVC. L, familyhistoryoverweight. L. NCP, familyhistoryoverweight. L. CAEC, familyhistoryoverweight. L. Y, NCP. CAEC, TUE. M, familyhistoryoverweight1, I (Y. interaction. full [,1] ^2). The best model selected by backward method select 8 variables. The variables include Age, FAV.L, SCC.L, familyhistorywithoverweight. L, FCVC, CAEC, FAF, CALC. According to the ridge regression, none of the coefficients are zero.

4. Discussion

We used six methods to analyze the data, namely linear regression, polynomial regression, interactions, forward stepwise selection, backward stepwise selection and Ridge region. Next, the results of each data analysis method are discussed separately. For linear regression, Age, FAVC, and SCC are the most significant, followed by family history with overweight. TUE and NCP are also significant. SMOKE is not significant. For polynomial regression, Age, FAVC.L, SCC.L, CAEC, CALC, poly (FCVC,2)1, and FAF are the most significant, followed by poly (FCVC, 2)2, CH20, poly (MTRANS, 2)1 and poly (MTRANS, 2)2. Family history with overweight and NCP are also significant. SMOKE.L and TUE are not significant. For interactions, Age, FAVC.L, SCC.L, CAEC, CH20, and poly (FCVC,2)1are the most significant, followed by poly (FCVC, 2)2, poly (MTRANS, 2)2 and I(CAEC*CALC). Family history with overweight, NCP, I(FAF*MTRANS), TUE and poly (MTRANS, 2)2 are also significant. SMOKE.L and TUE are not significant. For forward stepwise selection, the variable includes NCP, family history with overweight. L. FAVC. L, family history with overweight. L. NCP, family history with overweight. L. CAEC, family history with overweight. L. Y, NCP. CAEC, TUE. M, family history with overweight, I (Y. interaction. full [,1] ^2). For backward stepwise selection the variables include Age, FAV.L, SCC.L, familyhistorywithoverweight. L, FCVC, CAEC, FAF, CALC. For ridge regression, as expected, none of the coefficients are zero-ridge regression does not perform variable selection. According to data analysis, most of our hypothesis are reasonable. In other words, they have an association with obesity.

A small proportion of hypothesis are not reasonable, including SMOKE and TUE, which are negligible in relation to obesity.

Some of our results are the not only same as the existing literature, but also different as the literature [8]. We all get the results that the consumption of alcohol had an association in obesity. But, for transportation used, we got different results. We differ from the existing literature in that we have included the time people spend using cell phones in the study through the current social prevalence, although this has little relationship with the final results. In addition, transportation is so convenient nowadays that people can hardly travel without transportation and rarely choose to reach their destination on foot. Therefore, we also included people used of transportation in this study and found that it was also strongly associated with excessive obesity. We believe that the reason for this difference from the existing literature is that technological advances have influenced people's behavior in their daily lives. In a time when transportation was not as developed, it was not expected that transportation use would be associated with excessive obesity. There is a strong association between dietary aspects and excessive obesity, both before and now. The quality of life is much better than before, and it has also made excessive obesity more and more frequent.

5. Conclusion

In the past, obesity once accounted for a large proportion in developed countries. Nowadays, obesity is also gradually facing people in developing countries. [9] From the results obtained, it helps people understand what factors may be related to obesity since global trade, especially in recent years. With the rapid development of the world economy, people in developing countries gradually began to pursue more nutritious food, and people's living standards continued to improve. Therefore, it has become one of the hidden dangers of health. The fast pace of life indirectly changes people's way of life. The pressure of work leads to people's physical fatigue, which reduces the frequency of exercise. Based on an article, obesity can lead to many diseases, such as diabetes, cardiovascular disease, lungs disease and kidneys disease. [10] So how to help people effectively prevent obesity in today's era? First, the government should establish a relevant medical system to regularly check whether people are suffering from obesity. Secondly, governments in developing countries should build parks for people's leisure and fitness equipment for exercise. Relevant gyms are equipped in residential buildings. Finally, the government should encourage people to pursue a healthier lifestyle.

References

[1] MariaI.RodriguezMD,MPH,MeganSkyeMPH,MitraShokatBA,RachelLinzMPH,NisreenPedhiwalaMPH,Blair G.DarneyPhD, MPH,Expanded Access to Postabortion Contraception under Oregon's Reproductive Health Equity Act,2022, 32,Women's Health Issues

[2] SWASTIK SATPATHY, Overcoming Class Imbalance using SMOTE Techniques, October 6, 2020

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer: SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, Volume 16, pages 321-357, 2002

[4] Robert I. Kabacoff, R in Action - Data Analysis and Graphics with R, August 2011

[5] Hadley Wickham, Advanced R, 2014

[6] Yanchang Zhao, R and Data Mining Examples and Case Studies, 2012

[7] Nina Zumel, John Mount, Practical Data Science with R, 2019

[8] Samuel Dagne, Yalemzewod Assefa Gelaw, Zegeye Abebe, and Molla Mesele Wassie. Factors associated with overweight and obesity among adults in northeast Ethiopia: a cross-sectional study, 2019.

[9] Safia S Jiwani, Rodrigo M Carrillo-Larco, Akram Hernández-Vásquez, Tonatiuh Barrientos-Gutiérrez, Ana Basto-Abreu, Laura Gutierrez, Vilma Irazola, Ramfis Nieto-Martínez, Bruno P Nunes, Diana C Parra, J Jaime Miranda. The shift of obesity burden by socioeconomic status between 1998 and 2017 in Latin America and the Caribbean: a cross-sectional series study.

[10] Jenny C. Censin, Sanne A. E. Peters, Jonas Bovijn, Teresa Ferreira, Sara L. Pulit, Reedik Magl, Anubha Mahajan, Michael V. Holmes, Cecilia M. Lindgren. Causal relationships between obesity and the leading causes of death in women and men, 2019. PLoS Genet 15(10): e1008405.