The relation between Deaths caused by High BMI and Deaths Caused by Diet Factors

Yuhong Wang

University of Houston

12/22/2022

I. Introduction

The stubborn rise of obesity in all parts of the world has caused concern among high-income and low-income nations alike. The study and analysis of obesity have grown in tandem with its pervasiveness, as obesity can lead to life-ending complications and upend the well-being of communities around the globe. Unlike obesity in the past, where the ability to increase body size was restricted to those with food sources, wealth, and security, cheap, highly processed foods have democratized excessive weight gain. The general health of a population deteriorates as unabated weight gain causes potentially dangerous health complications, and a method of detection and treatment is a powerful tool to wield against such a threat.

This phenomenon prompted us to consider the following: is it possible to determine which risk factors, specifically diet factors, can be targeted to fight obesity-related deaths indirectly? For example, if a diet high in sodium causes high BMI, could reduce deaths related to high sodium decrease the prevalence of deaths tied to high BMI? This information could be used to understand the strength of the relationship between high BMI and other risk factors. These secondhand risk factors that influence BMI-related deaths could then be targeted through education or regulations to help combat deaths related to high BMI on multiple fronts.

We first needed to find the proper data set to answer this question. Our data set was found on Kaggle, an online source of datasets. The set we selected is titled "Worldwide Deaths by Country/Risk Factors," and the data is sourced from the World Health Organization. This dataset contains 6,840 observations with 31 different feature variables ranging from air pollution to drug use. Variables in the dataset are measured in deaths per year in a given country, which varies from observation to observation. For our study, we limited our variables to 8, including high BMI, and we only selected variables related to diet to have a more targeted approach to our analysis. All risk factors could have been analyzed to determine their relation to BMI-related deaths, but such as study is beyond the scope of our purpose to focus on diet factors. Therefore, the number of variables was cut down from 31 risk factors to 7 risk factors related to diet, excluding the variable target deaths caused by high BMI.

II. Methodology

The first step to deriving a model that can answer our questions is the prepare the data itself. This preprocessing will be done by adjusting and scaling the data set selected for this study. Then, exploratory methods will be employed to begin our understanding of the possible relationships between the variables. Such methods include a simple linear regression, a summary table, and an analysis of variance table to apply context to it. Correlation plots and variance inflation factors will also be applied to visualize how the features interact.

Once the primary linear model is made, partial F-tests and the sums of squared regression in tested models will be used to determine which feature variables are most and least relevant to understanding our question. The VIF will continually be used to verify that multicollinearity remains a nonissue whenever we derive a different model from explaining the data. When we have found the best models to explain the variation in the data, two methods will be used to select the best one: the AIC and SBC criteria. Both metrics determine which model best fits and adjust for adding more variables. The model with the lowest AIC and SBC scores has the better fit and will be selected as our final model to answer our questions.

III. Data Analysis

After the initial setup of the raw data set from Kaggle, it had to be edited to suit our needs. There are many different countries in our data set, and we limited our country selection to 8. Those countries are the following: Afghanistan, Bangladesh, Algeria, Angola, Argentina, Benin, Bolivia, and Botswana. These are relatively low-income countries, except Argentina, a middle-income country. The dataset itself was reduced to a subset of observations limited to these countries, and the features of each observation are the eight diet variables previously mentioned. Those eight diet-related features are as follows: diet high in sodium, diet low in whole grains, alcohol use, diet low in fruits, unsafe water source, diet low in nuts and seeds, diet low in vegetables, and high body mass index. High BMI will be the response variable in our study.

It was a concern that the number of deaths for a particular country would skew the results if one country had a more significant nominal number of deaths than others. To avoid this, we created individual sets for each country, which were scaled individually. Then, the sets were bound together to create the data set that would be used. The final dimensions of the data set are 240 observations by 31 features. Table 1 shows the regression results.

Though the overall F-statistic of the simple model is both high and statistically significant ($p < 2.2e-16$), there is already evidence that there are redundant variables in the model. Diet.low.in.whole.grains and Alcohol.use are highly suspect. This problem may be caused by high multicollinearity in the model. The VIF function was used to calculate the variance inflation factor of each variable.

As can be seen with the above results, there is high multicollinearity among our variables, with Diet.low.in.fruits and Diet.low.in.vegetables being the worst offenders. This amount of multicollinearity was already anticipated before starting the project. Any BMI results from an individual's diet, and the diet factors are elements of a diet. Naturally, any adjustment to diet habits will impact diet outcomes. This study aims to determine which elements have the most significant and least dependent effect on an individual's BMI. The correlation between variables is shown below in Figure 1 to illustrate the impact of multicollinearity on our model (see note about residuals in Appendix).

The correlation matrix and plots above show high levels of correlation between almost all variables. This problem will be remedied by reducing the number of variables in the model using partial F-tests, SSR tests, AIC tests, and SBC tests.

To explore further, I generated a variance analysis table to observe each feature variable's significance.

According to Table 3, Alcohol remains suspect, and lowgrains is no longer insignificant. lownutsseeds has become statistically insignificant now, and partial F-tests will be performed to determine if they are, in fact, insignificant to the model.

Tables 4 and 5 show the partial F-test results performed on two models without Alcohol and lownutsseeds. Table 4 shows that I fail to reject that the coefficient of Alcohol in the model at a 95% confidence level should be zero. Table 5, however, shows that there is sufficient evidence at a 95% level of confidence that lownutsseeds is statistically significant to the model lownutsseeds will be kept for now.

Another test that may determine which variables are insignificant to the model's power is the sequential sum of squares test. Table 6 displays two models: the full model with all variables on the left and a reduced model without Alcohol on the right.

This test compares the marginal change in the regression sum of squares when a feature variable is added to the model. The reduced model's SSR subtracted from the full model's SSR will show the marginal increase in SSR when Alcohol is added to the model. The difference is only 0.207, so it can be assumed that Alcohol adds very little to the model and can be discarded. The same was done for lownutsseeds, and the marginal difference when adding lownutsseeds to the full model was only 0.119. Lownutsseeds was also removed from the model as a result. The mathematics and illustrations for the remainder of this test can be seen in the Appendix.

The current model is HighBMI = highsodium + lowgrains + lowfruits + Unsafewater + lowvegetables. The F-statistic jumped from 429.7 to 590.5, with a very slight improvement in the adjusted R-squared value. However, multicollinearity is still a problem, as shown in Table 7.

Unsafewater will be kept, but there are still significant problems with the remaining variables. In an attempt to transform these variables, new variables of highsodium, lowgrains, lowfruits, and lowvegetables squared were added to see if this transformation could help the variables fit into the model. The results of the new model are shown in Table 8.

Observing the results from the summary(Table 8), ANOVA(Table 9), and VIF tables(Table 10) of the new model, we were not surprised with the performance of both lowfruits and lowvegetables. They still had large VIF values and low SSR values. Lowvegetables was the worst performer, with VIF of lowvegetables = 51, and the SSR of both lowvegetables and lowvegetables ^2 are less than 0.5. The decision was made to remove lowfruits and lowvegetables from the model due to their poor results.

The current model has been further reduced to HighBMI = highsodium + highsodium_2 + lowgrains + lowgrains_2 + Unsafewater. Though this model performed better than any model devised, tests were still conducted to see whether there was multicollinearity. The variable lowgrains was put to the test, and lowgrains may need to be removed from the final model. The VIF of this model is shown in Table 11.

Multicollinearity is still an issue, and the variable lowgrains is the most suspicious. This time, multiple tests were applied to determine the validity of the reduced model, as it is suspected that the model that comes out of these tests will be our complete and final model.

For the final test, the VIF of the full model(HighBMI = highsodium + highsodium_2 + lowgrains + lowgrains_2 + Unsafewater) and the reduced model (HighBMI = highsodium+highsodium_2+Unsafewater) are displayed, along with the AIC, SBC, and prediction sum of squares criteria. The results are displayed in Table 12 and Table 13.

With the AIC, SBC, and PRESS criteria, the models with the lowest values are those with the best fit. The AIC moderately favors the full model, whereas the SBC and PRESS criterion slightly favors the reduced model. However, a significant result is the VIF scores of the full and reduced models. The reduced model shows greater independence among the variables. The two models' summaries are shown in Table 14.

With all of the methods in consideration, it is evident that the reduced model, HighBMI = highsodium+highsodium_2+Unsafewater, is the superior model. This model will be the model used to answer the questions initially proposed. Thus, the regression model derived from our analysis is as follows:

$$Y = (0.52)\text{ highsodium} + (0.10717)\text{ highsodium}\,\hat{}2 – (0.50111)\text{ Unsafewater}$$

IV. Conclusions

At the beginning of our study, we wanted to know if it is possible to determine which diet factors could be targeted to fight against death related to high BMI. There are, in fact, two questions: (1) can we determine which factors are most significantly related to deaths caused by high BMI, and (2) what is the relationship between these factors and high BMI deaths? Our model attempts to solve both questions with the refined regression model we constructed.

To answer the first question, two of the seven variables we started with survived until the end. The two surviving features are deaths related to a diet high in sodium and deaths related to unsafe water sources. The multicollinear effect of all the variables together made it difficult to know which features to target; the final model removes much of the noise and presents a clear guideline for what is most related to deaths caused by high BMI.

The second question is just as important as the first. The relationship between deaths influenced by a high sodium diet and deaths influenced by high BMI is positive. For every death increase in the death count of high sodium-related deaths, the number of deaths attributed to high BMI increases by 0.633. The relationship between deaths influenced by unsafe water sources and deaths related to high BMI is negative. For every death increase in the death count caused by unsafe water sources, the number of deaths attributed to a high BMI decreases by 0.501.

The implications of these results imply that deaths caused by diets high in sodium are contributing most to deaths related to high BMI. The consumption of highly processed foods may cause this relationship, and their easy accessibility may be the culprit. An important note is that deaths attributed to diets high in sodium contribute most to high BMI deaths among the features sampled. If diets high in fat or high in sugar were also in the original data set, it is presumed that they also would play a significant role in high BMI-related deaths. However, the original data set did not include these, so they were not included in the model. In spite of this, it is essential to note the relation that high sodium diets have on overall health, and restricting their intake is an excellent method of combatting obesity-related deaths.

One final comment is the negative relationship between unsafe water sources and deaths related to high BMI. Although it would be highly unethical and counterproductive to give a population unsafe water to reduce BMI intentionally, the relationship was still there and is still interesting. The purpose of the project was to combat high BMI. However, in parts of the world where low BMI is a life-endangering reality, water purification seems to be an essential factor when considering not only high BMI deaths but deaths related to unhealthy BMI levels in general, high or low.

V. Appendix

Table 1. Regression results of the original model

term	estimate	std.error	statistic	p.value
(Intercept)	1.15E-16	0.017272386	6.65E-15	1
Diet.high.in.sodium	0.711346345	0.066568218	10.68597545	6.19E-22
Diet.low.in.whole.grains	-0.031993988	0.057420897	-0.557183699	0.577939071

Alcohol. use	-0.096456772	0.056798212	-1.698236064	0.090803936
Diet.low.in.fruits	-0.389692816	0.09385774	-4.151951833	4.63E-05
Unsafe.water.source	-0.480205399	0.027484267	-17.47201035	3.42E-44
Diet.low.in.nuts.and.seeds	-0.054825682	0.025991063	-2.109405171	0.035981429
Diet.low.in.vegetables	0.358278343	0.106813351	3.354246827	0.000929436

Table 2. VIF of each variable

	VIF
Diet.high.in.sodium	14.35839567
Diet.low.in.whole.grains	10.68346175
Alcohol.use	10.45301049
Diet.low.in.fruits	28.54382945
Unsafe.water.source	2.4475996
Diet.low.in.nuts.and.seeds	2.188870794
Diet.low.in.vegetables	36.96775961

Table 3. ANOVA table of the original model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
highsodium	1	177.033053	177.033053	2472.512106	1.03E-125
lowgrains	1	0.63427297	0.63427297	8.858501682	0.003226214
Alcohol	1	0.215679502	0.215679502	3.012263358	0.08396461
lowfruits	1	3.188650508	3.188650508	44.53392657	1.82E-10
Unsafewater	1	33.42106307	33.42106307	466.7714963	1.82E-57
lownutsseeds	1	0.090394673	0.090394673	1.26248698	0.262342173
lowvegetables	1	0.805574977	0.805574977	11.25097178	0.000929436
Residuals	232	16.61131131	0.07160048		

Table 4. F-test for the model without Alcohol

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	232	16.611				

2	233	16.817	-1	-0.20649	2.884	0.0908

Table 5. F-test for the model without lownutsseeds

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	232	16.611				
2	233	16.92	-1	-0.31859	4.449	0.0359

Table 6. Full model & model without Alcohol

Full model	Df	Sum Sq	Reduced model without Alcohol	Df	Sum Sq
highsodium	1	177.033	highsodium	1	177.033
lowgrains	1	0.634	lowgrains	1	0.634
Alcohol	1	0.215			
lowfruits	1	3.188	lowfruits	1	3.404
Unsafewater	1	33.421	Unsafewater	1	33.323
lownutsseeds	1	0.090	lownutsseeds	1	0.048
lowvegetables	1	0.805	lowvegetables	1	0.738
Residuals	232	16.611	Residuals	233	16.817

Table 7. VIF of model delete Alcohol

	VIF
highsodium	6.909049541
lowgrains	9.68586244
lowfruits	24.98329131
Unsafewater	1.775564327
lowvegetables	33.39587843

Table 8. Summary of the new model

term	estimate	std.error	statistic	p.value
(Intercept)	-0.119671986	0.023558363	-5.079809144	7.81E-07
highsodium	0.64654652	0.059744921	10.82178213	2.50E-22
highsodium_2	0.052454672	0.02924269	1.793770394	0.074163353
lowgrains	-0.114908615	0.067339389	-1.706410125	0.08928151
lowgrains_2	0.178716629	0.040317684	4.432710645	1.44E-05
lowfruits	0.177044115	0.098978156	1.788719073	0.074976093
lowfruits_2	-0.023606511	0.051042532	-0.462487063	0.6441693
Unsafewater	-0.514027432	0.021416158	-24.00185106	1.44E-64
lowvegetables	-0.236974859	0.111950469	-2.116783077	0.03535263
lowvegetables_2	-0.083766184	0.054941448	-1.524644626	0.128721592

Table 9. ANOVA table of the new model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
highsodium	1	177.033053	177.033053	3105.360349	1.49E-135
highsodium_2	1	5.439081395	5.439081395	95.40765081	4.49E-19
lowgrains	1	0.335917	0.335917	5.892364819	0.015975727
lowgrains_2	1	0.452709006	0.452709006	7.941028938	0.005253292
lowfruits	1	1.365703599	1.365703599	23.95598863	1.85E-06
lowfruits_2	1	0.088568107	0.088568107	1.553584959	0.213874948
Unsafewater	1	33.85903044	33.85903044	593.9257602	1.15E-65
lowvegetables	1	0.181380252	0.181380252	3.181615142	0.075789534
lowvegetables_2	1	0.132519446	0.132519446	2.324541235	0.128721592
Residuals	230	13.11203777	0.05700886		

Table 10. VIF of the new model

	VIF
highsodium	14.52605266
highsodium_2	5.201569649
lowgrains	18.45372196
lowgrains_2	9.863510748
lowfruits	39.86799039
lowfruits_2	13.72837089
Unsafewater	1.866503249
lowvegetables	51.00320496
lowvegetables_2	18.38416215

Table 11. VIF of the further reduced model

	VIF
highsodium	13.49361364
highsodium_2	4.584315085
lowgrains	15.89998212
lowgrains_2	5.86320275
Unsafewater	1.763341092

Table 12. Comparison of two models in AIC, SBC, and PRESS

Method	Full Model	Reduced Model	Full - Reduced
AIC	-667.5303	-662.4801	-5.050214
SBC	-646.6465	-648.5575	1.911064
PRESS	15.70204	15.58144	0.1206016

Table 13. Comparison of two models in VIF

Full Model	VIF	Reduced Model	VIF
highsodium	13.49361364	highsodium	1.700178
highsodium_2	4.584315085	highsodium_2	1.063176
lowgrains	15.89998212		
lowgrains_2	5.86320275		
Unsafewater	1.763341092	Unsafewater	1.672117

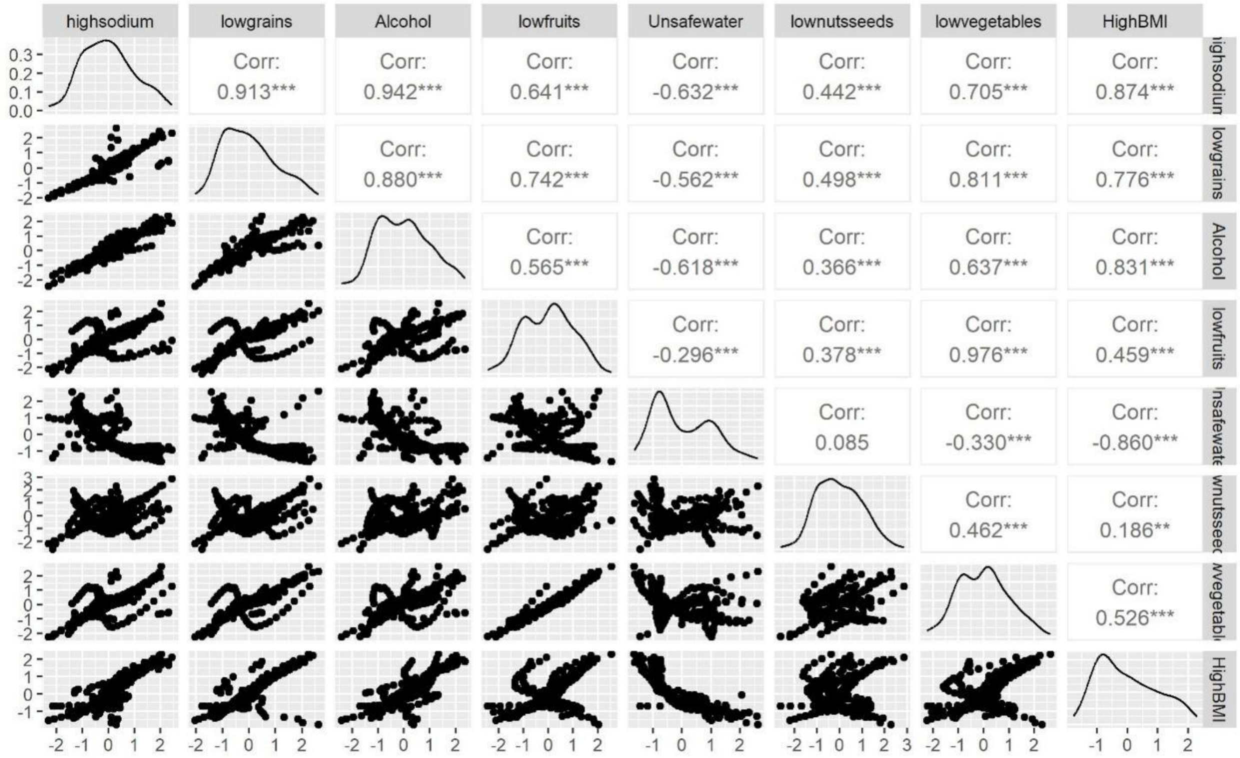
Table 14. Comparison of two models' summary

	Residual standard error	DF	Multiple R-squared	Adjusted R-squared	F-statistic	p-value
Full Model	0.2459	234	0.939	0.9377	720.9	< 2.2e-16
Reduced Model	0.2495	236	0.9367	0.9359	1164	< 2.2e-16

Table 15. New Variable Names for Reference

New Variable Names for Reference	
X1	Diet High in Sodium
X2	Diet Low in Whole Grains
X3	Alcohol Use
X4	Diet Low in Fruits
X5	Unsafe Water Source
X6	Diet Low in Nuts and Seeds
X7	Diet Low in Vegetables
Y	High Body Mass Index

Figure 1. The figure of variables' correlation



Notes on Sequential Sum of Squares Tests For Reduction Test:

R^2Y3|124567 = SSR (X3|X124567) / SSE (X124567) = 0.207 / 16.818 = 0.01230824

SSR (X3|X124567) =SSR (X3, X1, X2, X4, X5, X6, X7) – SSR (X1, X2, X4, X5, X6, X7) = 0.096

SSR (X3, X1, X2, X4, X5, X6, X7) = 177.033 + 0.634 + 0.216 + 3.189 + 33.421 + 0.090 + 0.806 = 215.389

SSR (X1, X2, X4, X5, X6, X7) = 177.033 + 0.634 + 3.404 + 33.323 + 0.049 + 0.739 = 215.182

SSR (X3|X124567) = SSR (X3, X1, X2, X4, X5, X6, X7) – SSR (X1, X2, X4, X5, X6, X7)

SSR (X3|X124567) = 0.207

We think X3 could be removed from the model

For X6 Reduction Test

R^2Y6|123457 = SSR (X6|X123457) / SSE (X123457) = 0.119 / 13.624 = 0.008734586

SSR(X6|X123457) = SSR (X3, X1, X2, X4, X5, X6, X7) – SSR (X1, X2, X3, X4, X5, X7) = 0.119

SSR (X3, X1, X2, X4, X5, X6, X7) = 176.924 + 0.580 + 0.235 + 3.081 + 36.915 + 0.008 + 0.752 = 218.495

SSR (X1, X2, X3, X4, X5, X7) = 176.924 + 0.58 + 0.235 + 3.081 + 36.915 + 0.641 = 218.376

SSR (X6|X123457) = SSR (X3, X1, X2, X4, X5, X6, X7) – SSR (X1, X2, X3, X4, X5, X7)

SSR (X6|X123457) = 0.119

We think X6 could be removed from the model