Factors affecting Weight and Diabetes

Section 2
Shengyuan Wang

Introduction

Introduction to Topic

Instructions: Introduce and motivate the general topic that you will be investigating through your data analysis. This can be brief, but it should hook the reader, provide relevant background, and motivate the Research Questions section that follows.

With increasing living standards and elevating life quality, people gradually have more food options and life-styles to choose from. Meanwhile, too much consumption of junk food and fast pace of life lead to higher probability of being overweight or getting diabetes. With the help of data science and modern technology, datasets shoulder the mission to inform correlation between variables and to point out factors impacting our weight and factors leading to diabetes. Scientists keep exploring methods to help people to predict diabetes-related diseases to better assist people staying healthy.

Research Questions

Instructions: In paragraph form, introduce and motivate the two research questions that you will be investigating (one involving a quantitative outcome, the other involving a related binary outcome). Explain why you think these are interesting questions and how they are related to one another (e.g., what is the overall goal?). For each research question, provide justification for all variables (including at least two explanatory variables per question) that you will be considering in your analysis.

The overall goal of the two research questions is to find out factors affecting body weight and the complications of diabetes and to provide warnings for the potential risks of getting diabetes-related diseases.

The first research question is whether combined systolic blood pressure and physical activity have a relationship with the weight of US individuals aged from 15 to 70? I choose to focus on people in this age range because individuals outside this age range are more likely to have physiological differences. Combined systolic blood pressure is always a necessary medical measurement in medical examinations, so I wonder if combined systolic blood pressure relates to weight. I introduce physical activity into this research question to find out whether having moderate or vigorous-intensity fitness will affect body weight or not. Since it is common sense that as people grow taller, their body weight will increase correspondingly, I plan to include the variable height into the research question. And another point I plan to point out is whether there exists a difference among weight patterns of different races.

The second research question is that if combined systolic blood pressure and BMI are associated with whether people have diabetes or not? Since diabetes has been tagged as a high-rate disease recently, it is urgent to figure out what factors influence diabetes. To make an expansion from the first research

question, I introduce BMI, which is calculated with both height and weight, to test whether obesity accompanies a higher rate of getting diabetes. Although flaw measurements may exist in BMI [1] (Brock, 2019, p28), we have to admit its ability to rate obesity class. In hospitals, doctors always test the combined systolic blood pressure for diabetes patients, so I wonder whether combined systolic blood pressure has a relationship with diabetes. And I also include physical activity and race in the research question as precision variables.

Data

Context

Instructions: Introduce the data you will be using in your analysis. Make sure to describe all relevant details of the data context, including:

- *Who (including total number of cases in dataset, what each case represents)*
- *What (including total number of variables in dataset, general summary of what the variables in the dataset represent, detailed description of any variables included in your final visualizations and models and their values [e.g., provide the categories for categorical variables and the units and range of values for quantitative variables]) --- note: when describing variables, use descriptive names rather than the R variable name*
- *Where*
- *When*
- *Why*
- *By whom*
- *How (including study design and sampling methods)*

Also provide a link and/or description of how readers can access the data. This section should just describe *the data; save your discussion of the implications/limitations of this context for the Limitations section at the end of your report.*

I will introduce the NHANES dataset (accessible from NHANES R-package) into the research project. The dataset is survey data and examination results collected by the US National Center for Health Statics (NCHS) and the Center for Disease Control (CDC) in the US. In this dataset, scientists used an observational study for the data collected by questionnaires administered at home and followed by a standardized health examination in specially equipped mobile examination centers. The dataset is representative and convincing for measuring the overall public health and finding valuable relationships between different variables because the sample design is a cluster design and incorporates differential probabilities of selection. In the dataset, the observation units are 10,000 people of all ages in different places around the US through questionnaires and standardized health examinations in mobile examinations centers (MECs) between 2009 and 2012. It incorporates 76 variables in the dataset, including demographic variables, physical measurements, health variables, lifestyles variables, and weighting variables.

In the variables table below, I introduce variables that will be used in later research questions about their description names, types of variables, units, and range or categories.

Variable Name	Description Name	Type	Unit	Range / Categories
Weight	Weight	Numerical	kg	[2.8, 230.0]
Height	Standing height	Numerical	cm	[83.6, 200.4]
PhysActive	Participants does moderate or vigorous-intensity sports, fitness or not	Categorical	—	[Yes, No]
BPSysAve	Combined systolic blood pressure	Numerical	mmHg	[76, 226]
Race3	Race of participants(including non-Hispanic Asian Category)	Categorical	—	[Mexican, Hispanic, White, Black, Asian, Other]
SleepHrsNight	Sleep hours per night	Numerical	hours	[2, 12]
Gender	Gender	Categorical	—	[male, female]
BMI	Body mass index	Numerical	kg/m^2	[12.88, 81.25]
BMI_WHO	Body mass index category	Categorical	—	[12.0_18.4, 18.5_24.9, 25.0_29.9, 30.0_plus]
Diabetes	Diabetes	Categorical	—	[Yes, No]
Age	Age (subjects 80 years or older were recorded as 80)	Numerical	year	[0, 80]

Link to dataset: https://rdrr.io/cran/NHANES/man/NHANES.html (*NHANES: NHANES 2009-2012 With Adjusted Weighting in NHANES: Data From the US National Health and Nutrition Examination Study*, n.d.)

Cleaning

Instructions: Describe any changes you made to your dataset. If you made any changes to your variables (e.g., mathematical transformations like log-transforming, creating a categorical variable out of a quantitative one, combining categories of a categorical variable) describe and justify those changes here. If you removed any cases from your analysis (e.g., individuals with missing values, outliers), describe and justify that filtering here, and make sure to mention how many cases are left after filtering.

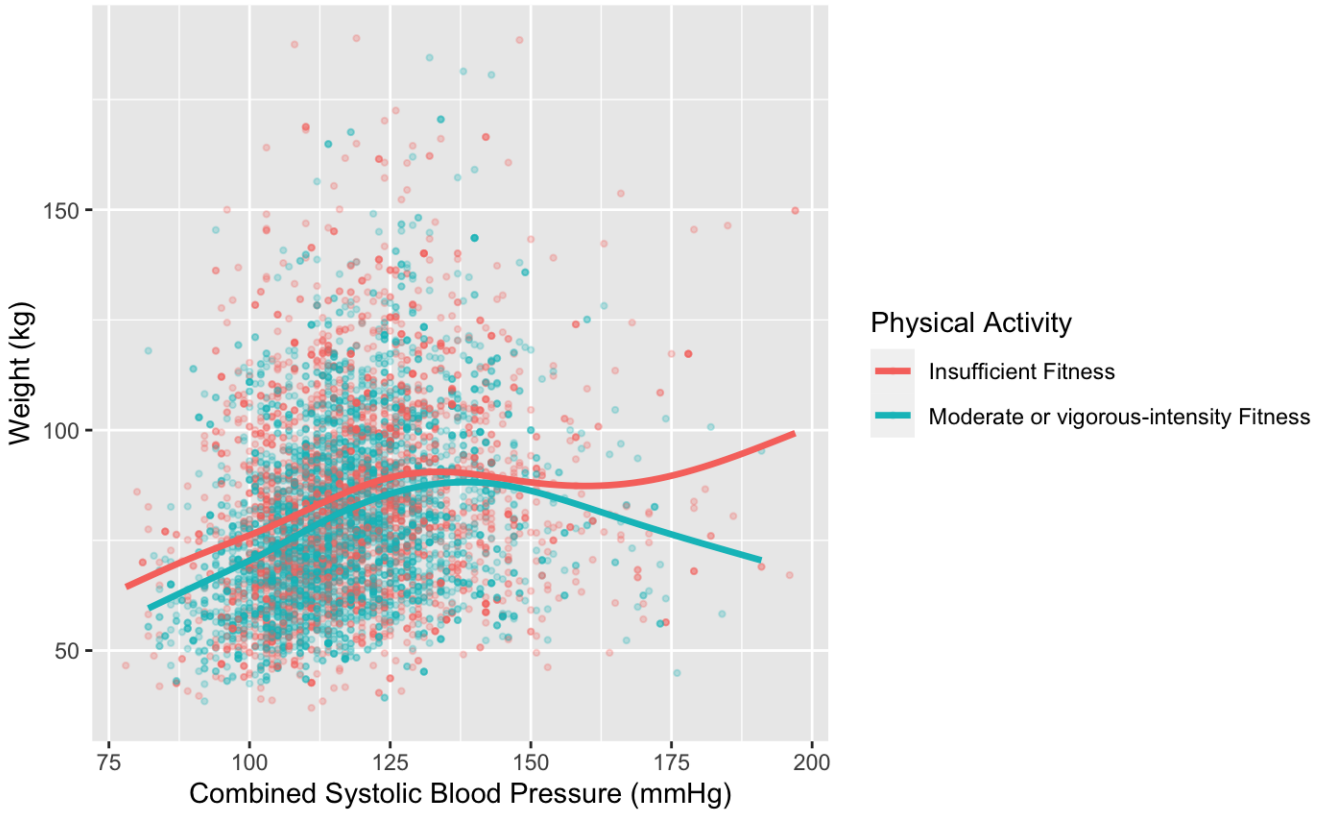
I first removed cases whose age is lower than 15 or higher than 70 because the objects we investigate are US individuals aging from 15 to 70. For individual cases not in this age range, there might be physiological differences among them. After this modification, 7101 cases are left in the dataset. Then I removed cases with no physical activity, diabetes, or BMI range information. Since we have to promise that all the data variables I plan to analyze must be included in each case of the dataset, I exclude the cases without enough information on these three variables. Because it shows some outliers with weight

higher than 190kg and combined systolic blood pressure higher than 200mmHg, I decide to remove cases whose weight is higher than 190kg or combined systolic pressure is higher than 200mmHg. After this modification process, 6735 cases are left in the dataset.

First Research Question: Whether systolic blood pressure and physical activity have a relationship with the weight of the US individuals aged from 15 to 70?

Exploratory Data Analysis

Instructions: Present a visualization that helps address your first research question (using 2-3 variables of interest) and thoroughly describe what information you gain from the visualization. You may also want to use numerical summaries in your paragraph to fully describe your visualization. Note: you do not need to (and should not) include all variables that are involved in your final linear regression model in this visualization; just focus on the primary variables of interest. If you feel that two visualizations would be more effective, that is ok too.



The visualization above illustrates relationships among three variables, combined systolic blood pressure, weight, and physical activity. There is a big cluster, the data-intensive area, in the visualization. We can observe that weight mostly ranges from 50kg to 120 kg, and most average

systolic blood pressure data ranges from 95 mmHg to 140 mmHg. In the data-intensive area, it shows a weak, positive, nonlinear association between the average systolic blood pressure and weight. That is to say, with the same fitness level, people having higher combined systolic blood pressure tend to weigh more. Plus, the plot shows that holding combined systolic blood pressure constant, people having moderate or vigorous-intensity fitness weigh less than those with insufficient fitness.

Model Creation

Instructions: Describe the final regression model that you chose and explain how you chose it. Present your final model statement in appropriate notation, using short descriptive variable names (not X, Y, or variable names from R) to represent the variables so that someone who has never seen your dataset can understand. In justifying your choice of model, explain why you included these variables, why you fit this type of regression, what other models you considered and how you ruled them out, etc..

$$E[Weight \mid Height,\ BPSysAve,\ PhysActive,\ Race3\ ] = \beta_0 + \beta_1 * Height + \beta_2 * BPSysAve + \beta_3 * PhysActiveYes + \beta_4 * Race3Black + \beta_5 * Race3Hispanic + \beta_6 * Race3Mexican + \beta_7 * Race3White + \beta_8 * Race3Other$$

It is acknowledged that there exists a positive relationship between a person's height and weight, thus height is necessary in this model. Combined systolic blood pressure is always a necessary medical measurement in medical examinations, so I introduce the combined systolic blood pressure in the model. I include combined systolic blood pressure and physical activity into the model because these two variables are what my research question is about. Since the percentage of muscular and body fat are not the same for each race or each gender, I wonder if race and gender have some effects on weight *[2](Bell and Blackman Carr, 2020, p973)*. So I introduce them as precision variables. I decided not to add any interaction terms between them because I assume there is the same relationship between physical activity and weight, regardless of one's race. As I do not think physical activity can have a different effect on weight depending on different races, I prefer not to add interaction terms in this model.

From the p-value table in the appendix, we see that the p-value of height, combined systolic blood pressure, physical activity and race ($p < 0.0001$) are lower than 0.05 threshold. Since p-values here are the probability of obtaining results at least as extreme as the observing result, assuming there truly were no relationship between weight and these variables, we have enough evidence to reject the null hypothesis and keep height, physical activity, combined systolic blood pressure and race in the model and exclude gender from the model. Plus, if we exclude race, combined blood pressure, physical activity and height separately from the model, we see the adjusted R-squared will decrease. It means that the larger models including race, combined blood pressure, physical activity and height are better for predictions. Therefore, our final model uses height, physical activity, combined systolic blood pressure, and race to predict weight.

Fitted Model

Instructions: Present the fitted model (estimates, confidence intervals, p-values) in a table format. All numerical values should be rounded to a reasonable number of digits. Use the same shortened descriptive variable names (not R variable names) to represent the variables as you used in your model statement above.

Model Coefficient	Estimate	95% Confidence Interval (LB, UB)	P-Value
Intercept	-106.17	(-117.74, -94.59)	< 0.0001
Height	0.89	(0.83, 0.95)	< 0.0001
Combined systolic blood pressure	0.27	(0.23, 0.31)	< 0.0001
PhysActiveYes	-4.64	(-5.91, -3.37)	< 0.0001
Race3Black	11.62	(8.47, 14.78)	< 0.0001
Race3Hispanic	8.75	(5.23, 12.27)	< 0.0001
Race3Mexican	9.77	(6.47, 13.08)	< 0.0001
Race3White	7.90	(5.21, 10.59)	< 0.0001
Race3Other	10.88	(6.36, 15.40)	< 0.0001

Model Interpretation

Instructions: Describe what you learn from the model about your research question. Use the estimates, 95% confidence intervals, and p-values for the coefficient(s) of interest to support your description. Note: you do not need to (and should not) interpret all coefficients in this model; just focus on the coefficient(s) that relate most directly to your research question. Make sure to provide a takeaway message describing what you learn from this model with respect to answering your research question.

We find that holding height, race, and combined systolic blood pressure constant, it is estimated that people with moderate or vigorous-intensity fitness weigh 4.64kg less than individuals without sufficient sports or fitness on average. The confidence interval for physical activity indicates that we are 95% confident that holding combined systolic blood pressure, race, and height constant, the true difference in weight between individuals with moderate or vigorous-intensity sport and individuals without enough fitness is a decrease between 3.37kg and 5.91kg on average, with physically active individuals having lower weight. In this part, 95% confident means the expectation that 95% of samples will generate confidence intervals that contain the true population value of the difference in weight between people with and without moderate or vigorous-intensity fitness. Since the interval does not contain 0, we have enough evidence that taking combined systolic blood pressure, race and height into account, it suggests a true negative relationship between physical activity and weight(since values are under 0)

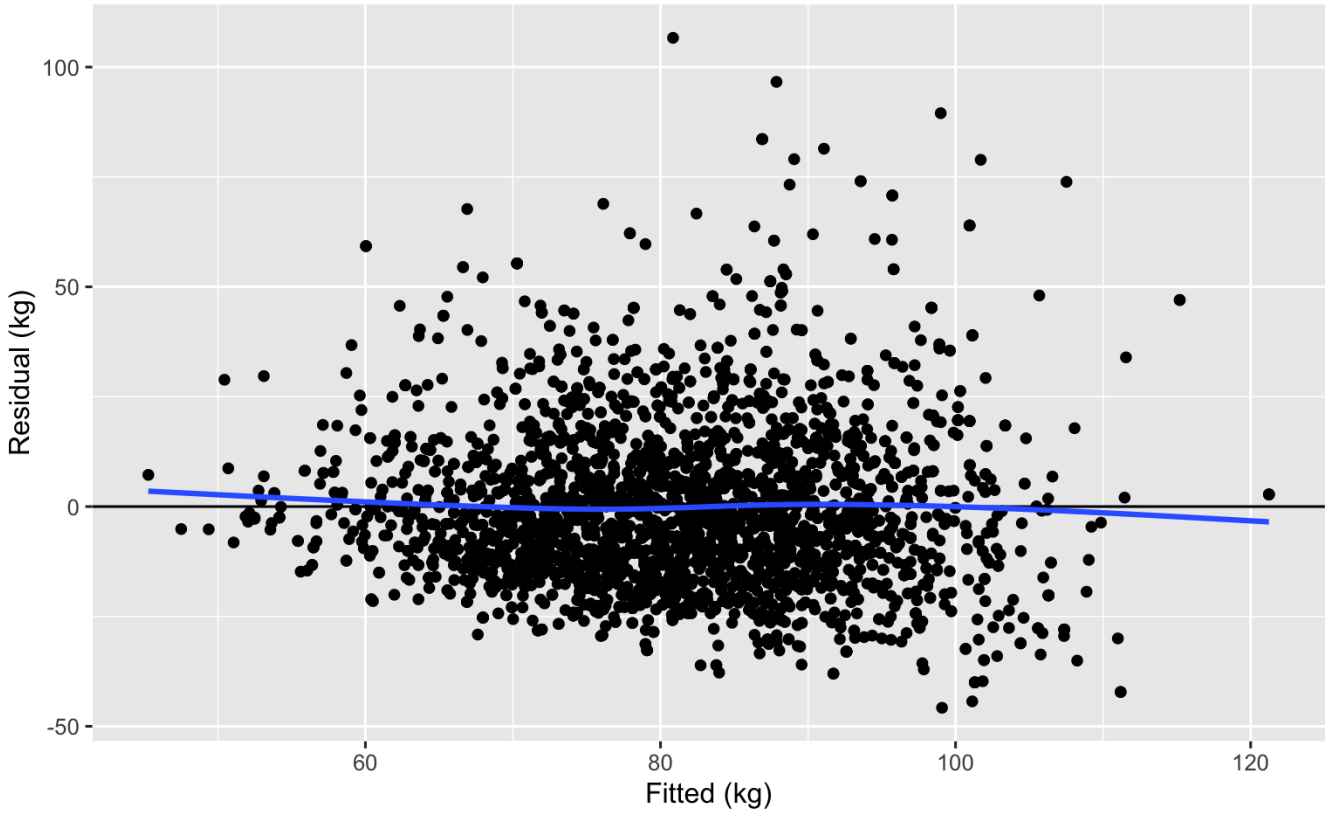
that people with moderate or vigorous-intensity fitness (PhysActiveYes) weigh less than people with insufficient fitness (PhysActiveNo). Plus, the p-value($p < 0.0001$) is incredibly small, meaning that the probability we would see such a difference in weight between people with and without moderate or vigorous-intensity fitness after adjusting for combined systolic blood pressure, height and race is quite minute if physical activity truly does not have a relationship with weight. It shows that we have enough evidence to reject the null hypothesis and conclude that there is an association between physical activity and weight, after accounting race, height and combined systolic blood pressure in the model.

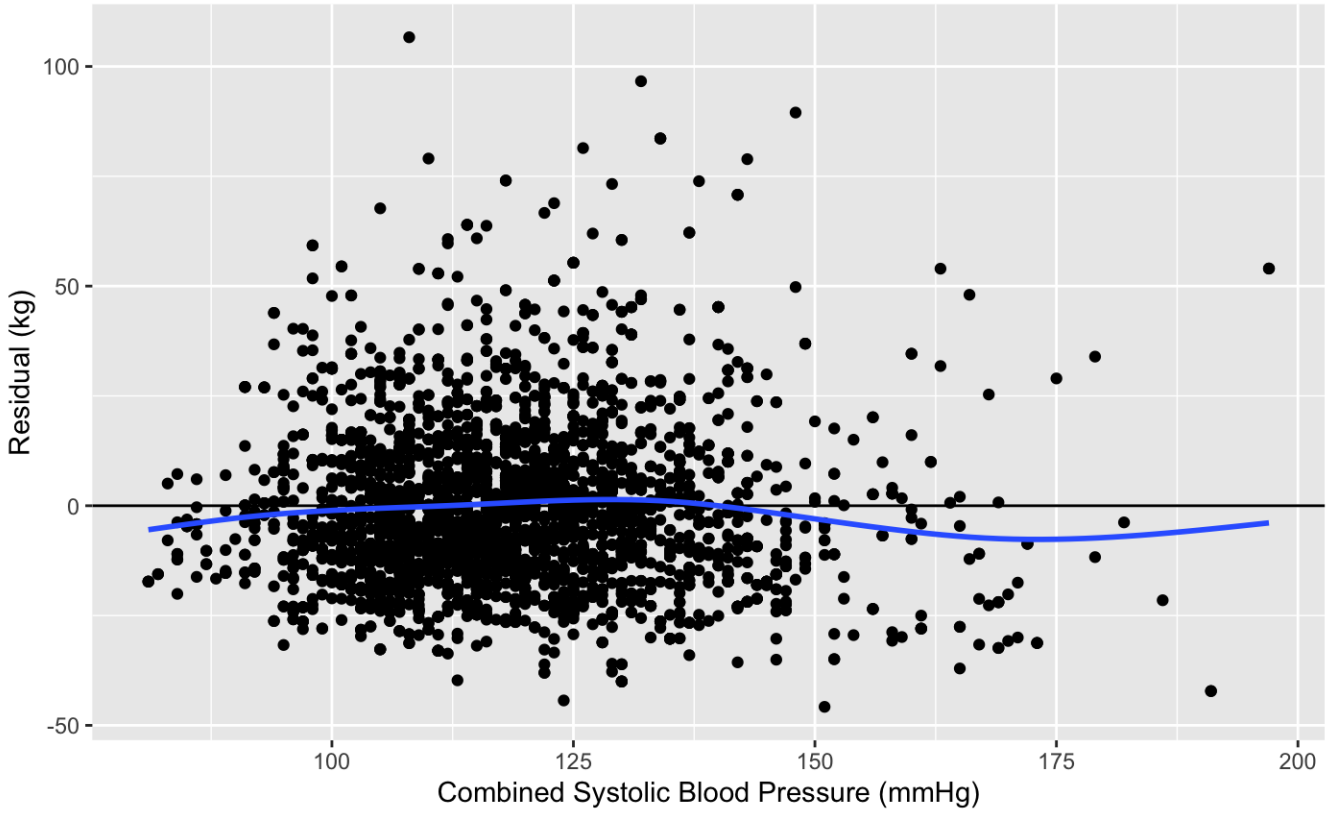
Holding physical activity, height and race constant, it is estimated that the weight increases about 0.27kg for a 1mmHg increase in combined systolic pressure on average. And we are 95% confident that holding physical activity, height and race constant, the true difference in weight associated with the 1mmHg increase in combined systolic blood pressure is an increase between 0.23kg and 0.31kg on average. Since 0 is not in the interval, we have enough evidence to conclude that there is a genuinely positive relationship (since values are above 0) between combined systolic blood pressure and weight after taking physical activity, race, and height into account. Moreover, the p-value($p < 0.0001$) is incredibly small, meaning that after adjusting for physical activity, height and race, the probability we would see such a difference in weight for people with higher combined systolic blood pressure is quite minute if combined systolic blood pressure truly does not have a relationship with weight. It shows that we have enough evidence to reject the null hypothesis and conclude that there is an association between combined systolic blood pressure and weight, after accounting for the other variables in the model.

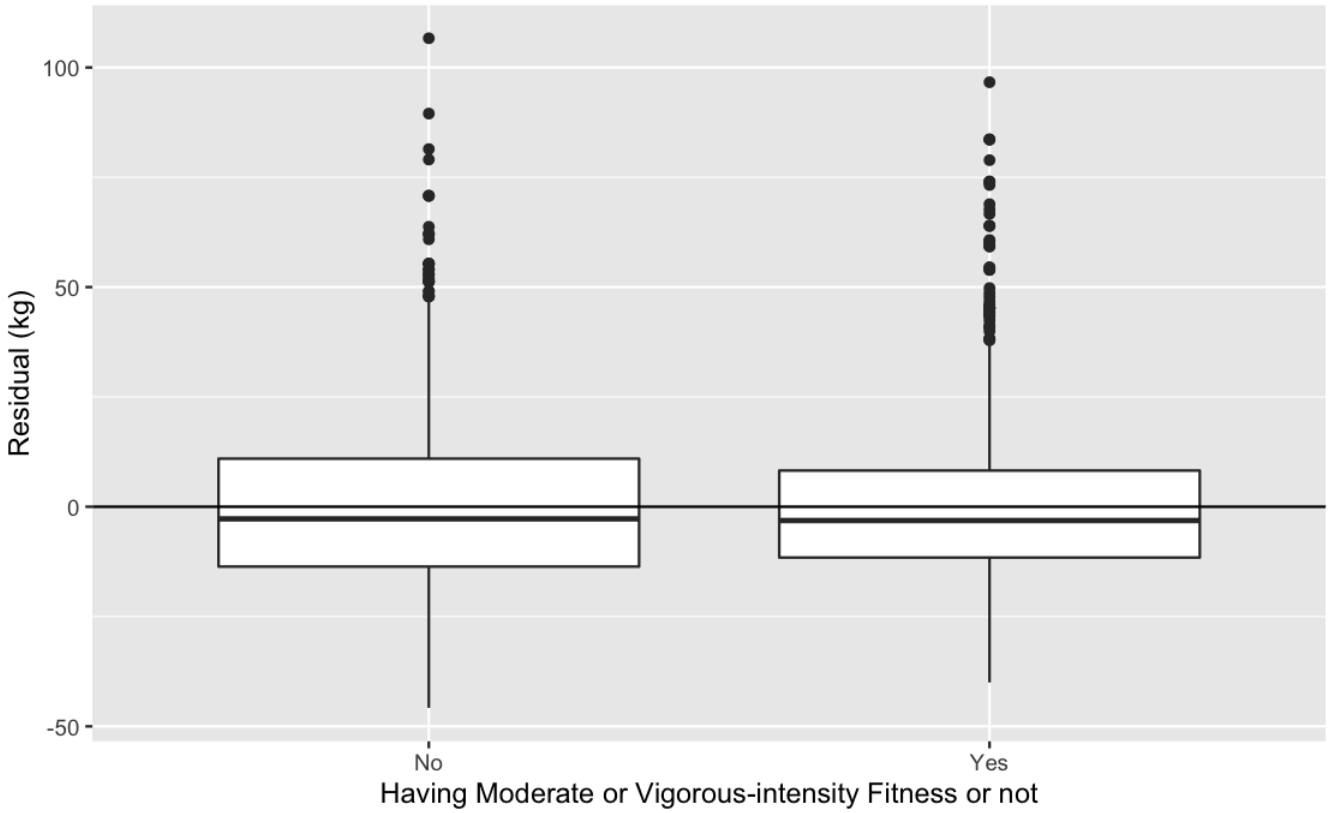
A similar positive relationship and significance (small p-value < 0.0001) appears for different races after accounting for physical activity, combined systolic blood pressure and height. However, the model predicts that Asian people are associated with lower weight in comparison with other races when keeping combined systolic blood pressure, physical activity and height constant(one can see the coefficient values increase in this order: Asian > Hispanic > Mexican > Others > Black, with the estimation that the Black weigh more in comparison to Asian when holding other variables in the model constant).

Model Evaluation

Instructions: Put your conclusions in context by providing a description of how well the conditions (straight enough, equal spread, no extreme outliers) of the final model are satisfied and how "good" (R-squared, residual standard error, unnecessary or missing variables) the final model is. Provide graphical and numerical evidence to support your response.


100
50
0
-50
Residual (kg)
60
80
100
120
Fitted (kg)


100
50
0
-50
Residual (kg)
100
125
150
175
200
Combined Systolic Blood Pressure (mmHg)

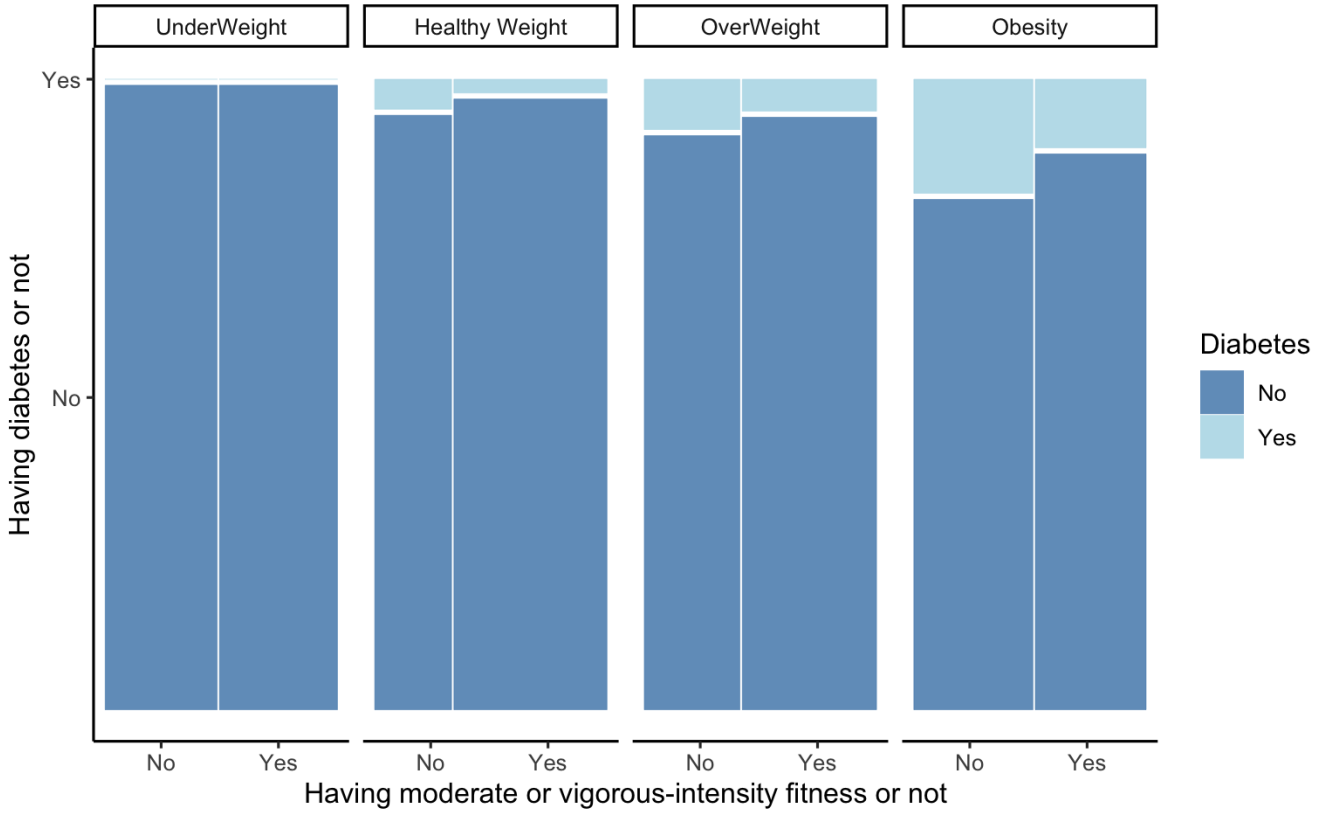


In the plot of residuals versus fitted values, it is straight enough and equal spread, thus not systematically overestimating or underestimating weight. For the residual plots for combined systolic blood pressure, the fitting line is not straight enough. For people with combined blood pressure lower than 100mmHg or higher than 140 mmHg, the model is more likely to underpredicting their weight because the fitting line is lower than residual = 0 line. When we look at the residual boxplot for physical activity, we find that residual = 0 line is slightly higher than the median for each box plot, but it is acceptable. Since I have already excluded the extreme outliers before, it is clear that there are no extreme outliers. The R-Squared indicates that 25.8% of total variations in weight can be explained by the model with height, combined systolic blood pressure and race. Also, the residual standard error is 18.24kg, which means that our model can predict the outcome within 36.48kg. In the model, race might be unnecessary as a redundancy variable although it slightly increases the adjusted R-squared. And age might be an additional variable that should be added as a confounding variable because it both affects the combined systolic blood pressure, physical activity, and weight.

Second Research Question: If BMI, and physical activeness are associated with whether people have diabetes or not?

Exploratory Data Analysis

Instructions: Present a visualization that helps address your second research question (using your 2-3 variables of interest) and thoroughly describe what information you gain from the visualization. You may also want to use numerical summaries in your paragraph to fully describe your visualization. Note: you do not need to (and should not) include all variables that are involved in your final linear regression model in this visualization; just focus on the primary variables of interest. If you feel that two visualizations would be more effective, that is ok too.



The visualization above includes three variables, BMI status, diabetes condition and physical activity. Within each BMI status, it shows that people with insufficient fitness have more possibility to get diabetes than people with moderate or vigorous-intensity fitness (a larger proportion of people with insufficient fitness have diabetes). For both people with and without moderate or vigorous-intensity fitness, people in higher BMI status (overweight or obesity) are more likely to get diabetes. In the BMI status of healthy weight and overweight, it is obvious that a larger proportion of individuals have moderate or vigorous-intensity fitness. Overall, it seems that individuals with higher BMI and insufficient fitness are more likely to have diabetes.

Model Creation

Instructions: Describe the final regression model that you chose and explain how you chose it. Present your final model statement in appropriate notation, using short descriptive variable names (not X, Y, or variable names from R) to represent the variables so that someone who has never seen your dataset can understand. In justifying your choice of model, explain why you included these variables, why you fit this type of regression, what other models you considered and how you ruled them out, etc..

$$log(Odds[Diabetes \mid BMI,\ BPSysAve,\ PhysActive]) = \beta_0 + \beta_1 * BMI + \beta_2 * BPSysAve$$

$$+ \beta_3 * PhysActiveYes$$

When I first created the logistic regression model, I included BMI, gender, combined systolic blood pressure, physical activity and race in the model to predict the diabetes condition. Since BMI and physical activity are two variables I plan to analyze in this research question, I introduce them in this model. I add combined systolic blood pressure and race into the model as precision variables to shrink confidence intervals and reduce standard errors. And I decided not to include interaction terms since I failed to see any interaction between BMI and physical activity. People with high BMI might be a bodybuilder or a person with little fitness. Also people with low BMI can be a long-distance runner or a person with little fitness. Thus, I hesitate to add any interaction terms in this model as I failed to clearly tell interactions between BMI and physical activity.

In this context, we may focus on maximizing the overall accuracy, and the model without gender and race can better make the positive result of the diabetes test accurately point out the subject who took the test has diabetes and a negative result of the diabetes test more certainly rule out the possibility of being diabetes. So I set the threshold as 0.075 here. When we exclude gender and race from the model, at the threshold of 0.075, both the specificity, sensitivity, and accuracy increases. In this context, specificity refers to the proportion of those who are estimated to have diabetes out of those who actually have diabetes; sensitivity refers to the proportion of those who are estimated not having diabetes out of those who actually have no diabetes; the accuracy refers to the proportions of those who are correctly diagnosed out of the whole subject population. Plus, the low p-value(< 0.0001) shows up in the nested hypothesis test of physical activity, BMI, and combined systolic blood pressure. The p-value here is the probability of obtaining results at least as extreme as the observed results, assuming the smaller models without the physical activity, BMI and combined systolic blood pressure are better. Thus, the low p-values indicate that we have enough evidence to reject the null hypothesis, and to choose the big model with physical activity, BMI, and combined systolic blood pressure for prediction.

Fitted Model

Instructions: Present the fitted model (exponentiated estimates & confidence intervals, p-values) in a table format. All numerical values should be rounded to a reasonable number of digits. Use the same shortened descriptive variable names (not R variable names) to represent the variables as you used in your model statement.

Model Coefficient	Estimate (exponentiated)	95% Confidence Interval (exponentiated)	P-Value
Intercept	0.0002	(0.00005, 0.00047)	< 0.0001
BMI	1.096	(1.078, 1.115)	< 0.0001
BPSysAve	1.032	(1.024,1.041)	< 0.0001
PhysActiveYes	0.537	(0.412, 0.698)	< 0.0001

Model Interpretation

Instructions: Describe what you learn from the model about your research question. Use the odds ratio interpretations, 95% confidence intervals, and p-value for the exponentiated coefficient(s) of interest to support your description. Note: you do not need to (and should not) interpret all coefficients in this model; just focus on the coefficient(s) that relate most directly to your research question. Make sure to provide a takeaway message describing what you learn from this model with respect to answering your research question.

Holding combined systolic blood pressure and physical activity constant, we estimate that 1kg/m^2 increase in BMI is associated with a multiplicative change in the odds of diabetes of 9.6% higher(1.096 times as high). Comparing individuals who differ in BMI by 1 kg/m^2, we are 95% confident that holding combined systolic blood pressure and physical activity constant, the true odds of diabetes is between 7.8% higher (1.078 times as high) and 11.5% higher (1.115 times as high) for those who are 1kg/m^2 higher in BMI. The phrase “95% confident” refers to our confidence in the interval construction process—the expectation that 95% of samples will generate confidence intervals that contain the true population value of the multiplicative increase in odds of diabetes associated with 1 kg/m^2 increase in BMI. Since 1 is not in this interval, after taking into account combined systolic blood pressure and physical activity, we can conclude that there is a genuinely positive relationship (since the values are above 1) between BMI and diabetes. In other words, people with higher BMI are more likely to get diabetes. Moreover, the p-value ($p<0.0001$) is incredibly small, meaning that the probability we would see such a difference in the odd of diabetes for one unit increase in BMI after adjusting for combined systolic blood pressure and physical activity is quite minute if BMI truly does not have a relationship with diabetes. Thus, we have enough evidence to reject the null hypothesis and conclude that there is a relationship between diabetes and BMI, holding other variables in the model constant. A similar positive relationship and significance ($p<0.0001$) also appears for combined systolic blood pressure after accounting for physical activity and BMI. It indicates that people with higher combined systolic blood pressure are more likely to get diabetes.

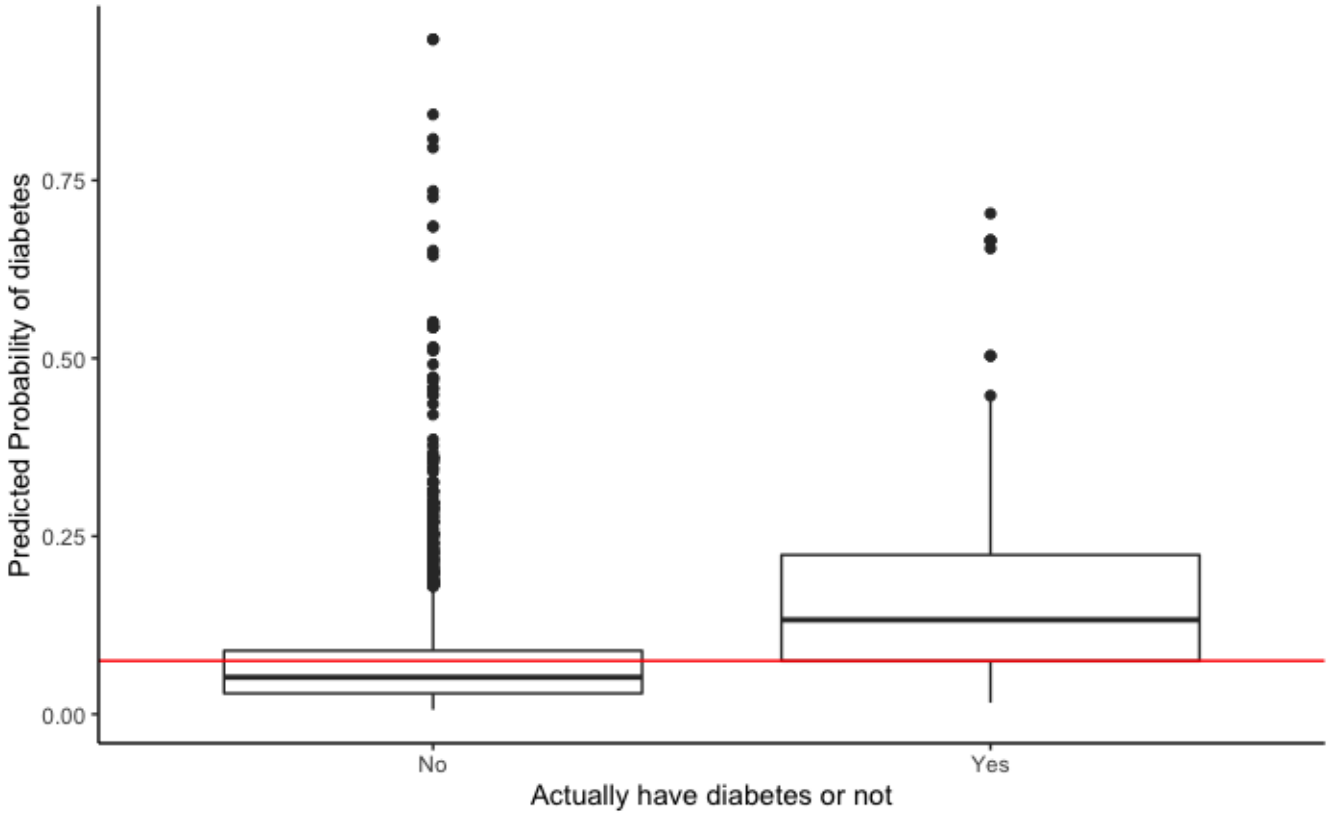
Plus, holding combined systolic blood pressure and BMI constant, we estimate that the odds of diabetes status for individuals who have moderate or vigorous-intensity fitness (PhysActiveYes) are 46.3% lower than (0.537 times as high) the odds of diabetes for those who have insufficient fitness. Comparing individuals who differ in physical activity, we are 95% confident that holding BMI and combined systolic blood pressure constant, the true odds of diabetes for individuals who have moderate or vigorous-intensity fitness (PhysActiveYes) is between 58.8% lower (0.412 as high) and

30.2% lower (0.698 as high) than the odds of diabetes for those who have insufficient fitness, with individuals with more fitness having lower odds of diabetes. Since 1 is not in the confidence interval, we have enough evidence to conclude that there is a truly negative relationship (since values are lower than 1) between physical activity and diabetes that people with moderate or vigorous-intensity fitness have less possibility to get diabetes than people with insufficient fitness. Plus, the p-value ($p < 0.0001$) is incredibly small, meaning that the probability we would see such a difference in the odd of diabetes between people with moderate or vigorous-intensity fitness (PhysActiveYes) and people without sufficient practice after adjusting for BMI and physical activity is quite minute if physical activity truly does not have a relationship with diabetes. Thus, we have enough evidence to reject the null hypothesis and conclude that there is a relationship between physical activity and diabetes, holding other variables in the model constant.

Overall, keeping other variables constant, increasing BMI or combined systolic blood pressure can separately increase the odds of diabetes, and by doing moderate or vigorous-intensity fitness, it can decrease the odds of diabetes.

Model Evaluation

Instructions: Put your conclusions in context by providing a description of how "good" (predicted probabilities, accuracy, specificity, sensitivity, false positive rate, false negative rate, unnecessary or missing variables) the final model is. Provide graphical and numerical evidence to support your response.



The boxplot shows a clear difference between the predicted probabilities of diabetes in the model for the two groups. The median of the predicted probability from the model is higher in the group of individuals who do actually have diabetes. It shows that people who actually have diabetes have a higher possibility of being diagnosed as having diabetes. False-positive rate and false-negative rate are expected to be low in the context because we want people with or without diabetes to be diagnosed correctly. Since these two cases are equally harmful, I systematically set the threshold to have similar sensitivity and specificity, and have higher accuracy. Thus, I set the threshold as 0.075. In the model, we find that the sensitivity, specificity, and total accuracy are about 70%. The sensitivity shows that we have a 74.6% probability of making correct predictions for those who actually have diabetes in this model. The specificity refers to a 68.3% probability of making correct predictions for those who actually have no diabetes in this model. The accuracy means that the proportion of those who are correctly diagnosed from the model out of all objects is 68.8% in this model. These are high numbers when we are predicting diabetes. Age might be an additional variable that should be added as a confounding variable, since aged people tend to have higher combined systolic blood pressure and less physical activity.

Conclusions

General Takeaways

Instructions: Make general conclusions that address your original research questions. This paragraph should describe takeaways from the two models in context. What have you learned about your first research question? What have you learned about your second research question?

Based on the linear regression model described above, holding physical activity, race and height constant, we estimate the difference in mean weight of US people aging from 15 to 70 for 1mmHg increase in combined systolic blood pressure is 0.27kg, with confidence interval from 0.23 to 0.31 kg. Plus, holding combined systolic blood pressure, race and height constant, the difference between physically active and inactive persons with the same combined systolic blood pressure, race and height is 4.64kg, with confidence interval from 3.37kg to 5.91kg(physically inactive ones weigh more). For the first research question, with small p-value of the combined systolic blood pressure and physical activity($p < 0.0001$), we can conclude that taller people and people with higher combined systolic blood pressure tend to weigh more, and people with moderate or vigorous-intensity fitness tend to weigh less.

Based on the logistic regression model described above, holding combined systolic blood pressure and physical activity constant, we estimate that 1kg/m^2 increase in BMI is associated with a multiplicative change in the odds of diabetes of 9.6% higher(1.096 times as high), with confidence interval from 7.8% higher (1.078 times as high) to 11.5% higher (1.115 times as high) for those who are 1kg/m^2 higher in BMI. Plus, holding combined systolic blood pressure and BMI constant, we estimate that the odds of diabetes status for individuals who have moderate or vigorous-intensity fitness (PhysActiveYes) are 46.3% lower than (0.537 times as high) the odds of diabetes for those who have insufficient fitness, with confidence interval from 58.8% lower (0.412 as high) to 30.2% lower (0.698 as high). For the second research question, with the small p-values($p < 0.0001$) of two variables, we can conclude that BMI and physical activity truly have a relationship with diabetes.

The modified dataset only includes the information of people aged from 15 to 70 in the US between 2009 and 2012. Facing the fact, I would hesitate to generalize the results to people of all ages considering the physiological differences for children or old people in their eighties and nineties. Also, since the dataset was collected and incorporated differential probabilities of selection to be representative for US citizens, the dataset is hard to fit in other countries with more considerations upon the different distribution of races or different age distribution. Moreover, the dataset was collected 10 years before, people's lifestyle and food preferences might have changed. Thus, we limit the analysis to 2009-2012 to best approximate the research results at the time period of data collection.

Limitations

There are some limitations to our analyses and dataset should be mentioned. Our NHANES dataset incorporated differential probabilities of selections for US people from 2009 to 2012, so the dataset is certainly representative for US people of all ages. However, to increase the health knowledge of the health status of older Americans, it over-sampled persons 60 or older. The older the individual, the more extensive the examination. So, the sample might be less and less representative of the population aged lower than 60 due to the sampling scheme of the dataset. This might make the research question investigating disease exist more deviation with the imbalanced probability selection. Also, non-response bias might arise from the part of the survey in the study. Since the invitation of the survey is sent through email, people who do not check the emails may become non-responders and be left off from the list of individuals to draw from. For those who take surveys, if they meet the questions such as whether you have moderate or vigorous-intensity fitness and physical practice, they are likely to choose "no" or "prefer not to answer" because different people have different standards and degrees towards "moderate" and "vigorous-intensity".

In the linear regression model, the categorical variable physical activity is not precise enough. That is to say, it fails to provide exact criteria for what is moderate or vigorous-intensity fitness and what is insufficient fitness. If this variable changes into fitness hours per week, it will be more effective and validate. Also in the logistic regression model, the choice of BMI as its variable might affect the validity of the result since using the same BMI range to define obesity for different races and different genders may have certain flaws *[1] (Brock, 2019, p28)*. To address limitations, we should add more accurate numerical variables such as body fat rate and muscle mass to better predict weight and diabetes status. Plus, some confounding variables should be added into the model. The variable age, for example, can be added as an interaction term or a confounding variable since aged people tend to have higher combined systolic blood pressure and less physical activity. Besides the variable limitation, since the dataset was collected about 10 years ago, we can not ensure different factors affecting the variables in the two research questions keep stable. Thus, the conclusions from these data would be hard to generalize with much has changed within 10 years. To address the problem, newly-updated dataset should be analyzed later to promise the validity of generalizing the conclusions.

Ethical Considerations

For participants, they received a more comprehensive health examination and well-analysis report about their body health. Through this data-collecting process, they will be aware of their potential diseases and pay more attention to their body health. Meanwhile, it raises the risks of data privacy to

participants that their health data might be collected to be shared by other people for varied purposes. Since the consent for sharing information was signed for a time limit, after that time limit, the dataset might also be shared without the volunteers' consent, which is a potential risk. Moreover, participants would be in psychological harm after answering some sensitive questions such as poverty, living situations or receiving an examination report informing potential disease, which might trigger negative emotions such as shame or anxiety.

Appendix

Tables

P-values Table

Variable Name	P-value
Height	< 0.0001
PhysActiveYes	< 0.0001
BPSysAve	< 0.0001
Gendermale	0.718
Race3Black	< 0.0001
Race3Hispanic	< 0.0001
Race3Mexican	< 0.0001
Race3White	< 0.0001
Race3Other	< 0.0001

Adjusted R-squared Table

Variables in the model	Adjusted R-Squared
Height + PhysActive + BPSysAve + Gender +Race3	0.256

Height + PhysActive + BPSysAve + Gender	0.244
PhysActive + BPSysAve + Gender +Race3	0.171
Height + PhysActive + Gender +Race3	0.223
Height + PhysActive + BPSysAve+Race3	0.256
Height + BPSysAve + Gender +Race3	0.245

Works Cited

Instructions: Include references to any works cited above. You are welcome to use any citation style of your choosing (MLA, APA, Chicago, etc.), as long as you are consistent.

[1] Brock, A. (2019). *Is BMI an Accurate Way to Measure Body Fat?* Scientific American. https://www.scientificamerican.com/article/is-bmi-an-accurate-way-to-measure-body-fat/

[2] Bell, C.N. and Blackman Carr, L.T. (2020), The Role of Weight Perception in Race Differences in BMI Among College Graduate and Non–College Graduate Women. Obesity, 28: 970-976. https://doi.org/10.1002/oby.22765

[3] *NHANES: NHANES 2009-2012 with adjusted weighting in NHANES: Data from the US National Health and Nutrition Examination Study*. (n.d.). Rdrr.io. Retrieved December 9, 2021, from https://rdrr.io/cran/NHANES/man/NHANES.html

R Code

Instructions: All of the R code used to produce this final report should be included in an R Markdown file and submitted to Moodle. Please submit both the original RMD and the knitted version (saved as a PDF). Make sure that this R Markdown document meets all requirements listed in the Final Grading Rubric (well-commented, no unnecessary output or error messages, etc.).