Technical Report | DASC 240

Group A | Hunter Heffernan, Inoa Nakahara, Adam Ng, Brody Kallman

Abstract

The goal of this project was to utilize data from NHANES from 2021-2023 to create a multiple linear regression model in an attempt to make predictions on a person's pulse pressure based on a risk factor, alcohol consumption as well as a person's BMI, physical activity and age. The data was wrangled and filtered such that we drop non-respondents and keep only adults (adult being defined as anyone 18+). After testing the initial model conditions with plots, we see that variance was an issue, to remedy this we used a log transformation to normalize the data. While the model only explains around 21% of the variability, the model can still provide meaningful inferences about the relationship between variables. Notably, we found that BMI is associated with a small decrease in pulse pressure while age & vigorous activity were associated with a slight increase in pulse pressure. These predictions may be helpful to understand how risk factors and lifestyle choices impact our cardiovascular health. This model may be particularly useful for our group, as a group member's father's cardiovascular health has declined greatly (heart attack and stroke in the past two years), so this model may serve to assist in predicting his pulse pressure.

Introduction

Every 33 seconds, one person dies from cardiovascular disease in the United States,[1] and since 1950 heart disease has been the leading cause of death in the United States.[2] One way that heart disease is assessed and predicted is by looking a patient's pulse pressure. Pulse pressure is the difference between the highest and lowest blood pressure readings during one cardiac cycle; it can be computed by subtracting the diastolic from the systolic blood pressure (systolic - diastolic). Our project aims to predict pulse pressure by looking at alcohol consumption, age, BMI, and frequency of vigorous physical activity, to potentially better understand how

[1] "Heart Disease Facts," *Heart Disease in the United States*, Centers for Disease Control and Prevention, accessed May 19, 2025, https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html.

[2] "Heart Disease Deaths - Health, United States," *Health, United States*, Centers for Disease Control and Prevention, accessed May 19, 2025, https://www.cdc.gov/nchs/hus/topics/heart-disease-deaths.htm#ref1.

certain risk factors and lifestyle choices may affect pulse pressure, and thus better understand how one could change habits as a preventative measure to combat heart disease.

Data

Our data comes from The National Health and Nutrition Examination Survey (NHANES) which is a program under the CDC's National Center for Health Statistics (NCHS).[3] NHANES typically runs in two year cycles for their data collection, but due to the pandemic, they released an August 2021-August 2023 data set, this is the cycle/data set we used. NHANES selects its subjects via a stratified multistage probability sampling, this ensures subjects are randomly selected and thus the data collected via examination of subjects and questionnaires can be used to make generalizations about the entire (U.S.) population. NHANES data is modular in the sense that their data is separated into multiple files such as demographic, dietary, examination, etc. that can be downloaded individually. This feature is nice for researchers such as ourselves as it reduced time and energy to filter out a lot of unnecessary data, and we could look for specific variables that were of interest to the project. Initially the data sets, before wrangling or filtering consisted of 43084 observations and 78 variables. From those, we only needed a handful, systolic and diastolic to find pulse pressure (quantitative) as our response variable, BMI (quantitative), alcohol amount (quantitative), and the frequency of vigorous activity per year (quantitative). With these we created the variables: pulse_pressure, BMI, alcohol_amount, vigorous_per_year and renamed other columns for better understanding.

Load Data

First we load the data in from the data file (.zip attached). Each of these files are downloaded from the NHANES website for the 2021-2023 cycle. Even though the NHANES data is module, there still is a few things we need to change in the data and a lot of unnecessary data we need to drop.

[3] "August 2021–August 2023 Laboratory Data – Continuous NHANES," *National Health and Nutrition Examination Survey*, Centers for Disease Control and Prevention, accessed May 19, 2025, https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Laboratory&Cycle=2021-2023.

Wrangle & Filter Data

First, we dropped everything from the demographic data set other than age, but only for subjects that are 18+ (effectively making the data adults only) and SEQN, which is a code that uniquely identifies each person (row) of our data - we keep this column for each data set no matter what wrangling or filtering we do. Next, we dropped every row from the alcohol data except the row that noted who was a drinker and how often they drank, we dropped anyone who did not answer, and for alcohol amount, if it was left blank we assumed that meant 0 as the only options were 1-14. For the physical data set we needed to compute how many times a year each person participated in vigorous activity. To do this we used two columns, one that denoted the frequency and the other the unit (2 times a week, 1 time a month, etc.) and multiplied each value by a constant depending on the unit for time. Each value denoted with a 'D' was multiplied by 365, 'W' by 52, and 'M' by 12. Any null values were set to 0 as there was no option for no vigorous activity. We then put these products in a new column 'vigorous_per_year' to get each subjects yearly rigorous activity frequency for a year. The last meaningful data adjustment came when we computed the pulse pressure by subtracting the diastolic column from the systolic column and saving the difference into a new column, 'pulse_pressure' and once again dropped anyone who did not get a reading.

Combine Data & Fit Model

After wrangling and filtering the data we downloaded, all that was left to tidy everything up was to join our filtered data to create one table relating to our model. We also create our first model, mlr4.

Check Conditions

Based on the VIF test we can clearly see that we do not have to worry about multicollinearity as the highest value is still < 2. However, based on the Q-Q plot, we have issues with the normality of residuals, as the right tail pulls off the fitted line heavily and the left pulls off a tad. This is a violation of the normality condition. Our solution to this violation was to use a log-transformation in order to normalize the data.

Transformation of Data (log)

The log-transformation of pulse pressure, alcohol amount, vigorous per year and BMI greatly improved the Q-Q Plot, showing a normalized plot with residuals following the fitted line almost perfectly, our new model now meets the condition for normality of residuals. The VIF conclusion remains unchanged. The purpose for adding the constants was that log(0) is undefined, so to retain the zeroes in the data we add one as $\log(1) = 0$, and 0.1 to the pulse pressure to further normalize the data, as it was skewed to the right. We still have three more conditions to check before we start interpreting any variables.

Finnish Checking Conditions (L.I.N.E)

L

Based on the Residual vs. Fitted plot, we can see that both tails pull a bit away from the fitted line and a small dip around 3.8, however we would still consider the linearity condition to be met as the curvature is minor and the residuals are what we would expect for a linear model.

I

Independence cannot be checked with a plot, but this condition is met as the rows for our data, as previously mentioned, are randomly selected individuals.

N

Normality was shown to be met via the Q-Q Plot of the transformed data.

E

Equality of variance can be checked with a Residual vs. Fitted plot and a Residual vs. Leverage plot. The Residual vs. Fitted plot shows the points are very evenly distribute across the plot and above/below the fitted line. The Residual vs. Leverage plot shows no data points are outside of Cook's distance. Thus, there are no extreme outliers with high leverage, allowing us to conclude that the equality of variance condition is met. With that, all L.I.N.E conditions have been met.

Univariate Visualizations

We used histograms to plot each variable for univariate visualizations, with this we highlight how using a log transformation normalized the data.

Results

From our model summary we can make the following interpretations of the coefficients: Holding other variables constant, a 1% increase in alcohol consumption is associated with an approximate 0.0019% decrease in pulse pressure on average. This effect is small and only not significant, as the p-value = $0.73 > 0.05$.

Holding other variables constant, a 1% increase in the frequency of vigorous activity is associated with an approximate 0.0034% increase in pulse pressure on average. Again, the effect is very small, however it is significant as the p-value = $0.0313 < 0.05$.

Holding all else constant, a 1% increase in BMI is associated with an approximate 0.1963% decrease in pulse pressure on average. This is very statistically significant as the p-value $<$ 2e-16 < 0.05.

Holding other variables constant, each additional year of age is associated with an approximate 0.693% increase in pulse pressure on average. Again, this is highly statistically significant as the p-value $<$ 2e-16 < 0.05.

Something interesting we found was that alcohol appears to not be a significant predictor for pulse pressure. This was very surprising as we hypothesized it would be one of the most significant. We could have created a new model without alcohol amount, however we felt it would be nice to keep in the model as it allows us to hold alcohol amount constant while interpreting out other coefficients. It is not surprising that age and BMI are the best predictors for pulse pressure as they are both closely related to cardiovascular health - as we age our systems begin to degenerate, and as BMI gets to unhealthy levels, visceral fat builds up around organs and plaque in the arteries. So, it is no surprise that these variables predicted pulse pressure the best, however it is important to note that age's influence on pulse pressure was much larger than BMI's influence.

Conclusion

Based on NHANES 2021–2023 data, in this study, we developed a multiple linear regression model to explore the relationship of age, exercise, BMI, and alcohol intake with pulse pressure, a commonly used measurement relating to cardiovascular health. Utilizing a log transformation, we were able to normalize our data and meet all the conditions (L.I.N.E) for drawing conclusions from our model. Additionally, our model explains approximately 21% of the variance in pulse pressure. At face value this seems like a low percentage, however an "R2 of >15% is a generally a meaningful value in clinical research."[4]

Our findings identified BMI and age as the best predictors of pulse pressure, the latter having strongest positive correlation. Surprisingly, alcohol consumption had no impact on pulse pressure. Additionally, Vigorous exercise was associated with a small positive correlation, which supports the established cardiovascular benefits of exercise.

In spite of its weaknesses, such as dependency on self-report and a relatively low explanatory power, we feel our model can potentially provide the foundation for better understanding how pulse pressure is affected by risk factors and behavior. Future iterations of this project could look into other risk factors, such as smoking, utilizing perhaps a different NHANES cycle that would not be filtered to as low of observations as our 2021-2023 cycle when wrangling/filtering the data. This model is also specific to the U.S. population, so if similar data is collected for other countries we could look into how similar the results are to this model. Generally, our model helps to shed light on cardiovascular risk assessment and could be a valuable tool for individuals to better understand and possibly reduce their risk of heart disease.

Bibliography

"August 2021-August 2023 Laboratory Data - Continuous NHANES." n.d. https://wwwn.cdc.gov/nchs/nhanes/s 2023.

"Heart Disease Deaths - Health, United States." August 05, 2024. https://www.cdc.gov/nchs/hus/topics/heart-disease-deaths.htm#ref1.

"Heart Disease Facts." 2024. Heart Disease. October 24, 2024. https://www.cdc.gov/heart-disease/data-research/facts-stats/index.html.

Gupta, Avi, et al. "Determining a Meaningful R-Squared Value in Clinical Medicine: Published in Academic Medicine & Surgery." Academic Medicine & Surgery, University Medical Press, 27 Oct. 2024, https://academic-med-surg.scholasticahq.com/article/125154-determining-a-meaningful-r-squared-value-in-clinical-medicine

[4] Gupta, Avi, et al., "Determining a Meaningful R-Squared Value in Clinical Medicine," *Academic Medicine & Surgery*, University Medical Press, 27 Oct. 2024, https://academic-med-surg.scholasticahq.com/article/125154-determining-a-meaningful-r-squared-value-in-clinical-medicine