STAC67 Case Study

Factors Affecting Systolic Blood Pressure

Mustafa Mughal 1004108075

Vithushan Thevakesaran 1004390018

University of Toronto

April 10th, 2022

Abstract

Systolic blood pressure is one of the most important measurements that an individual should be aware of as it can be a telling sign of related health issues. It measures the force that the heart exerts on the walls of the arteries upon each time is beats and can be categorized as normal (<120 mg in Hg), elevated (120-129 mm in Hg), stage 1 high (130-139 mm in Hg) and stage 2 high blood pressure (>140 mm in Hg) (Blood pressure chart: What your reading means, 2022). There are many factors that can affect this measure and we will be considering which ones are the most related and have the greatest impact on systolic blood pressure. This process will be conducted through several regression models on a sample of 500 individuals ranging in factors such as smoking status, alcohol usage, stress levels and more, which will be compared and validated to determine what has the greatest effect on systolic blood pressure.

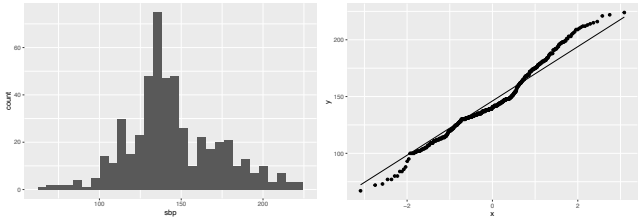
Background and methods

Systolic blood pressure readings can be determinant signs of heart related diseases and other terminal issues. High ratings (hypertension) can increase the risk of heart disease which can induce heart attacks, kidney failure and overall mortality. However, for low blood pressure ratings (hypotension), it usually does not cause as much of a serious threat as high ratings, although severe hypotension can be extremely dangerous as it causes a lack of blood flow to the brain. This results in symptoms such as dizziness, blackouts, nausea, vomiting and can even be life threatning in some cases (Kumar, 2021). Common practices to reduce high blood pressure are to take prescribed medicine by a licensed professional and keeping a healthy and regular diet and lifestyle. This means to reduce alcohol consumption and avoiding smoking, although individuals who are heavy smokers may sufferer from very low blood pressure. Regular exercise and healthy diets consisting of low sodium, high potassium and high magnesium foods can help to relive high systolic blood pressure greatly. Other external factors also have an impact on the systolic blood pressure such as gender, weight, stress, childbearing capabilities, and income. In the given data set, we will be leaving out variables bmi as it does not have high associativity with blood pressure and its rather casually associated (Linderman et al., 2018). Also, for age which does not have consistent results in one's blood pressure across the board as the systolic blood pressure on average can increase and decrease over the years, so that will also not be included in the model. We will examine the model through backwards stepwise selection and include interaction terms which have been proven in previous studies as will be explained throughout this report. We will fit various regression models and compare them using anova to determine the best model, then validate that specific model.

Exploratory Data Analysis

The analysis of systolic blood pressure is based on a dataset consisting of 500 observations (humans). There are 18 variables present in the data set, including the response variable systolic blood pressure (sbp) and 17 predictor variables. As mentioned before, the objective of the study is to determine which of these predictors have an impact on systolic blood pressure based on the data.

Systolic Blood Pressure: The response variable of the study and is recorded as a numerical value. The mean and standard deviation of SBP are 144.95 and 27.99 respectively.



The normality assumption of SBP is withheld from the histogram and normal quantile plot.

Gender: A binary variable where M = Male and F = Female. There are 264 females and 236 males in the study.

Marital Status: A binary variable where Y = Married and N = Not Married. There are 239 married individuals and 61 unmarried individuals.

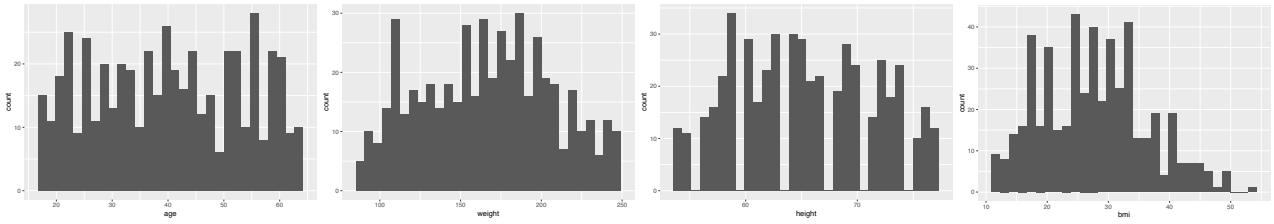
Smoking Status: A binary variable where Y = Smoker and N = Non-Smoker. There are 266 smokers and 234 non-smokers.

Age: A continuous variable measuring the subjects' age given in years. The mean and standard deviation of age are 40.20 and 13.30 years respectively.

Weight: A continuous variable measuring the subjects' weight given in pounds. The mean and standard deviation of weight are 166.64 and 40.90 pounds respectively.

Height: A continuous variable measuring the subjects' height given in inches. The mean and standard deviation of height are 65.33 and 6.19 inches respectively.

Body Mass Index(BMI): A continuous variable measuring the subjects' BMI. The CDC states the Body mass index(BMI) is a person's weight in kilograms divided by the square of height in meters. The observations will be multiplied by 703 to account for the conversion of kilograms to pounds and meters to inches. The mean and standard deviation of BMI are 27.66 and 8.56 respectively.



- Distribution of Quantitative variables Against SBP

Overweight: A categorical variable where subjects are separated into 3 groups based on obesity levels. Subjects are placed into the groups: 1 = Normal, 2 = Overweight and 3 = Obese. There are 187 normal weight, 109 overweight and 204 obese individuals.

Race: A categorical variable where subjects are separated into 4 groups based on race. While the races are not provided, each racial group takes a value of 1, 2, 3 or 4. Of these races; 355 belong to race 1, 99 belong to race 2, 25 belong to race 3 and 21 belong to race 4.

Exercise Level: A categorical variable where subjects are separated into 3 groups based on exercise levels, where 1 = Low, 2 = Medium and 3 = High. There are 195 low, 136 medium and 169 high exercise level subjects.

Stress Level: A categorical variable where subjects are separated into 3 groups based on stress levels, where 1 = Low, 2 = Medium and 3 = High. There are 151 low, 175 medium and 174 high stress level subjects.

Salt Intake Level: A categorical variable where subjects are separated into 3 groups based on salt intake levels, where 1 = Low, 2 = Medium and 3 = High. There are 166 low, 157 medium and 177 high salt level subjects.

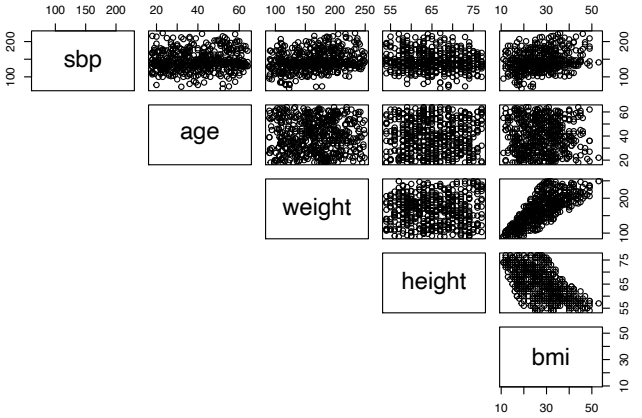
Childbearing Potential: A categorical variable where subjects are separated into 3 groups based on ability to childbear, where 1 = Male, 2 = Able Female and 3 = Unable Female. There are 236 males, 143 able females and 121 unable females.

Income Level: A categorical variable where subjects are separated into 3 groups based on income level, where 1 = Low, 2 = Medium and 3 = High. There are 176 low, 167 medium and 157 high income level subjects.

Education Level: A categorical variable where subjects are separated into 3 groups based on education level, where 1 = Low, 2 = Medium and 3 = High. There are 171 low, 159 medium and 170 high education level subjects.

Treatment (for hypertension): A binary variable for whether or not subjects have received treatment for hypertension recorded as 0 = Treated and 1 = Untreated. There are 399 treated and 101 untreated individuals.

Correlation of Continuous Variables



From the correlation plots, it looks as if there are correlations between SBP and weight, and SBP and BMI. There seems to be correlations between the predictor variables height and BMI, and weight and BMI as well.

Model Selection

Initially, the full regression model with 17 predictors is selected and fitted against SBP. The R^2 value is 0.2298 and the adjusted R^2 value is 0.1874. At an attempt to increase the R^2 values for the model, various models were fitted using different methods.

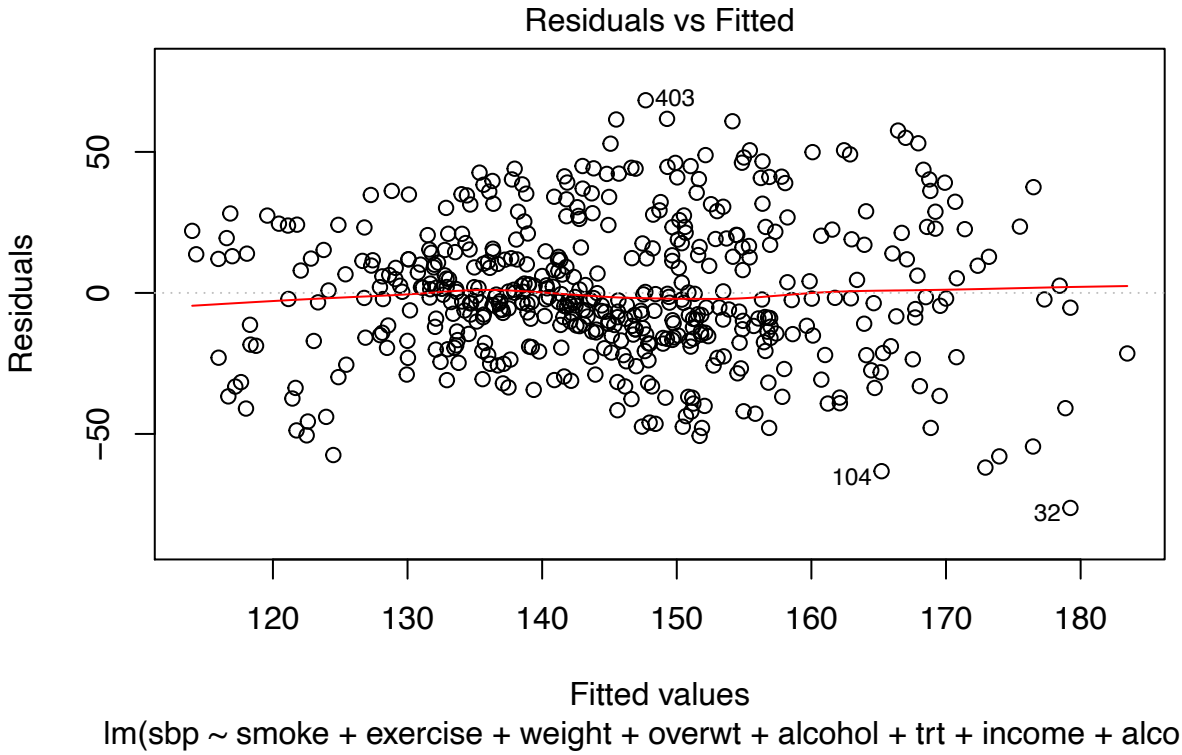
The first step in the process involved performing a backwards elimination using the built-in function in R. The R^2 value and the adjusted R^2 value are now 0.2036 and 0.189 respectively. Through this step, 10 predictors are eliminated. ANOVA F-test is used to compare the two models, the backwards elimination model is chosen to progress for further inspection.

The second step in the process involved creating a model that included interaction terms for all remaining 7 predictors. This model had an R^2 and adjusted R^2 value of 0.2818 and 0.214. It was suspected that this difference in R^2 values may be caused by having 28 predictors in the model. The model was then shortened by eliminating predictors and interaction terms that were considered to not have an impact on SBP. These terms were carefully selected to be removed through research provided from previous studies. The model now has 13 predictors with an R^2 and adjusted R^2 value of 0.2562 and 0.1914.

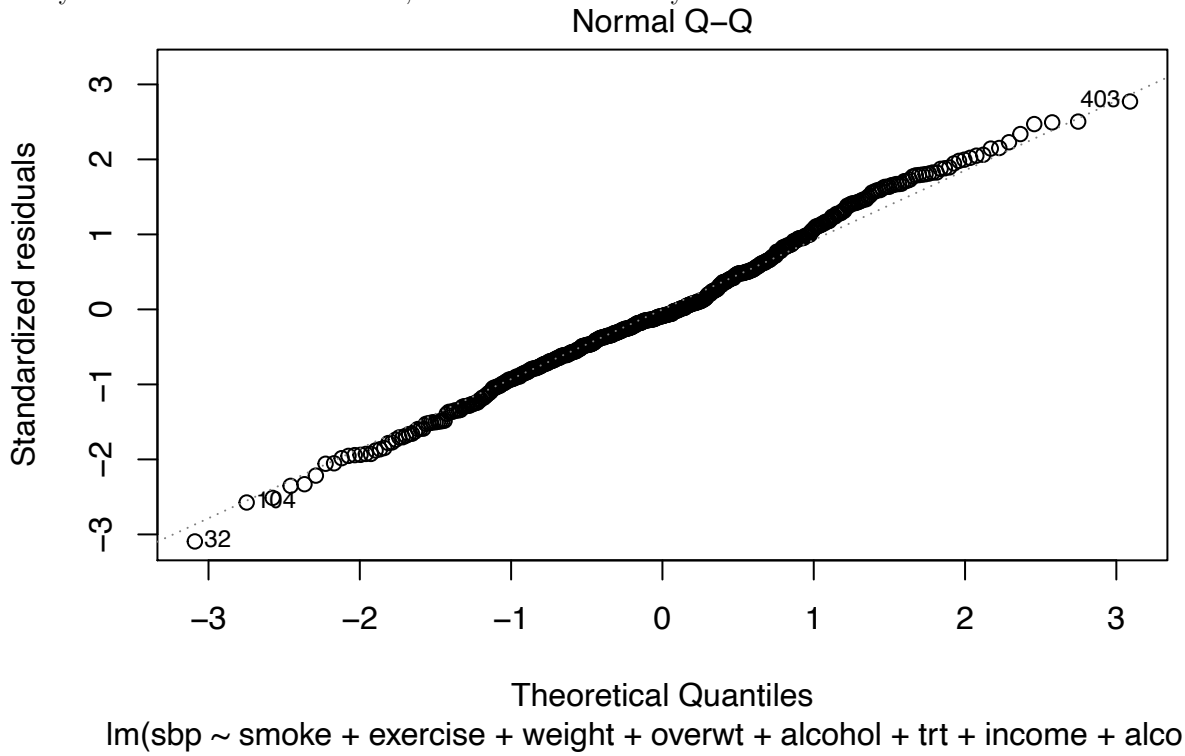
Finally, backwards elimination is performed on the current model and the resulting model is lm(sbp ~ smoke + exercise + weight + overwt + alcohol + trt + income + alcohol:trt + smoke:trt). The R^2 and adjusted R^2 values are 0.229 and 0.2067. An ANOVA F-Test is ran to compare this model with the previous model and therefore the simpler model is concluded to be significantly better than the complex model.

```
##
## Call:
## lm(formula = sbp ~ smoke + exercise + weight + overwt + alcohol +
##     trt + income + alcohol:trt + smoke:trt, data = BP)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.239 -15.444  -2.164  15.382  68.311
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   116.90856    6.29601  18.569  < 2e-16 ***
## smokeY         13.99240    2.54539   5.497 6.25e-08 ***
## exercise2     -10.89435    2.82645  -3.854 0.000132 ***
## exercise3     -11.09583    2.64611  -4.193 3.27e-05 ***
## weight          0.08842    0.04025   2.197 0.028489 *
## overwt2         7.28980    3.51367   2.075 0.038542 *
## overwt3        13.09303    3.69226   3.546 0.000429 ***
## alcohol2        1.97767    3.05760   0.647 0.518065
## alcohol3       16.05752    3.13670   5.119 4.43e-07 ***
## trt1            3.45874    6.25814   0.553 0.580739
## income2         1.84193    2.72646   0.676 0.499631
## income3         5.81837    2.79106   2.085 0.037624 *
## alcohol2:trt1  -6.15426    7.37458  -0.835 0.404397
## alcohol3:trt1 -16.87328    6.70117  -2.518 0.012125 *
## smokeY:trt1   -15.87276    5.71837  -2.776 0.005720 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.93 on 485 degrees of freedom
## Multiple R-squared:  0.229,  Adjusted R-squared:  0.2067
## F-statistic: 10.29 on 14 and 485 DF,  p-value: < 2.2e-16
```

Diagnostics

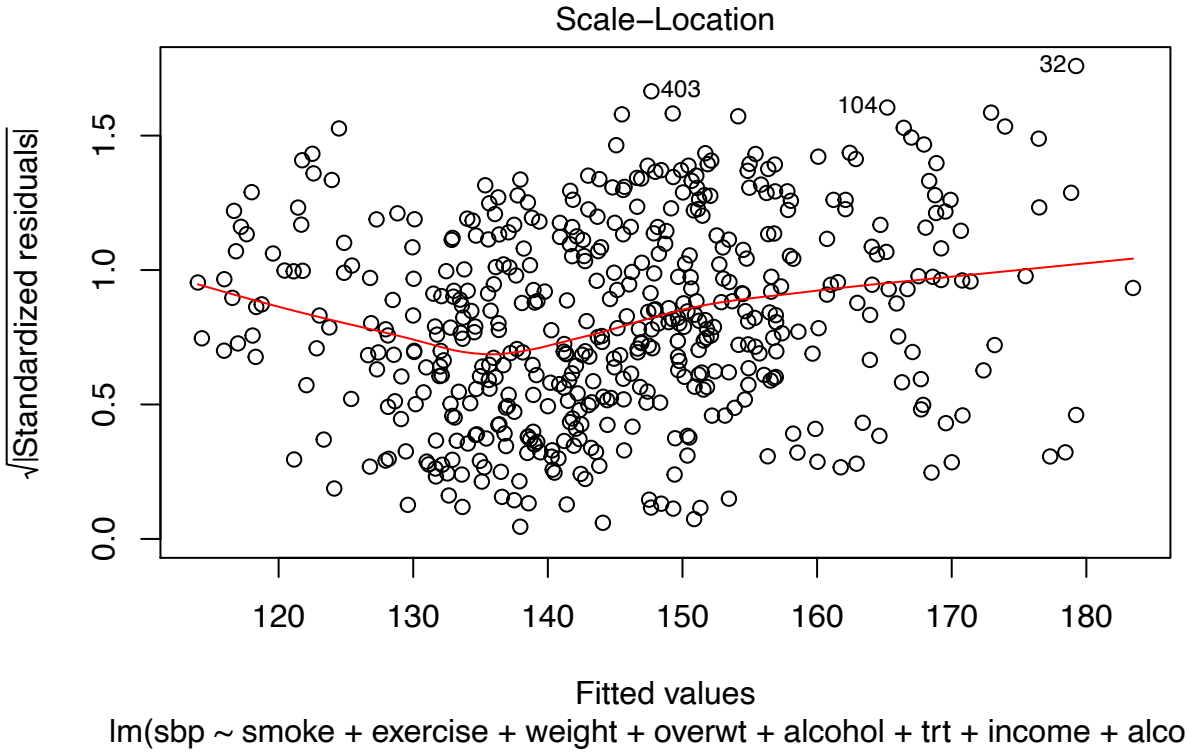


This residual and fitted value plot shows that there does seem to be a linear relations as points are roughly evenly distributed across the 0 line, which is also shown by the line in red as it is almost horizontal along 0.



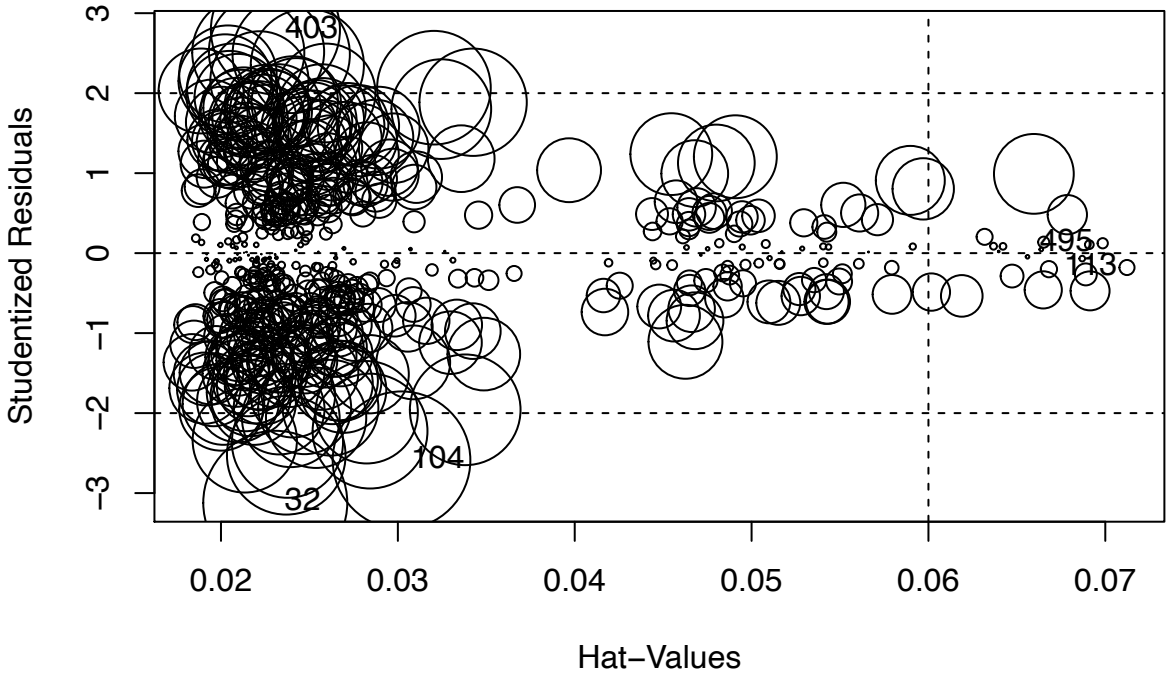
This QQ plot helps to show normality in the model as they are linearly distributed on the line with a few

outliers visible in 403, 32 and 104. We will further evaluate these outliers with VIF.



Although the red line is not completely horizontal, we can still see random and evenly spread distribution of the points along the red line indicating that homoscedasticity holds in this model.

Model Validation




```
##        StudRes        Hat        CookD
## 32  -3.1212528 0.02306969 0.0150655917
## 104 -2.5894496 0.03020752 0.0137619829
## 113 -0.1797701 0.07122630 0.0001655546
## 403  2.7911371 0.02307678 0.0120989473
## 495  0.1234678 0.06985459 0.0000764791
```

Our computed DFFITS and DFBETAS show that there are several obeservations which are influential in predicting the sbp in our model indicating that these points will have a great affect on the slope.

For cooks distance, none of the points are greater than $F(p', n - p')$ so there are no influential points based on cooks distance.

Here our max VIF is 5.077175, which is less than 10 and our $\overline{VIF}$ is 1.78549 which is not considerably larger than 1. Therefore we have evidence indicating that there is no serious multicollinearity present in this model.

Conclusion

The purpose of the study was to determine which of the 17 predictors had an impact on systolic blood pressure. The final model obtained consisted of 7 predictors and 2 interaction terms. However, the final R^2 and adjusted R^2 values are 0.229 and 0.2067 are very similar to the full model's R^2 and adjusted R^2 values of 0.2298 and 0.1874. This improvement does not seem important because the R^2 values did not change much but it is important to note that 10 predictors were eliminated in the process.

That being said, it is also important to comment on the final R^2 and adjusted R^2 values of 0.229 and 0.2067 because these values are not high enough to show a great relationship between final predictors and systolic blood pressure. Further research should be completed to help improve R^2 values. One way to improve R^2 values would be to include predictors that are not included in the current study. For example, high SBP is known to be more common in individuals whose parents have had it. Another way to possibly improve this model would be to use the final model for another data set. In conclusion, it is evident from the analysis that there was a significant improvement in the final model compared to the full model but not enough evidence to conclude a significant relationship selected predictors and the response variable.

References

Bosely, S. (2012, December 13). *High blood pressure, smoking and alcohol are biggest health risks*. The Guardian. Retrieved April 7, 2022, from https://www.theguardian.com/society/2012/dec/13/health-risks-high-blood-pressure-smoking

Centers for Disease Control and Prevention. (2021, August 27). *About adult BMI*. Centers for Disease Control and Prevention. Retrieved April 7, 2022, from https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html

Diverse Populations Collaborative Group (2005, August 15). *Weight-height relationships and body mass index: Some observations from the diverse populations collaboration*. American journal of physical anthropology. Retrieved April 7, 2022, from https://pubmed.ncbi.nlm.nih.gov/15761809/

Kumar, K. (2021, May 6). *Which is more important: Systolic or diastolic blood pressure?* MedicineNet. Retrieved April 6, 2022, from https://www.medicinenet.com/importance_systolic_vs_diastolic_blood_pressure/article.htm

Linderman, G. C., Lu, J., & Lu, Y. (2018, August 17). *Association of Body Mass index with blood pressure among 1.7 million Chinese adults*. JAMA Network Open. Retrieved April 6, 2022, from https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2696872#:~:text=Body%20mass%20index%20(BMI)%20is%20positively%20associated%20with%20both%20systolic,diastolic%20blood%20pressure%20(DBP).&text=Weight%20loss%20significantly%20reduces%20blood,BP%20but%20is%20causally%20associated.

Mayo Foundation for Medical Education and Research. (2021, March 18). *Stress and high blood pressure: What's the connection?* Mayo Clinic. Retrieved April 7, 2022, from https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/stress-and-high-blood-pressure/art-20044190#:~:text=Your%20reaction%20to%20stress%20may,your%20blood%20vessels%20to%20narrow

Mayo Foundation for Medical Education and Research. (2022, March 18). *Blood pressure chart: What your reading means*. Mayo Clinic. Retrieved April 6, 2022, from https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/blood-pressure/art-20050982

Primatesta, P., Falaschetti, E., Gupta, S., Marmot, M. G., & Poulter, N. R. (2001, February 1). *Association between smoking and Blood Pressure*. Hypertension. Retrieved April 7, 2022, from https://www.ahajournals.org/doi/10.1161/01.hyp.37.2.187#:~:text=Smoking%20causes%20an%20acute%20increase,be%20associated%20with%20malignant%20hypertension.&text=Nicotine%20acts%20as%20an%20adrenergic,possibly%20the%20release%20of%20vasopressin

Ramezankhani, A., Azizi, F., & Hadaegh, F. (2019). Associations of marital status with diabetes, hypertension, cardiovascular disease and all-cause mortality: A long term follow-up study. *PloS one*, *14*(4), e0215593. https://doi.org/10.1371/journal.pone.0215593

Richardson, C. (2021, September 14). *How does alcohol affect blood pressure?* Medical News Today. Retrieved April 7, 2022, from https://www.medicalnewstoday.com/articles/alcohol-and-blood-pressure

Smith, M. (2021, May 7). *Isolated systolic hypertension: Causes of high systolic blood pressure*. WebMD. Retrieved April 7, 2022, from https://www.webmd.com/hypertension-high-blood-pressure/isolated-systolic-hypertension