Predicting Systolic Blood Pressure[*]

Ensar Pajtesa

27/04/2022

Abstract

Each time you visited the doctor as a kid and even as an adult, one of the first things measured is blood pressure. This is an indicator of overall cardiovascular conditions and health. We are interested in using other characteristics of individuals and run a regression model to predict the blood pressure reading. This will give us information on the characteristics that affect the rate of blood pressure and how this can be foreseen or treated.

Contents

[*] Code and data can be found at: https://github.com/ensarpajtesa1/bloodpressure

Introduction

In the medical field, statistical analysis is imperative. Although every patient's condition differs slightly from the next, there is a general consensus among the things that are detrimental to our health and the rate at which they are affecting us is very important to study and analyze. As per a report conducted by Medical News Today the number one cause of death worldwide in 2020 was not cancer, not accidents, not even COVID-19. The number one cause of death is cardiovascular disease (Cronkleton, n.d.). Cardiovascular disease however is a broad term and relates to many different conditions had by people. Systolic blood pressure measures the force the heart exerts on the arteries each time it pumps (Staff, n.d.). High or low blood pressure can lead to a number of problems and most notably due to the thickening of the arteries to be able to withstand the force, this condition sometimes in combination with cholesterol leads to higher risk of heart attacks and strokes (Health n.d.)

Our hearts are the beat that keeps us alive. It is important to assess heart health and be proactive on measures. In this analysis our attempt is to create a model that best predicts systolic blood pressures of patients, this is important because with these measurements it can be possible to detect early signs of regressing cardiovascular health.

Data

For this analysis we are using `R` (Team 2020), `tidyverse` (Wickham et al. 2019) and `dplyr` (Wickham et al. 2021) functions. For the creation of figures and tables we will use `ggplot2` (Wickham 2016), `kableextra` (Zhu 2020) and `reshape2` (Wickham 2007). The package `knitr` (Xie 2021) is used to generate the R markdown report.

The data used for this analysis was extracted from the 2011-2012 NHANES dataset (Disease Control and (CDC) (n.d.)). NHANES is the acronyms for Natural Health and Nutrition Examination Survey. This is conducted yearly in the United States and among other information collects data from the general physical attributes and mental/lifestyle attributes from those involved in the Survey. The NHANES dataset features over 70 variables, however for this analysis we have extracted some relevant variables which are believed to have something to do with the Systolic Blood Pressure reading. Working with minimized set of variables allows us to better assess our Exploratory Data Analysis.

We further clean the data removing further variables based on their relationship with others in an attempt to remove redundancy and/or variables that are obvious to have a strong correlation to one another.
- An example of this is BMI, this is heavily correlated to weight and height as this is what is used in its calculation. We can use BMI instead of Weight and Height and cover this category while remaining a simpler model
- Another example used of this is Income and Poverty, the poverty level uses the income of the household against the benchmark and produces a numerical value. This can be used instead of income as the redundancy and correlation is clear.

The variables that remain are the ones we will use to start this analysis, below is a description of each one, categorical or numerical and what the outputs could be in each variable:

- Gender (Categorical): Gender (sex) of study participant as male or female
- Age (Numerical): Age in years at screening of study participant. Subjects 80 years or older were recorded as 80.
- Race3 (Categorical): Reported race of study participant: Mexican, Hispanic, White, Black, Asian, or Other.
- Poverty (Numerical): A ratio of family income to poverty guidelines. Smaller numbers indicate more poverty
- BMI (Numerical): Body mass index (weight/height2 in kg/m2). Reported for participants aged 2 years or older.
- BPSysAve (Desired Outcome Variable): Combined systolic blood pressure reading.

- DirectChol (Numerical): Direct HDL cholesterol in mmol/L. Reported for participants aged 6 years or older.
- Depressed (Categorical): Self-reported number of days where participant felt down, depressed or hopeless. Reported for participants aged 18 years or older. One of None, Several, Majority (more than half the days), or Almost All.
- SleepHrsNight (Numerical): Self-reported number of hours study participant usually gets at night on weekdays or workdays. Reported for participants aged 16 years and older.
- PhysActive (Categorical): Participant does moderate or vigorous-intensity sports, fitness or recreational activities (Yes or No). Reported for participants 12 years or older.
- SmokeNow (Categorical): Study participant currently smokes cigarettes regularly. Reported for participants aged 20 years or older as Yes or No

Figure 1 displays a density histogram of our outcome variable which is systolic blood pressure. It is important to assess a distribution of this model. We see that the bulk of our data lies between 80 an 160 reading which is quite a large range. Distribution is skewed to the left as is expected because more people tend to be within normal ranges with some people who have abnormally high blood pressure skewing the distribution.

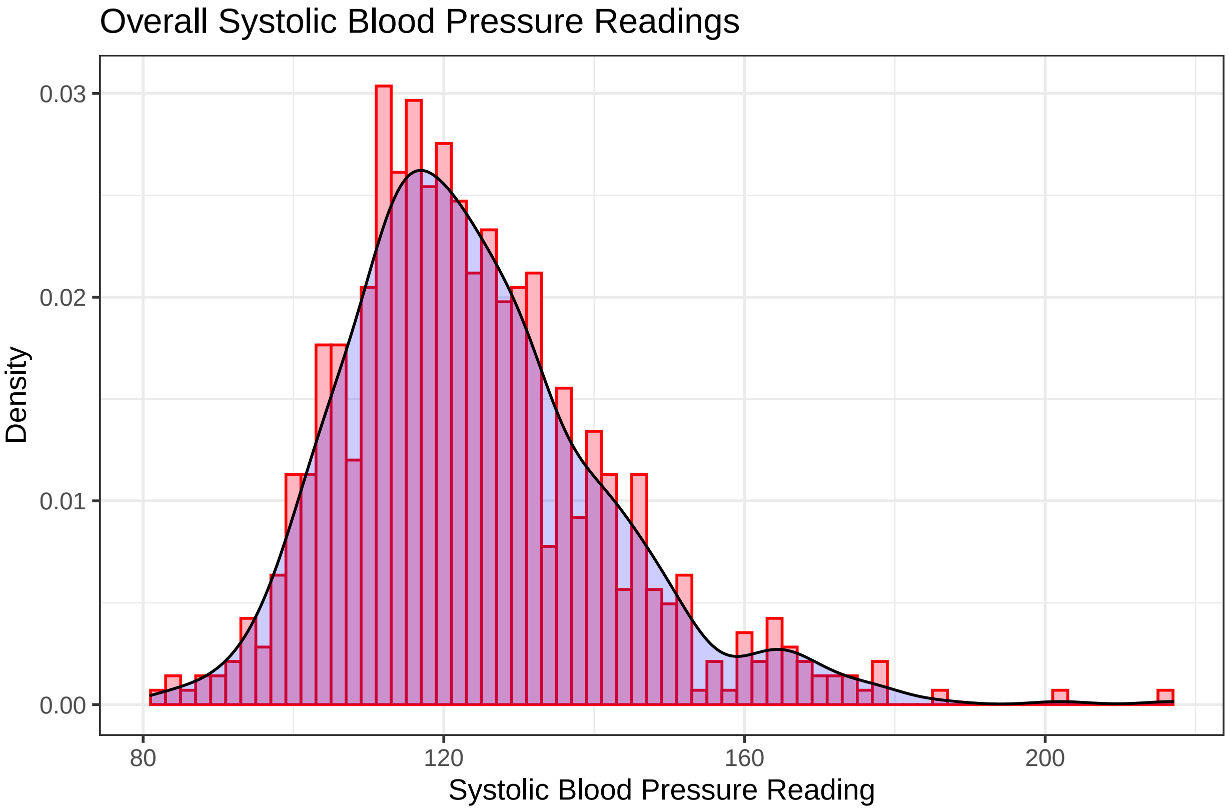


*Figure 1

Model

We are going to work through and create a linear regression model that in the end provides us with coefficients that we hope will allow us to predict the blood pressures to some degree of significance. We must first ensure that all steps and assumptions or linear regression are met and hold.

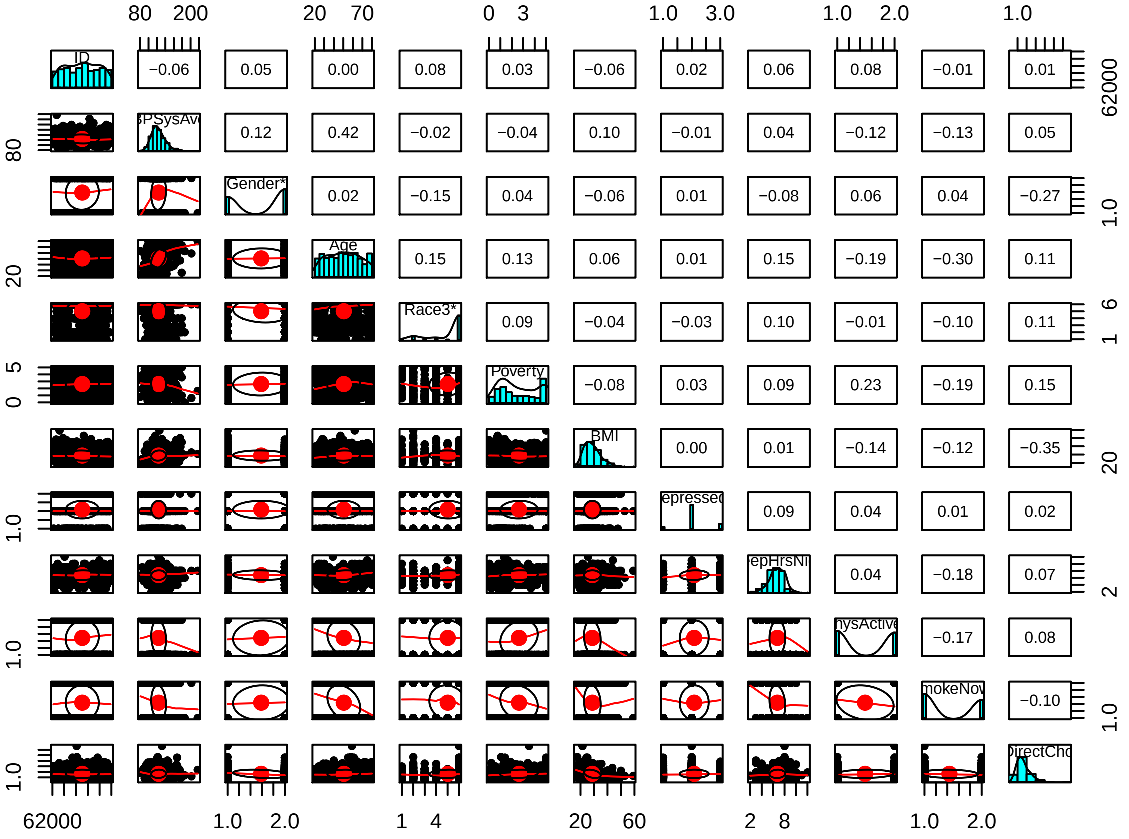
This linear regression was done as a backward elimination method where we selected all variables in the data set first and then used linear regression assumptions and model diagnostics to remove one variable at a time until the model does not get any better or gets worse.

In order to validate our model and ensure reproducibility it is important that we subset the data into a train and test dataset and specify a seed for these samples. We will fit our model into the train data set and then use the model created to validate on the test set.

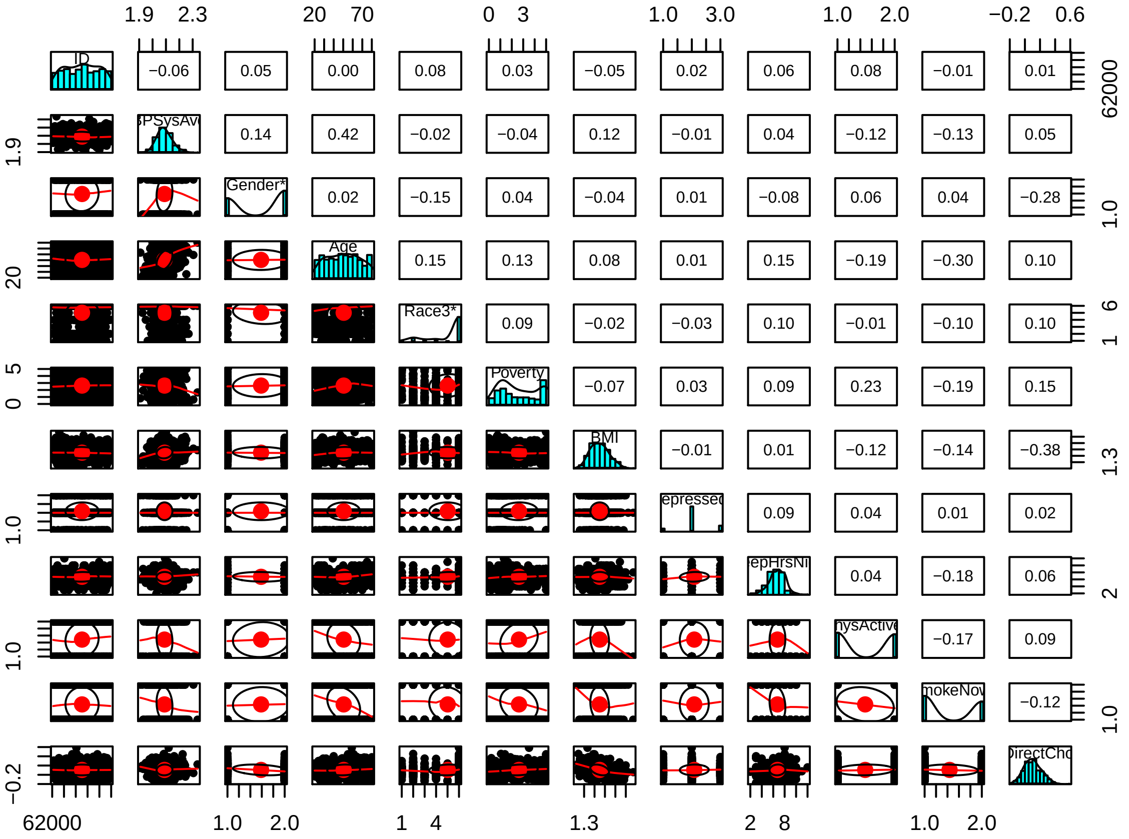
Assessing Linear Regression Assumptions and Variable Preparation:

Normality of Residuals and Independence of Predictors:

We run a pairwise function of all the variables to see how our variables are distributed and in general what the makeup of our predictors will be. On the main diagonal displays distributions of the variables, to the right the entire triangle calculates correlation between the perpendicular variables and the correlations plots of the perpendicular variables fill the bottom right of these graphs. Our main points of focus when doing this is assessing our linear regression assumptions. 2 main assumptions can be seen from this pairwise function, Normality of Residuals and Independence of Residuals. As foreseen due to the selection that we made and described above there is little correlation between our predictor variables. However for numerical values there are some variables which have skewed distributions and we will try to fix this using methods such as log and square root and applying this to the data column in its entirety.

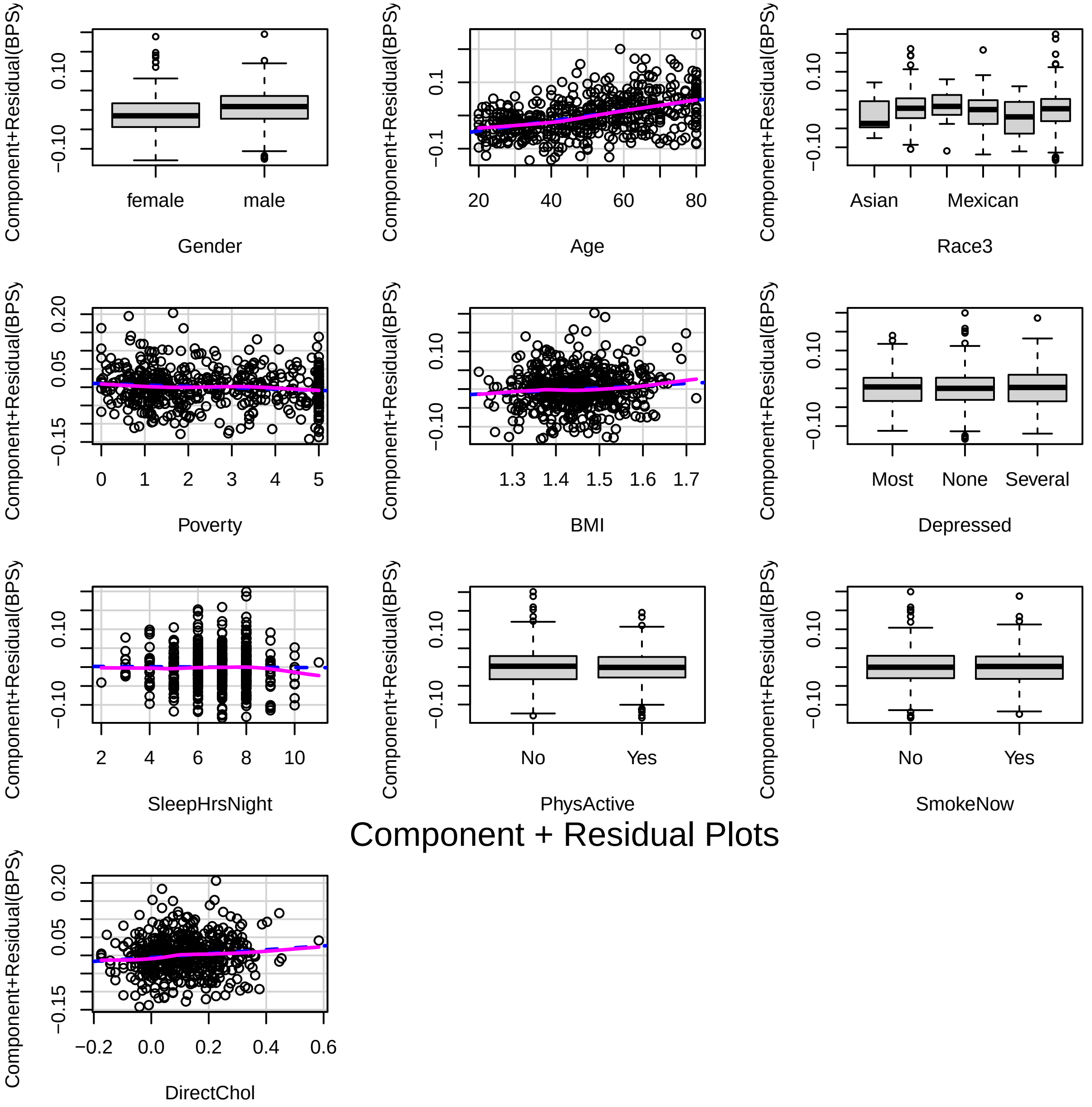


After these we run the pairwise function again and have the following results, Blood Pressure, BMI, Cholesterol all now have adequate normal distributions allowing all of our numerical variables to satisfy this regression assumption except for one of them. That is Poverty rate, this variable did not respond to any manipulation techniques and therefore will be left as is. This will be discussed in a model limitation portion of this report.



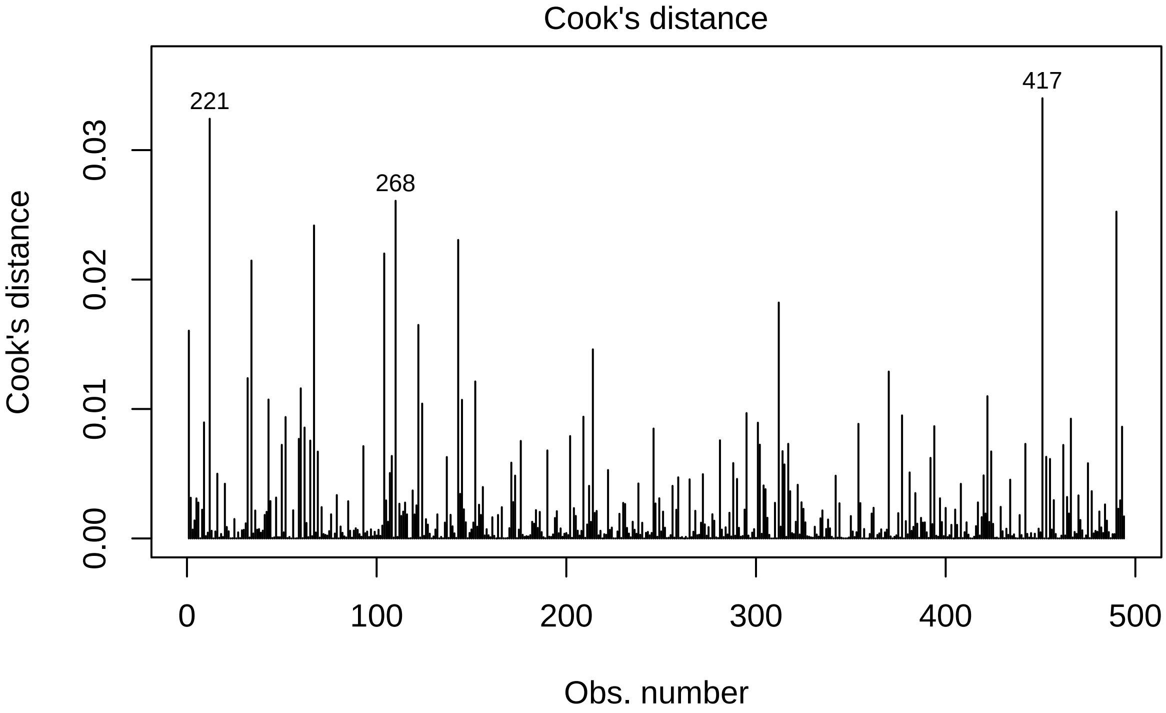
Assumption of Linearity and Constant Variance

We display in the graphs below of all the predictors the residuals in relation to our blood pressure reading and we see that the pink line in all numerical variables closely follows the dotted regression line. Assumptions of linearity and constant variance hold.



Extreme Values

It is important that we attempt to find data points in our set that may be unnecessarily skewing our regression line
To do this we use the Cooks Distance which find the data points influence on the fitted response values. The cutoff for Cook's distance is set to 4 times. We run the Cook's distance multiple times and we remove the 3 most influential points from the data that are severely outside our cutoff. Our goal is to ensure that our model gets better when assessing the diagnostics that are discussed above. We do not want to remove too many variables and run the risk of over-fitting our model to this specific training data set.
The final cooks distance chart looks as below and the values do not show any extremities in our regression set.



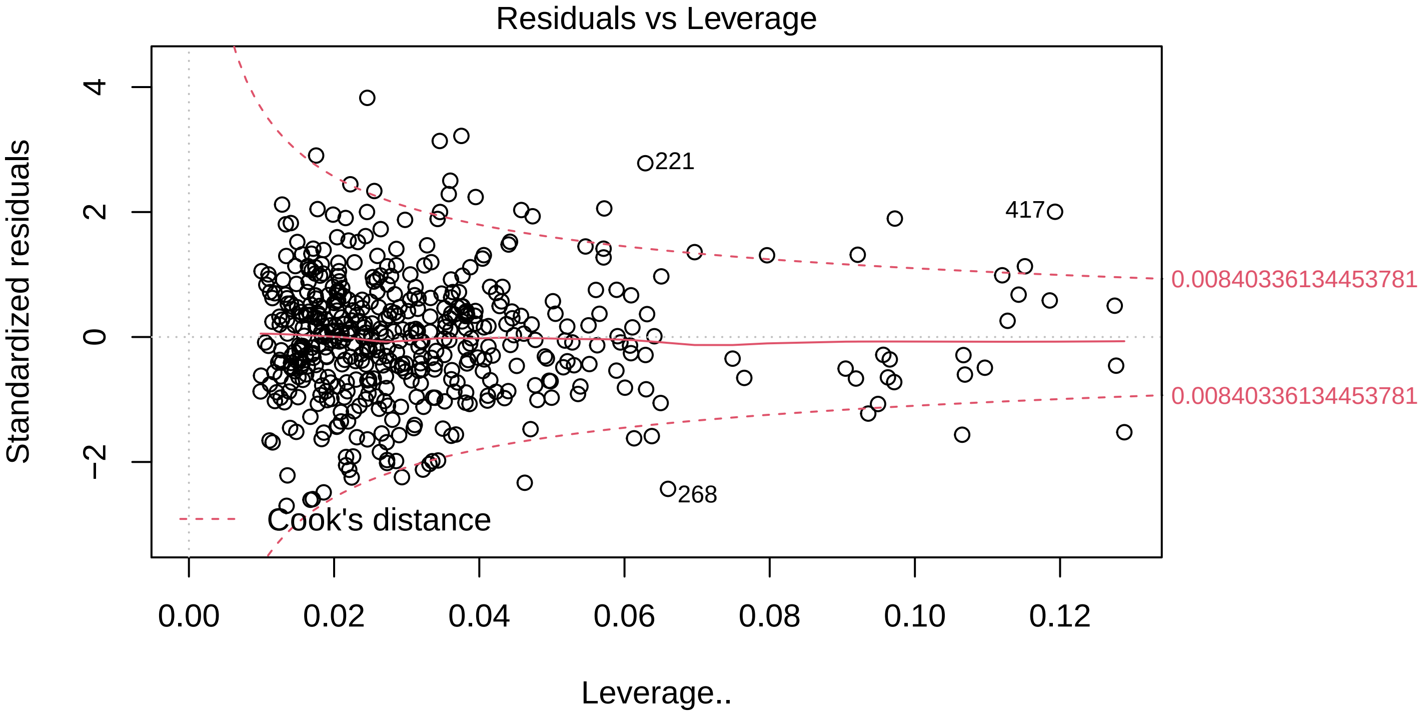

Residuals vs Leverage
Standardized residuals
4
2
0
−2
0.00
0.02
0.04
0.06
0.08
0.10
0.12
Leverage..
221
417
268
Cook's distance
0.00840336134453781
0.00840336134453781

Table 1: Multicollinearity of Predictors

Predictor Variable	GVIF Value
Gender	1.080
Age	1.133
Race	1.019
Poverty	1.109
BMI	1.122
Depressed	1.030
Hours Slept at Night	1.035
Physically Active	1.105
Smoker	1.144
Direct Cholesterol	1.621

Multicollinearity

We use the VIF function which is a measure of how easy one variable can be predicted using the other predictor variables. Since some of the variables in our regression are categorical variables there are different degrees of freedom and our GVIF takes into account all degrees of freedom. Any VIF value above 5 must be examined. However we see in our model that all GVIF values are well below this mark (Table 1). We can proceed assuming no multicollinearity

Table 2: Overview of Linear Regression Model

Statical Test	Value
Multiple R-Squared	0.2505
Adjusted R-Squared	0.2350
F Test Statistic	16.1400
P-Value	0.0000

Results

Throughout the analysis of our model, the diagnostic tests used to assess whether the model is improving or regressing are Adjusted R-Squared which displays the percentage of the residuals that the model explains and the P-Values of both the model and the predictors. These p-values test the null hypothesis that states that there is no correlation. A low p-value shows that we reject the null as it is shown that there exists a correlation.

Final Model

To obtain our final model we remove variables one at a time based on the highest p-value until we see the model not improving in terms of Adjusted R-Squared and if the model begins to perform worse we then stop the elimination

Our final model summary is above in Table 2.

Interpretation

We have an adjusted R-Squared of 0.233 which means our model is able to explain 23.3% of the residuals in this dataset. All predictor variables have a low P-Value and P-Value of our entire model is extremely low meaning we reject the null and we have reason to believe the model created does have correlation to the Systolic Blood Pressure.

Our interpretation of the model is the coefficients that are produced are made into a formula where the values of the predictors can be plugged in and the systolic blood pressure can predicted to the level of efficiency of Multiple R-Squared.

Linear Regression formula:

$$log(SystolicBloodPressure) = 1.902{+}0.022[Gender(1ifMale, 0ifFemale)]{+}0.0014[Age]{+}0.0282[Race[1ifBlack, 0ifnot)]+ 0.0406[Race(1ifHispanic, 0ifnot)] + 0.0254[Race(1ifMexican, 0ifnot)] + 0.0262[Race(1ifWhite, 0ifnot)) - 0.0163[Race(1ifOther, 0ifnot)] - 0.0034[Poverty] + 0.0525[BMI] + 0.063[DirectCholesterol(mmol/L)]$$

The variables which were log transformed to maintain normality must be displayed in our formula. We plug in values of the predictor variables to get an answer for Systolic Blood Pressure.

Table 3: Predicting Values of Test Set

Type of Result	Train Set	Test Set
R-Squared	0.25	0.1849
Root Mean Square Error	15.00	16.0000
Mean Absolute Error	11.00	12.0000

Validate the Model

In order to properly assess the effectiveness of our model we must run our model on the test dataset we created and see how our prediction works.

We see that our correlation squared is slightly lower. Mean squared error and Mean absolute error are slightly higher but given the size of the test data set these are still within reasonable agreeance with eachother.

Discussion

What can we learn from the analysis? The main drivers in our predictions for systolic blood pressure are age, gender and cholesterol. This is something that could have been inferred prior to beginning as females and males have quite different physiological make-ups as well as age is the single greatest factor on the changing of our bodies. Cholesterol blocks arteries and therefore making it difficult for blood and oxygen to flow through. However what may be less intuitive is the impact that variables such as Race and Poverty. These variables made a difference in our prediction model and some races such as Hispanic have greater impact on our predictions.

More so than in terms of predictions what this model has allowed us to do is create a coefficient to each characteristic of an individual that is significant for Blood Pressure. By knowing the relationship between these variables medical staff can investigate possible future problems earlier on in a patients life and recommend proper preventative course of action.

Limitations of the Model

As mentioned earlier in the analysis there were some data distributions such as poverty which were not normal and could have violated normality assumptions. Considering a dataset of 708 observations we can consider this a small population metric and therefore data may be skewed or heavily in favor of a certain direction not indicative of the general population. Many statistics were taken on the basis of a reporting system and people who did not report or missed information could have skewed data sets.

Works Cited

Cronkleton, Emily. n.d. "The Biggest Causes of Death in 2020." https://www.medicalnewstoday.com/articles/death-statistics-by-cause-2020.

Disease Control, Centers for, and Prevention (CDC). n.d. "NHANES 2011-2012 Examination Data." Accessed 2011. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&CycleBeginYear=2011.

Health, Mount Sinai. n.d. "Hypertensive Heart Disease." Accessed 2022. https://www.mountsinai.org/health-library/diseases-conditions/hypertensive-heart-disease.

Staff, Mayo Clinic. n.d. "Blood Pressure Chart: What Your Reading Means." https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/blood-pressure/art-20050982.

Team, R Core. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.%20tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R.* https://yihui.org/knitr/.

Zhu, Hao. 2020. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax.* https://CRAN.R-%20project.org/package=kableExtra.