A Case Study on the Influence of Lifestyle Factors on Systolic Blood Pressure

2024-04-08

Authors

Adam Badar (1007965338)
Ashwin Mallik (1008329729)
Alex Zeng (1008064381)
Francis Ayyad (1008091985)

Background and Significance

Systolic Blood Pressure (SBP) has emerged as a significant medical concern across various age groups, posing serious health risks and mortality rates, particularly in adulthood (Theodore, R. F., et al., 2015). It is closely linked with cardiovascular health and is influenced by numerous factors, including age, weight, and height. Given the pressing medical issues associated with high blood pressure, our study aims to employ statistical models and inferences to identify the most significant predictors of SBP. By analyzing measurements from 500 patients, we seek to develop a model that accurately represents the relationship between SBP and various factors such as age, weight, height, and other lifestyle choices.

Understanding the determinants of SBP is crucial for guiding interventions and policies aimed at improving heart health. Lifestyle factors such as diet, exercise, and alcohol consumption play a significant role in influencing SBP levels. Through exploratory data analysis and model building, we aim to isolate the effects of these factors and their interactions on SBP. Our study's significance lies in its potential to highlight modifiable lifestyle factors that individuals can address to proactively improve their cardiovascular health.

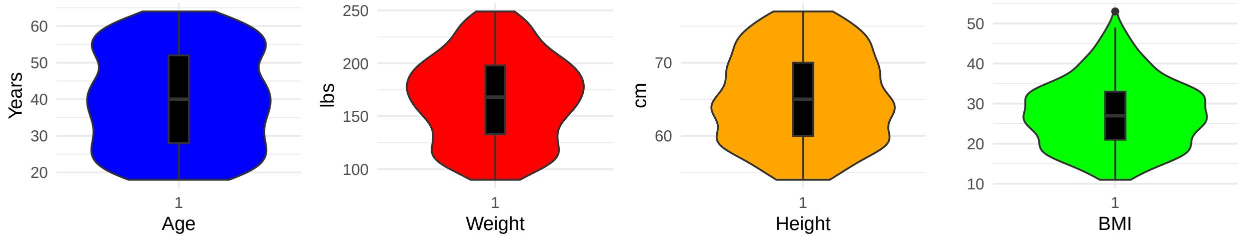
The dataset used in our analysis comprises various lifestyle and demographic variables related to individuals' choices and characteristics. Through descriptive statistics and graphical visualization, we explore the distribution of these variables and their effects on SBP. Model building involves constructing main effect and interaction models, refined using backward elimination and Akaike Information Criterion (AIC). Our analysis identifies significant interactions among categorical variables, indicating the nuanced effects of lifestyle factors across different demographic groups.

Model diagnostics, including Cook's Distance, help identify and remove outliers, enhancing the reliability of our models. Validation is performed using a training-testing split, with findings indicating the superior predictive power of the Interaction Model. Our study underscores the importance of considering both individual behaviors and demographic contexts in addressing SBP determinants.

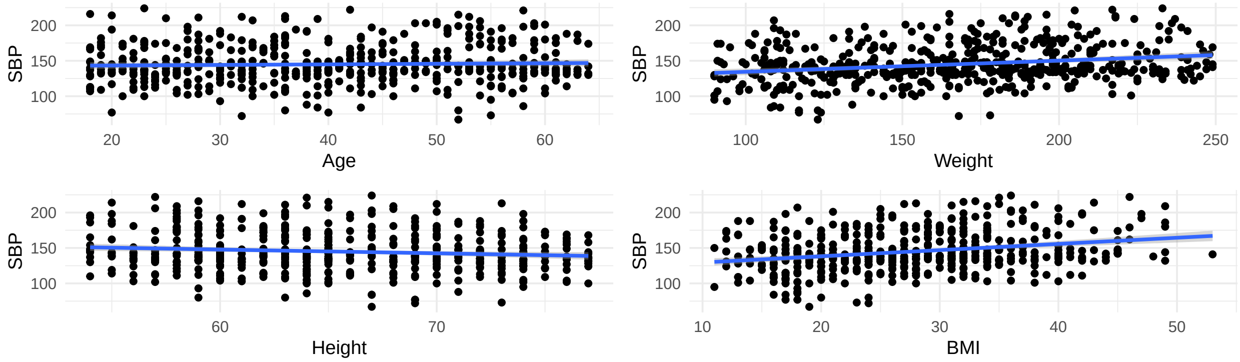
In conclusion, our study contributes to the understanding of SBP determinants and serves as a foundational step towards personalized healthcare interventions targeting modifiable lifestyle factors. While our analysis provides valuable insights, future research could explore genetic factors and additional lifestyle variables to further enhance our understanding of SBP dynamics.

Exploratory Data Analysis

To begin our investigation, we can start off by examining our given data set of 500 observations. The continuous variables we have are age, weight, height, and body mass index (BMI). Plotting out the violin chart for each variable's distribution, it is clear that these variables are mostly evenly distributed with the exception of BMI, which has a noticeably lower density towards the upper tail.



Upon further investigation using scatter plots with relation to the systolic blood pressure, it can be concluded that, holding all other factors constant, there exist a significant linear relationship between for weight, height, and BMI. This is explained by the incline in the fitted line for weight and BMI data points, and decline for height data points. Age is suspected to not have a significant linear relationship with SBP due to the flat fitted and can be confirmed to be the case by examining the summary of the model. With a P-value of 0.4032, we can conclude that we failed to reject the null hypothesis of the slope being zero, which suggests that there is indeed no significant relationship with SBP. In brief, age will not be a good predictor on its own as compared height, weight, and BMI.



```
##
## Call:
## lm(formula = SBP ~ Age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.883 -15.310  -4.804  17.496  80.404
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 141.78196    3.99072  35.528   <2e-16 ***
## Age           0.07886    0.09427   0.837    0.403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28 on 498 degrees of freedom
## Multiple R-squared:  0.001404,   Adjusted R-squared:  -0.0006017
## F-statistic: 0.6999 on 1 and 498 DF,  p-value: 0.4032
```

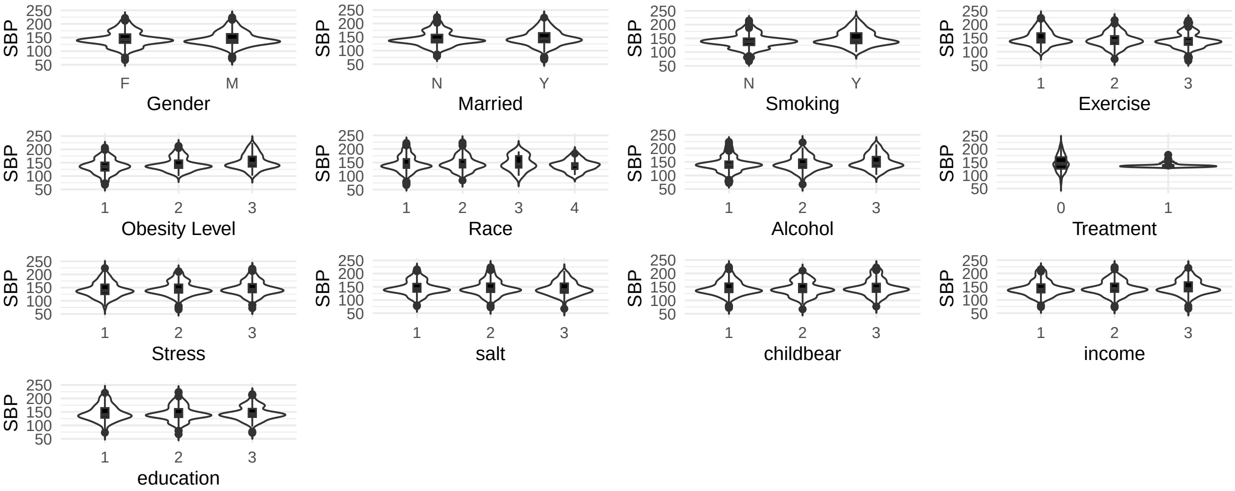
Exploring the categorical variables, we can compare the distributions across different categories for each variable to get a sense of how the SBP differs for each category of the same factor to determine if this variable is significant or remains constant throughout different categories. We will divide these variable into three categories from not influential to moderate to very influential.

The variables that seem to not have a significant effect of difference in distribution across different categories are gender, ability to bear children, levels of stress, levels of salt intake, and levels of income. This suggest that these candidates could potentially be dropped for the model training as it may not contribute as a significant factor when it comes to determining SBP.

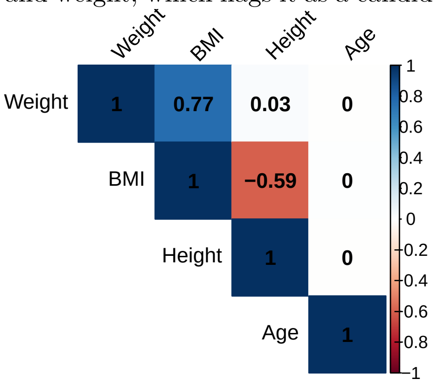
The variables that have a moderate change in distribution across different categories are being married, smoking levels, exercise levels, obesity level, alcohol intake levels, and levels of education.

The remaining variables of treatment for hypertension and race pose a high influence as it shows a significant change in distribution of SBP across different categories. Having received treatment for hypertension, the distribution of SBP is highly concentrated near the mean with an extremely low variance, while no treatment seems to have an inverse distribution as the variance is extremely high and data points are not concentrated at any point. The other factor of race shows a significant difference in categories 3 and 4 between 1 and 2, this not is surprising as race is a dominant genetic factor for SBP (Cené, C. W., et al., 2009).

```
grid.arrange(plot1, plot2, plot3, plot4, plot5, plot6, plot7, plot8, plot9, plot10,
    plot11, plot12, plot13, ncol = 4)
```



After obtaining some insight on the relationship between SBP and the factors, we can examine the relationship between all the factors by inspecting the correlation between them. First, we can use a heat map to view the correlation between our four continuous variables. It is shown that BMI is relatively correlated with both height and weight, which flags it as a candidate to be dropped for model training.



Furthermore, revealing the variance inflation factor of the obesity levels, we can see that it shows some moderate multicollinearity with height and weight. Which flags it as a candidate to be left out.

That is also the case with gender and the ability to bear children, it goes without saying that being a biological male, it would be impossible to bear a child. This means that there is a strong negative correlation between these two factors and one can be considered to be dropped

```
## Loading required package: carData
```

```
##                             GVIF Df GVIF^(1/(2*Df))
## weight                  3.724762  1        1.929964
## height                  2.424356  1        1.557034
## factor(data$Obesity) 5.115950  2        1.503944
```

Model

The group has decided that more than one model will be built, and a process of model selection will take place in order to find the more effective mode. To achieve this, the two models that will be tested upon will be a main effect model of all the cleaned data from the previous step (a model with no correlated variables), and a model with interaction.

Main Effect Model

To begin with, we considered a model with only interactions with the following structure:

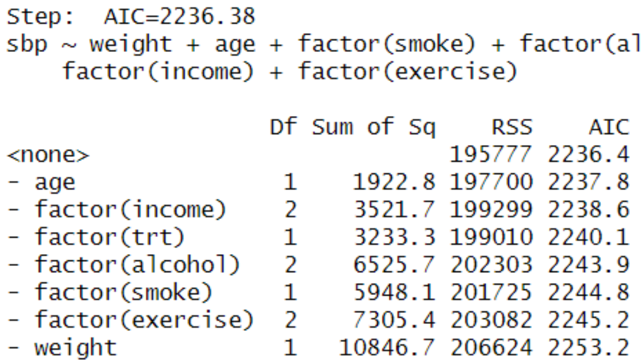
```r
full_additive <- lm(sbp ~ weight + height + age + factor(gender) + factor(married) +
    factor(smoke) + factor(race) + factor(alcohol) + factor(trt) + factor(stress) +
    factor(salt) + factor(income) + factor(educatn) + factor(exercise), data = data)
```

After identifying the model, we were able to run a summary of the model and gain some information from it.

```r
summary(full_additive)
```

```
##
## Call:
## lm(formula = sbp ~ weight + height + age + factor(gender) + factor(married) +
##     factor(smoke) + factor(race) + factor(alcohol) + factor(trt) +
##     factor(stress) + factor(salt) + factor(income) + factor(educatn) +
##     factor(exercise), data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -64.05 -14.05  -2.01  12.05  64.14
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       117.59852   15.97519   7.361 1.49e-12 ***
## weight              0.16010    0.03499   4.575 6.76e-06 ***
## height             -0.33220    0.22614  -1.469  0.14279
## age                 0.18280    0.09656   1.893  0.05921 .
## factor(gender)M    -1.19423    2.79404  -0.427  0.66935
## factor(married)Y    1.97841    2.69065   0.735  0.46269
## factor(smoke)Y      7.05890    2.67172   2.642  0.00864 **
## factor(race)2      -2.89404    3.41899  -0.846  0.39792
## factor(race)3       7.84772    5.91544   1.327  0.18555
## factor(race)4     -11.64628    7.62725  -1.527  0.12775
## factor(alcohol)2    0.29977    3.17422   0.094  0.92482
## factor(alcohol)3    9.84969    3.30310   2.982  0.00308 **
## factor(trt)1       -7.33715    3.21952  -2.279  0.02331 *
## factor(stress)2    -0.89377    3.21991  -0.278  0.78151
## factor(stress)3     4.33353    3.23761   1.338  0.18166
## factor(salt)2       2.06239    3.31582   0.622  0.53438
## factor(salt)3       3.76182    3.22145   1.168  0.24376
## factor(income)2     1.93544    3.19341   0.606  0.54489
## factor(income)3     7.73242    3.19092   2.423  0.01592 *
## factor(educatn)2    3.57300    3.19025   1.120  0.26355
## factor(educatn)3    2.66460    3.20777   0.831  0.40677
## factor(exercise)2  -9.92764    3.32205  -2.988  0.00302 **
## factor(exercise)3  -9.94157    3.09287  -3.214  0.00144 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24 on 327 degrees of freedom
## Multiple R-squared:  0.1732, Adjusted R-squared:  0.1176
## F-statistic: 3.114 on 22 and 327 DF,  p-value: 5.492e-06
```

Next as part of an attempt to build a better model, a model based AIC approach was used, using the backward elimination algorithm.

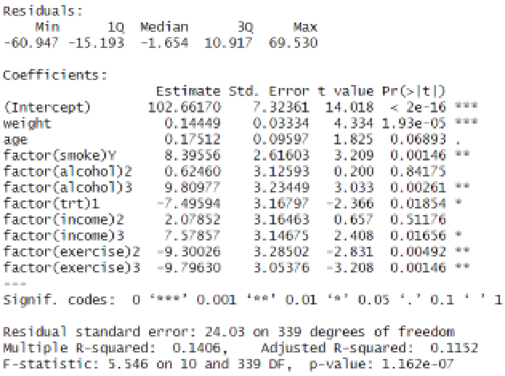


```
Step:  AIC=2236.38
sbp ~ weight + age + factor(smoke) + factor(alcohol) + factor(trt) +
    factor(income) + factor(exercise)

                   Df Sum of Sq    RSS    AIC
<none>                          195777 2236.4
- age               1    1922.8 197700 2237.8
- factor(income)    2    3521.7 199299 2238.6
- factor(trt)       1    3233.3 199010 2240.1
- factor(alcohol)   2    6525.7 202303 2243.9
- factor(smoke)     1    5948.1 201725 2244.8
- factor(exercise)  2    7305.4 203082 2245.2
- weight            1   10846.7 206624 2253.2
```

And based on the result of the AIC method we were able to find a reduced additive model with the following form with the following summary:

```
reducedAdditive <- lm(sbp ~ weight + age + factor(smoke) + factor(alcohol) + factor(trt) +
    factor(income) + factor(exercise), data = data)
```



```
Residuals:
    Min      1Q  Median      3Q     Max
-60.947 -15.193  -1.654  10.917  69.530

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       102.66170    7.32361  14.018  < 2e-16 ***
weight              0.14449    0.03334   4.334 1.93e-05 ***
age                 0.17512    0.09597   1.825  0.06893 .
factor(smoke)Y      8.39556    2.61603   3.209  0.00146 **
factor(alcohol)2    0.62460    3.12593   0.200  0.84175
factor(alcohol)3    9.80977    3.23449   3.033  0.00261 **
factor(trt)1       -7.49594    3.16797  -2.366  0.01854 *
factor(income)2     2.07852    3.16463   0.657  0.51176
factor(income)3     7.57857    3.14675   2.408  0.01656 *
factor(exercise)2  -9.30026    3.28502  -2.831  0.00492 **
factor(exercise)3  -9.79630    3.05376  -3.208  0.00146 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.03 on 339 degrees of freedom
Multiple R-squared:  0.1406,    Adjusted R-squared:  0.1152
F-statistic: 5.546 on 10 and 339 DF,  p-value: 1.162e-07
```

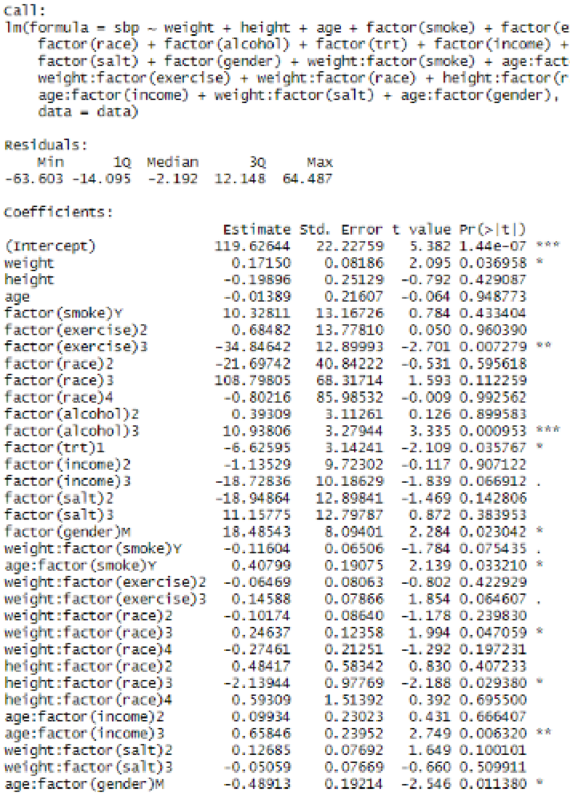
As it can be observed the reduced model has a lower R^2 value however a lower AIC. This lead the group to think that there might some interaction terms that would make the model much more powerful.

Interaction Model

When observing the data we believed that age, weight and height could have interactions with other variables, and effects on it. We saw that these continuous variables, when interacting with the categorical variables may allow the model to grow in predictive power. Thus our interactive model looked as follows:

```
interactive_model <- lm(sbp ~ weight + height + age + +factor(married) + factor(married) *
    weight + factor(married) * height + factor(married) * age + factor(married) +
    factor(smoke) + factor(smoke) * weight + factor(smoke) * height + factor(smoke) *
    age + factor(smoke) + factor(exercise) + factor(exercise) * weight + factor(exercise) *
    height + factor(exercise) * age + factor(race) + factor(race) * weight + factor(race) *
    height + factor(race) * age + factor(race) + factor(alcohol) + factor(alcohol) *
    weight + factor(alcohol) * height + factor(alcohol) * age + factor(trt) + factor(trt) *
    weight + factor(trt) * height + factor(trt) * age + factor(income) + factor(income) *
    weight + factor(income) * height + factor(income) * age + factor(educatn) + factor(educatn) *
    weight + factor(educatn) * height + factor(educatn) * age + factor(salt) + factor(salt) *
    weight + factor(salt) * height + factor(salt) * age + factor(gender) + factor(gender) *
    weight + factor(gender) * height + factor(gender) * age, data = data)
```

Following the same steps as before, a model based AIC approach was taken where the backward elimination algorithm was being implemented. Here is the output of the following algorithm:



```
Call:
lm(formula = sbp ~ weight + height + age + factor(smoke) + factor(exercise) +
    factor(race) + factor(alcohol) + factor(trt) + factor(income) +
    factor(salt) + factor(gender) + weight:factor(smoke) + age:factor(smoke) +
    weight:factor(exercise) + weight:factor(race) + height:factor(race) +
    age:factor(income) + weight:factor(salt) + age:factor(gender),
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.603 -14.095  -2.192  12.148  64.487 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              119.62644   22.22759   5.382 1.44e-07 ***
weight                     0.17150    0.08186   2.095 0.036958 *  
height                    -0.19896    0.25129  -0.792 0.429087    
age                       -0.01389    0.21607  -0.064 0.948773    
factor(smoke)Y            10.32811   13.16726   0.784 0.433404    
factor(exercise)2          0.68482   13.77810   0.050 0.960390    
factor(exercise)3        -34.84642   12.89993  -2.701 0.007279 ** 
factor(race)2            -21.69742   40.84222  -0.531 0.595618    
factor(race)3            108.79805   68.31714   1.593 0.112259    
factor(race)4             -0.80216   85.98532  -0.009 0.992562    
factor(alcohol)2           0.39309    3.11261   0.126 0.899583    
factor(alcohol)3          10.93806    3.27944   3.335 0.000953 ***
factor(trt)1              -6.62595    3.14241  -2.109 0.035767 *  
factor(income)2           -1.13529    9.72302  -0.117 0.907122    
factor(income)3          -18.72836   10.18629  -1.839 0.066912 .  
factor(salt)2            -18.94864   12.89841  -1.469 0.142806    
factor(salt)3             11.15775   12.79787   0.872 0.383953    
factor(gender)M           18.48543    8.09401   2.284 0.023042 *  
weight:factor(smoke)Y     -0.11604    0.06506  -1.784 0.075435 .  
age:factor(smoke)Y         0.40799    0.19075   2.139 0.033210 *  
weight:factor(exercise)2  -0.06469    0.08063  -0.802 0.422929    
weight:factor(exercise)3   0.14588    0.07866   1.854 0.064607 .  
weight:factor(race)2      -0.10174    0.08640  -1.178 0.239830    
weight:factor(race)3       0.24637    0.12358   1.994 0.047059 *  
weight:factor(race)4      -0.27461    0.21251  -1.292 0.197231    
height:factor(race)2       0.48417    0.58342   0.830 0.407233    
height:factor(race)3      -2.13944    0.97769  -2.188 0.029380 *  
height:factor(race)4       0.59309    1.51392   0.392 0.695500    
age:factor(income)2        0.09934    0.23023   0.431 0.666407    
age:factor(income)3        0.65846    0.23952   2.749 0.006320 ** 
weight:factor(salt)2       0.12685    0.07692   1.649 0.100101    
weight:factor(salt)3      -0.05059    0.07669  -0.660 0.509911    
age:factor(gender)M       -0.48913    0.19214  -2.546 0.011380 *  
```

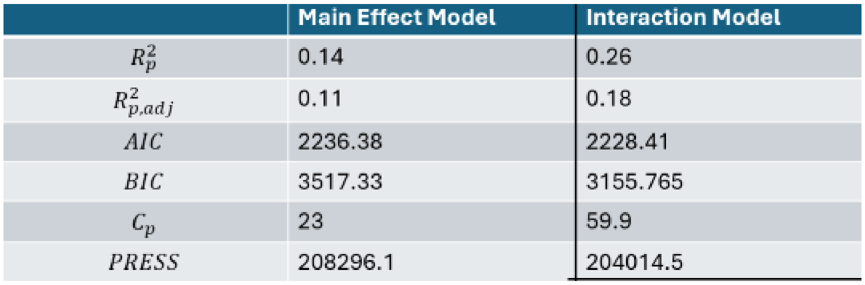
And thus the reduced model looked as follows:

```
modelAIC2 <- lm(sbp ~ weight + height + age + factor(smoke) + factor(exercise) +
    factor(race) + factor(alcohol) + factor(trt) + factor(income) + factor(salt) +
    factor(gender) + weight:factor(smoke) + age:factor(smoke) + weight:factor(exercise) +
    weight:factor(race) + height:factor(race) + age:factor(income) + weight:factor(salt) +
    age:factor(gender), data = data)
```

The model has now a the largest adjusted R-Squared so far and a higher R-squared than the main effect model. And thus the backtracking based AIC approach was beneficial in this situation, while it got rid of many of the factors reducing the number of features.

Model Selection

The next important step after building the following models would be to preform model selection, a process of identifying the model based on the following factors: AIC, BIC, adjusted R^2, multiple R^2 and PRESS. Below is a comprehensive table that summarizes the findings:



	Main Effect Model	Interaction Model
R_p^2	0.14	0.26
$R_{p,adj}^2$	0.11	0.18
AIC	2236.38	2228.41
BIC	3517.33	3155.765
C_p	23	59.9
$PRESS$	208296.1	204014.5

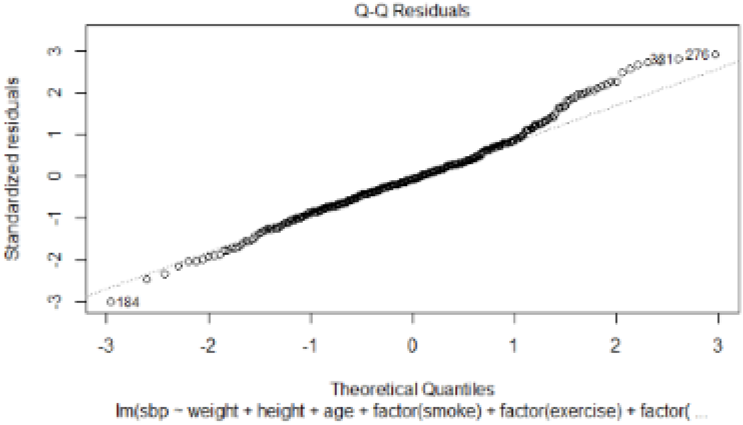
The result is clear, the interaction model has a higher multiple and adjusted R^2, a lower AIC and BIC, a C_p closest to the p' value, and a smaller PRESS. By preforming better on all the following criteria it becomes clear that the interaction model is the better preforming model.

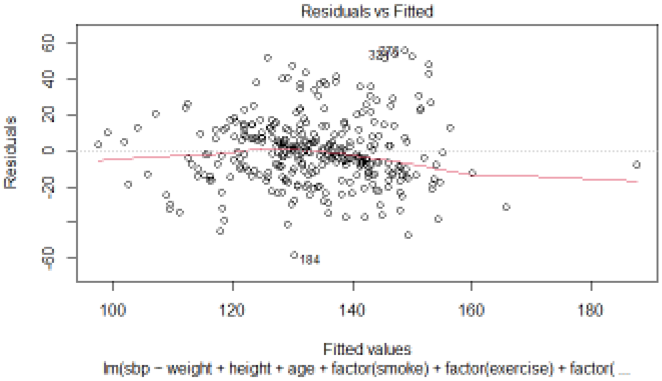
Influential Points

Cleaning the model is an essential process that involves finding outlines that may effect the model and deleting them from the model. To complete this process, we used cooks distance to identify influential points, and had a threshold value. If Cooks distance had a larger value than the threshold we would eliminate it from the model.

Diagnostics

After cleaning the data diagnostics were run on the model to check for model assumptions. The following graphs would give us the best feedback on the model assumptions:





The first graph showcases normality, while the residuals vs fitted graph showcases homoscedasticity with constant variance around the y = 0 line.

Model Validation

Finally validation as used to determine weather the model has similar performances on new training data. To accommodate for this, the models above were trained only on 70% of the data while the remaining 30%

where kept aside for validation. And upon running validation on the model, the standard error on the validation set was approximately 35 opposed to the 20 from our training set. While having larger error in the testing set makes sense, this value is extremely large in relation to blood pressure since in the context of blood pressure 35 units is a significant amount. This can be explained by the small predictive power of the model as described by the small R^2 value from the training model, and thus this model isn't a strong predictive model. However, in the context of the research, the model is able to identity factors that have the highest influence on the model from the provided data set.

Goal of the Study

The primary goal of this study was to identify significant factors of Systolic Blood Pressure (SBP), a critical health marker associated with numerous cardiovascular risks and mortality. Utilizing statistical models and inferences, this research aimed to elucidate the relationships between SBP and various demographic, behavioral, and biological factors in a sample of 500 patients.

Key Findings

Our exploratory data analysis revealed that among the continuous variables—age, weight, height, and BMI—weight, height, and BMI exhibited significant linear relationships with SBP, whereas age did not significantly predict SBP when considered independently. Further examination of categorical variables showed that treatment for hypertension and race were highly influential in predicting SBP variations, pointing to both genetic and lifestyle factors as key determinants.

The development and comparison of two statistical models—a main effect model and an interaction model—underscored the complexity of the relationships affecting SBP. The interaction model, which included terms for potential interactions between continuous and categorical variables, demonstrated superior performance over the main effect model across several statistical metrics, including adjusted R-squared and AIC values. This suggests that SBP prediction accuracy benefits significantly from considering the nuanced interplay between various factors.

Our analysis revealed that weight, height, and BMI are significantly associated with SBP, establishing them as key predictors. This finding suggests a clear link between physical characteristics and cardiovascular health, emphasizing the role of body composition in blood pressure regulation. Interestingly, age did not emerge as a significant predictor of SBP when analyzed independently. This outcome challenges the conventional understanding of aging as a dominant factor in the development of hypertension. It suggests that while age is undoubtedly related to increased SBP, its effects are possibly mediated through other factors or conditions that accrue or worsen with age, such as arterial stiffness, rather than age per se being a direct contributor.

The moderate changes in SBP distribution across categories of lifestyle factors—such as marital status, smoking habits, exercise frequency, alcohol intake, and educational levels—illuminate the multifactorial nature of blood pressure regulation. These findings highlight the potential of lifestyle modifications in SBP management and the importance of incorporating behavioral health interventions into comprehensive treatment plans.

Impact on the Field

Our findings contribute valuable insights into the multifaceted nature of SBP determinants, highlighting the importance of both genetic factors (such as race) and lifestyle factors (such as treatment for hypertension). By demonstrating the enhanced predictive power of models that incorporate interactions between variables, this study encourages a more nuanced approach to researching and understanding cardiovascular health risks.

Our study offers clinicians a more sophisticated framework for assessing cardiovascular risk. Understanding the interplay between weight, race, lifestyle factors, and treatment for hypertension can lead to more personalized patient care. For instance, the predictive value of weight and BMI on SBP emphasizes the importance of nutritional counseling and weight management in clinical settings. Similarly, recognizing the differential impact of race on SBP could guide more tailored screening and intervention strategies, acknowledging genetic susceptibilities and social determinants of health.

The limitations of our study, particularly the challenges in model validation, highlight the need for further research into additional variables and their interactions that may influence SBP. The pursuit of more comprehensive models necessitates large-scale, multi-center studies that include a wide range of demographic groups and more granular data on genetic factors, environmental exposures, and psychosocial stressors. Such research could unveil previously unrecognized predictors of SBP, offering new targets for intervention and prevention.

Limitations and Future Research

While our study has provided meaningful insights, it is not without limitations. The interaction model, despite its statistical robustness, showed a significant prediction error when validated against a subset of the data. This discrepancy underscores the model's limited predictive power and suggests that other unexamined variables or complex interactions might play a crucial role in SBP variation.

Furthermore, our study's focus on a specific sample population limits the generalizability of our findings across different demographics and geographical regions. The models developed here are based on available data and may not fully account for other influential factors such as genetic predispositions, environmental factors, or the impact of specific medications on SBP.

For future research, there is a clear need to explore additional variables and potentially more complex modeling approaches that can accommodate the multifactorial nature of SBP. Longitudinal studies might offer insights into how the relationships between SBP and its predictors evolve over time. Investigating the genetic basis of SBP variation, especially in the context of race and ethnicity, could also provide deeper understanding and lead to more personalized approaches to managing and predicting cardiovascular risk.

In conclusion, this study advances our understanding of the significant predictors of SBP, leveraging statistical models to highlight the complex interactions at play. The insights gained lay a foundation for future research, emphasizing the need for more nuanced models and broader datasets to enhance our predictive capabilities and ultimately inform better cardiovascular health interventions.

References

Theodore, R. F., Broadbent, J., Nagin, D., Ambler, A., Hogan, S., Ramrakha, S., Cutfield, W., Williams, M. J. A., Harrington, H., Moffitt, T. E., Caspi, A., Milne, B., & Poulton, R. (2015). Childhood to Early-Midlife Systolic Blood Pressure Trajectories: Early-Life Predictors, Effect Modifiers, and Adult Cardiovascular Outcomes. Hypertension (Dallas, Tex. 1979), 66(6), 1108–1115. https://doi.org/10.1161/HYPERTENSIONAHA.115.05831

Cené, C. W., Roter, D., Carson, K. A., Miller, E. R., & Cooper, L. A. (2009). The Effect of Patient Race and Blood Pressure Control on Patient-Physician Communication. Journal of General Internal Medicine : JGIM, 24(9), 1057–1064. https://doi.org/10.1007/s11606-009-1051-4