Analysis of Systolic Blood Pressure (SBP) and Insight Into The Factors that Affect it

Aly Abdelwahed, Manish Suresh

Name of the Data Set: Data Set 2 - Factors Affecting Systolic Blood Pressure (SBP)

Jobs Done By Each Member

Aly Abdelwahed

Job	Description
Introduction	Did the introduction section of the case study
References	Added the APA Citation at the end of the document
Description	Added the Description of the data Set section of the Case Study
Model Validation	Performed the validation on the model to make sure it is sound/effective
Model Diagnostics - Regression Assumptions	Performed the verification that the model satisfies all the regression assumptions
Limitations	Discussed the limitations of the model we chose on how we can improve it further
Conclusion	Gave the final conclusive statement of the case study

Manish Suresh

Job	Description
Data Cleaning	Performed the cleaning process of the data
Model Building	Built the models used in the case study
Model Selection	Performed model selection process for the final model used in the case study
Model Diagnostics - Influential Observations	Performed model diagnostics on the final model chosen

Introduction

Describe your research question/hypothesis/study aims

Our aim of this study is to prove that there is a relationship between the Systolic Blood Pressure (SBP) and the following:

- **Age**: As a person's age increases, his/her SBP increases **(Singh, 2020)**
- **Gender**: Males generally tend to have a higher SBP than females **(Reckelhoff, 2001)**
- **BMI**: BMI is positively associated with SBP. Weight loss means a lower BMI significantly reduces the SBP **(George C. Linderman, 2018)**
- **Stress**: Stressful situations can cause your blood pressure to spike **(Mayo Clinic Staff, 2021)**. So the higher the stress level, the higher the SBP. And the lower the stress level, the lower the SBP
- **Salt Intake Level**: There is a direct correlation between salt and blood pressure: the consumption of salt raises blood pressure **(McCabe, 2019)**. So the more salt intake a person gets, the higher his/her SBP becomes, and the lower the salt intake a person gets, the lower his/her SBP becomes
- **Smoking**: The nicotine in tobacco products affects the blood vessels in a way which makes a person's blood pressure get higher which means the SBP gets higher too **(American Family Physician, 2004)**. So in summary, a smoking person has a higher blood pressure and as such a higher SBP than a non-smoking person
- **Alcohol Use**: Drinking too much alcohol can raise one's blood pressure **(American Heart Association, 2016)**. So the higher the alcohol level in a person, the higher his/her blood pressure becomes which means the higher his/her SBP becomes, and the lower the alcohol levels in a person, the lower the person's blood pressure becomes and which his/her SBP becomes lower as well
- **Race**: People from different races have different blood pressures at matched ages **(Lackland, 2014)**
- **Treatment (for hypertension)**: There are treatments that can reduce a person's blood pressure **(American Heart Association, 2017)**
- **Exercise**: Exercise helps in reducing a person's blood pressure which means the SBP also gets reduced the more a person exercises because the more exercise a person does, the more weight he/she burns. Similarly, lack of exercise contributes to an increase in the person's blood pressure which his/her SBP increases with the lack of exercise **(Mayo Clinic, 2021)**

Brief background about the topic

The topic of this research is that a person's **Blood pressure** is measured using two values: The first value is called the **systolic blood pressure (SBP)**, it measures the **pressure** in a person's arteries when his/her heart beats. The second value is called **diastolic blood pressure (DSP)**, it measures the **pressure** in a person's arteries when his/her heart rests between beats. **(Centers for Disease Control and Prevention, 2020)** The goal of this research is to prove the effect of certain factors such as age, race, BMI, alcohol level...etc have on the **systolic blood pressure (SBP)** and how their variations can either decrease or increase a person's **systolic blood pressure (SBP)**.

Describe how your data was cleaned

First, we verified if there is any missing data. Afterwards, we verified that there are no duplicate records in the given data. Then, we checked if there are any corrupted data (e.g. data record of a person who comes from a different race like race 5 for example, or a stress level that does not belong to levels 1,2,3,....etc). Afterwards, we verified that the BMI values are correct for the given height and weight. Then, we changed the categorical values into dummy variables (i.e. converted some of the categorical data into usable format). Lastly, we checked multicollinearity between the data. And finally, we removed the columns of the data we wouldn't need

Brief description of what analyses you will conduct in the paper

Briefly, we will begin by checking some statistics if there are any . Afterwards, we begin the data cleaning process. Afterwards, we begin the model building and model selection process. Then, after choosing the model we will use, we will perform model validation and model diagnostics on it and verify that it satisfies all the regression assumptions. Afterwards, we will discuss the potential limitations of the model and lastly, we will come to the conclusion we get to.

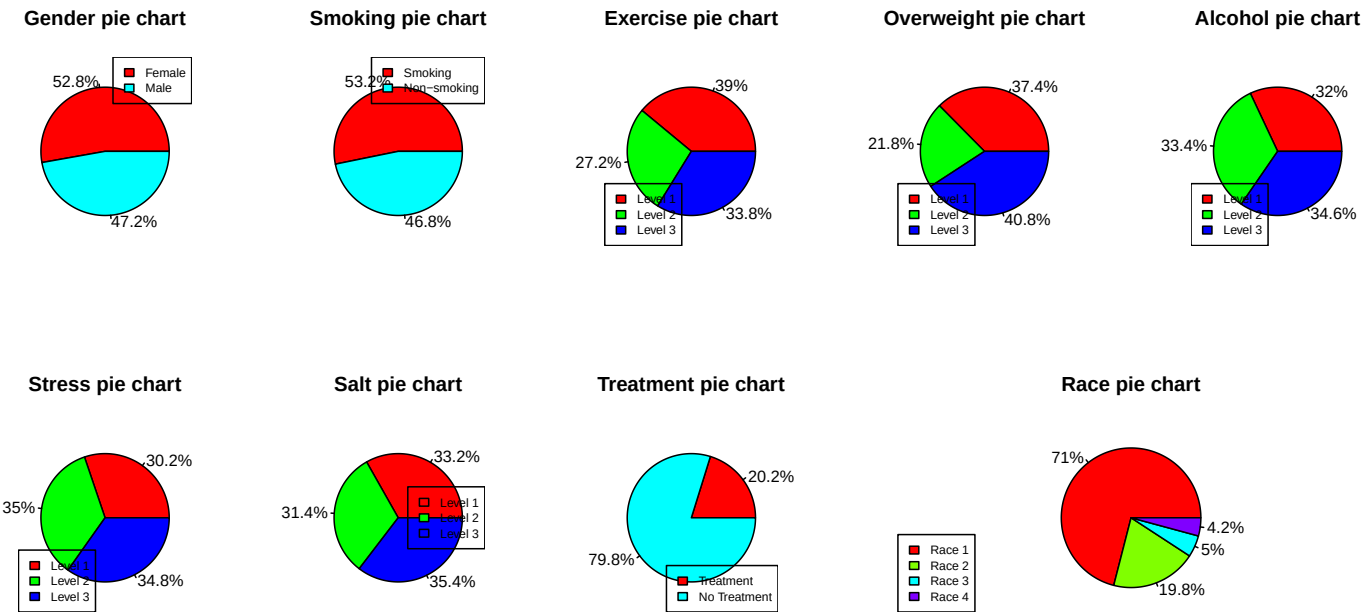
Explain why some variables will be left out of the model (if you decide not to use all the variables listed)

- **Height and Weight**: **Height** and **weight** are used in calculating the BMI because **BMI** is a measure of body fat based on **height and weight** that applies to adult men and women, hence using height and weight is not important in seeing how it affects a person's **SBP** since we are already using BMI as part of our calculation. In other words, their effect on a person's **SBP** is already considered through using **BMI** as part of our analysis
- **Overweight**: Similar situation to height and weight, it is already accounted for in our model with the inclusion of BMI
- **Martial Status, Education Level, income, and Childbearing Potential**: They do not provide enough information that can help in determining a person's SBP

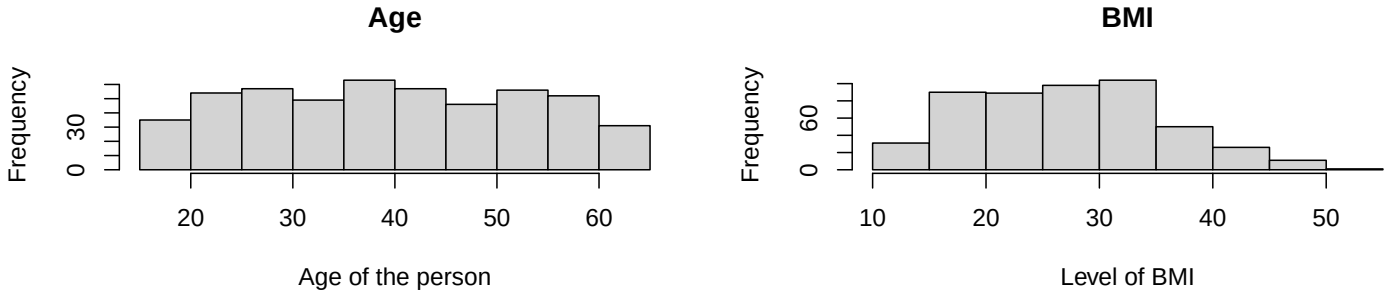
Description of the Data Set

Include descriptive statistics for each one of the variables that you will be analyzing

```
bloodPressureData <- read_excel("BloodPressure.xlsx")
actual.bloodPressureData <- bloodPressureData
bloodPressureData %>%
  mutate(married = ifelse(married == "N", 0, 1)) %>%
  mutate(gender = ifelse(gender == "F", 0, 1)) %>%
  mutate(smoke = ifelse(smoke == "N", 0, 1)) -> actual.bloodPressureData
```

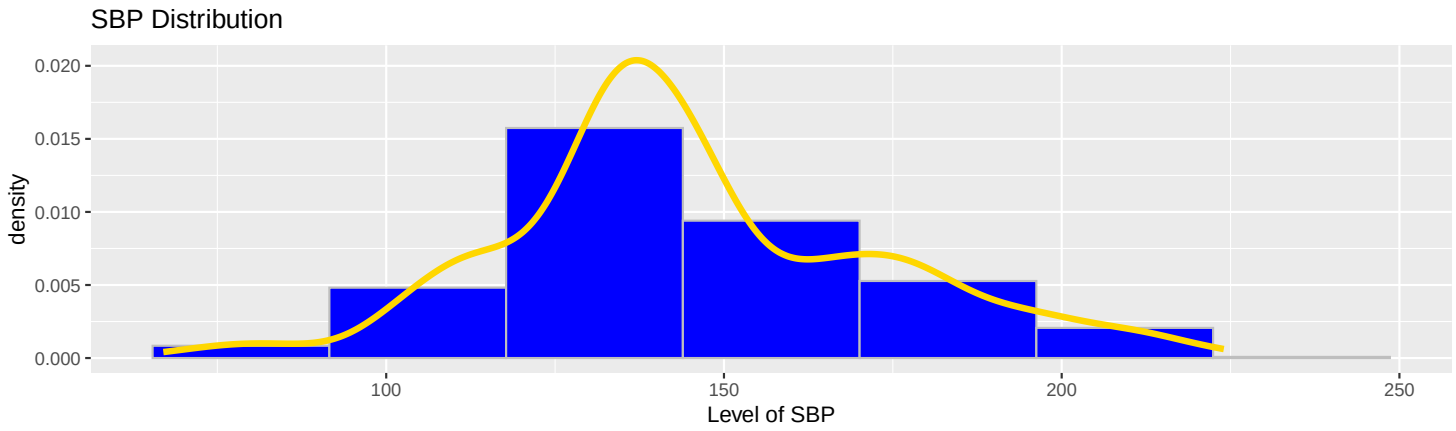


As we can see, almost all the pie charts look evenly split, except for the Race and Treatment pie charts.



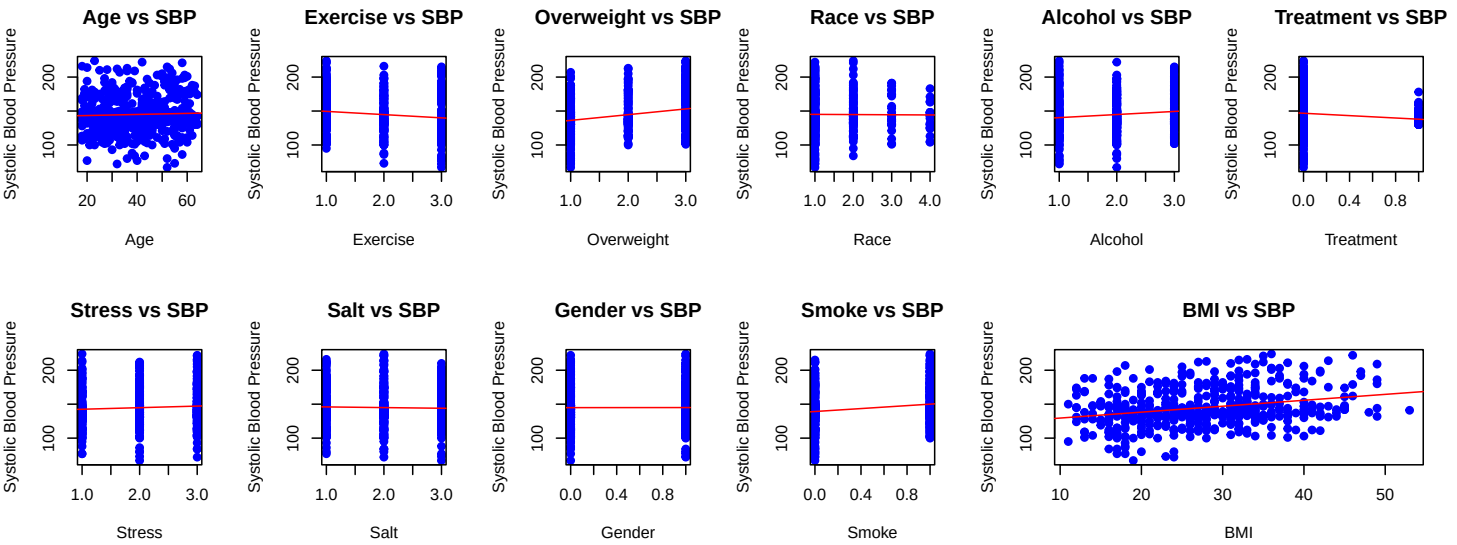
As we can see from the two histograms, both age and BMI are normally distributed but BMI is slightly skewed to the right

— **The distribution of the response variable**



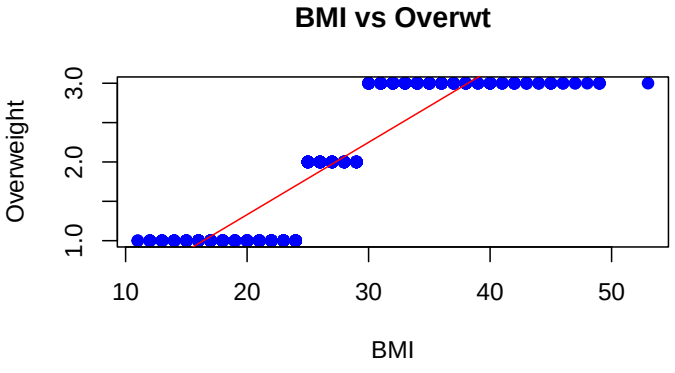
As we can see from the graph above, the response variable is normally distributed.

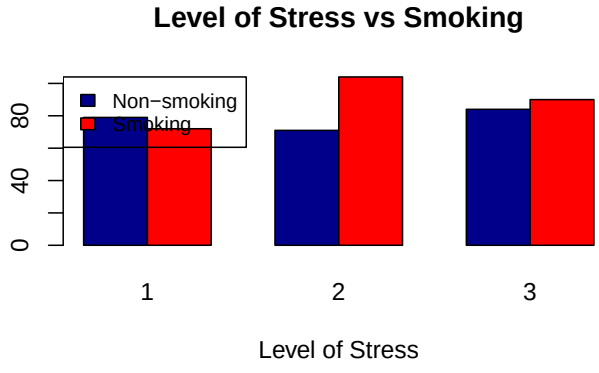
— **The relationship between the response variable and the exploratory variables**



We can see from the above that there is a direct linear relationship between a person's SBP and each of Stress Level, BMI, Alcohol Level, and whether a person is smoking or not.

— **The relationship between the exploratory variables themselves**





As we can see from these plots, we notice that the higher BMI means the more overweight a person is. We also notice that a person's stress level increases, the amount of people smoking increases as well.

Data Cleaning

1. Check if there is any missing data

There is no missing data in the model, hence we can move forward with further cleaning of the data

2. Check if there is any duplicate data

As we can see, there are no duplicate records, hence we can move forward with further cleaning of the data

3. Check if there is any corrupted data

Basically what we did is check whether the categorical variables are as identified as mentioned in the sbp data description.

4. Check if the BMI is correct The BMI formula is $\frac{Weight}{Height^2} * 703$. So lets confirm whether the data we have been given is correct

Hence, the BMI numbers are correct

5. We need to convert the data into usable format and removed the unnecessary columns

```
bloodPressureData %>%
  mutate(gender = ifelse(gender == "F", 0, 1)) %>%
  mutate(married = ifelse(married == "N", 0, 1)) %>%
  mutate(smoke = ifelse(smoke == "N", 0, 1)) %>%

  dplyr::select(sbp:age,race:salt) -> cleaned.bloodPressureData
```

6. Checking multicollinearity

```
correlations<-cor(cleaned.bloodPressureData)
correlations[correlations < 0.6 | correlations ==1] <- ""
correlations
```

```
##          sbp gender married smoke exercise age race alcohol trt bmi stress salt
## sbp      ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
## gender   ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
## married  ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
## smoke    ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
## exercise ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
## age      ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
## race     ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
## alcohol  ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
## trt      ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
## bmi      ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
## stress   ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
## salt     ""  ""     ""      ""    ""       ""  ""   ""      ""  ""  ""     ""
```

As we can see from the correlation matrix above, every " " sign indicates that the correlation between our predictors is less than 0.6, which means there is no sign of multicollinearity.

Building Model and Model Validation

Build the best multiple regressions model that you can find

Now, we split our data into a training set and a test set

```
set.seed(1004269365)
samples <- sample(1:500, 400, replace = FALSE)
training.bloodPressureData <- cleaned.bloodPressureData[samples,]
validate.bloodPressureData <- cleaned.bloodPressureData[-samples,]
```

Since we are dealing with a lot of categorical predictors, we fit an additive as follows:

```
## # A tibble: 6 x 5
##   term              estimate std.error statistic  p.value
##   <chr>                <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)       110.       7.45      14.8   2.72e-39
## 2 bmi                 0.914    0.157      5.82  1.24e- 8
## 3 age                 0.105    0.0978     1.08  2.82e- 1
## 4 factor(gender)1     0.770    2.63       0.293 7.70e- 1
## 5 factor(married)1    1.14     2.64       0.433 6.66e- 1
## 6 factor(smoke)1     11.4      2.64       4.30  2.17e- 5
```

As we can see from the model above, the additive model improves the adjusted R^2 from 0 to 0.1778739

Now we would like to improve this model by running backwards regression

```
stepAIC(fit.additive, direction = "both")
```

```
fit.add.optimal <- lm(formula = sbp ~ bmi + factor(smoke) + factor(exercise) + factor(alcohol) + factor(trt), d

head(tidy(fit.add.optimal))
```

```
## # A tibble: 6 x 5
##   term               estimate std.error statistic  p.value
##   <chr>                 <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)         119.        5.06    23.6    2.82e-77
## 2 bmi                   0.889     0.155    5.72   2.10e- 8
## 3 factor(smoke)1       12.1       2.59     4.67   4.18e- 6
## 4 factor(exercise)2   -11.9       3.23    -3.69   2.56e- 4
## 5 factor(exercise)3    -9.87      3.00    -3.29   1.08e- 3
## 6 factor(alcohol)2      0.108     3.18     0.0341 9.73e- 1
```

```
adjustedRAdditiveOptimal <- summary(fit.add.optimal)$adj.r.squared
```

Now, as we can see, after optimizing and removing certain predictors, we improved our adjusted R^2 from 0.1778739 to 0.1812653

Now, we would like improve our model even further because both our multiple R^2 and adjusted R^2 are pretty low. As such, we will use interactions between the predictor variables and optimize it using backwards regression.

```
fit.interact <- lm(formula = sbp ~ bmi * factor(smoke) * factor(exercise) * factor(alcohol) * factor(trt), data

stepAIC(fit.add.optimal, direction = "both", scope = list(upper = fit.interact, lower = fit.add.optimal))
```

Now, from all the optimization we just performed we reached what we believe is a good model which is;

```
fit.interact.optimal <- lm(formula = sbp ~ bmi + factor(smoke) + factor(exercise) + factor(alcohol) + factor(tr

summary(fit.interact.optimal)
```

```
##
## Call:
## lm(formula = sbp ~ bmi + factor(smoke) + factor(exercise) + factor(alcohol) +
##     factor(trt) + bmi:factor(trt) + bmi:factor(alcohol) + factor(alcohol):factor(trt) +
##     factor(smoke):factor(trt), data = training.bloodPressureData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.036 -14.428  -0.712  15.020  64.983
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     93.9500     8.2926  11.329  < 2e-16 ***
## bmi                              1.7500     0.2925   5.983 5.01e-09 ***
## factor(smoke)1                  13.7799     2.8296   4.870 1.63e-06 ***
## factor(exercise)2              -11.6137     3.1966  -3.633 0.000318 ***
## factor(exercise)3              -10.4366     2.9387  -3.551 0.000431 ***
## factor(alcohol)2                22.5669    11.1547   2.023 0.043753 *
## factor(alcohol)3                43.2910    10.5607   4.099 5.06e-05 ***
## factor(trt)1                    29.0770    12.6448   2.300 0.022008 *
## bmi:factor(trt)1                -0.9777     0.3885  -2.517 0.012251 *
## bmi:factor(alcohol)2            -0.8005     0.3886  -2.060 0.040063 *
## bmi:factor(alcohol)3            -1.0246     0.3753  -2.730 0.006615 **
## factor(alcohol)2:factor(trt)1   -2.8878     8.3952  -0.344 0.731053
## factor(alcohol)3:factor(trt)1 -16.3190     7.5403  -2.164 0.031059 *
## factor(smoke)1:factor(trt)1   -11.5970     6.3535  -1.825 0.068729 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.95 on 386 degrees of freedom
## Multiple R-squared:  0.2455, Adjusted R-squared:  0.2201
## F-statistic: 9.663 on 13 and 386 DF,  p-value: < 2.2e-16
```

```
adjustedRInteractOptimal <- summary(fit.interact.optimal)$adj.r.squared
```

Now, as we can see, after optimizing and removing certain predictors, we improved our adjusted R^2 from 0.1812653 to 0.2201242

Interpretation of the model

From the estimates that are provided by the output summary of our final model above, we can interpret it as follows:

1. **BMI:** It plays a significant role and its' estimate suggests that every unit of increase of BMI, we get a 1.7500106 increase in the SBP
2. **Smoke:** It also plays a significant role and its' estimate suggests that the average sbp of a person smoking is 13.7799381 higher than a non-smoking person at any BMI
3. **Exercise:** It also plays a significant role and its' estimate suggests that on average a person whose exercises is 11.0251546 lower than a non-exercising person
4. **Alcohol:** It plays a significant role and its' estimate suggests that on average a 32.9289437 which is a very considerable change. Hence, don't drink alcohol, stay healthy
5. **(BMI)*(Alcohol):** This interaction term suggests that while alcohol increases our SBP, the slope of BMI reduces. In other words, a person who drinks alcohol, and controlling for other variables, with a single increase of a unit of BMI, only increases on 0.83745 on average

Clearly indicate your final (selected) regression equation based on your output from R

$SBP = 93.95 + 1.75(BMI) + 13.78(Smoke) - 11.61(Exercise2) - 10.44(Exercise3) + 22.57(Alcohol2) + 43.29(Alcohol3) + 29.08(trt) - 0.98(bmi)(trt) - 0.8(bmi)(Alcohol2) - 1.02(bmi)(Alcohol3) - 2.89(Alcohol2)(trt) - 16.32(Alcohol3)(trt) - 11.6(Smoke)(trt)$

Validate your final selected model

As we can see, the MSPR is 634.9406388 and the MSE is 622.633222. When comparing both values in terms of their ratios, we will see that $\frac{622.633222}{634.9406388} = 0.9806164$ which means they are very close and hence we can validate the model. This is also confirmed by performing cross validation:

```
pressStat <- PRESS(fit.interact.optimal)
pressStat
```

```
## [1] 254096.8
```

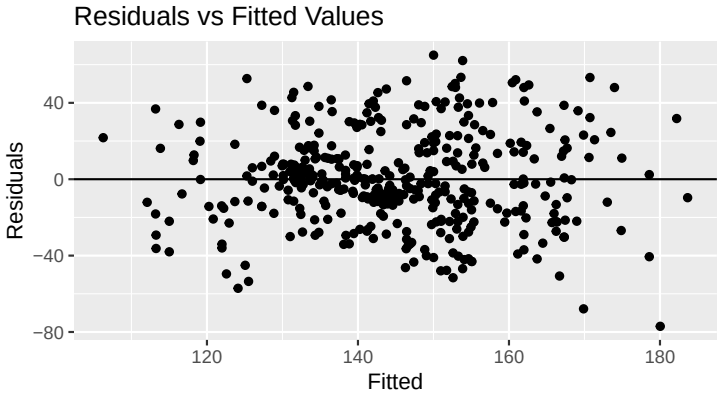
```
SSE <- anova(fit.interact.optimal)$`Sum Sq`[10]
anova(fit.interact.optimal)
```

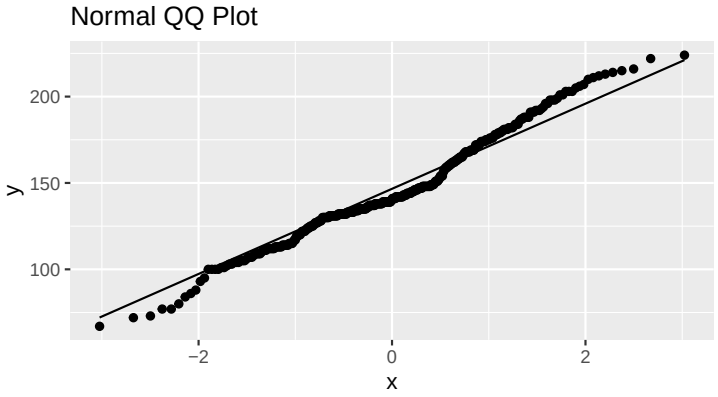
```
## Analysis of Variance Table
##
## Response: sbp
##                               Df Sum Sq Mean Sq F value    Pr(>F)
## bmi                            1  20045 20044.9 32.1938 2.742e-08 ***
## factor(smoke)                  1  11175 11174.5 17.9472 2.843e-05 ***
## factor(exercise)               2   9694  4847.0  7.7848 0.0004848 ***
## factor(alcohol)                2   9772  4885.9  7.8471 0.0004566 ***
## factor(trt)                    1  11633 11632.6 18.6830 1.967e-05 ***
## bmi:factor(trt)                1   5376  5376.2  8.6346 0.0034965 **
## bmi:factor(alcohol)            2   5438  2719.1  4.3671 0.0133214 *
## factor(alcohol):factor(trt)    2   3008  1504.2  2.4159 0.0906346 .
## factor(smoke):factor(trt)      1   2074  2074.4  3.3317 0.0687292 .
## Residuals                    386 240336   622.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we can see, the PRESS statistic's value is 240336.4 and the SSE of the model is 240336.4 which are also fairly close as their ratio is $\frac{240336.4}{254096.8} = 0.9458458$. Hence, we can further confirm the validation of our model using cross validation. Hence, we can confirm that our model is sound and effective for the purpose for which it was intended.

Model Diagnostics

Check if regression assumptions are valid using residual diagnostics & make sure to interpret what the plots mean in relation to your model

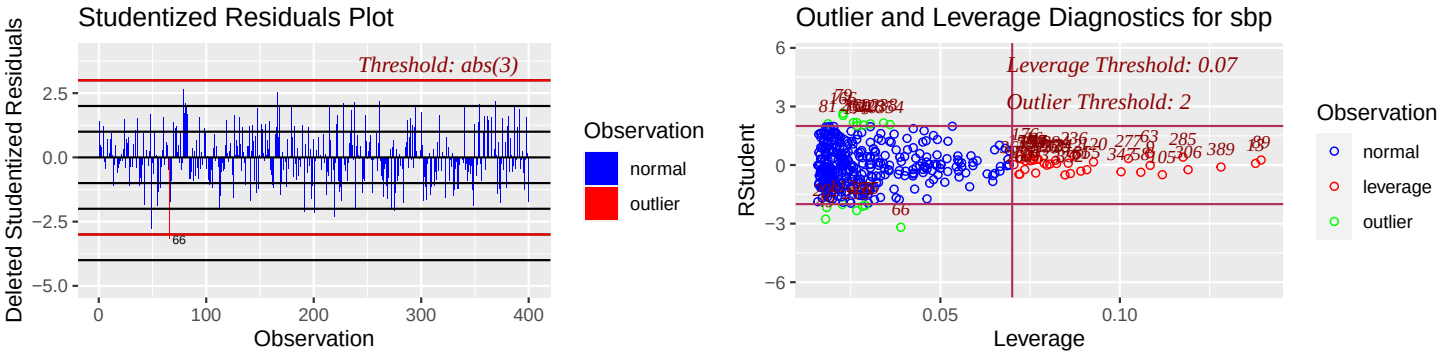




```
##
##  Shapiro-Wilk normality test
##
## data:  Residuals
## W = 0.9934, p-value = 0.07762
```

We can see from the summary of the model that $R^2 = 0.2455337$ which means that the predictors of this model explain almost 24% of the variance of our response variable **(SBP)**. Also, based on the residual plot, we cannot see any patterns, they are randomly scattered. Hence, the error variance seems to be constant, and the relationship is linear. In addition to that, Based on how closely the Normal Q-Q plot is forming a shape of a line y=x and the Shapiro-Wilk's test's p-value which is 0.07762 which is greater than 0.05 as required. Hence, we can conclude that the residuals are coming from the normal population. Thus all regression assumptions are satisfied.

Check if there any outlying and influential points



Studentized Deleted Residuals - Outlying Y observations:

As we can see, from the studentized residual plot, the threshold value is abs(3) which is fitted for large data sets. However, since we are working with medium data sets, the threshold is 3.8756131. So, as we can see, the output that we received shows that there are no outlying Y observations.

Leverage - Outlying X observations: As we can see, the output that we received from the Leverage shows that the values failed the guideline 1, but satisfy the guideline 2 which indicates that there are no outlying X observations.

Now since there are no outlying observations, therefore there are no influential observations either. Hence, our diagnostics is done!

Conclusion

Summarize findings

- Our analysis do indeed support our hypothesis in the sense that smoking, exercise, and alcohol level do indeed play a role in determining a person's SBP

Address limitations of your study

- Even though our research shows that people from different races have different blood pressures at matched ages, when looking at the distribution of the races, we see that the data is very biased towards people from Race 1. Hence, the model is not as good at predicting the SBP of people who come from races 2,3,4 as it is for predicting the SBP for people from Race 1. Hence is the reason why it has been excluded out of our model after performing optimization
- Our Multiple R^2 is 0.2455337, and our Adjusted R^2 is 0.2201242 which are both pretty low. In terms of Multiple R^2, this means that our model is not great at predicting the behavior of a person's SBP. Since adjusted R^2 depends on both Multiple R^2 and the number of predictors, all we can gather from the low adjusted R^2 is that our model is not good because we know almost all our predictors are significant

- The treatment predictors in our model do not make sense. As we know, the underlying reason for taking treatment is if the person's SBP is high which SHOULD reduce the SBP to a healthy level. However, in our model, treatment does the opposite. Controlling for all other predictors, treatment on it's own increases a person's SBP by 29.1 which does not make sense. In addition to that, controlling for all other predictors, the smoke predictor on its' own increases the person's SBP by 13.8 which makes sense. However, controlling for all other predictors, if a person does smoke and does take treatment then his SBP increases by 31.3 which is higher than if the person smokes and doesn't take treatment, which also does not make sense

References

- Singh, J. (2020, September 03). Physiology, blood pressure age related changes. Retrieved March 30, 2021, from https://www.ncbi.nlm.nih.gov/books/NBK537297/#
- Reckelhoff, J. F. (2001). Gender differences in the regulation of blood pressure. Hypertension, 37(5), 1199-1208. doi: 10.1161/01.hyp.37.5.1199
- George C. Linderman, B. (2018, August 17). Association of body mass index with blood pressure Among 1.7 million Chinese adults. Retrieved March 30, 2021, from https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2696872#
- Mayo Clinic Staff. (2021, March 18). Stress and high blood pressure: What's the connection? Retrieved March 30, 2021, from https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/stress-and-high-blood-pressure/art-20044190#
- McCabe, B. (2019, November 26). Can salt help improve low blood pressure? - Cardiac / heart HEALTH, FEATURED, health topics. Retrieved March 30, 2021, from https://www.hackensackmeridianhealth.org/HealthU/2019/11/22/can-salt-help-improve-low-blood-pressure/#
- American Heart Association. (2016, October 31). Limiting alcohol to manage high blood pressure. Retrieved March 30, 2021, from https://www.heart.org/en/health-topics/high-blood-pressure/changes-you-can-make-to-manage-high-blood-pressure/limiting-alcohol-to-manage-high-blood-pressure#
- American Heart Association. (2017, October 31). Types of blood pressure medications. Retrieved March 30, 2021, from https://www.heart.org/en/health-topics/high-blood-pressure/changes-you-can-make-to-manage-high-blood-pressure/types-of-blood-pressure-medications
- A.D.A.M. Inc. (n.d.). High blood pressure - medicine-related: Medlineplus medical encyclopedia. Retrieved March 30, 2021, from https://medlineplus.gov/ency/article/000155.htm
- American Family Physician. (2004, October 15). High blood pressure. Retrieved March 30, 2021, from https://www.aafp.org/afp/2004/1015/p1542.html#
- Lackland, D. (2014, August). Racial differences in hypertension: Implications for high blood pressure management. Retrieved March 30, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4108512/#
- Centers for Disease Control and Prevention. (2020, May 19). High blood pressure symptoms and causes. Retrieved March 30, 2021, from https://www.cdc.gov/bloodpressure/about.htm#
- Mayo Clinic. (2021, January 16). High blood pressure (hypertension). Retrieved April 02, 2021, from https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/symptoms-causes/syc-20373410