STAC67 Final Project

Mingjia Chen, Kuei-Sheng Hou, Raymond Moy, Ruichen Rachel Zhou

2023-04-10

Mingjia Chen: Evaluate influential and outlying observations, Abstract, Conclusion, and formatting final report and slides

Kuei-Sheng Hou: Data Cleaning, checked Multicollinearity before model selection, conduct model selection, and tailored the final report.

Raymond Moy: Examining correlations in data, checking LINE assumptions, model diagnostics and remedial measures.

Ruichen Rachel Zhou: Data cleaning and summarizing into tables, creating histograms, checking data validation

Significant and Influential Factors of Systolic Blood Pressure
Group 6

Library Used in Case Study:

```
library(dplyr)
library("GGally")
library("ggplot2")
library(ggpubr)
library(knitr)
library(leaps)
library(lmtest)
library(olsrr)
library("readxl")
library(tidyverse)
library(xtable)
library(corrplot)
library("MASS")
```

Background and Significance

Abstact

Systolic blood pressure is a medical measure that indicates how much pressure blood is exerting against artery walls when the heart beats. The measure is considered as a major risk factor for cardiovascular disease for people over fifty-five years old (American Heart Association, 2023). There may be numerous internal and external factors influencing the volume or level of systolic blood pressure. From our analysis, we have found these significant factors: smoking status, exercise level, height, alcohol use, treatment status, body mass index (BMI), interaction between smoke status and alcohol, and interaction between treatment and BMI.

This case study aims to better understand the relationship between various factors and the blood pressure by speculating and analyzing possible key predicting factors.

The research question is: What Factors Play Significant Role in Influencing Systolic Blood Pressure?

Variable description

- Gender: gender of the participant (Female = F, Male = M)
- Marital status: (Married = Y, Not Married = N)
- Smoking status: (Smoker = Y, Non-Smoker = N)
- Age: Age of the participants in years
- Weight: in lbs
- Height: in inches
- Body Mass Index (BMI) = (weight/height^2) * 703
- Overweight: Normal = 1, Overweight = 2, Obese = 3
- Race: 1, 2, 3, or 4 (Categorical)
- Exercise level: Low = 1, Medium = 2, High = 3
- Alcohol Use: Low = 1, Medium = 2, High = 3
- Stress level: Low = 1, Medium = 2, High = 3
- Salt (NaCl) Intake Level: Low = 1, Medium = 2, High = 3
- Childbearing Potential: Male = 1, Able Female = 2, Unable Female = 3
- Income Level: Low = 1, Medium = 2, High = 3
- Education Level: Low = 1, Medium = 2, High = 3
- Treatment (for hypertension): Treated = Y, Untreated = N
- Systolic Blood Pressure (SBP): continuous measure

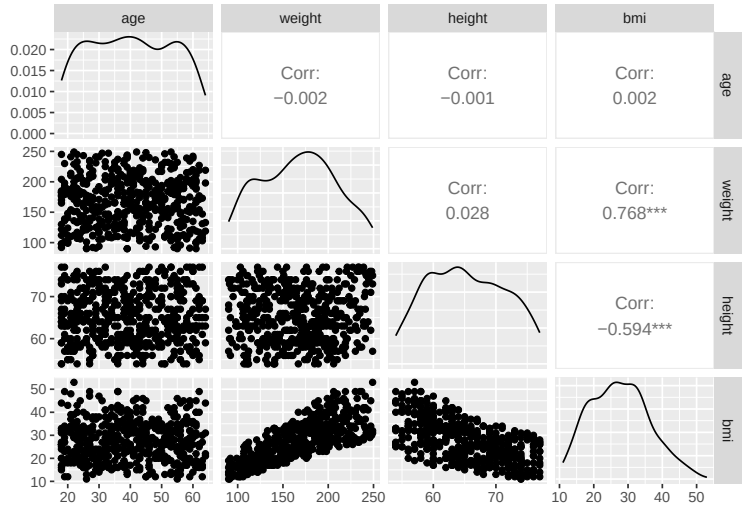
Split data

Let's use 70% of dataset as training set and 30% as testing set

```
sample = sample(seq_len(nrow(data)), size = floor(0.7 * nrow(data)), replace = FALSE)
data.train  <- data[sample, ]
data.test   <- data[-sample, ]
```

Exploratory Data Analysis / Data visualization

MultiCollinearity



From the correlation table, BMI with weight and height have the highest correlation and possible multi-collinearity. We also suspect that weight, height, and BMI have some correlation with categorical variable overweight status. In our final models, we only have height and BMI as those 2 predictors are sufficient in capturing the data of these 4 related variables. Additionally, BMI is a linear function of weight, but a non-linear function of height. Taking the height instead of weight predictor helps us avoid multicollinearity as much as possible.

Find best model

```
step(fit, direction = "backward")$anova
best_fit_backward = lm(sbp ~ smoke + exercise + height + alcohol + trt + bmi, data = data.train)
```

After testing, best model for backward direction is sbp ~ smoke + exercise + height + alcohol + trt + bmi
AIC = 2288.4

```
step(fit_simple, scope=list(upper = fit, lower = fit_simple),direction = "forward")$anova
best_fit_forward = lm(formula = sbp ~ bmi + exercise + smoke + trt + height + alcohol,
                      data = data.train)
```

After testing, best model for forward direction is sbp ~ bmi + exercise + smoke + trt + height + alcohol, which is the same model as backward direction. AIC = 2288.4

Check interaction in both direction

```
fit_inter <- lm(sbp ~ (smoke*exercise*height*alcohol*trt*bmi), data = data.train)
step(fit_inter, direction = "backward")$anova
best_fit_inter_backward = lm(sbp ~ smoke + exercise + height + alcohol + trt +
                               bmi + smoke:alcohol + trt:bmi, data = data.train)
```

AIC = 2277.34
Best model with interaction term is using backward selection: sbp ~ smoke + exercise + height + alcohol + trt + bmi + smoke:alcohol + trt:bmi

```
step(fit_simple, scope=list(upper = fit_inter, lower = fit_simple),direction = "forward")$anova
best_fit_inter_forward = lm(sbp ~ bmi + trt + smoke + exercise + alcohol +
                              height + bmi:trt + smoke:alcohol, data = data.train)
```

AIC=2277.34
Best model with interaction term is using forward selection: sbp ~ bmi + trt + smoke + exercise + alcohol + height + bmi:trt + smoke:alcohol which is the same model as backward selection

Test models with interaction and without

```
anova(best_fit_backward, best_fit_inter_backward, test = "Chisq")
```

```
## Analysis of Variance Table
##
## Model 1: sbp ~ smoke + exercise + height + alcohol + trt + bmi
## Model 2: sbp ~ smoke + exercise + height + alcohol + trt + bmi + smoke:alcohol +
##     trt:bmi
##   Res.Df    RSS Df Sum of Sq  Pr(>Chi)
## 1    341 229778
## 2    338 218831  3     10947 0.0007381 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
best_model = lm(sbp ~ bmi + trt + smoke + exercise + alcohol + height +
                  bmi:trt + smoke:alcohol, data = data.train)
```

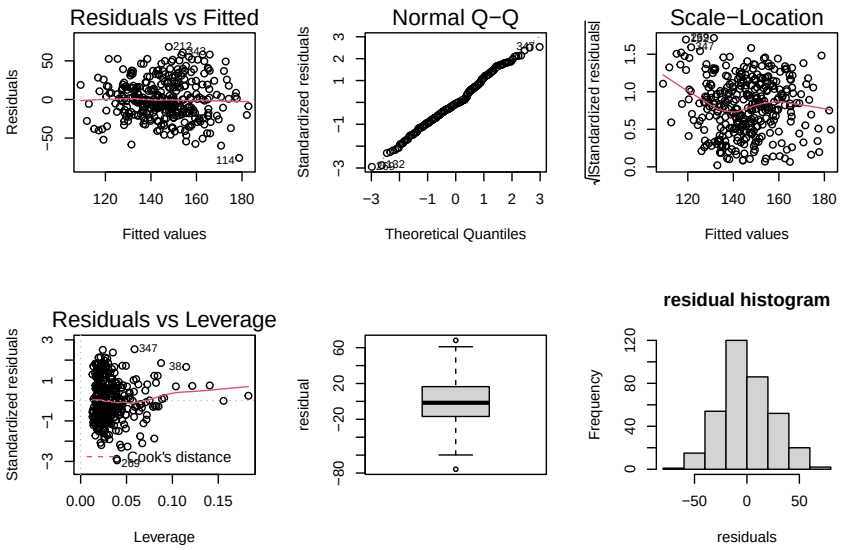
from the anova test, the complex model is better best_model = sbp ~ smoke + exercise + height + alcohol + trt + bmi + smoke:alcohol + trt:bmi

Checking LINE assumptions

- Residuals vs fitted plot: A band around 0 indicates the linearity assumption holds. The points also seem randomly scattered, suggesting the independent errors assumption also holds.
- QQplot is mostly aligned with the expected normal distribution. There are some outliers at the tails.
- 2 outlying leverage points
- histogram hist of semi-studentised resids looks vaguely normal, slightly skewed to the left
- boxplot shows one very apparent outlier
- Shapiro–Wilk test: p-value is > 0.05 we fail to reject the null hypothesis, would indicate that the resids are normal
- Scale-Location plot line is very bent at for fitted values below 150. Shows that variances of errors are not equal.
- Breusch-Pagan Test: reject the null hypothesis that the resid variances are equal we have evidence that the residual variances are NOT equal! Thus, we make a weighted least squares model.

We also tried to create the boxcox transformed model, but that did not make the variances of errors equal.

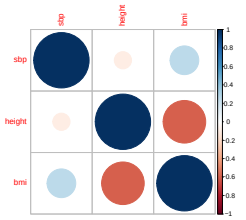
Weighted least squares

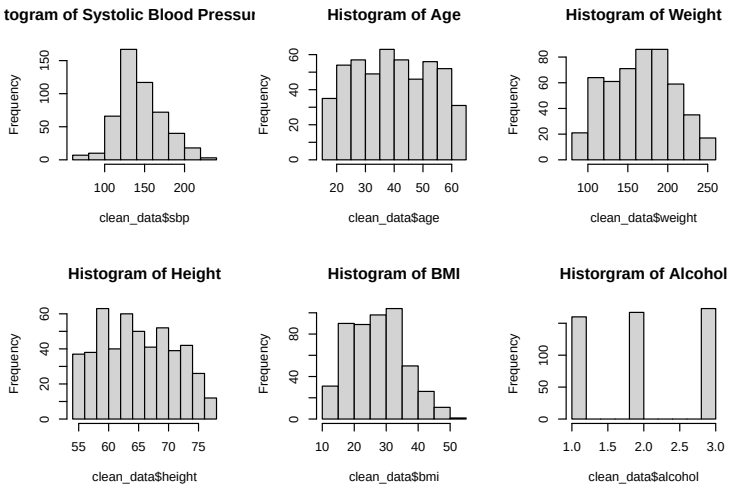


The weighted least squares model passed the BP-test and has somewhat more convincing model diagnostic plots than the OLS model.

MultiCollinearity

Correlation Matrix:





Graph for SBP-> slightly skewed, closest to normal distribution. No particular observation can be made for the remaining graphs.

By calculating the summary and sd from cleaned data, we obtain two summary tables:

Variables	Descriptions	Min	Q1	Median	Mean	Q3	Max	SD
sbp	Systolic Blood Pressure(SBP)	67	130	140.5	145	162.2	224	28
age	years	18	28	40	40	52	64	13.3
weight	lbs	90	133	168	166.6	198	249	40.9
height	inches	54	60	65	65.33	70	77	6.2
bmi	Body Mass Index(BMI)	11	21	27	27.66	33	53	8.6

Variables	Categories	n
Gender	Female	264
	Male	236
Smoking	yes	266
	no	234
Exercise	low	195
	medium	136
	high	169
Alcohol	low	160
	medium	167
	high	173
Treatment	yes	101
	no	399

```
fit1 <- lm(sbp ~ . , data = clean_data)
fit2 <- lm(sbp ~ smoking + exercise + alcohol + trt + bmi, data = clean_data)
full <- lm(sbp ~ . , data = data)
fit.null <- lm(sbp ~ 1, data = data)
model <- step(fit.null,direction = "forward", scope = list("lower" = fit.null, "upper" = full), trace =

# multiple regression model
AICs = c(AIC(fit1),AIC(fit2),AIC(model))
AICs
```

```
## [1] 4667.657 4667.885 4660.987
```

fit1 and fit2 have similar AIC values

Model Validation / Diagnostics

```
##### Obtain validation sets
set.seed(12345)
n = nrow(clean_data)
cv.samp <- sample(1:n, round(0.5*n),replace = FALSE)
cv.in <- clean_data[cv.samp,]
cv.out <- clean_data[-cv.samp,]
```

```
##### fit model for training set
fit.cv.in <- lm(sbp ~ smoking+exercise+alcohol+trt+bmi, data=cv.in)
anova(fit.cv.in)
```

```
## Analysis of Variance Table
##
## Response: sbp
##            Df Sum Sq Mean Sq F value    Pr(>F)
## smoking     1   7739  7738.8 11.4017 0.0008537 ***
## exercise    1   5792  5791.8  8.5332 0.0038141 **
## alcohol     1   1365  1365.0  2.0111 0.1574255
## trt         1   6342  6341.5  9.3431 0.0024872 **
## bmi         1  15729 15728.5 23.1733 2.598e-06 ***
## Residuals 244 165611   678.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##### Obtain Predicted  values and prediction errors for validation sample
##### Regression is based on same predictors
##### Compute MSPR
pred.cv.out <- predict(fit.cv.in, cv.out[,c(3,4,8,9,10)])
delta.cv.out <- clean_data[-cv.samp,]-pred.cv.out
n.star = dim(cv.out)[1]
MSPR <- sum(delta.cv.out)^2/n.star
MSPR
```

```
## [1] 255260891
```

```
#Fit Model on Validation Sample and Compare regression coefficients with model for Training Sample
fit.cv.out <- lm(sbp ~ smoking+exercise+alcohol+trt+bmi, data=cv.out)
anova(fit.cv.out)
```

```
## Analysis of Variance Table
##
## Response: sbp
##            Df Sum Sq Mean Sq F value    Pr(>F)
## smoking     1   6941  6940.5 10.9000 0.0011056 **
## exercise    1   3930  3930.4  6.1726 0.0136458 *
## alcohol     1   7845  7844.9 12.3204 0.0005334 ***
## trt         1   2970  2969.6  4.6637 0.0317808 *
## bmi         1  11446 11445.5 17.9751 3.178e-05 ***
## Residuals 244 155366   636.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	R^2_{adj}
Training	117.05 (8.45)	10.96 (3.38)	-5.85 (1.94)	4.76 (2.03)	-15.28 (4.17)	0.99 (0.21)	0.1657
Validation	114.32 (7.67)	10.66 (3.21)	-4.47 (1.89)	6.84 (2.00)	-9.57 (4.03)	0.77 (0.18)	0.1589

Since the regression results for the training and validation data sets are similar, we can conclude that the model is valid and we can make statistical reference based on the model

Outlying & Influential Points

```
influences = influence.measures(best_model)
# Y outlying points
best_model = lm(sbp ~ bmi + trt + smoke + exercise + alcohol + height + bmi:trt
                + smoke:alcohol, data = data)
# studentized deleted residuals
t = rstudent(best_model)
#standardized residuals
r = rstandard(best_model)
rt.table = cbind(r,t)

alpha = 0.05
n = dim(data)[1]
p.prime = length(coef(best_model))
t.crit = qt(1-alpha/(2*n), n - p.prime - 1)
t.crit
```

```
## [1] 3.922859
```

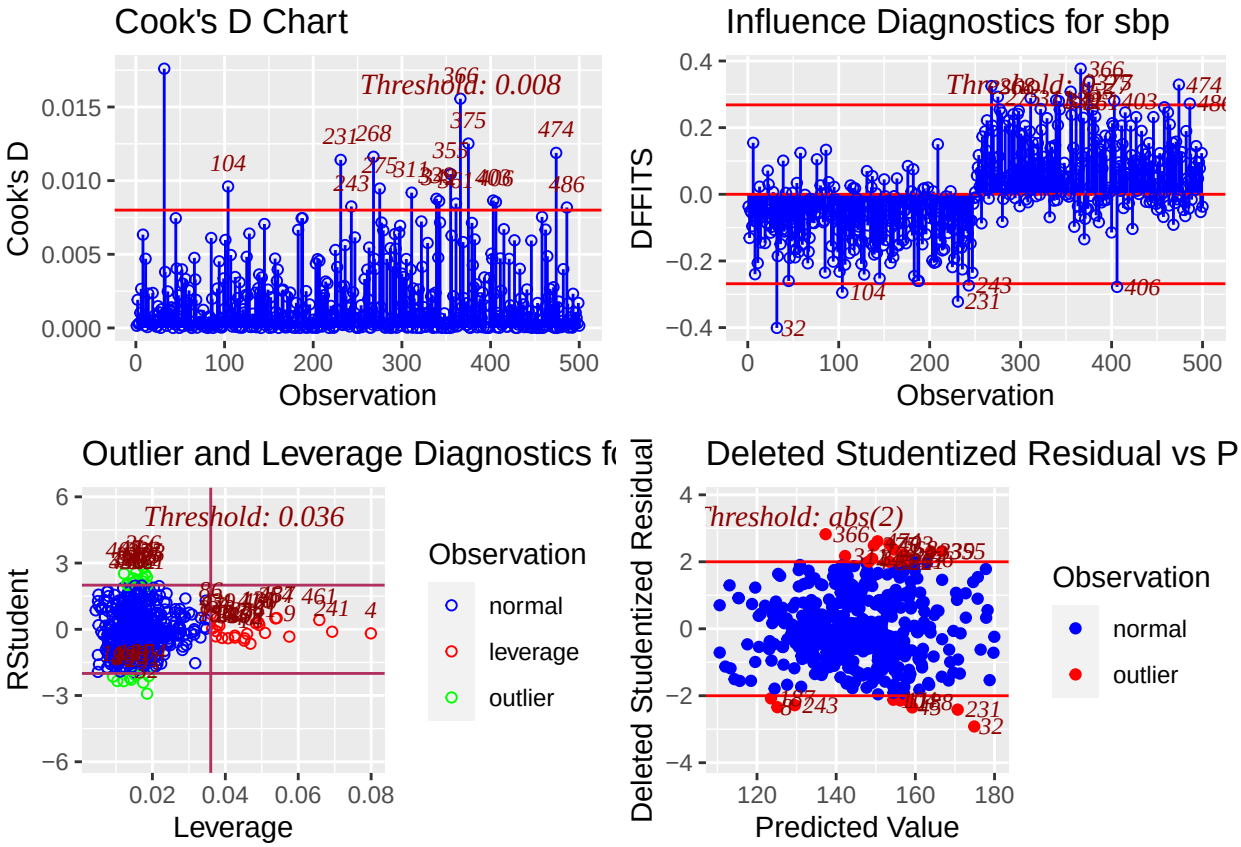
```
# X outlying points
hii = hatvalues(best_model)
which(hii > 2*p.prime/n)
which(hii > 0.05)
```

```
#Influential observations
DFFITTS = dffits(best_model)
#Cook's Distance
D = cooks.distance(best_model)

DFBETAS = dfbetas(best_model)
c(which(abs(t) > t.crit),which(DFFITTS >1),which(D>qf(0.2,p.prime,n-p.prime)),which(DFBETAS > 1))
```

```
## integer(0)
```

We got named integer(0) for result, meaning there are no outlier observations of Y according to studentized deleted residual.



According to BFFITS, D and DFBETAS, we do not have influential observations in the data.
Even though there are outliers of X observations, we decide to keep them as they are. Outliers are not too rare in real life scenarios. As long as there are no influential observations in the model, we will stick to the original X dataset. This may also help us avoid the problem of overfitting.

Discussion/Conclusion

We investigated measures that could play significant roles in predicting SBP. From the result of the case study, with a large dataset of 500 observations, we see the relationship between SBP and numerous contributing factors such as Smoking status, Exercise level, Height, Alcohol use, Treatment, Body Mass Index (BMI) and 2 interactions (smoking:alcohol, treatment:BMI). In conclusion, the study could provide insight on how we could improve our SBP and predict it using these measures that a family doctor could obtain. Despite of the strength of the study, there are some limitations as well. For instance, where the population dataset was from is unclear, we cannot know how accurate the data is. The study may face challenges representing any population in real life. During the study, outlying observations of X variables are detected. There could be confounding factors. For future directions, new samples could be obtained from different types of populations and cultures. We can also reuse the model for on datasets to see if the model is explaining real life scenarios in general or only specific to this case.

Reference

Understanding blood pressure readings. www.heart.org. (2023, February 2). Retrieved April 2, 2023, from https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings