# C67 Case Study Report - DETERMINANTS OF THE SYSTOLIC BLOOD PRESSURE

Group 18: Yunrong Liu(1005732735), Yulun Wu(1004912785), Yuanxin Su(1005754015)

2022-4-8

**Suprrvisor: Dr. Sohee Kang (sohee.kang@utoronto.ca)**
University of Toronto Scarborough
Toronto, Ontario, Canada
{ yunrong.liu, yulunbblythe.wu, yuanxin.su }@mail.utoronto.ca

**Yunrong Liu** (1005732735): Schedule the case study, help coding, Consulting the supervisor, write the report.
**Yulun Wu** (1004912785): Do the presentation of the case study, do the coding, writing the report.
**Yuanxin Su** (1005754015): Do the coding.

## RESEARCH BACKGROUND DESCRIPTION

Systolic blood pressure is the pressure exerted when the heartbeats and blood is ejected into the arteries. Researchers want to find out what determines systolic blood pressure. Researchers want to find out what determines the systolic blood pressure, so we can prevent diseases such as hypertension.

## GOAL OF STUDY

The aim of this study is to determine which factors have impact on systolic blood pressure (SBP). We will analyze the relationship between SBP and gender, marital status, smoking status, age, weight, height, Body Mass Index (BMI), overweight status, race, exercise level, alcohol use, stress level, salt (NaCl) intake level, childbearing potential, income level, education level, treatment (for hypertension) status. We will come up with a model to predict SBP given selected predictors.

## RESEARCH OBJECTIVE

Analyze the factors that determine the systolic blood pressure of the human body. Building a statistical model. Then optimal the model uses *AIC(1974)*[2]. Do the model diagnostic about outlying points, influential points, multicollinearity problems and regression assumptions. The final model should give an acceptable accuracy rate.

## DATA DESCRIPTION

Based on the dataset *Factors Affecting Systolic Blood Pressure (SBP)* [1]:

**Categorical Data:**

Gender: sex of the observee (M=Male, F=Female) Marital/Smoking Status: whether the observee is married/smokes

Overweight: whether the observee is overweight (1=Normal, 2=Overweight, 3=Obese)

Race: observee's race Exercise level: how much exercise the observee does (1=low, 2=medium, 3=high) Alcohol Use: alcohol drinking level of the observee (1=low, 2=medium, 3=high). Stress Level: the stress level of the observee (1=low, 2=medium, 3=high) Salt(NaCl) Intake Level: the level of salt intake of the observee in diet (1=low, 2=medium, 3=high) Childbearing Potential: whether the observee is capable to pregnant (1=Male, 2=Able Female, 3=Unable Female) Income/Education Level: the level of observee's income/education (1=low, 2=medium, 3=high)

Treatment: whether the is treated for hypertension

**Quantitative Data:**

SBP: Systolic Blood Pressure. This is response variable Age : age of the observee (in years) Weight : the observee's weight (in lbs) Height: the height of the observee (in inches) BMI : the Body Mass Index of the observee, is calculated by (weight/$height^2$) x 703

## GENERAL RESEARCH DESIGN

1. Cleaning the data. 2. Fit the regression model with main effects only and do validation. 3. Consider interaction and use *AIC(1974)*[2] method to find the optimal model. 4. Model Selection Decision. 5. Do the model diagnostic about outlying points, influential points, multicollinearity problem and regression assumptions. 6. Decide final Model. 7. Test the accuracy of our final model. 8. Discuss result, limitations and the future work can be done.

## STEP 1: CLEANING DATA

We need to clean the data before we start the actual model building process. We create dummy variables for 13 categorical variables in our dataset. We also separate the data into training set and validation set for the validation later on.

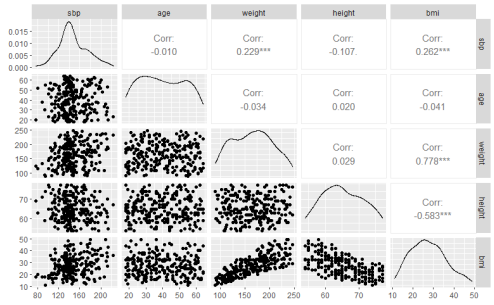
```
# Read Excel table
BloodPressure = read_excel("BloodPressure.xlsx")
```

```
##   mean.BloodPressure.sbp. mean.BloodPressure.height. mean.BloodPressure.weight.
## 1                 144.952                    65.334                     166.636
##   mean.BloodPressure.bmi. mean.BloodPressure.age.
## 1                  27.658                 40.196
```

```
##   sd.BloodPressure.sbp. sd.BloodPressure.height. sd.BloodPressure.weight.
## 1              27.99495                 6.191212                  40.90273
##   sd.BloodPressure.bmi. sd.BloodPressure.age.
## 1              8.559414             13.29854
```

We generate the Scatter plot and correlation matrix for response variable and quantitative variables in training

set. We want to see if we need to consider polynomial regression and if we can find any high correlation variable to drop before fitting the model.




There is a linear relationship between bmi and height, bmi and weight, but their correlations are -0.583 and 0.778 which are less than +-0.9, we don't regard them as highly correlated terms, so we keep them.
Response variable v.s. any quantitative variables don't show evidence of the polynomial relationship, so we don't need to consider polynomial regression in our model.

## STEP 2: FIT THE REGRESSION MODEL WITH THE MAIN EFFECT ONLY AND DO VALIDATION

First, we fit all main effects into the model, this is definitely not the best model, we will use stepwise elimination to eliminate main effects later.

We apply stepwise elimination to our model. This is our best model with main effects only.

```
##
## Call:
## lm(formula = sbp ~ smoke + exercise + height + alcohol + trt +
##     bmi, data = BloodTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.311 -16.846  -0.462  14.790  55.851
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.2316    24.8170   2.629 0.009127 **
## smokeY       10.3478     3.0436   3.400 0.000789 ***
## exercise2   -12.5962     3.8150  -3.302 0.001107 **
## exercise3    -7.7189     3.5599  -2.168 0.031116 *
## height        0.7125     0.3125   2.280 0.023482 *
## alcohol2      0.2737     3.7816   0.072 0.942355
## alcohol3     16.0933     3.8020   4.233 3.28e-05 ***
## trt1        -18.3465     3.6833  -4.981 1.21e-06 ***
## bmi           1.2453     0.2239   5.562 7.06e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.68 on 241 degrees of freedom
## Multiple R-squared:  0.2484, Adjusted R-squared:  0.2234
## F-statistic: 9.954 on 8 and 241 DF,  p-value: 5.794e-12
```

The following are the validation of this model.

```
# Note that validation set is called: BloodValid
# Validation process
anova(Blood_Main_Best)
pred.cv.out = predict(Blood_Main_Best,BloodValid[,c(4,5,8,11,12,13)])
delta.cv.out =BloodPressure$sbp[-BloodPressure.cv.samp]-pred.cv.out
n.star=dim(BloodValid)[1]
```

This picture shows the MSE of this model.

```
## Residuals 241 135180   560.9
```

```r
# MSPR
MSPR = sum((delta.cv.out)^2)/n.star
MSPR
```

```
## [1] 744.0471
```

```r
# Calculate the ratio between MSPR and MSE
MSPR/560.9
```

```
## [1] 1.326524
```

The ratio between MSPR and MSE is less than 3, we say MSE is close to MSPR, so validate this model.

```r
# PRESS statistic
PRESS(Blood_Main_Best)
```

```
## [1] 144822.3
```

### STEP 3: ADD INTERACTION TERMS INTO THE MODEL, USE *AIC(1974)*[2] AND DO VALIDATION

Then, we add interaction terms to the best model we find in step 2, use stepwise elimination to eliminate the predictors and do validation for the eliminated model.

```
## Call:
## lm(formula = sbp ~ bmi + trt + alcohol + smoke + exercise + height +
##     bmi:trt + trt:alcohol + trt:smoke, data = BloodTrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.167 -15.961   0.299  11.471  54.497
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   55.9378    24.4048   2.292   0.0228 *
## bmi            1.4557     0.2323   6.265 1.74e-09 ***
## trt1          32.7893    15.0746   2.175   0.0306 *
## alcohol2       1.1796     4.0754   0.289   0.7725
## alcohol3      20.5858     4.2117   4.888 1.88e-06 ***
## smokeY        14.1230     3.3855   4.172 4.25e-05 ***
## exercise2    -11.8557     3.7547  -3.158   0.0018 **
## exercise3     -9.5503     3.5105  -2.720   0.0070 **
## height         0.7182     0.3063   2.345   0.0199 *
## bmi:trt1      -1.0607     0.4559  -2.327   0.0208 *
## trt1:alcohol2 -8.1004     9.7371  -0.832   0.4063
## trt1:alcohol3 -21.3505    8.7192  -2.449   0.0151 *
## trt1:smokeY   -15.1259    7.3355  -2.062   0.0403 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.12 on 237 degrees of freedom
## Multiple R-squared:  0.2954, Adjusted R-squared:  0.2597
## F-statistic: 8.279 on 12 and 237 DF,  p-value: 5.296e-13
```

The following are the validation of this model.

This picture shows the MSE of this model.

```
## Residuals   237 126728   534.7
```

```
# MSPR
MSPR.i = sum((delta.cv.out.i)^2)/n.star
MSPR.i
```

```
## [1] 730.9055
```

```
# Calculate the ratio between MSPR and MSE
MSPR.i/534.7
```

```
## [1] 1.366945
```

The ratio between MSPR and MSE is less than 3, we say MSE is close to MSPR, so validate this model.

```
# PRESS statistic
PRESS(Blood_Main_Inter_Best)
```

```
## [1] 137213.9
```

**STEP 4: MODEL SELECTION DECISION**

We calculate PRESS statistic, Cp, and *AIC(1974)*[2] for the best model with main effects only and best model adding interaction terms.

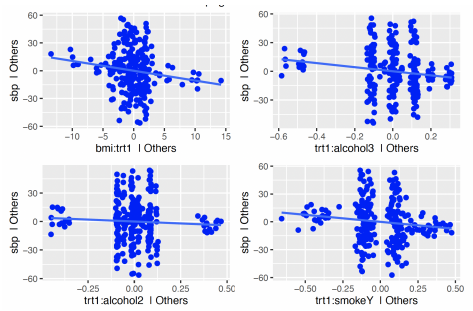
```
##                           PRESS         Cp p'      AIC
## Blood_Main_Best       144822.3 -5.8692066  9 2302.694
## Blood_Main_Inter_Best 137213.9  0.9395469 13 2294.553
```

From the table above, both models have Cp far from p', Blood_Main_Inter_Best has the minimum PRESS statistic and AIC, so we select Blood_Main_Inter_Best for model diagnosis. Notice that we don't know if this model violates any of the regression assumptions, so we can't say this is our final model, it is just the best model we have so far.

**STEP 5: MODEL DIAGNOSIS**

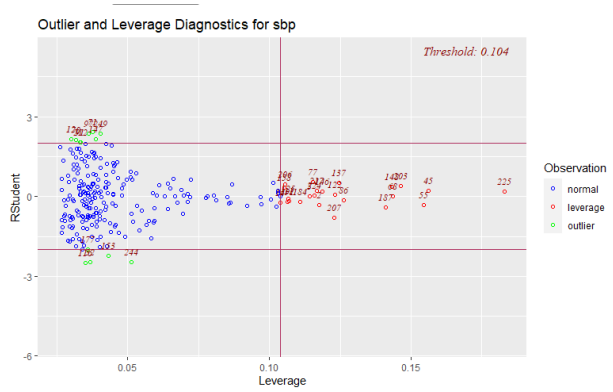
**OLS Added Variable Plot**

Since we have 13 predictors in our model, we just put some of the plots here.



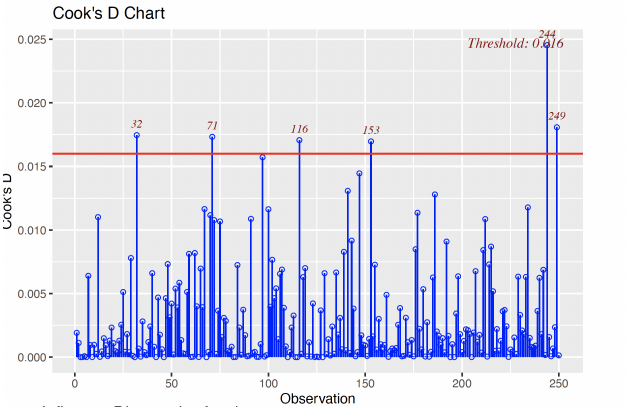

No polynomial relationship was observed between the response variable and predictors, no need to consider polynomial regression.

**Outlying Points**




We find 37 outlying points.

**Influential Points**




From Cook's distance chart, we can see that observations 32, 71, 116, 153, 244, and 249 cause major changes to the fitted regression model.

**Former Multicollinearity Check**

```
# Calculate *VIF*^[1]^ for each predictors
VIF = vif(Blood_Main_Inter_Best)
VIF
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## smoke        1.337559  1        1.156529
## exercise     1.117492  2        1.028161
## height       1.607146  1        1.267733
## alcohol      1.736600  2        1.147955
## trt         18.468019  1        4.297443
## bmi          1.796452  1        1.340318
## smoke:trt    2.882535  1        1.697803
## alcohol:trt  5.592193  2        1.537785
## trt:bmi     15.698227  1        3.962099
```

```
# Calculate VIF_bar
VIFbar = mean(VIF)
VIFbar
```

```
## [1] 2.950817
```

We can ignore the high VIF of interaction term, because high VIF of interaction term is expected. Since trt is a categorical variable, we can ignore its high VIF as well. The rest terms all have VIF $< 10$, and $\bar{VIF} < 10$, therefore, no multicollinearity problem detected in this model.

### Checking Regression Assumption

```
## Check Normal Population Assumption uses *shapiro wilk test*^[3]^
shapiro.test(Blood_Main_Inter_Best$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Blood_Main_Inter_Best$residuals
## W = 0.98991, p-value = 0.08001
```

p-value = 0.08001 > 0.05, so we fail to reject the null hypothesis that the sample is from normal population.

```
# Check Equal Variance Assumption
bptest(Blood_Main_Inter_Best)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Blood_Main_Inter_Best
## BP = 35.32, df = 12, p-value = 0.0004164
```

p-value = 0.0004164 < 0.05, so we reject the null hypothesis that the residuals are distributed with equal variance. We need to apply weighted least square (WLS)

```
# Validation of this WLS model
anova(Blood_Main_Inter_Best_WLS)
```

```
   ## Residuals   237 181347   765.2
```

```
pred.cv.out.ib = predict(Blood_Main_Inter_Best_WLS,BloodValid[,c(4,5,8,11,12,13)])
delta.cv.out.ib =BloodPressure$sbp[-BloodPressure.cv.samp]-pred.cv.out.ib
```

```
# MSPR
MSPR.ib = sum((delta.cv.out.ib)^2)/n.star
MSPR.ib
```

```
## [1] 732.2964
```

```
# Calculate the ratio between MSE and MSPR
765.2/MSPR.ib
```

```
## [1] 1.044932
```

Then we apply WLS to our best model, we hide the code due to page limitations. The ratio between MSE and MSPR is less than 3, we say MSE is close to MSPR, so validate this model.

```
##                 WLS model   OLS model
## (Intercept)    58.5623547  55.9377634
## bmi             1.4439952  14.1229808
## trtl           36.3371303 -11.8556869
## alcohol2        2.2694815  -9.5503246
## alcohol3       21.8288751   0.7182113
## smokeY         13.9982302   1.1795602
## exercise2      -7.4228826  20.5857678
## exercise3      -5.2519764  32.7893415
## height          0.6244442   1.4557450
## bmi:trtl       -1.1158174 -15.1259452
## trtl:alcohol2  -8.1979981  -8.1003582
## trtl:alcohol3 -23.3849594 -21.3504663
## trtl:smokeY   -15.3644836  -1.0607169
```

The coefficients of Blood_Main_Inter_Best_WLS is not significantly different from the coefficients of Blood_Main_Inter_Best.

```
# Run bptest to see if we have equal variance right now
bptest(Blood_Main_Inter_Best_WLS)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Blood_Main_Inter_Best_WLS
## BP = 35.32, df = 12, p-value = 0.0004164
```

The bptest remains the same for two model. We will talk about this problem in limitation section.

**STEP 6: FINAL MODEL**

We compare the best model with OLS with model with WLS

```
## Multiple R-squared:  0.2954, Adjusted R-squared:  0.2597
```

```
## Multiple R-squared:  0.3455, Adjusted R-squared:  0.3123
```

$R^2$ for Blood_Main_Inter_Best is 0.2954, and $R^2$ for Blood_Main_Inter_Best_WLS is 0.3455, we see a significant improvement in $R^2$ after applying WLS.

```
##                               PRESS      AIC
## Blood_Main_Inter_Best     137213.9 2294.553
## Blood_Main_Inter_Best_WLS 135834.8 2185.162
```

From the table above, model with WLS has the minimum PRESS statistic and AIC, so we select model with WLS as our final model.

**CONCLUSION**

The equation of this model is

```
lm(formula = sbp ~ bmi + trt + alcohol + smoke + exercise + height +
    bmi:trt + trt:alcohol + trt:smoke, data = BloodTrain, weights = 1/var.s)
```

**Interpretation**

We round all numbers to two decimal places.

$\beta_0$: When the all variables equal to 0, we expect the observee's sbp to be 58.56.

$\beta_1$: The observee's sbp is expected to be increased by 1.44 if observee's bmi is increase by 1 with holding other variables unchanged.

$\beta_2$: The observee who receives treatment's sbp is expected to have sbp 36.33 bigger than the observee who doesn't receive treatment, with holding other variables unchanged.

$\beta_3$: For observee who drink a lot of alcohol, the sbp is expected to be 21.82 higher than observee who drinks little alcohol, with holding other variables unchanged. For observee who drink meduim amount of alcohol, the sbp is expected to be 2.27 higher than observee who drinks little alcohol, with holding other variables unchanged.

$\beta_4$: The observee who smokes is expected to have sbp 36.33 bigger than the observee who doesn't smoke, with holding other variables unchanged.

$\beta_5$: For observee who does medium amount of exercise, the sbp is expected to be 7.42 lower than observee who does low amount of exercise, with holding other variables unchanged. For observee who does high amount of exercise, the sbp is expected to be 5.25 lower than observee who does low amount of exercise, with holding other variables unchanged.

$\beta_6$: The observee's sbp is expected to be increased by 0.62 if observee's height is increased by 1 inches with holding other variables unchanged.

$\beta_7$: The observee's sbp will have additional decrease of 1.12 if the observee's bmi increased by 1 for observee who takes treatment compared to observee who doesn't take treatment with holding other variables unchanged.

$\beta_8$: The observee's sbp will have additional decrease of 8.20 if the observee drinks meduim amount of alcohol and takes treatment compared to observee who drinks low amount of alcohol and doesn't take treatment with holding other variables unchanged. The observee's sbp will have additional decrease of 23.38 if the observee drinks high amount of alcohol and takes treatment compared to observee who drinks low amount of alcohol and doesn't take treatment with holding other variables unchanged.
$\beta_9$: The observee's sbp will have additional decrease of 15.36 if the observee smokes and takes treatment compared to observee who doesn't smoke and doesn't take treatment with holding other variables unchanged. This suggest that people who doesn't take treatment for hypertension, doesn't smoke, do medium amount of exercise and drink low amount of alcohol with lower bmi and shorter height tend to have lower sbp.

**LIMITATION**

The bptest result for our OLS model and WLS model is exactly the same. We did some research and asked TA, it turns out that probably the error is skew distributed, so WLS can't sufficiently adjust the variance of our model. We tried another approach of WLS and get the same result. But applying WLS regression does significantly increase our $R^2$.
The $R^2$ of our final model is 0.3455, which is not really high.

**FUTURE WORK**

Researching on what makes us get the same bptest result and solve this problem.
Try other regression methods and see if we can improve the R^2.
Lack of predictor can also cause low R^2, so collecting more data and find out more potential predictor that may affect sbp.

**REFERENCES**

[1]VIF: Snee, Ron (1981). Origins of the Variance Inflation Factor as Recalled by Cuthbert Daniel (Technical report). Snee Associates.
[2] AIC: Akaike, Hirotsugu. A new look at the statistical model identification. IEEE Transactions on Automatic Control. 1974, 19 (6): 716–723.
[3] Hosmer-Lemeshow test: Hosmer, David W.; Lemeshow, Stanley (2013). Applied Logistic Regression. New York: Wiley. ISBN 978-0-470-58247