Systolic Blood Pressure:
How High Can You Go?

Arib Shaikh, Joseph Wang, Shruti Sood, Zhe Fan Li

Overview

High SBP Rates

SBP measures the force your heart exerts on the arteries

High SBP Rates play a factor in determining cardiovascular disease

The results are interpreted for measures to prevent or lower high SBP.

Goals and Hypothesis

Goals

The goal is to find the most **significant health and social factors** that affect Systolic Blood Pressure in the BloodPressure.xlsx dataset.

Hypothesis

Based on initial data analysis:

It is predicted that smoking, alcohol use, bmi, stress and exercise have a significant relationship with SBP.

Data Cleaning Process

Step 1

Multi-collinear terms

- The alias() function was used to detect perfectly multicollinear terms.

Step 2

VIF

- VIF was used to identify multicollinearity, the BMI variable is multicollinear to height and weight.

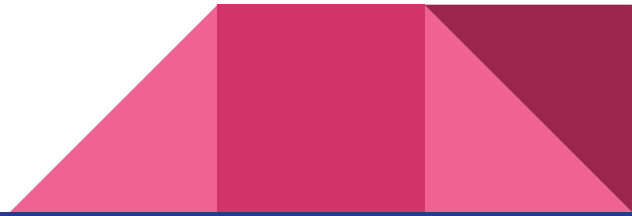
Step 3

Binary Categorical Variables

- Categorical Variables such as Gender, Married and Smoke were changed into binary values.
- NA values were checked for.

Unused Variables

- **Height, Weight and Overweight** were unused as they are already considered in the BMI variable.
- The **childbearing** factor was also unused as it is a **linear combination** of the gender variable.

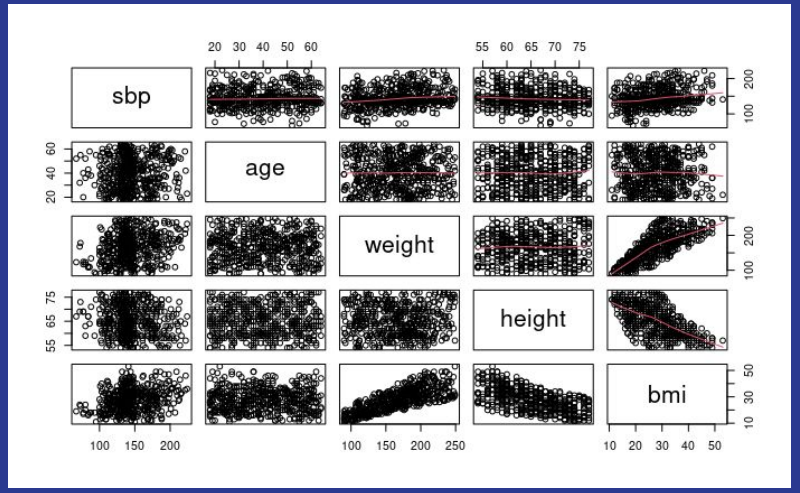


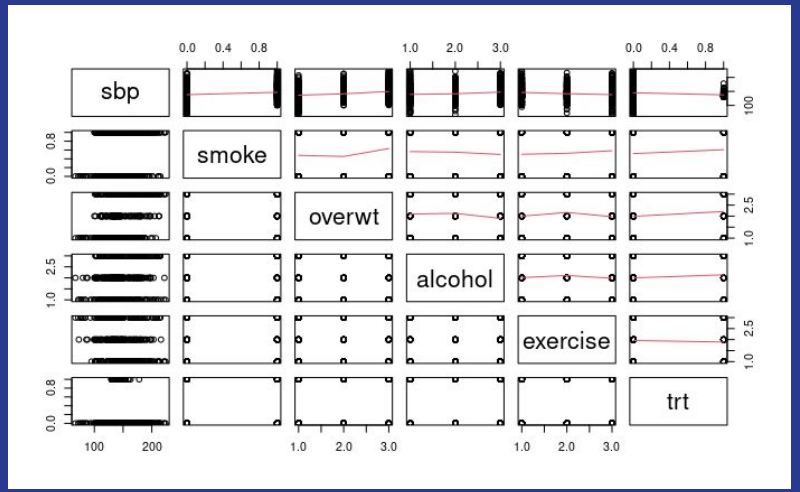
Scatterplot

It can be concluded that:

1. The factors: **smoke, alcohol, BMI** have a **positive** relation with SBP.

2. The factor **exercise and hypertension treatment** have a **negative** relation with SBP.

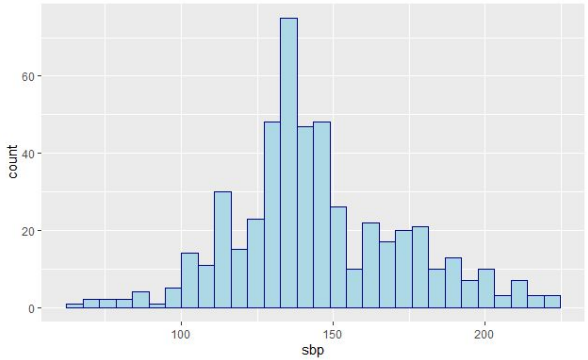


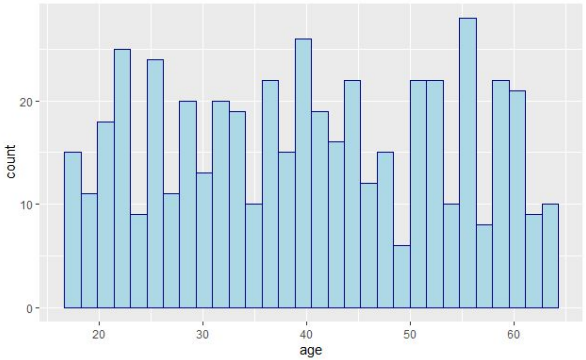


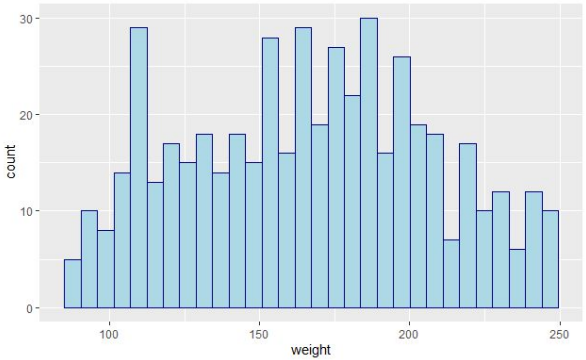
Correlation Matrix

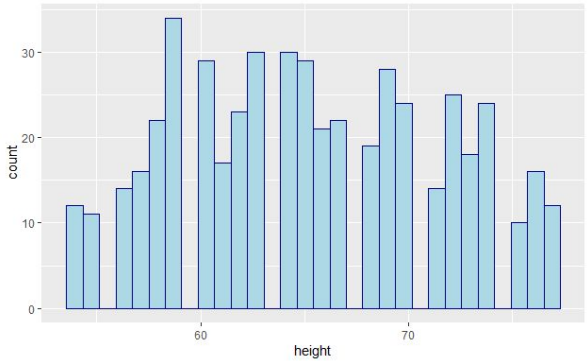
	sbp	gender	married	smoke	exercise	age	weight	height	overwt	race	alcohol	trt	bmi	stress	salt	chldbear	income	educatn
sbp	1.000	0.002	0.061	0.193	-0.145	0.037	0.230	-0.117	0.267	-0.008	0.133	-0.126	0.267	0.067	-0.029	0.025	0.046	-0.009
gender	0.002	1.000	-0.039	-0.045	-0.032	0.005	0.236	0.293	0.004	0.020	-0.089	0.063	0.000	0.031	0.009	-0.895	0.044	-0.109
married	0.061	-0.039	1.000	0.031	-0.036	-0.017	-0.081	0.017	-0.064	-0.069	0.073	-0.043	-0.077	-0.080	-0.054	0.039	-0.019	0.036
smoke	0.193	-0.045	0.031	1.000	0.060	-0.004	0.079	-0.069	0.122	-0.040	-0.049	0.063	0.106	0.029	-0.053	0.040	-0.088	-0.007
exercise	-0.145	-0.032	-0.036	0.060	1.000	0.048	0.025	0.045	-0.008	0.012	-0.012	-0.028	-0.018	-0.017	0.038	0.029	0.084	-0.031
age	0.037	0.005	-0.017	-0.004	0.048	1.000	-0.002	-0.001	0.050	-0.022	-0.094	0.035	0.002	0.040	0.052	-0.023	0.034	-0.020
weight	0.230	0.236	-0.081	0.079	0.025	-0.002	1.000	0.028	0.717	0.032	-0.105	0.121	0.768	0.057	-0.020	-0.207	0.010	-0.006
height	-0.117	0.293	0.017	-0.069	0.045	-0.001	0.028	1.000	-0.514	0.031	-0.062	0.014	-0.594	0.058	0.079	-0.269	0.039	0.033
overwt	0.267	0.004	-0.064	0.122	-0.008	0.050	0.717	-0.514	1.000	0.023	-0.084	0.093	0.889	0.034	-0.050	0.003	-0.023	-0.030
race	-0.008	0.020	-0.069	-0.040	0.012	-0.022	0.032	0.031	0.023	1.000	0.075	-0.076	0.003	0.075	-0.024	-0.052	0.054	0.080
alcohol	0.133	-0.089	0.073	-0.049	-0.012	-0.094	-0.105	-0.062	-0.084	0.075	1.000	0.063	-0.040	-0.029	-0.081	0.093	0.038	-0.039
trt	-0.126	0.063	-0.043	0.063	-0.028	0.035	0.121	0.014	0.093	-0.076	0.063	1.000	0.093	0.058	-0.025	-0.035	0.036	-0.011
bmi	0.267	0.000	-0.077	0.106	-0.018	0.002	0.768	-0.594	0.889	0.003	-0.040	0.093	1.000	0.003	-0.060	0.016	-0.006	-0.018
stress	0.067	0.031	-0.080	0.029	-0.017	0.040	0.057	0.058	0.034	0.075	-0.029	0.058	0.003	1.000	-0.029	-0.039	0.015	-0.009
salt	-0.029	0.009	-0.054	-0.053	0.038	0.052	-0.020	0.079	-0.050	-0.024	-0.081	-0.025	-0.060	-0.029	1.000	-0.025	0.010	-0.085
chldbear	0.025	-0.895	0.039	0.040	0.029	-0.023	-0.207	-0.269	0.003	-0.052	0.093	-0.035	0.016	-0.039	-0.025	1.000	0.002	0.115
income	0.046	0.044	-0.019	-0.088	0.084	0.034	0.010	0.039	-0.023	0.054	0.038	0.036	-0.006	0.015	0.010	0.002	1.000	-0.027
educatn	-0.009	-0.109	0.036	-0.007	-0.031	-0.020	-0.006	0.033	-0.030	0.080	-0.039	-0.011	-0.018	-0.009	-0.085	0.115	-0.027	1.000

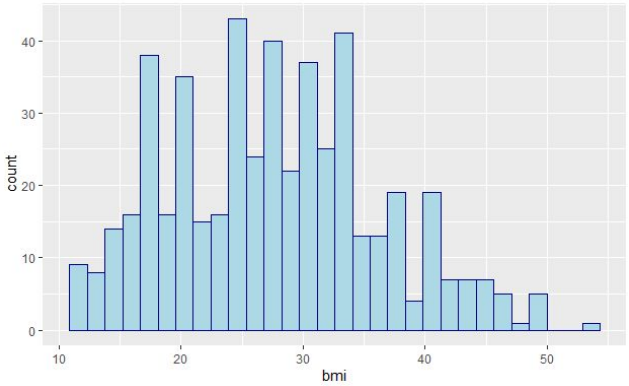
Quantitative Distributions





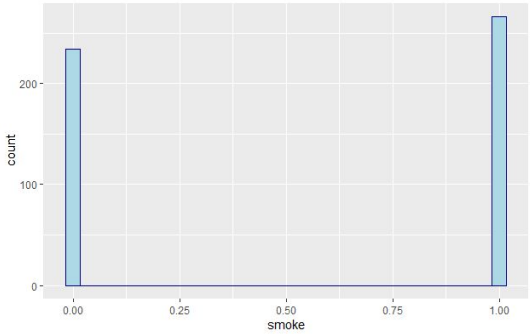


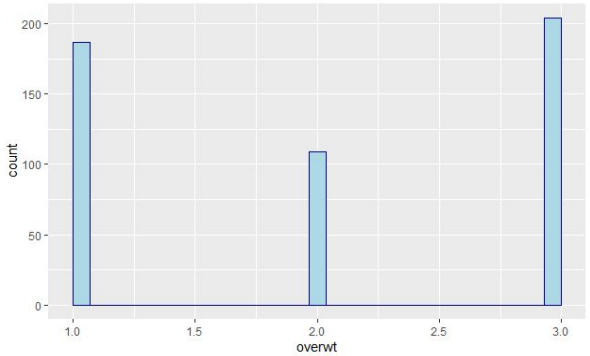


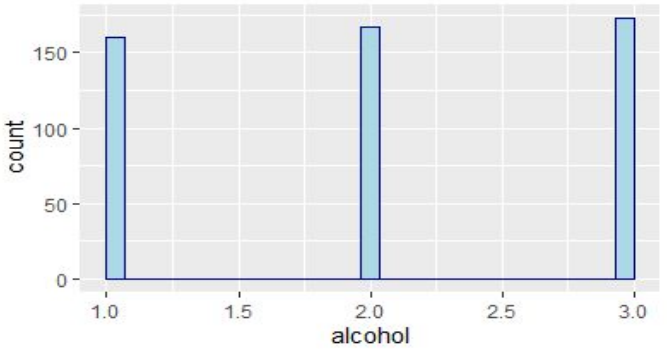


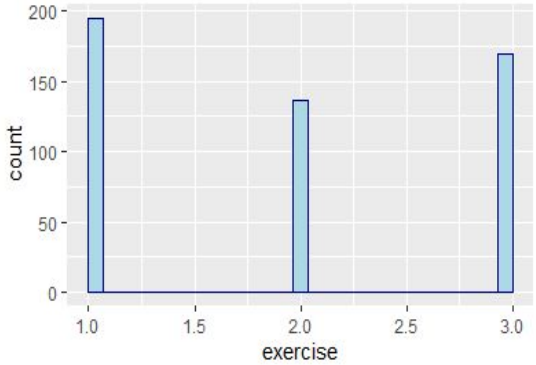
- Sbp, bmi normal
- Everything else roughly uniform

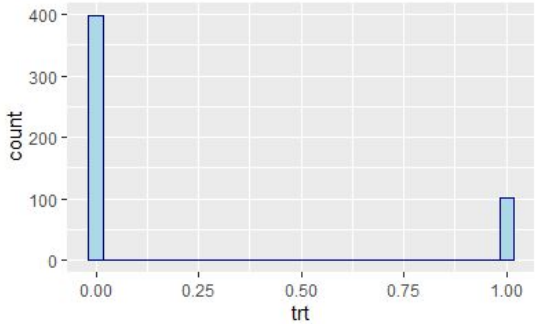
Categorical Distributions











- Everything evenly distributed / trt
- More overweight
- Not obvious bias within data

Model Selection

- Model selection
- Model significance
- Model interpretation
- Model validation

Model Selection

- Used the backwards StepAIC algorithm
- Interact quantitative variables with all qualitative variables
- Quantitative variables used:
  - Age
  - BMI (== Weight/Height^2)
- All categorical variables used
- Further selection done based on p-value significance

```
```{r echo=FALSE , include=FALSE}
fit <- lm(data = dat, formula = sbp ~ age*(factor(exercise)+factor(stress)+factor(salt)+factor(alcohol)+factor(smoke)+factor(gender)+factor(trt)+factor(married)+factor(race)+factor(income)+factor(educatn)) + bmi*(factor(exercise)+factor(stress)+factor(salt)+factor(alcohol)+factor(smoke)+factor(gender)+factor(trt)+factor(married)+factor(race)+factor(income)+factor(educatn)))

stepAIC(fit, direction="backward")
```
```

AIC is a statistic, based on SSE, used for model selection that penalizes a large amount of variables

Model Analysis

```
fit <- lm(formula = sbp ~ age + factor(exercise) + factor(stress) +
    factor(salt) + factor(alcohol) + factor(smoke) + factor(gender) +
    factor(trt) + factor(married) + bmi + age:factor(stress) +
    age:factor(gender) + factor(exercise):bmi + factor(salt):bmi +
    factor(trt):bmi, data = dat)
```

- Initial model from stepAIC

```
Residual standard error: 24.76 on 477 degrees of freedom
Multiple R-squared:  0.2525,    Adjusted R-squared:  0.218
F-statistic: 7.324 on 22 and 477 DF,  p-value: < 2.2e-16
```

- Once fitted, model overall very significant (p-val)
- Does not fit the data well (adjR^2)
  - Noise, seen in scatter-plot
  - Does not mean model is worthless

Coefficients using 100% of data

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	104.43132	10.97588	9.515	< 2e-16	***
age	0.09029	0.16896	0.534	0.593318	
factor(exercise)2	-14.39095	9.45903	-1.521	0.128823	
factor(exercise)3	-31.89574	9.08799	-3.510	0.000491	***
factor(stress)2	-0.37803	8.81681	-0.043	0.965819	
factor(stress)3	-10.61952	8.76562	-1.211	0.226305	
factor(salt)2	-0.87597	9.41027	-0.093	0.925874	
factor(salt)3	17.10853	9.06762	1.887	0.059798	.
factor(alcohol)2	1.61011	2.79974	0.575	0.565500	
factor(alcohol)3	11.42735	2.79787	4.084	5.18e-05	***
factor(smoke)1	11.24000	2.27519	4.940	1.08e-06	***
factor(gender)1	14.58584	7.15606	2.038	0.042076	*
factor(trt)1	20.35289	10.30989	1.974	0.048945	*
factor(married)1	3.72367	2.25750	1.649	0.099710	.
bmi	1.05592	0.26810	3.939	9.42e-05	***
age:factor(stress)2	0.06159	0.20926	0.294	0.768618	
age:factor(stress)3	0.40295	0.20846	1.933	0.053830	.
age:factor(gender)1	-0.32344	0.16838	-1.921	0.055333	.
factor(exercise)2:bmi	0.18961	0.31789	0.596	0.551145	
factor(exercise)3:bmi	0.79870	0.32112	2.487	0.013216	*
factor(salt)2:bmi	0.08413	0.32125	0.262	0.793522	
factor(salt)3:bmi	-0.59753	0.31476	-1.898	0.058253	.
factor(trt)1:bmi	-1.18829	0.34229	-3.472	0.000564	***

Model Significance

```
age                0.09029    0.16896    0.534 0.593318
factor(stress)2   -0.37803    8.81681   -0.043 0.965819
factor(stress)3  -10.61952    8.76562   -1.211 0.226305
```

The main effect of age, stress have high p-values.

- An f-test was run
  - H0: Beta_age = Beta_stress = 0
  - Ha: At least one not 0
- Unable to reject H0
- Slight improvement in adjR^2 after removal(0.02)

```
Model 1: sbp ~ factor(exercise) + factor(salt) + factor(alcohol) + factor(smoke) +
    factor(gender) + factor(trt) + factor(married) + bmi + age:factor(stress) +
    age:factor(gender) + factor(exercise):bmi + factor(salt):bmi +
    factor(trt):bmi
Model 2: sbp ~ age + factor(exercise) + factor(stress) + factor(salt) +
    factor(alcohol) + factor(smoke) + factor(gender) + factor(trt) +
    factor(married) + bmi + age:factor(stress) + age:factor(gender) +
    factor(exercise):bmi + factor(salt):bmi + factor(trt):bmi
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    479 293497
2    477 292328  2    1169.2 0.9539  0.386
```

Model Interpretation

- There are still high p-values
  - Ex. Medium use of alcohol(p = 0.5522)
  - Doesn't mean alcohol not significant
  - High level of alcohol(p = 0.0000476)
- Some important covariates(alpha<0.1):
  - High levels of exercise □
  - High levels of alcohol use □
  - Smoking □
  - Salt □
  - BMI □
- Results match with hypothesis

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          101.4905     9.6296  10.539  < 2e-16 ***
factor(exercise)2    -14.9008     9.4509  -1.577 0.115536
factor(exercise)3    -31.9906     9.0500  -3.535 0.000448 ***
factor(salt)2         -0.8487     9.4026  -0.090 0.928113
factor(salt)3         17.3068     9.0656   1.909 0.056851 .
factor(alcohol)2       1.6577     2.7869   0.595 0.552244
factor(alcohol)3      11.4361     2.7863   4.104 4.76e-05 ***
factor(smoke)1        11.2001     2.2727   4.928 1.15e-06 ***
factor(gender)1       13.4850     7.1093   1.897 0.058455 .
factor(trt)1          20.9926    10.2851   2.041 0.041790 *
factor(married)1       3.7263     2.2567   1.651 0.099349 .
bmi                    1.0465     0.2680   3.905 0.000108 ***
age:factor(stress)1    0.1633     0.1239   1.318 0.188209
age:factor(stress)2    0.2145     0.1200   1.788 0.074390 .
age:factor(stress)3    0.3277     0.1197   2.736 0.006444 **
factor(gender)1:age   -0.3009     0.1675  -1.796 0.073129 .
factor(exercise)2:bmi  0.2035     0.3177   0.640 0.522203
factor(exercise)3:bmi  0.8011     0.3204   2.501 0.012732 *
factor(salt)2:bmi      0.0855     0.3211   0.266 0.790175
factor(salt)3:bmi     -0.6026     0.3147  -1.915 0.056099 .
factor(trt)1:bmi      -1.2029     0.3416  -3.522 0.000470 ***
```

Model Validation

- Data is split
  - **60**% testing
  - **40**% validation
- MSE = **627**
- MSPR = **717.2913**

Since MSE ≈ MSPR, the model is sound.

Model is not overfit and can generalize.

```
set.seed(123)
n = length(dat$sbp)*0.6
datIdx <- sample(1:length(dat$sbp), n, replace=FALSE)

trainingDat <- dat[datIdx,]
validationDat <- dat[-datIdx,]

fitTrain <- lm(formula = sbp ~ factor(exercise) +
    factor(salt) + factor(alcohol) + factor(smoke) + factor(gender) +
    factor(trt) + factor(married) + bmi + age:factor(stress) +
    age:factor(gender) + factor(exercise):bmi + factor(salt):bmi +
    factor(trt):bmi, data = trainingDat)

fitValidate <- lm(formula = sbp ~ factor(exercise) +
    factor(salt) + factor(alcohol) + factor(smoke) + factor(gender) +
    factor(trt) + factor(married) + bmi + age:factor(stress) +
    age:factor(gender) + factor(exercise):bmi + factor(salt):bmi +
    factor(trt):bmi, data = validationDat)

anova(fitTrain)

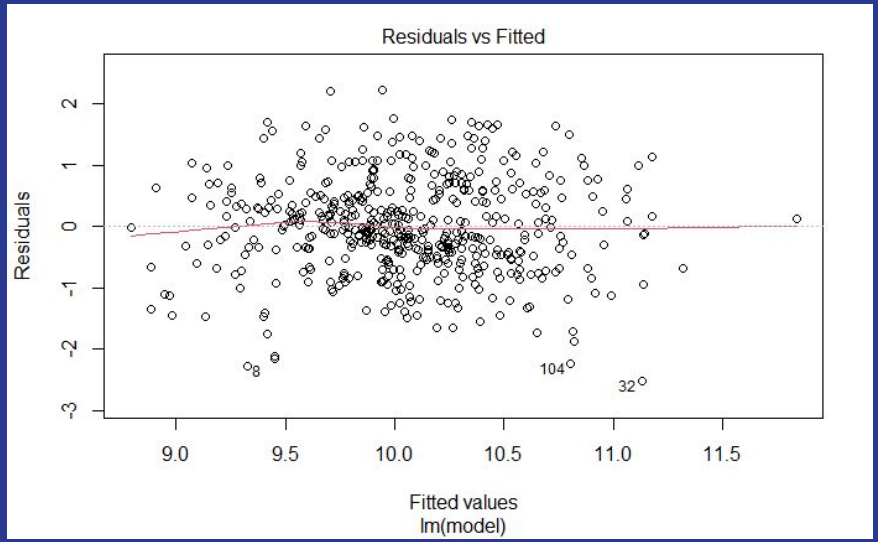
predictTrainWVal <- predict(fitValidate, trainingDat)
MSPR <- sum((trainingDat$sbp - predictTrainWVal)^2)/n
MSPR
```

Model Assumptions

- Linearity
- Normality
- Independence
- Constant Variance

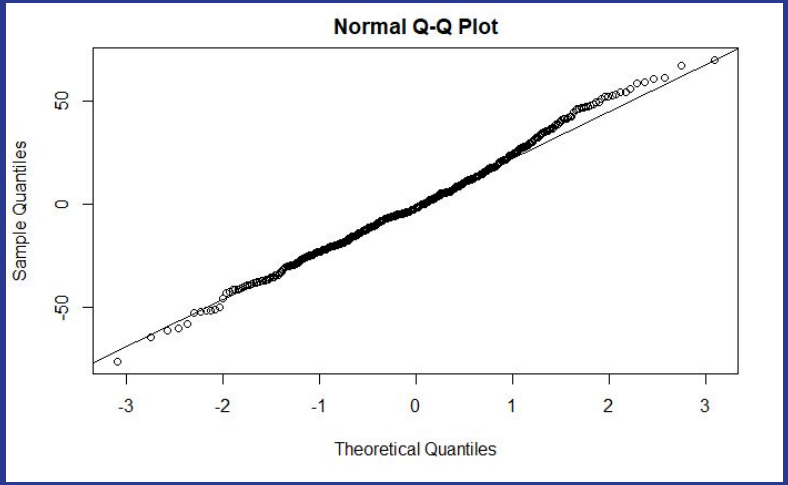
Linearity

- Relationship seems linear.
- **Cannot Reject** linearity.



Normality

- Normality Assumption seems to be satisfied
- P-value > 0.05
- QQ Plot




```
        Shapiro-Wilk normality test

data:  fit$residuals
W = 0.99446, p-value = 0.0672
```

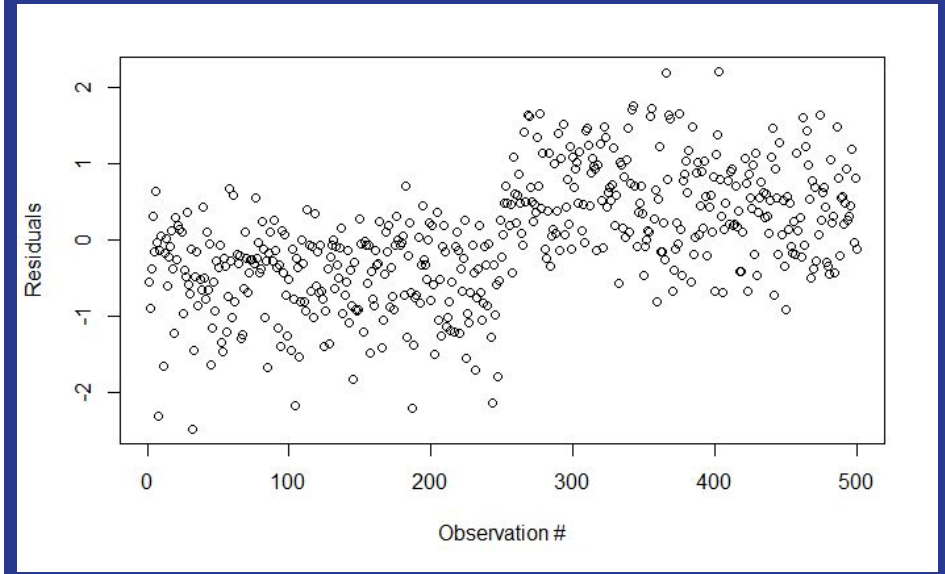
After Boxcox transformation

```
        Shapiro-Wilk normality test

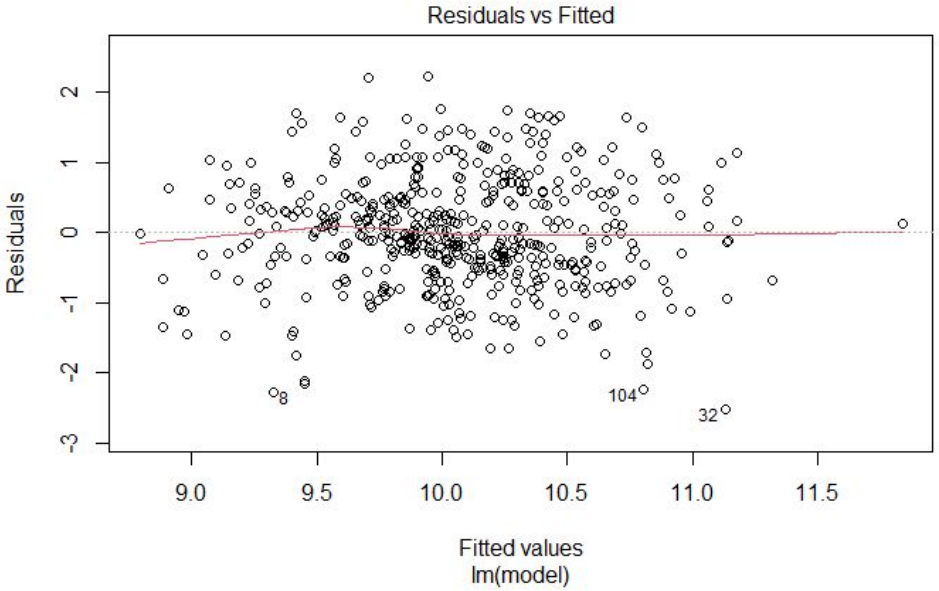
data:  fit$residuals
W = 0.99587, p-value = 0.2144
```

Independent Errors

- No obvious deviate pattern
- **Cannot reject** independent error terms



Model Assumptions - Heteroskedasticity




```
Breusch Pagan Test for Heteroskedasticity
-----------------------------------------
Ho: the variance is constant
Ha: the variance is not constant

                  Data
-----------------------------------------
Response : sbp^lambda
Variables: fitted values of sbp^lambda

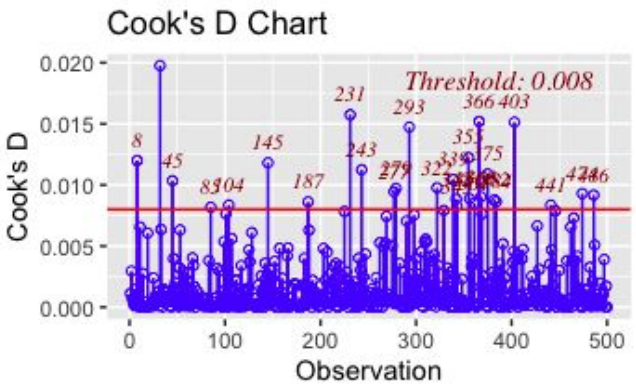
        Test Summary
-----------------------------------------
DF            =    1
Chi2          =    1.990896
Prob > Chi2   =    0.1582472
```

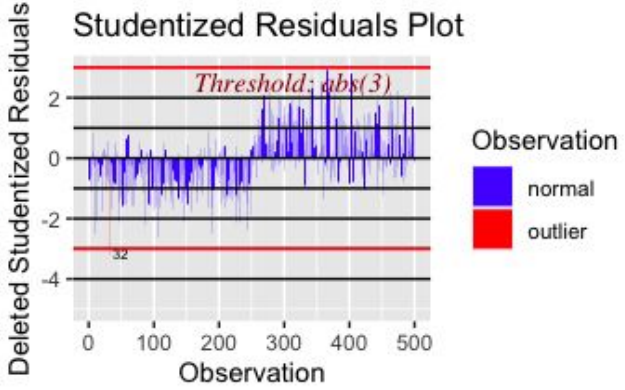
Cannot reject Homoscedasticity

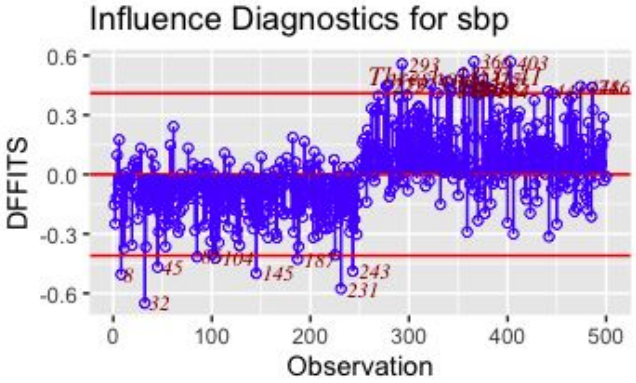
Outlier Detection

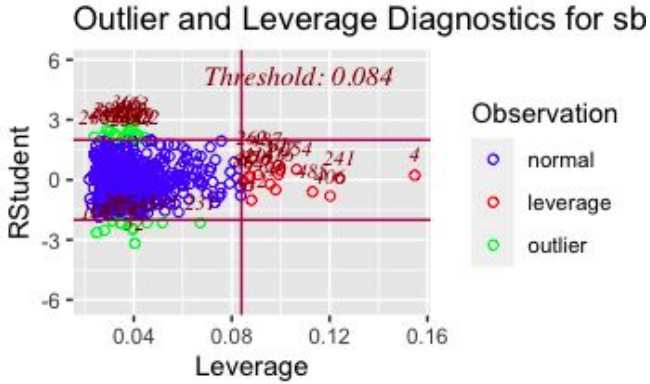
- Cook’s Distance
- Deleted Studentized Residuals
- DIFFITS
- Leverage

Outlier Detection Charts









Outliers

```
#deleted_stu_resid_outliers == NULL
#leverage_outliers == NULL
all_outliers=intersect(cook_outliers, dffits_outliers)
all_outliers
```

```
  [1]   8  32  45  85 104 145 187 231 243 277 279 293 322 339 342 343 355 356 361 366 368 375 382 384 403 441 474 486
```

Influential Points

```
striped_data = data[-all_outliers, ]
data = striped_data
summary(lm(formula = model, data = striped_data))$adj.r.squared
```

```
[1] 0.2825429
```

- Adjusted R-Squared is significantly higher than previous model (0.21)
- These points can be noted as **Influential Points**

Conclusion - Model Significance/Assumptions

- Low p-value, model is significant
- Low adjusted R^2, unknown variables
- The 627.0 = **MSE** ≈ **MSPR** = 717.29, not overfitting model
- The model also satisfies all the assumptions :
  - Constant Variance
  - Error Normality
  - Model Linearity

Conclusion - Model Interpretation

SBP are correlated with the following variables:

- High levels of exercise □
- High levels of alcohol use □
- Smoking □
- Hypertension Treatment □ (expected □)
- BMI □
- Age:Stress □
- Exercise:BMI □ (expected □)
- Hypertension Treatment:BMI □

Ways to Lower SBP

- When your SBP is high:
  - You are 4x more likely to die from a stroke
  - You are 3x more likely to die from heart disease

- Negative correlation with SBP:
  - Exercise
  - Healthy diet
  - Reduce smoking and Alcohol
  - Lower Weight

Questions?

Thank you :D