Determining the Factors Affecting Systolic Blood Pressure Predictions

by

John Ma
Ben Mak
Micheal Chen

STAC67
April 10, 2020

Introduction

Studies have shown that systolic blood pressure (SBP) is an important marker in the risk of all-cause mortality and chronic diseases, some of which include heart attacks, strokes, and diabetes (Böhm, 2020). When reading blood pressure with a machine, there are two readings. One which is the higher number which is the SBP and diastolic pressure is the lower number. SBP measures the force of blood being pushed around the body when your heart contracts. (Bhyan, 2018) In this analysis, the objective is to determine which factors have an impact on systolic blood pressure (SBP) and to create a model to predict the systolic blood pressure for a person. The analysis uses the process described in the textbook Applied Linear Regression Models, on page 344.

The continuous covariates in the data set are age, bmi, height, and weight. BMI is dependent on height and weight as BMI is calculated as weight/height2. Height is in centimeters, age is in years, weight is in kilograms. The categorical variables are alcohol usage as alcohol, childbearing potential as chldbear, education level as educatn, exercise level as exercise, gender, income level as income, marital status as married, overweight status as overwt, race, salt level as salt, if someone is a smoker as smoke, stress level as stress and hypertension treatment as trt. Alcohol usage, childbearing potential

The data has been cleaned to change some categorical variables into numerical variables. The data for married status had to be converted into 0 for not married and 1 for married. Similarly, for the smoke column, Y is 1 and 0 is No,. Data type of the columns had to be changed from char to double value to work with. Also, for gender status, 0 for Female, and 1 for Male.

Graphs of SBP and covariates

Figure 1. Continuous variables vs SBP

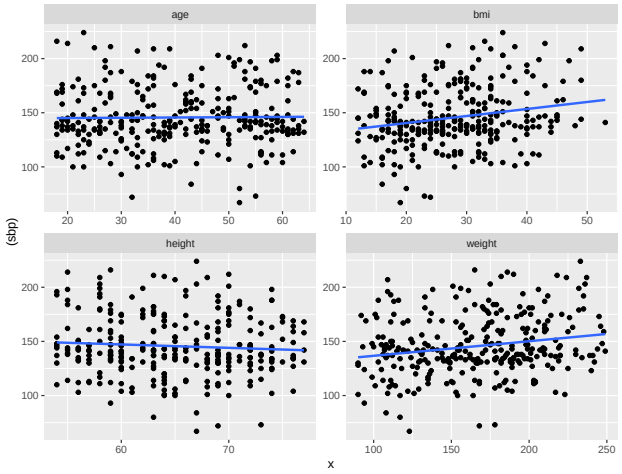
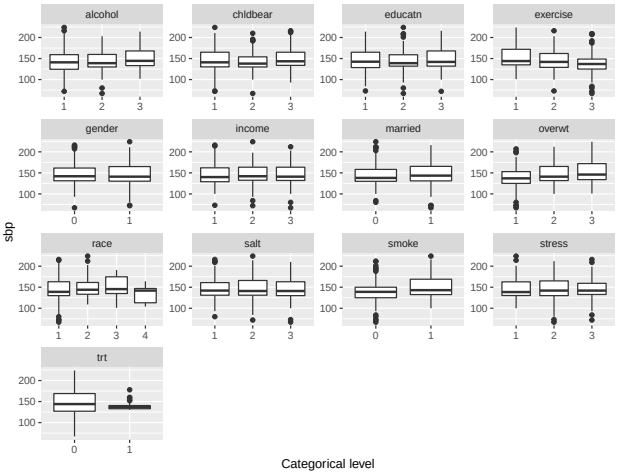
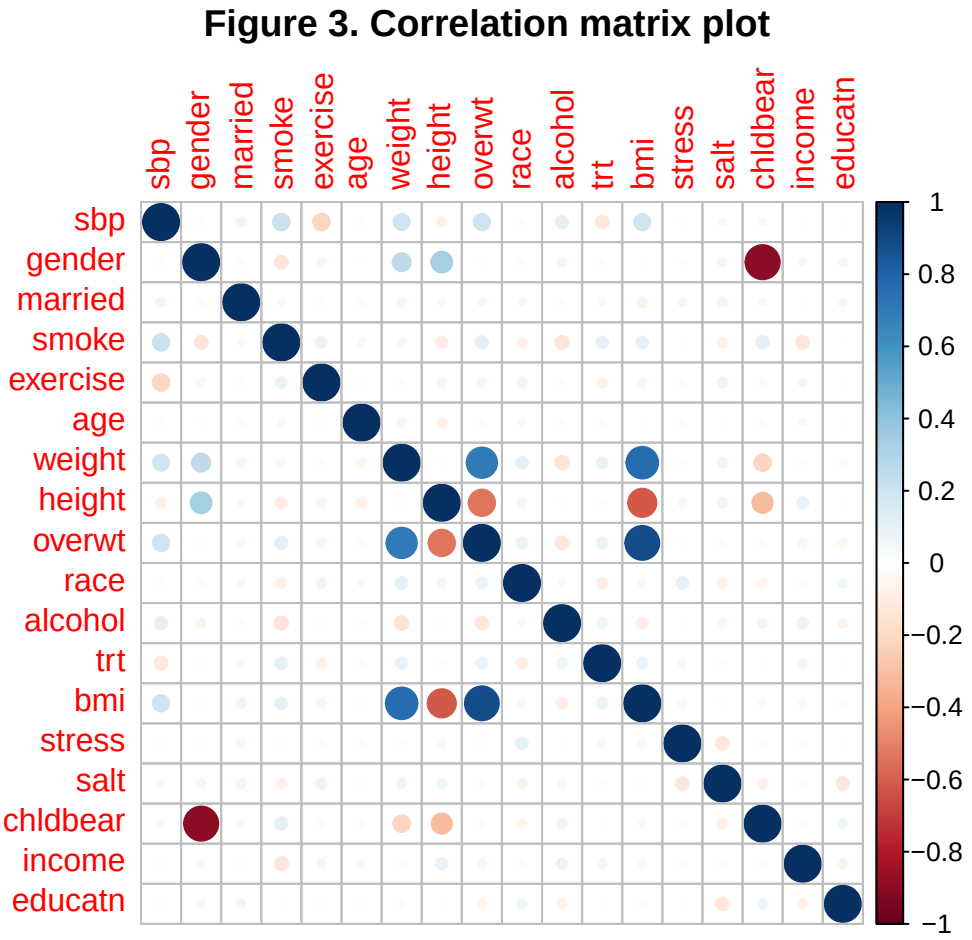


Figure 2. Categorical variables vs SBP



To analyze the correlation between all variables with Systolic Blood pressure, we constructed separate scatter plots and their fitted lines for the continuous predictors and box plots for the categorical. For the continuous set (Figure 1.), we noticed that for the fitted line, there is a positive correlation with SBP, for Weight and BMI and a negative correlation for height. For each fitted model, we also noticed that the data for each correlation was randomly and symmetrically distributed along each regression line. This indicates that by model assumptions, that the data is selected from a normally distributed sample. For the box plots, we noticed that there are some box plots that have noticeable change in mean and median values. For each categorical level, there is an increase in sbp for overweight level and smoking while there is a decrease in level of exercise. Smoking causes blood vessels to clot, which increases heartbeats per second. Exercising strengthens the heart, which results in it pumping blood with less effort.

Relationships between covariates



Using Figure 3, it is seen that the covariates that have the strongest relationship with SBP are smoke, exercise, weight, height, overweight, alcohol, TRT, and BMI. The other covariates do not seem to have that strong of a relationship with SBP. BMI is strongly correlated with height, weight and overweight which raises a multicollinearity problem if it were to be added to the model. Ideally would be better to remove it.

Main Effect Full Model

Figure 4. VIF values of all covariates

```
##    gender   married     smoke  exercise       age    weight    height    overwt
##  5.443472  1.025684  1.109423  1.041756  1.026504 22.872725 14.952780  5.109263
##      race   alcohol       trt       bmi    stress      salt  chldbear    income
##  1.084261  1.079912  1.060084 34.883669  1.046019  1.071238  5.276208  1.068000
##   educatn
##  1.044346
```

Variance Inflation Factor

Looks like there is a multicollinearity problem with weight, height, and BMI. Gender and chldbear seems to have a VIF about 5~ and could pose as a problem. Removing BMI could be beneficial to get a more accurate

model. In addition, BMI = $weight/height^2$, and that a higher BMI depending on the gender, determines if a person is overweight or not. (CDC,2021) The data seems like it is using BMI to determine someone is overweight or not.

Doing the ANOVA test, shows that reduced model has the same effect as the full model as the p-value $>$ 0.05. Thus we can continue to do some of the residual diagnostics with this model.

Figure 4. VIF values of all covariates witout BMI

```
##   gender  married    smoke exercise      age   weight   height   overwt
## 5.326526 1.023309 1.107292 1.041521 1.026344 3.629298 2.616850 5.057393
##     race  alcohol      trt   stress     salt chldbear   income  educatn
## 1.083517 1.074089 1.059565 1.045077 1.071100 5.166147 1.066463 1.043999
```

As we can see, there are no VIF values greater than 10 anymore. There is no indication of serious multicollinearity anymore.

This model with coefficients in Figure 4. includes all predictors, excluding BMI. Since BMI could cause a multicollinearity issue, it was removed. The model here is significant with p-value $<$ 0.05. Meaning that not all values of the coefficients are 0. The coefficients for trt, smoke, exercise and alcohol seems to be significant factors for the model with p-values $<$ 0.05. The intercept standard is quite high as well, at 58.8365, in comparison to the intercept itself at 69.08773. Using the F-test to compare the model with BMI and without BMI, it results in a p-value of 0.7715. Which means that BMI does not make a significant impact on the model predicting SBP.

Residual diagnostics of the main effect model

Plots for the residual values for the full model

Figure 5. Normal Q−Q Plot

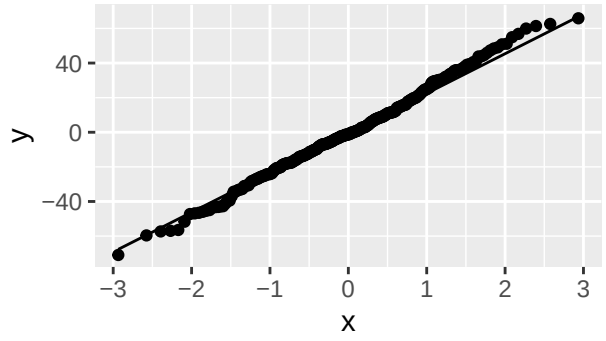


Figure 6. Frequency of Residuals

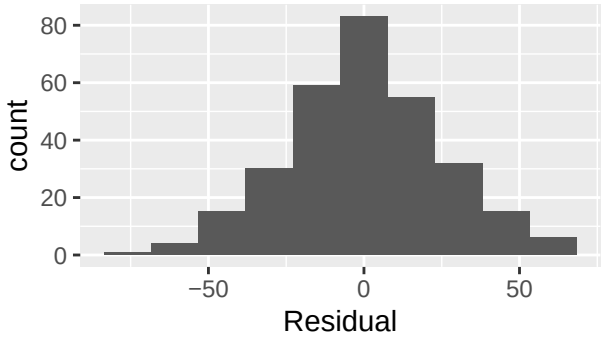
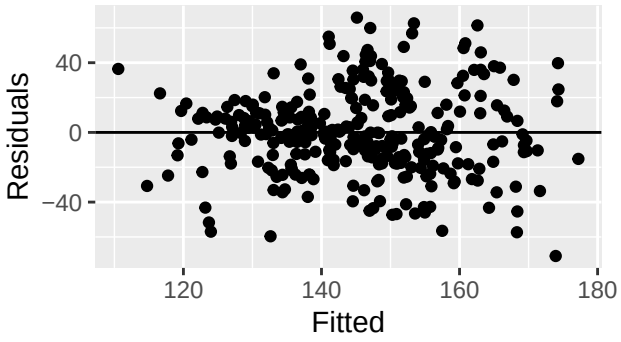


Figure 7. Fitted vs. Residuals



The residuals values of this model look normally distributed in reference to Figure 5. and Figure 6.. The Shapiro-Wilk test gives a P-Value of 0.5881, therefore failing to reject the null hypothesis of normality. However, there is a concern with the variance of the residuals, which they are not constant. There seems to be some resemblance of a trumpet shape for figure 7. with the points, which puts the constant variance of the error terms into question. To diagnose this, the Breusch-Pagan test, gives a P-value of 0.01755 therefore rejecting the null hypothesis that the variance of the residual is constant. As a remedial measure, a box cox transformation is done below.

```
shapiro.test(bp_full2$residuals)
bptest(bp_full2, studentize = F)
```

Transformed full model

Using a box-cox transformation, let $\lambda = 0.5454545$ onto SBP values.

Residual Diagnosis for the Transformed model

Plots for the residual values for the model transformed full model

Figure 9. Normal Q−Q Plot

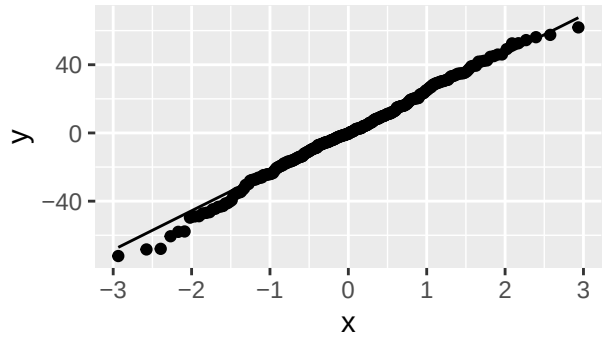


Figure 10. Frequency of Residuals

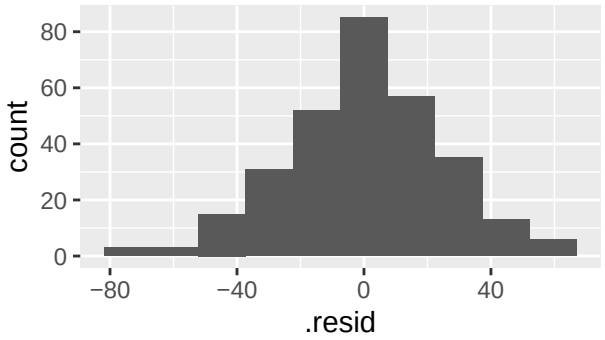
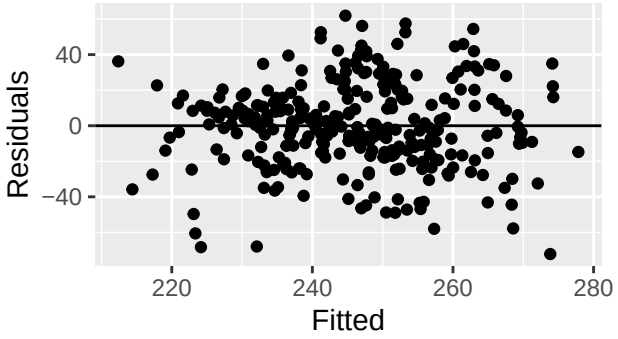


Figure 11. Residuals vs. Fitted



Visually, using Figure 9, and Figure 10, it shows that the residuals are likely from a normally distributed sample. Verifying speculations of the normality of the error terms, Shapiro-Wilk test gives a p-value of 0.4015 therefore failing to reject the null hypothesis where the residuals are not from a normally distributed sample with a significance level of 0.05. Providing strong evidence that the residuals are from a normally distributed sample. Although figure 11. shows that the variance of the error terms may not look constant with the resemblance of a trumpet shape with the plots the Breusch-Pagan test, gives us a p-value of 0.06193. Thus, failing to reject the null hypothesis that the variance of the error terms are constant with a significance level of 0.05 providing evidence that the residuals have constant variance.

Backwards elimination

AIC/BIC

Using backwards elimination as described in the lecture with significance level of =0.20, (through the function stepAIC() in R) for the procedure, gives us a new model dropping married, age, height, overwt, race, stress, salt ,income, educatn. The AIC for the new model is 1936.76, in comparison to 1950.57 for the full model. Using the F-test to compare both models, with a p-value of 0.9117, which results in a failure to reject the null hypothesis that the reduced model and full model have the same effect. In other words, the reduced model is as effective as the full model.

Residual Diagnostics for Reduced model

Plots for the residual values for the reduced model

Figure 12. Normal Q−Q Plot

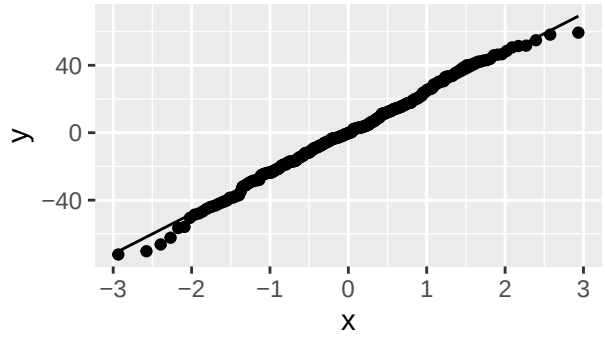


Figure 13. Frequency of Residuals

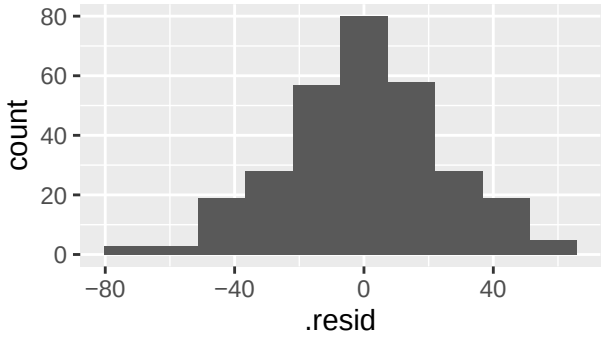
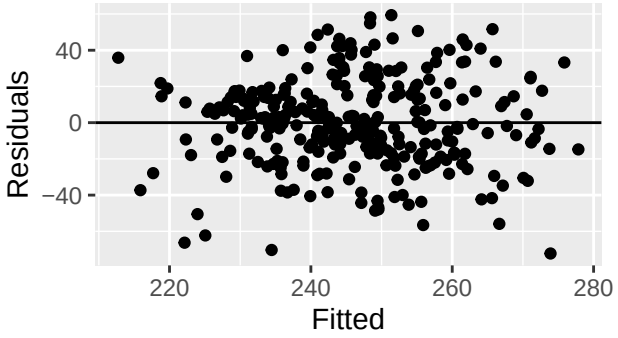


Figure 14. Residuals vs. Fitted



The residuals values of this model look normally distributed in reference to Figure 12. and Figure 13.. The Shapiro-Wilk test gives a P-Value of 0.5881, therefore failing to reject the null hypothesis of normality. However, there is a concern with the variance of the residuals, which are not constant. There seems to be some resemblance of a trumpet shape for figure 7. with the points, which puts the constant variance of the error terms into question. To diagnose this, the Breusch-Pagan test, gives a P-value of 0.01755 therefore rejecting the null hypothesis that the variance of the residual is constant. As a remedial measure, a box cox transformation is done below.

Residual diganostics for Transformed reduced model

Figure 15.

Transformed reduced model

Again, using a box-cox transformation, = 1.030303. Figure 15. show the box cox plot for the reduced model, in which the maximum lambda is 1.030303. Using the equation from Figure 8b. to apply the transformation on the SBP values in the data set. Then use this new transformed data to regress over the same covariates.

Residual diagnosis for transformed reduced model

Plots for the residual values for the transformed reduced model

Figure 16. Normal Q−Q Plot

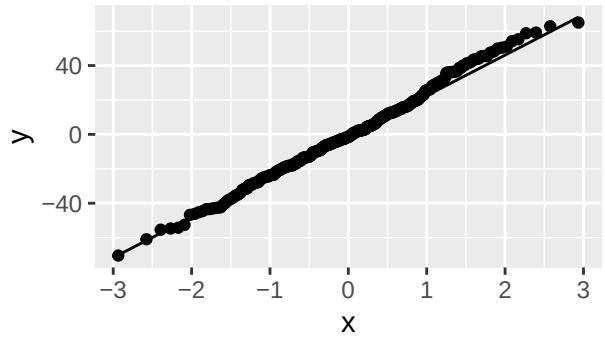


Figure 17. Frequency of Residuals

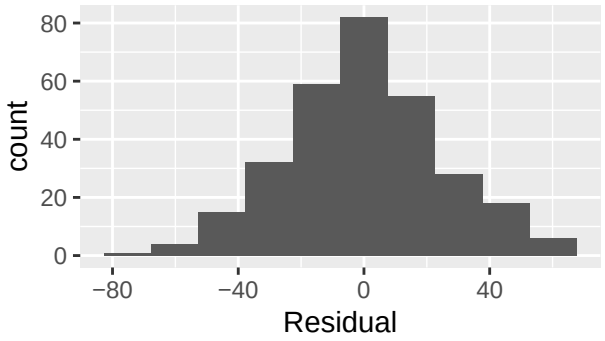
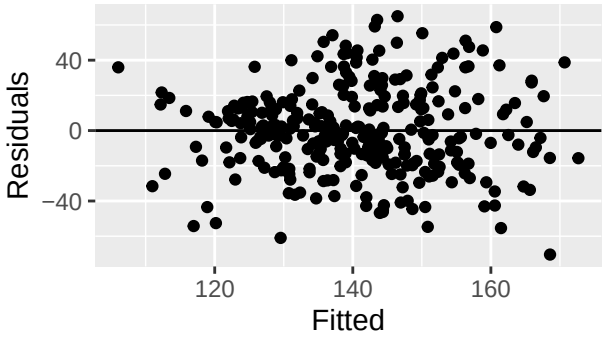


Figure 18. Fitted vs. Residuals



Similarly to the past residual diagnosis, the graphs give the same result, figure 16 and figure 17, showing that the residuals are from a normally distributed sample. However, figure 18. shows that there could be some issue with error variance being constant. Using Shapiro-Wilk test to verify the normality, since P-value as 0.4873 is greater than 0.05, therefore failing to reject the null hypothesis that the residuals are from a normal sample. With the Breusch-Pagan test, the p-value is 0.002354, which is less than 0.05, thus rejecting the null hypothesis that the error terms have constant variance.

Remedial measures: Weighted least squares

Figure 19. Residuals vs Weight

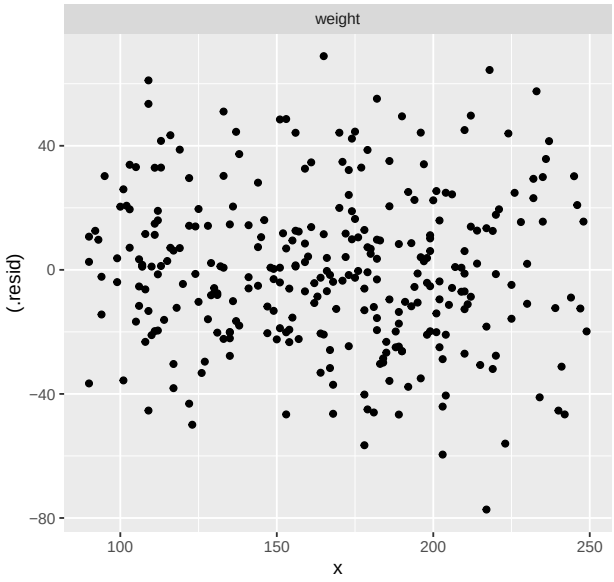
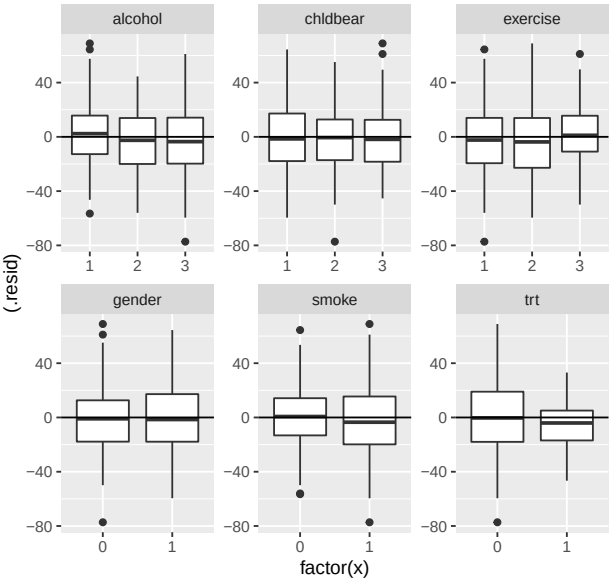


Figure 20. Residuals vs Categorical Variables

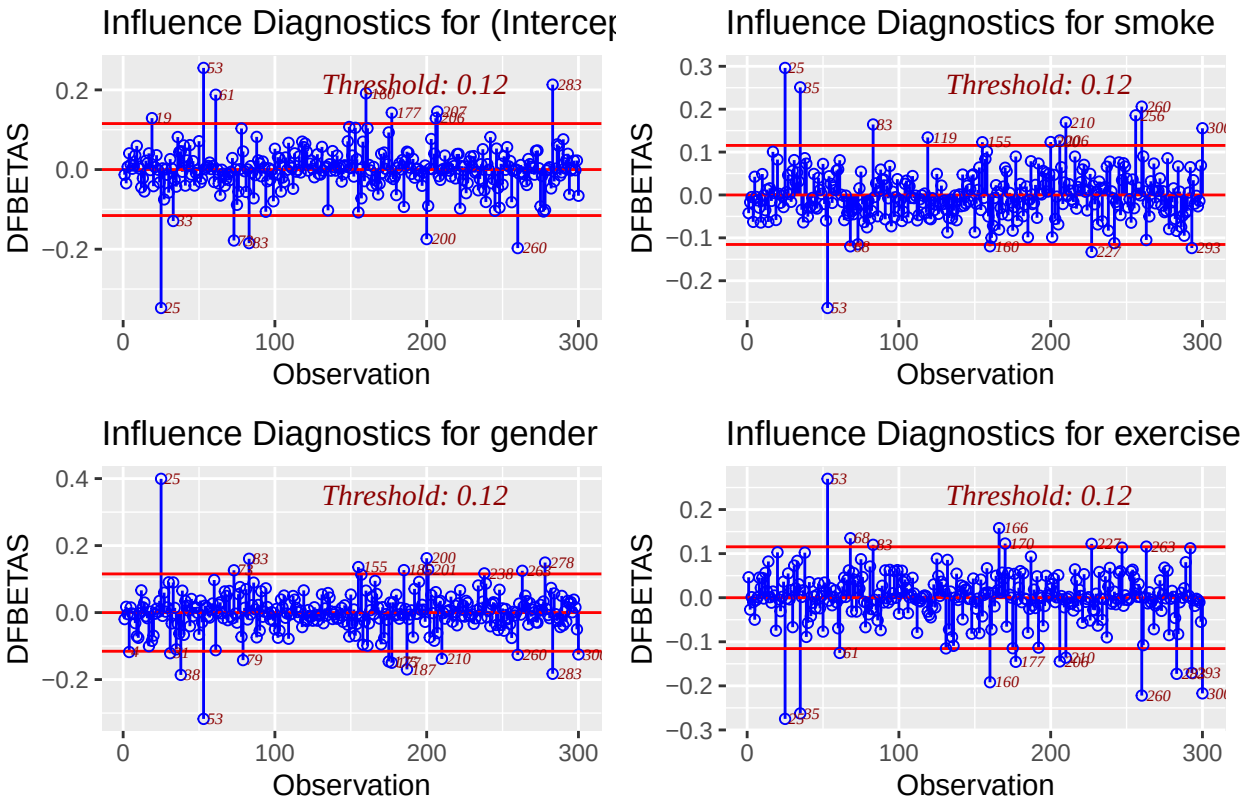


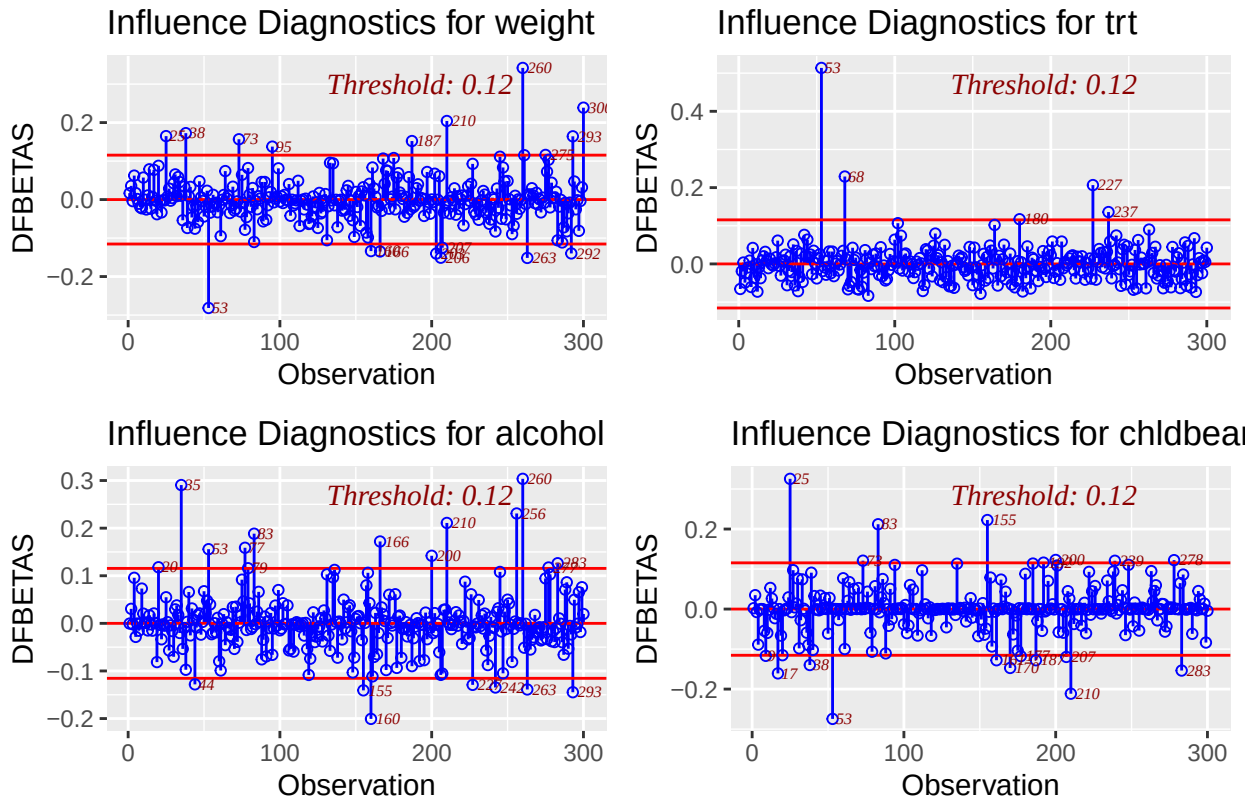
In figure 19. residual values of the fitted weight least squares model are plotted against weight. There seems to be a random scatter of points on this plot, verifying that the residuals do not depends on this covariate. In Figure 20. the whiskers of the box-plot for treatment, seems to give the residuals more variance one with, than one without. Chldbear, follows a similar pattern where the whiskers are slowly shortening but slightly. In addition, the rest of the box-plots shows more of a constant variance of the residuals.

Graphical Diagnostics for influence

```
## NULL
```

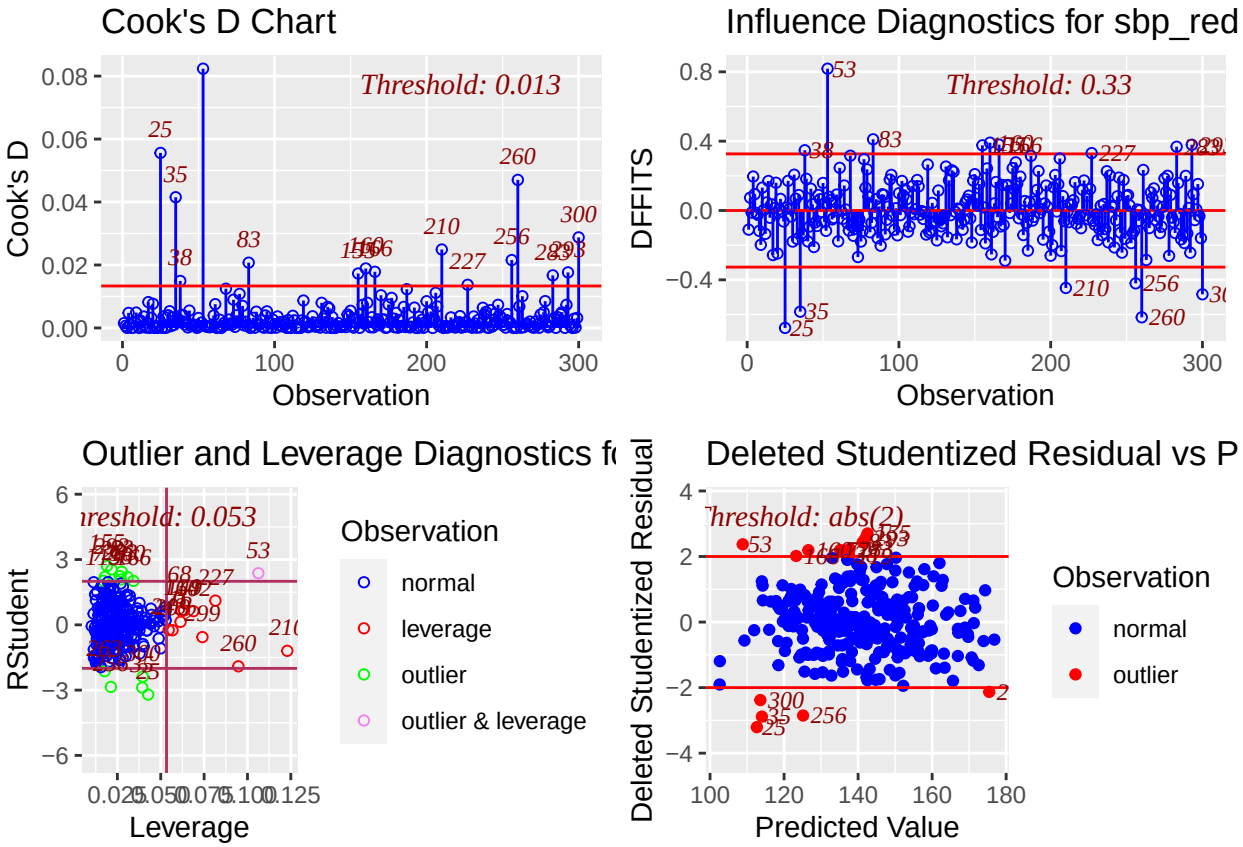
```
ols_plot_dfbetas(mod9, print_plot =T)
```


Influence Diagnostics for (Intercep
Threshold: 0.12
DFBETAS
Observation
Influence Diagnostics for smoke
Threshold: 0.12
DFBETAS
Observation
Influence Diagnostics for gender
Threshold: 0.12
DFBETAS
Observation
Influence Diagnostics for exercise
Threshold: 0.12
DFBETAS
Observation



The DFBETAS plots show that observations 53 seems to give the influence above the threshold on all the covariates. In addition, trt has the least amount of observations above the threshold for DFBETAS and exercise has the most above the threshold. Observations 53 for exercise seems to affect the covariate trt a lot more than the other observations. Alcohol seems to be the least affected by observations 53.

```
ggarrange(p2, p3,p4,p5)
```



Observing Cooks D chart and the influence diagnostic plots, it can be seen that the most influential points are 25,35, 53. However, overall there are many influential observations such that if these influential points were to be gotten rid of, there is a possiblity would be another set of them after with the way they are scattered.

The outlier and leverage diagnostics plot. shows that observation 53 has leverage and is an outlier. Outlier values in this figure shows that it is in a triangular shape. When ridding the data of these influential points there is a possibility of another set of outliers being present. The deleted studentized residual vs. Predictor plot also has an abundance of outliers. Again, removing these outliers could create another set of outliers here. Therefore, removing these outliers would probably not have much effect on the data set.

Cross validation for main effect model

```
##       WLS Model Reduced Model Full Model
## MSE      1.6670       628.9000   630.8000
## MSPR   745.4301       720.4247   674.2833
```

The WLS model seems very likely to be over fitting the data, given a very low MSE of 1.667, in comparison to the MSPR of 745.4301. In addition, the reduced model has a MSPR of 720.4247 and a MSE of 628.9. While the full model has an MSE of 630.8 and a MSPR of 674.2833. The reduced model would be considered less valid than the full model since there is a greater difference between MSPR and MSE with the reduced model than the full model.

Conclusion

	WLS Model	Reduced Model	Full Model
R squared	0.2800	0.2100	0.2200
R Squared adj	0.2600	0.1900	0.1700
AIC	160.7800	1941.0800	1950.5700
PRESS	196808.2000	193144.4100	199432.8000
MSE	1.6670	628.9000	630.8000
MSPR	745.4301	720.4247	674.2833

To conclude, the transformed full model is the most reliable model since it is the only model out of the three that has passed the residual diagnostics, in addition having MSE and MSPR closer than the other two models. Both the reduced model and WLS model have issues with variance of the error terms being constant.

The full model equation:

$$Y = 182.19336 + 9.71851(gender) + 2.96943(married) + 14.99813(smoke)$$

$$-7.81869(exercise) + 0.08189(age) + 0.09049(weight) + 0.16125(height) +$$
$$4.08000(overwt) + 0.38382(race) + 6.80274(alcohol) - 13.55571(trt) +$$
$$0.46647(stress) + 0.10333(salt) + 6.38980(chldbear) +$$

$$1.00487(income) + 0.57094(educatn)$$

Using the table above, it shows that the R^2 for all models are quite low. In addition it also shows that MSE and MSRP have a high difference for the WLS model. With these in mind, it can be considered that there could be a possibility of this model missing key predictors. SBP seems to be mostly correlated to smoking status, exercise, alcohol consumption, child bearing potential, treatment status, weight and gender.

References

Bhyan, Poonam, et al. “ASSOCIATIONS OF SYSTOLIC BLOOD PRESSURE (SBP) <120 (VERSUS 120-139) MMHG WITH OUTCOMES IN PATIENTS WITH HEART FAILURE AND PRESERVED EJECTION FRACTION (HFPEF) WITHOUT HYPERTENSION (HTN).” Journal of the American College of Cardiology, vol. 71, no. 11, Elsevier Inc, 2018, pp. A919–A919, https://doi.org/10.1016/S0735-1097(18)31460-8.

Böhm, Michael, et al. “Heart Failure and Renal Outcomes According to Baseline and Achieved Blood Pressure in Patients with Type 2 Diabetes: Results from EMPA-REG OUTCOME.” Journal of Hypertension, vol. 38, no. 9, Copyright Wolters Kluwer Health, Inc. All rights reserved, 2020, pp. 1829–40, https://doi.org/10.1097/HJH.0000000000002492.

Centers for Disease Control and Prevention. (2021, August 27). About adult BMI. Centers for Disease Control and Prevention. Retrieved April 10, 2022, from https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html#Interpreted