Covariates

When testing the correlation between each of the observational values, we observed that there some variables that had strong correlations with one another. To observe the correlations of each of the predictor variables, we constructed a correlation matrix and compared the correlation coefficients between each predictor. We found that there was a high correlation between weight, height, overweight and the BMI predictor variables.

For the BMI, the formula of calculating BMI is weight/height^2. So, BMI is correlated to weight and height because BMI depends on weight and height.

Secondly, we observed that BMI is positively correlated with the overweight level. This does make sense because BMI is an indicator for being overweight. This would also imply that overweight levels are also highly correlated with weight and height

Thirdly, we found that there is a high correlation between childbearing and gender. This does make sense because women can only bear children.

There exists a multicollinearity issue because overweight levels are highly correlated with weight and height, two different independent observations.

Overall, we found 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. The significance of these observations is that the higher correlated variables can heavily influence fits of the model, which indicates that they may be dropped from the model.

Fits With 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, we noticed that for the fitted line, there is a positive correlation with SBP, for Weight and BMI and a negative correlation for height. This does make sense because of the formula for BMI. When height increases, the BMI decreases, while when weight increases, BMI Increases.

For each fitted model, we also noticed that the data for each correlation was randomly and symmetrically distributed along each regression line. This indicated 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.

Residual Diagnostics

To check the normality of the dataset, we set up a linear model and used the residuals of the fitted line to create another scatter plot with a fitted line. To do this, we set up the fitted model and check LINE assumptions for the model. Examine that the Residuals seem normal, since there are no visible outliers, and residual plot forms a line close to y=x. In addition, we also conducted a Shapiro-Wilk test, where we obtained a p-value of 0.6039, which implies that the distribution is from a normal sample.

We then applied a box cox transformation on the systolic blood pressures and obtained a lambda of 0.545454. Again, we compared the new residuals using the LINE conditions. Similarly, just look more normal. Similarly, the densities form a normal distribution, and the residuals are randomly and symmetrically scattered. So, for the Shapiro-Wilk test we see that the p-value = 0.4033, so we reject the null hypothesis that the residuals are not from a normally distributed sample. This expected of the intention of the box-cox transformation.

DFFITS model

To further examine the model, we created a DFFITS and Cook’s distance Model to examine the possible outliers of the data set. Unlike the observation for the residual plots, the outliers in the dataset were more evident. The outlier that has the most influence is 263, with others outside the threshold closer to it

By inspection, there looks like there are a couple of outliers here, closer to the threshold.

For the t-values of the DFFITS data, we find there are no values that are abs(t) > t.crit. So there are no values where hii>2\*(p’/n) and that there are no significant correlated datasets with SBP. This would mean that when removing the outliers affect the data set that there would still be outliers. So, there are no outliers that necessarily need to be removed.