STATS 506 Project

Tian Wang

Core Analysis

As we have got a nice cleaned dataset, we want to visualize our data with pair plots first. As shown in the pair plots for the systolic blood pressure, specifically from the scatter plots, we can find that there are weak relations or almost no relations between the variables. Moreover, from the histograms, we can find that the all the variables seem not to be normally distributed. Athough the variable systolic blood pressure and the variable bmi approach to be normally distributed, they are still right skewed. In addition, as shown in the pair plots for the diatolic blood pressure, from the scatter plots, again, we can find that there are weak relations or almost no relations between the variables. Also, from the histograms, except that the variable diatolic blood pressure seems to be almost normally distributed (only a little bit left skewed), all the other variables are not normally distributed.

After exploring the data through vizualizing with pair plots, we fit our linear regression models.

For the systolic blood pressure, we fit the model:

$$y = \beta_0 + \beta_{workhrs} * x_{workhrs} + \beta_{gender} * x_{gender} + \beta_{age} * x_{age} + \beta_{bmi} * x_{bmi} + \beta_{sleep} * x_{sleep} + \beta_{smoke} * x_{smoke} + \beta_{avg\_alcohol\_freq\_wk} * x_{avg\_alcohol\_freq\_wk} + \epsilon$$

And for diatolic blood pressure, we fit the model:

$$y = \beta_0 + \beta_{workhrs} * x_{workhrs} + \beta_{gender} * x_{gender} + \beta_{age} * x_{age} + \beta_{bmi} * x_{bmi} + \beta_{sleep} * x_{sleep} + \beta_{smoke} * x_{smoke} + \beta_{avg\_alcohol\_freq\_wk} * x_{avg\_alcohol\_freq\_wk} + \epsilon$$

From the summary results of the linear regression model for the systolic blood pressure, we can find that although the coefficient of the working hours is positive, which implies that working overtime increases the risk of high systolic blood pressure, this effect is not significant ($\hat{\beta_{workhrs}} = 1.58$, p = 0.557). As a result, we can conclude that working overtime cannot lead to abnormal systolic blood pressure. Besides, from the summary results of the linear regression model for the diatolic blood pressure, we can find that although the coefficient of the working hours is positive as well, which implies that working overtime increases the risk of high diatolic blood pressure too, this effect is not significant as well ($\hat{\beta_{workhrs}} = 0.2246$, p = 0.905). As a result, we can conclude that working overtime also cannot lead to abnormal diatolic blood pressure. Therefore, in conclusion, working overtime also cannot lead to abnormal blood pressure.

After fitting the linear regression model, we want to check whether the following assumptions of the linear regression model are held:

1. Homoscedasticity: The variance of the residual is constant.
2. Linearity: The relationship between between the independent and dependent variables is linear.
3. No or little multicollinearity: There is no or little collinearity between independent variables.

1. Homoscedasticity

In order to check the homoscedasticity assumption, we plotted residual plots for both systolic blood pressure and diatolic pressure.

From the residual plot of the systolic blood pressure, we can find the mean of the residual is almost 0, and the variance seems to be almost constant. Also, we have the same finds for the diatolic blood pressure. Therefore, the assumption of homoscedasticity can be considered as satisfied.

2. Linearity

In order to check the linearity assumption, we plotted partial regression plots for both systolic blood pressure and diatolic pressure against each of the independent variables.

From the partial regression plots of the systolic blood pressure, we can find that for each of the independent variable, the expected value of the dependent variable (systolic blood pressure) is indeed a straight-line function of the independent variable, holding the others fixed. Also, from the partial regression plots of the diatolic blood pressurewe, we can get the same conclusion. Therefore, the assumption of linearity can be considered as satisfied.

3. No or little multicollinearity

In order to check the no/little-multicollinearity assumption, we computed the correlations between the continuous independent variables and the Pearson correlations between the binary independent variables.

As shown in the heatmap of the correlations between the continuous independent variables, we can find that there are very small or almost no correlations between the continuous independent variables. Also, as the table of the Pearson correlations between the binary variables shows, we can find that there is little collinearity between each pair of the binary variables.