Final Report

GROUP NUMBER 06

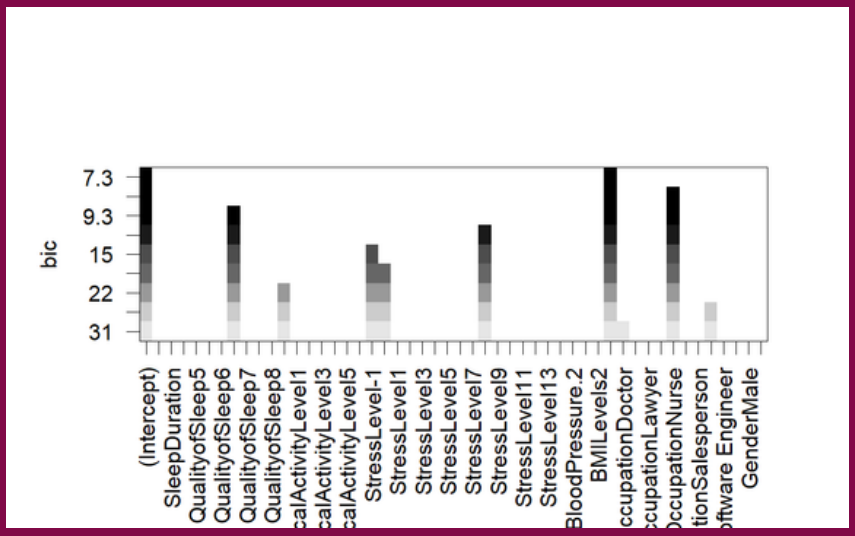
Interpretation of Results

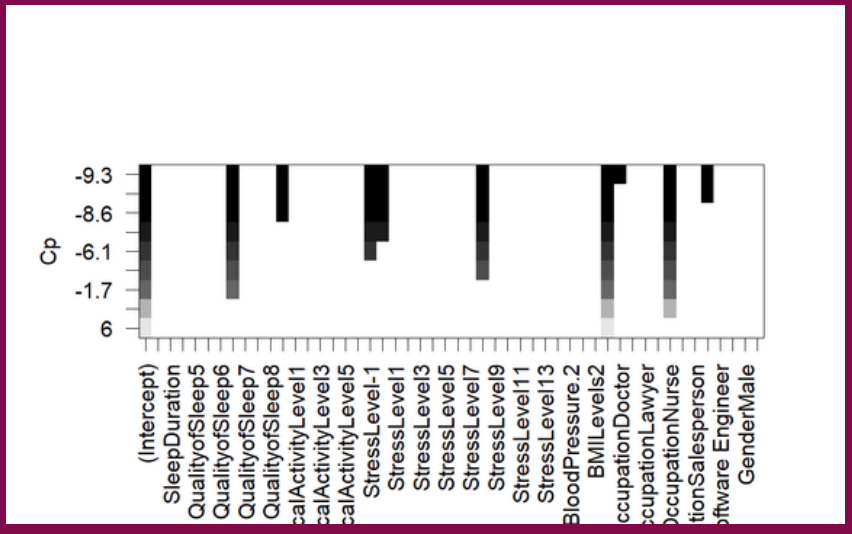
- First of all, the full model, which includes all the variables in the dataset, has been developed. Here, the intercept is significant under 5% significance level (p-value = 2.27e-06 < 0.05).
- Except the intercept, only two levels of the categorical variable "Occupation" was significant. Therefore, Occupation has been considered significant.
- All other variables are insignificant here in the full model.
- For the full model, the Residual Standard error is 10.09.
- That is the fitted value from the model can be deviated from the actual value by 10.09 units on average.
- The Multiple R squared value for this full model is 0.02147.
- That is only about 2.147% of the variability in the response variable is explained by the model. This indicates a weak fit.
- Adjusted R squared value for this model is 0.001484
- That value suggests that the model explains approximately 0.1484% of the variance in the response variable, considering the number of predictors.
- And finally, p value for the model is 0.3712 and it is greater than 0.05. Hence the conclusion was made as the overall model is not statistically significant.

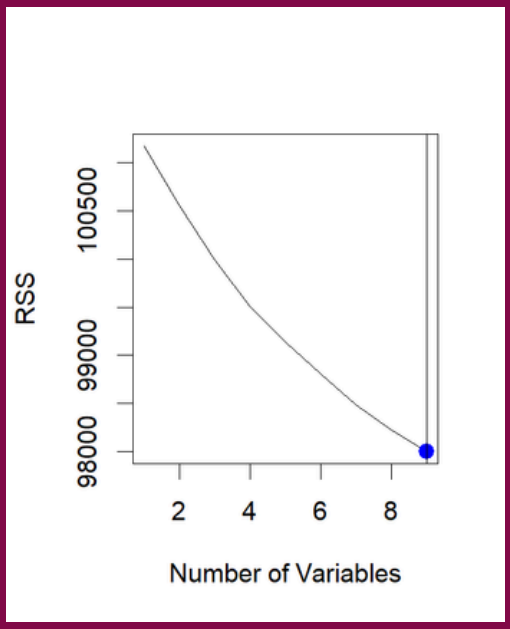
The procedure of choosing a better model

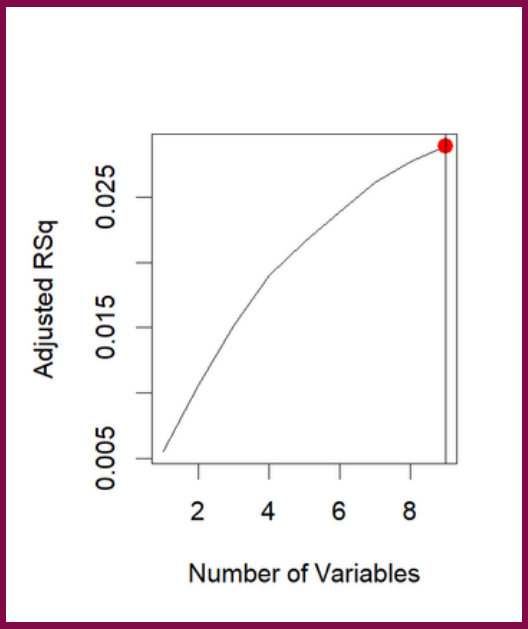
First of all the full model has been developed and checked for all the factors that interpretes the goodness of the model. As the full model is not significant, the necessity was to select a best model which is significant. For that purpose, the forward stepwise selection method has been used.
From that method 9 models were obtained and selcted the 9th model as the better model due to highest Adjusted R squared value and lowest CP & RSS values. Results were obtained as follows.

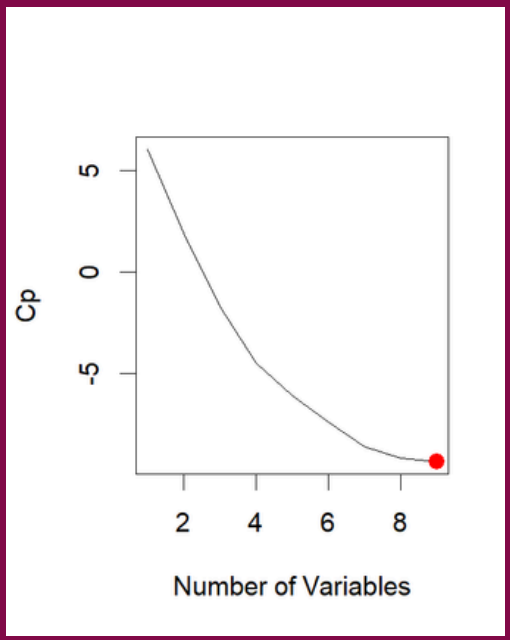
```
##       Adj.R2          CP       BIC       RSS
## 1 0.00553212   6.9526893  7.266529 101169.67
## 2 0.01060009   2.8416338  8.062582 100553.23
## 3 0.01518629  -0.7742675  9.320713  99986.75
## 4 0.01897007  -3.5740570 11.374430  99502.58
## 5 0.02155231  -5.1592779 14.641007  99140.94
## 6 0.02389377  -6.4976624 18.146312  98804.19
## 7 0.02610878  -7.7045485 21.774701  98480.70
## 8 0.02768260  -8.2648273 26.056569  98222.44
## 9 0.02887311  -8.4384464 30.729580  98003.19
```

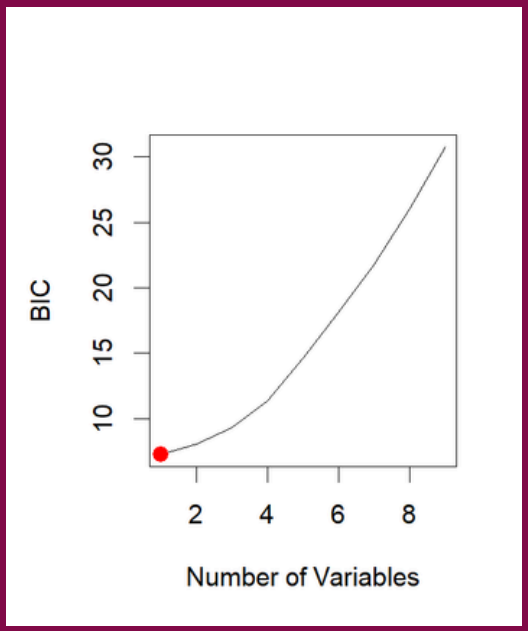





RSS
98000
99000
100500
2
4
6
8
Number of Variables

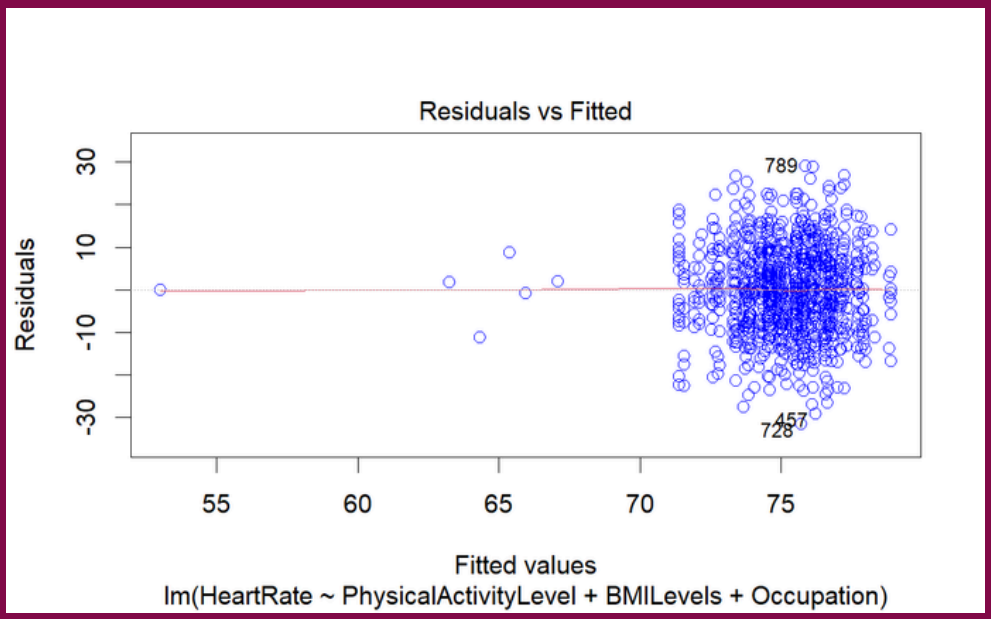

Adjusted RSq
0.005
0.015
0.025
2
4
6
8
Number of Variables


Cp
-5
0
5
2
4
6
8
Number of Variables

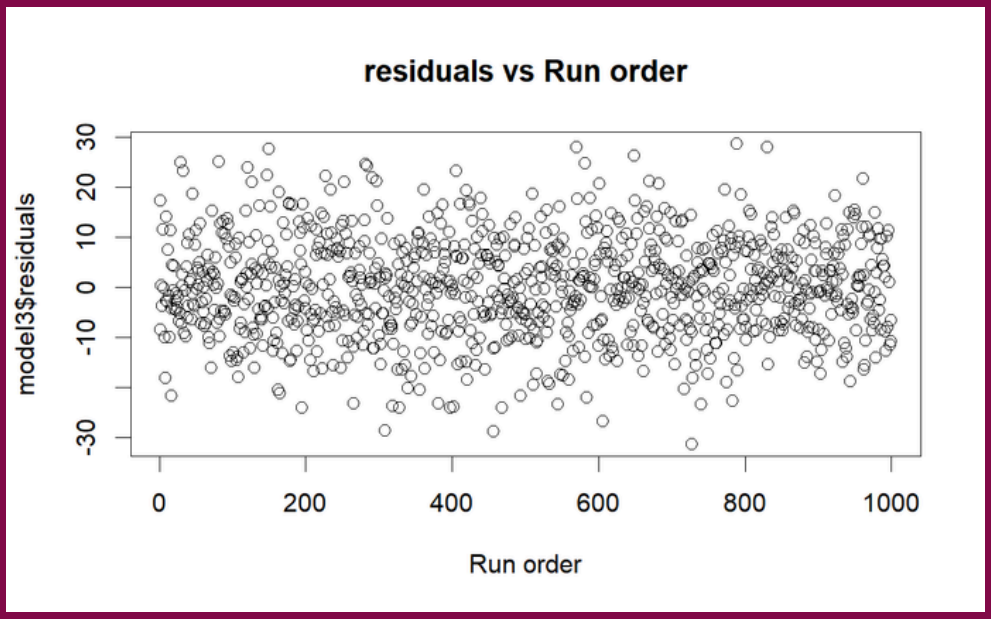

BIC
10
15
20
25
30
2
4
6
8
Number of Variables

Residual Analysis

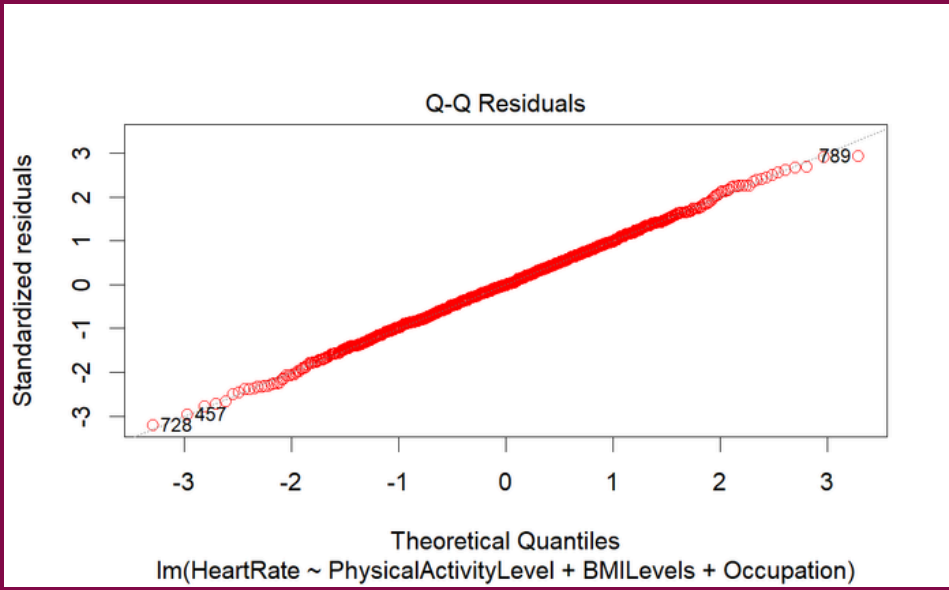
- Here the results we have obtained while checking the validity of the assumptions for the model 9 through the residual analysis.



- Residuals vs Fitted plot looks like all the residuals are randomly distributed. Hence variane of the residuals assumed to be constant. Model 9 would be Homoscedastic.



- It seems like residuals are randomly distributed along the time. Therefore residuals could assumed to be independent of each other.



- It seems to be most of the residuals fall approximately along the reference line. Therefore, residuals assumed to be normally distributed. Here few leverage points also can be observed.

- Then if look at the significance of the best model, the intercept is significant under 5% level of significance because its p value is lower than 0.05.
- The categorical variable "Physical Activity Level" is also significant under 5% level of significance because two factor levels of that variable are significant under 5% level of significance.
- "BMI level" is also significant here because it is also a categorical variable and it also belongs one factor level which is significant under 5% level of significance.
- The remaining variable of this model is "Occupation" and it is also a categorical variable with ten factor levels and from those factor levels, two levels are significant under 5% level of significance and therefore the variable "Occupation" is also significant here.
- That is, all three variables included in the model are significant with the intercept.
- Then, the Residuals Standard error of this model is 10.01.
- That is the fitted value from the model can be deviated from the actual value by 10.01 units on average.
- The Multiple R squared value for this full model is 0.03384.
- That is only about 3.384% of the variability in the response variable is explained by the model. This indicates also a weak fit.
- Adjusted R squared value for this model is 0.01611.
- That value suggests that the model explains approximately 1.611% of the variance in the response variable, considering the number of predictors.
- And the p value for the model is 0.01243 and it is a value lower than 0.05.
- Hence we can conclude that this model is statistically significant at 5% level of significant.

Discussion

- In our fitted model, the adjusted R-squared value is close to 2.9%. The CP and Sum of square residuals values are lower compared to other fitted models. However, the BIC value is higher than other fitted models, it represents the 2.9% percentage of variation of the full data set.
- According to fitted model, three factors affecting the heart rate could be identified.

  Physical Activity Level
  Occupation
  Body mass index (BMI)

Fitted Model

Heart Rate = 63.239+ 9.456(PhysicalActivityLevel 1)
+8.873(PhysicalActivityLevel 2)+8.327(PhysicalActivityLevel 3)
+ 9.532(PhysicalActivityLevel 4)+8.946(PhysicalActivityLevel 5)
-10.064(PhysicalActivityLevel 6)+2.271(BMI Levels 2)
+ 1.273(BMI Levels 3)+2.746(Doctor)+2.187(Engineer)+1.816(Lawyer)
-0.175(Manager)+2.958(Nurse)+1.726(Sales Representative)+3.860(Sales person)
+ 2.697(Scientist)+1.748(Software Engineer)+ 2.151(Teacher)

Team Members

PS/2020/029 - G.D.N.P.M.Gamage
PS/2020/053 - K.A.B.Chamudika
PS/2020/068 - W.W.G.V.Vinsura
PS/2020/097 - R.P.Y.P.Kumara
PS/2020/120 - A.H.P.Tarushika
PS/2020/168 - D.M.P.M.Dasanayaka
PS/2020/173 - H.S.K.Peiris
PS/2020/214 - M.S.Sendanayake
PS/2020/244 - W.D.T.D.P.Perera