

hw 4

Group

2024-12-03

```
setwd("/cloud/project")
sleepdata <- read.csv("Sleep_health_and_lifestyle_dataset (1).csv", header = TRUE)
attach(sleepdata)
#this will allow us to name variables just as they are
names(sleepdata)
```

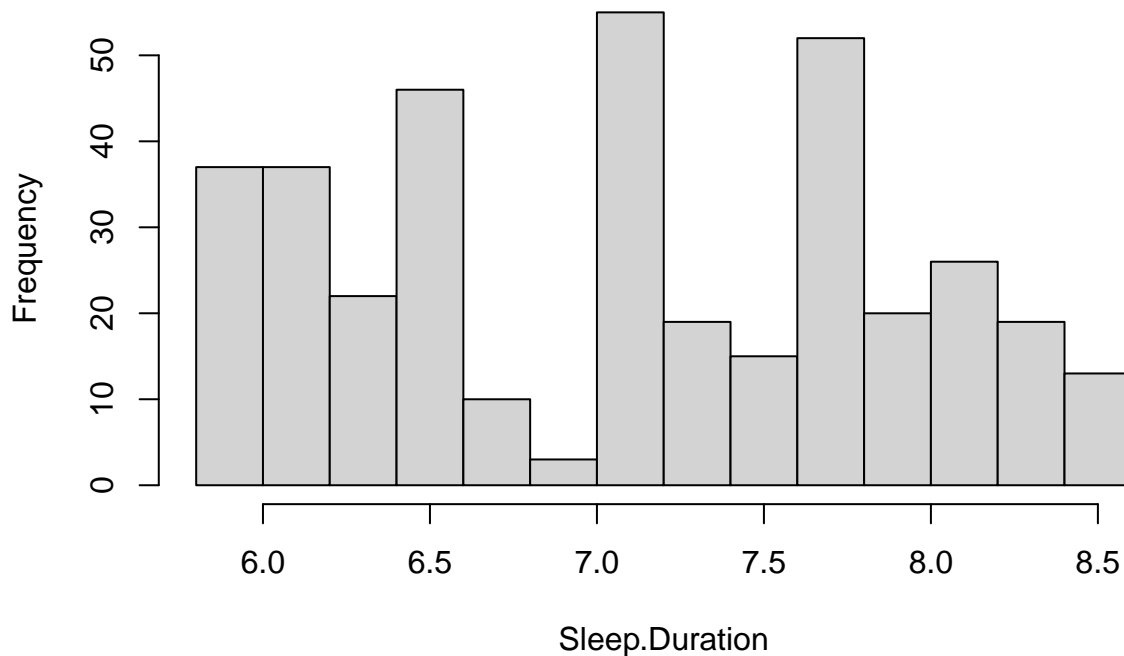
```
## [1] "Person.ID"          "Gender"
## [3] "Age"                "Occupation"
## [5] "Sleep.Duration"     "Quality.of.Sleep"
## [7] "Physical.Activity.Level" "Stress.Level"
## [9] "BMI.Category"       "Blood.Pressure"
## [11] "Heart.Rate"         "Daily.Steps"
## [13] "Sleep.Disorder"
```

```
shapiro.test(Sleep.Duration)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Sleep.Duration
## W = 0.93577, p-value = 1.268e-11
```

```
#Ho: data is not normal; ha: data is normal
#type I error set to 0.95
#reject null hypothesis that the outcome is not normal and conclude
#that the outcome is normal
hist(Sleep.Duration)
```

Histogram of Sleep.Duration



```
#creates a new variable systolic extracting the first 3 digits of
#the Bloodpressure
```

```
sleepdata$systolic = substr(Blood.Pressure, 1, 3)
sleepdata$systolic = as.numeric(sleepdata$systolic)
sleepdata$diastolic = substr(Blood.Pressure, 5, 6)
sleepdata$diastolic = as.numeric(sleepdata$diastolic)
install.packages("leaps")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(leaps)
```

```
#Now we run the regsubsets to find the best model
```

```
output <- regsubsets(Sleep.Duration ~ Gender + Age + Occupation +
                    Quality.of.Sleep + Physical.Activity.Level +
                    Stress.Level + BMI.Category + Heart.Rate + Daily.Steps +
                    Sleep.Disorder + systolic + diastolic, data=sleepdata,
                    nvmax=12)
```

```
summOut1 <- summary(output)
summOut1
```

```
## Subset selection object
```

```
## Call: regsubsets.formula(Sleep.Duration ~ Gender + Age + Occupation +
##   Quality.of.Sleep + Physical.Activity.Level + Stress.Level +
##   BMI.Category + Heart.Rate + Daily.Steps + Sleep.Disorder +
##   systolic + diastolic, data = sleepdata, nvmax = 12)
```

```
## 24 Variables (and intercept)
```

```
##
##               Forced in Forced out
## GenderMale      FALSE      FALSE
## Age             FALSE      FALSE
## OccupationDoctor FALSE      FALSE
```

```

## OccupationEngineer          FALSE      FALSE
## OccupationLawyer             FALSE      FALSE
## OccupationManager            FALSE      FALSE
## OccupationNurse              FALSE      FALSE
## OccupationSales Representative FALSE      FALSE
## OccupationSalesperson        FALSE      FALSE
## OccupationScientist          FALSE      FALSE
## OccupationSoftware Engineer  FALSE      FALSE
## OccupationTeacher            FALSE      FALSE
## Quality.of.Sleep             FALSE      FALSE
## Physical.Activity.Level       FALSE      FALSE
## Stress.Level                 FALSE      FALSE
## BMI.CategoryNormal Weight     FALSE      FALSE
## BMI.CategoryObese            FALSE      FALSE
## BMI.CategoryOverweight        FALSE      FALSE
## Heart.Rate                   FALSE      FALSE
## Daily.Steps                  FALSE      FALSE
## Sleep.DisorderNone           FALSE      FALSE
## Sleep.DisorderSleep Apnea     FALSE      FALSE
## systolic                     FALSE      FALSE
## diastolic                    FALSE      FALSE
## 1 subsets of each size up to 12
## Selection Algorithm: exhaustive
##      GenderMale Age OccupationDoctor OccupationEngineer OccupationLawyer
## 1  ( 1 ) " "      " " " "      " "      " "
## 2  ( 1 ) " "      " " "*"      " "      " "
## 3  ( 1 ) " "      " " "*"      "*"      " "
## 4  ( 1 ) " "      " " "*"      "*"      " "
## 5  ( 1 ) " "      " " "*"      "*"      " "
## 6  ( 1 ) " "      " " "*"      "*"      " "
## 7  ( 1 ) " "      " " "*"      "*"      " "
## 8  ( 1 ) " "      " " "*"      "*"      "*"
## 9  ( 1 ) " "      " " "*"      "*"      "*"
## 10 ( 1 ) " "      " " "*"      "*"      "*"
## 11 ( 1 ) " "      " " "*"      "*"      "*"
## 12 ( 1 ) " "      "*" "*"      "*"      "*"
##      OccupationManager OccupationNurse OccupationSales Representative
## 1  ( 1 ) " "      " "      " "
## 2  ( 1 ) " "      " "      " "
## 3  ( 1 ) " "      " "      " "
## 4  ( 1 ) " "      " "      " "
## 5  ( 1 ) " "      " "      " "
## 6  ( 1 ) " "      " "      " "
## 7  ( 1 ) " "      " "      " "
## 8  ( 1 ) " "      " "      " "
## 9  ( 1 ) " "      " "      " "
## 10 ( 1 ) " "      " "      "*"
## 11 ( 1 ) " "      " "      "*"
## 12 ( 1 ) " "      " "      "*"
##      OccupationSalesperson OccupationScientist OccupationSoftware Engineer
## 1  ( 1 ) " "      " "      " "
## 2  ( 1 ) " "      " "      " "
## 3  ( 1 ) " "      " "      " "
## 4  ( 1 ) " "      " "      " "

```

```

## 5 ( 1 ) " " " " " "
## 6 ( 1 ) " " " " " "
## 7 ( 1 ) "*" " " " "
## 8 ( 1 ) "*" " " " "
## 9 ( 1 ) "*" " " " "
## 10 ( 1 ) "*" " " " "
## 11 ( 1 ) "*" " " " "
## 12 ( 1 ) "*" " " " "
## OccupationTeacher Quality.of.Sleep Physical.Activity.Level
## 1 ( 1 ) " " "*" " "
## 2 ( 1 ) " " "*" " "
## 3 ( 1 ) " " "*" " "
## 4 ( 1 ) " " "*" " "
## 5 ( 1 ) " " "*" "*"
## 6 ( 1 ) "*" "*" "*"
## 7 ( 1 ) " " "*" "*"
## 8 ( 1 ) " " "*" "*"
## 9 ( 1 ) " " "*" " "
## 10 ( 1 ) " " "*" "*"
## 11 ( 1 ) " " "*" " "
## 12 ( 1 ) " " "*" "*"
## Stress.Level BMI.CategoryNormal Weight BMI.CategoryObese
## 1 ( 1 ) " " " " " "
## 2 ( 1 ) " " " " " "
## 3 ( 1 ) " " " " " "
## 4 ( 1 ) " " " " " "
## 5 ( 1 ) " " " " " "
## 6 ( 1 ) "*" " " " "
## 7 ( 1 ) "*" " " " "
## 8 ( 1 ) "*" " " " "
## 9 ( 1 ) "*" " " " "
## 10 ( 1 ) "*" " " "*"
## 11 ( 1 ) "*" " " "*"
## 12 ( 1 ) "*" " " " "
## BMI.CategoryOverweight Heart.Rate Daily.Steps Sleep.DisorderNone
## 1 ( 1 ) " " " " " "
## 2 ( 1 ) " " " " " "
## 3 ( 1 ) " " " " " "
## 4 ( 1 ) " " "*" " " "
## 5 ( 1 ) " " " " "*"
## 6 ( 1 ) " " " " " "
## 7 ( 1 ) " " "*" " " "
## 8 ( 1 ) " " "*" " " "
## 9 ( 1 ) " " "*" " " "
## 10 ( 1 ) " " "*" " " "
## 11 ( 1 ) " " "*" " " "
## 12 ( 1 ) " " " " "*" "
## Sleep.DisorderSleep Apnea systolic diastolic
## 1 ( 1 ) " " " " " "
## 2 ( 1 ) " " " " " "
## 3 ( 1 ) " " " " " "
## 4 ( 1 ) " " " " " "
## 5 ( 1 ) " " " " " "
## 6 ( 1 ) " " " " " "

```

```
## 7 ( 1 ) " " " " " "
## 8 ( 1 ) " " " " " "
## 9 ( 1 ) " " "*" "*"
## 10 ( 1 ) " " " " " "
## 11 ( 1 ) " " "*" "*"
## 12 ( 1 ) " " "*" "*"

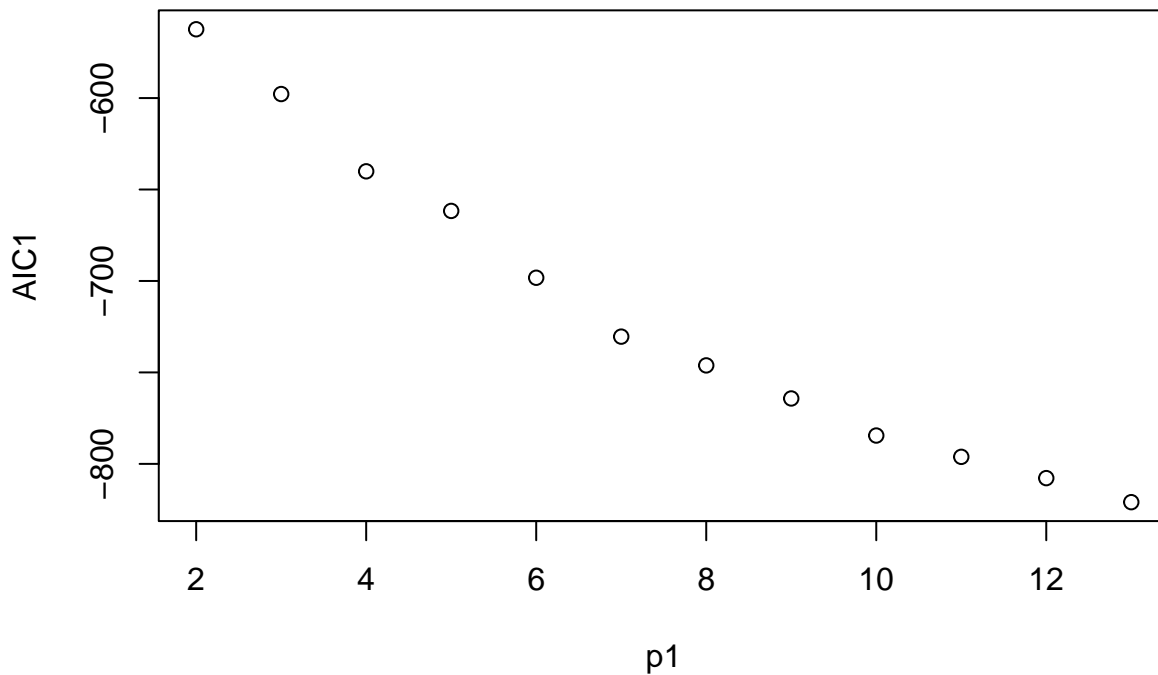
```

```
n1 <- length(Sleep.Duration)
n1
```

```
## [1] 374
```

```
p1 <- apply(summOut1$which, 1, sum)
aic1 <- summOut1$bic - log(n1) * p1 + 2 * p1
plot(p1, aic1, ylab = "AIC1")

```



```
summOut1
```

```
## Subset selection object
## Call: regsubsets.formula(Sleep.Duration ~ Gender + Age + Occupation +
##   Quality.of.Sleep + Physical.Activity.Level + Stress.Level +
##   BMI.Category + Heart.Rate + Daily.Steps + Sleep.Disorder +
##   systolic + diastolic, data = sleepdata, nvmax = 12)
## 24 Variables (and intercept)
##
##               Forced in Forced out
## GenderMale      FALSE      FALSE
## Age             FALSE      FALSE
## OccupationDoctor FALSE      FALSE
## OccupationEngineer FALSE      FALSE
## OccupationLawyer FALSE      FALSE
## OccupationManager FALSE      FALSE
## OccupationNurse  FALSE      FALSE
## OccupationSales Representative FALSE      FALSE
## OccupationSalesperson FALSE      FALSE
## OccupationScientist FALSE      FALSE

```

```

## OccupationSoftware Engineer      FALSE      FALSE
## OccupationTeacher                 FALSE      FALSE
## Quality.of.Sleep                  FALSE      FALSE
## Physical.Activity.Level            FALSE      FALSE
## Stress.Level                       FALSE      FALSE
## BMI.CategoryNormal Weight          FALSE      FALSE
## BMI.CategoryObese                  FALSE      FALSE
## BMI.CategoryOverweight             FALSE      FALSE
## Heart.Rate                        FALSE      FALSE
## Daily.Steps                       FALSE      FALSE
## Sleep.DisorderNone                 FALSE      FALSE
## Sleep.DisorderSleep Apnea          FALSE      FALSE
## systolic                          FALSE      FALSE
## diastolic                         FALSE      FALSE
## 1 subsets of each size up to 12
## Selection Algorithm: exhaustive
##      GenderMale Age OccupationDoctor OccupationEngineer OccupationLawyer
## 1  ( 1 )  " "      " " " "      " "      " "
## 2  ( 1 )  " "      " " "*"      " "      " "
## 3  ( 1 )  " "      " " "*"      "*"      " "
## 4  ( 1 )  " "      " " "*"      "*"      " "
## 5  ( 1 )  " "      " " "*"      "*"      " "
## 6  ( 1 )  " "      " " "*"      "*"      " "
## 7  ( 1 )  " "      " " "*"      "*"      " "
## 8  ( 1 )  " "      " " "*"      "*"      "*"
## 9  ( 1 )  " "      " " "*"      "*"      "*"
## 10 ( 1 )  " "      " " "*"      "*"      "*"
## 11 ( 1 )  " "      " " "*"      "*"      "*"
## 12 ( 1 )  " "      "*" "*"      "*"      "*"
##      OccupationManager OccupationNurse OccupationSales Representative
## 1  ( 1 )  " "      " "      " "
## 2  ( 1 )  " "      " "      " "
## 3  ( 1 )  " "      " "      " "
## 4  ( 1 )  " "      " "      " "
## 5  ( 1 )  " "      " "      " "
## 6  ( 1 )  " "      " "      " "
## 7  ( 1 )  " "      " "      " "
## 8  ( 1 )  " "      " "      " "
## 9  ( 1 )  " "      " "      " "
## 10 ( 1 )  " "      " "      "*"
## 11 ( 1 )  " "      " "      "*"
## 12 ( 1 )  " "      " "      "*"
##      OccupationSalesperson OccupationScientist OccupationSoftware Engineer
## 1  ( 1 )  " "      " "      " "
## 2  ( 1 )  " "      " "      " "
## 3  ( 1 )  " "      " "      " "
## 4  ( 1 )  " "      " "      " "
## 5  ( 1 )  " "      " "      " "
## 6  ( 1 )  " "      " "      " "
## 7  ( 1 )  "*"      " "      " "
## 8  ( 1 )  "*"      " "      " "
## 9  ( 1 )  "*"      " "      " "
## 10 ( 1 )  "*"      " "      " "
## 11 ( 1 )  "*"      " "      " "

```

```

## 12 ( 1 ) "*" " " " "
## OccupationTeacher Quality.of.Sleep Physical.Activity.Level
## 1 ( 1 ) " " "*" " "
## 2 ( 1 ) " " "*" " "
## 3 ( 1 ) " " "*" " "
## 4 ( 1 ) " " "*" " "
## 5 ( 1 ) " " "*" "*"
## 6 ( 1 ) "*" "*" "*"
## 7 ( 1 ) " " "*" "*"
## 8 ( 1 ) " " "*" "*"
## 9 ( 1 ) " " "*" " "
## 10 ( 1 ) " " "*" "*"
## 11 ( 1 ) " " "*" " "
## 12 ( 1 ) " " "*" "*"
## Stress.Level BMI.CategoryNormal Weight BMI.CategoryObese
## 1 ( 1 ) " " " " " "
## 2 ( 1 ) " " " " " "
## 3 ( 1 ) " " " " " "
## 4 ( 1 ) " " " " " "
## 5 ( 1 ) " " " " " "
## 6 ( 1 ) "*" " " " "
## 7 ( 1 ) "*" " " " "
## 8 ( 1 ) "*" " " " "
## 9 ( 1 ) "*" " " " "
## 10 ( 1 ) "*" " " "*"
## 11 ( 1 ) "*" " " "*"
## 12 ( 1 ) "*" " " " "
## BMI.CategoryOverweight Heart.Rate Daily.Steps Sleep.DisorderNone
## 1 ( 1 ) " " " " " "
## 2 ( 1 ) " " " " " "
## 3 ( 1 ) " " " " " "
## 4 ( 1 ) " " "*" " "
## 5 ( 1 ) " " " "*" " "
## 6 ( 1 ) " " " " " "
## 7 ( 1 ) " " "*" " "
## 8 ( 1 ) " " "*" " "
## 9 ( 1 ) " " "*" " "
## 10 ( 1 ) " " "*" " "
## 11 ( 1 ) " " "*" " "
## 12 ( 1 ) " " " " "*" " "
## Sleep.DisorderSleep Apnea systolic diastolic
## 1 ( 1 ) " " " " " "
## 2 ( 1 ) " " " " " "
## 3 ( 1 ) " " " " " "
## 4 ( 1 ) " " " " " "
## 5 ( 1 ) " " " " " "
## 6 ( 1 ) " " " " " "
## 7 ( 1 ) " " " " " "
## 8 ( 1 ) " " " " " "
## 9 ( 1 ) " " "*" "*"
## 10 ( 1 ) " " " " " "
## 11 ( 1 ) " " "*" "*"
## 12 ( 1 ) " " "*" "*"

```

```

#best model is the one with all the predictors as it has the lowest AIC
modell1 <- lm(Sleep.Duration ~ Gender + Age + Occupation + Quality.of.Sleep +
              Physical.Activity.Level +
              Stress.Level + BMI.Category + Heart.Rate + Daily.Steps +
              Sleep.Disorder + systolic + diastolic, data=sleepdata)
summary(modell1)

##
## Call:
## lm(formula = Sleep.Duration ~ Gender + Age + Occupation + Quality.of.Sleep +
##     Physical.Activity.Level + Stress.Level + BMI.Category + Heart.Rate +
##     Daily.Steps + Sleep.Disorder + systolic + diastolic, data = sleepdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71735 -0.14289 -0.03386  0.13013  0.97101
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      6.5255127   1.1566767    5.642 3.49e-08 ***
## GenderMale      -0.0163363   0.0852287   -0.192 0.848107
## Age              0.0274491   0.0065377    4.199 3.41e-05 ***
## OccupationDoctor  0.8323400   0.0861232    9.665 < 2e-16 ***
## OccupationEngineer 0.7772756   0.0867901    8.956 < 2e-16 ***
## OccupationLawyer   0.7294035   0.0989878    7.369 1.26e-12 ***
## OccupationManager  0.1027142   0.2601035    0.395 0.693160
## OccupationNurse    0.2427417   0.1128672    2.151 0.032187 *
## OccupationSales Representative 1.4483922   0.2358399    6.141 2.23e-09 ***
## OccupationSalesperson 0.6357357   0.1080311    5.885 9.35e-09 ***
## OccupationScientist 0.4568531   0.1701922    2.684 0.007614 **
## OccupationSoftware Engineer 0.6326634   0.1507845    4.196 3.45e-05 ***
## OccupationTeacher  0.2883608   0.0883850    3.263 0.001213 **
## Quality.of.Sleep  0.2860928   0.0561463    5.095 5.71e-07 ***
## Physical.Activity.Level 0.0092998   0.0015524    5.991 5.20e-09 ***
## Stress.Level     -0.1628751   0.0341770   -4.766 2.77e-06 ***
## BMI.CategoryNormal Weight -0.0338319   0.0682966   -0.495 0.620653
## BMI.CategoryObese  -0.6002143   0.1938760   -3.096 0.002121 **
## BMI.CategoryOverweight -0.3467385   0.1028555   -3.371 0.000832 ***
## Heart.Rate       0.0332898   0.0101959    3.265 0.001203 **
## Daily.Steps      -0.0001284   0.0000219   -5.863 1.05e-08 ***
## Sleep.DisorderNone -0.1020383   0.0602559   -1.693 0.091268 .
## Sleep.DisorderSleep Apnea -0.0549659   0.0674340   -0.815 0.415567
## systolic        -0.1212507   0.0164668   -7.363 1.30e-12 ***
## diastolic        0.1359929   0.0221031    6.153 2.09e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2442 on 349 degrees of freedom
## Multiple R-squared:  0.9119, Adjusted R-squared:  0.9058
## F-statistic: 150.4 on 24 and 349 DF,  p-value: < 2.2e-16
table(Occupation)

## Occupation

```



```
##           Accountant           Doctor           Engineer
##           37           71           63
##           Lawyer           Manager           Nurse
##           47           1           73
## Sales Representative           Salesperson           Scientist
##           2           32           4
## Software Engineer           Teacher
##           4           40
```

```
table(BMI.Category)
```

```
## BMI.Category
##           Normal Normal Weight           Obese           Overweight
##           195           21           10           148
```

```
table(Sleep.Disorder)
```

```
## Sleep.Disorder
##           Insomnia           None Sleep Apnea
##           77           219           78
```

```
#Integration of the significant variables from this model:
#Seep Duration increases significantly by 0.027 units for every unit
#Increase in age, adjusting for everything else
#Sleep Duration increases significantly by 0.027 units for every unit
#increase in age, adjusting for everything else

#Sleep Duration increases significantly by 0.83 units for Doctors
# vs accountants, adjusting for everything else

#Sleep Duration increases significantly by 0.78 units for Engineers
# vs. Accountants, adjusting for everything else

# Sleep Duration increases significantly by 0.73 units for Lawyers
# vs Accountants, adjusting for everything else

# Sleep Duration increases significantly by 0.24 units for Nurses
# vs Accountants, adjusting for everything else

# Sleep Duration increases significantly by 1.45 units for Sales Reps
# vs Accountants, adjusting for everything else

# Sleep Duration increases significantly by 0.63 units for SalesPerson
# vs Accountants, adjusting for everything else

# Sleep Duration increases significantly by 0.46 units for Scientists
#vs Accountants, adjusting for everything else

#Sleep Duration increases significantly by 0.63 units for SoftwareEngineers
#vs Accountants, adjusting for everything else

# Sleep Duration increases significantly by 0.29 units for Teachers
#vs Accountants, adjusting for everything else

# Sleep Duration increases significantly by 0.29 units for every unit
#increase in quality of sleep score, adjusting for everything else
```

```
#Sleep Duration increases significantly by 0.009 units for every unit
#increase in physical activity level score, adjusting for everything else
```

```
#Sleep Duration decreases significantly by 0.16 units for every unit
#increase in stress level score, adjusting for everything else
```

```
#Increase in daily steps, adjusting for everything else
```

```
#Sleep Duration decreases significantly by 0.121 units for every unit
#Increase in systolic reading, adjusting for everything else
```

```
#Sleep Duration increases significantly by 0.13 units for every unit
#increase in diastolic reading, adjusting for everything else
```

```
#we check for multicollinearity using vif and tolerance
install.packages("car")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.4'
## (as 'lib' is unspecified)
```

```
library(car)
```

```
## Loading required package: carData
```

```
vif(model1)
```

##		GVIF	Df	GVIF ^{1/(2*Df)}
## Gender	11.386299	1		3.374359
## Age	20.107382	1		4.484126
## Occupation	2933.052016	10		1.490617
## Quality.of.Sleep	28.246011	1		5.314698
## Physical.Activity.Level	6.539635	1		2.557271
## Stress.Level	23.003395	1		4.796185
## BMI.Category	110.831848	3		2.191681
## Heart.Rate	11.120014	1		3.334668
## Daily.Steps	7.851678	1		2.802085
## Sleep.Disorder	11.655508	2		1.847706
## systolic	101.805454	1		10.089869
## diastolic	115.998380	1		10.770254

```
#if the vif shows greater than 10, it implies that there is such a strong
#relationship between variables, such that these may be collinear
#If collinear, this will bias the results of the model from our results
#we see that systolic and diastolic may be collinear
#tolerance the inverse of vif; we run this as an extra check
1/vif(model1)
```

##		GVIF	Df	GVIF ^{1/(2*Df)}
## Gender	0.0878248485	1.0000000		0.29635257
## Age	0.0497329793	1.0000000		0.22300892
## Occupation	0.0003409418	0.1000000		0.67086327
## Quality.of.Sleep	0.0354032289	1.0000000		0.18815746
## Physical.Activity.Level	0.1529137380	1.0000000		0.39104186
## Stress.Level	0.0434718439	1.0000000		0.20849903
## BMI.Category	0.0090226773	0.3333333		0.45627069
## Heart.Rate	0.0899279456	1.0000000		0.29987989

## Daily.Steps	0.1273613137	1.0000000	0.35687717
## Sleep.Disorder	0.0857963433	0.5000000	0.54121172
## systolic	0.0098226564	1.0000000	0.09910932
## diastolic	0.0086208100	1.0000000	0.09284832

*#From these results we look for the last column to be >0.10, if it less than this
#it implies collinear, the two variables are systolic and diastolic*