Midterm Exam

Wendy Liang

Nov 2, 2020

Instruction

This is your midterm exam that you are expected to work on it alone. You may NOT discuss any of the content of your exam with anyone except your instructor. This includes text, chat, email and other online forums. We expect you to respect and follow the GRS Academic and Professional Conduct Code.

Although you may NOT ask anyone directly, you are allowed to use external resources such as R codes on the Internet. If you do use someone's code, please make sure you clearly cite the origin of the code.

When you finish, please compile and submit the PDF file and the link to the GitHub repository that contains the entire analysis.

Introduction

In this exam, you will act as both the client and the consultant for the data that you collected in the data collection exercise (20pts). Please note that you are not allowed to change the data. The goal of this exam is to demonstrate your ability to perform the statistical analysis that you learned in this class so far. It is important to note that significance of the analysis is not the main goal of this exam but the focus is on the appropriateness of your approaches.

Data Description (10pts)

Please explain what your data is about and what the comparison of interest is. In the process, please make sure to demonstrate that you can load your data properly into R.

1. Question: What to compare

I want to compare the sleep time of man and woman, among my close contacts.

In addition, I wonder which sex spends longer time on smart phone, since using smart phone might be an impact factor of sleep time.

2. Method: How to collect data

I make a online questionnaire survey by app *Questionnaire Star*, setting several choices and completions. Then, I publish it on my social media *Wetchat* and ask my friends and families to finish it. Finally, I manually fill these data into a csv file.

The following are my questions on the questionnaire:

- 1.Your gender?
- 2.Your age?
- 3.How many hours do you sleep at night per day?
- 4.What time do you fall asleep?
- 5.What time do you get up?

- 6.Do you take a nap after lunch? If you do, how many minutes do you sleep?
- 7.How many hours do you use smart phone per day?
- 8.What activities take most time when using smart phone?
- 9. Within an hour before sleep, how many minutes do you use smart phone?
- 10.Within an hour after getting up, how many minutes do you use smart phone?

3. My Dataset

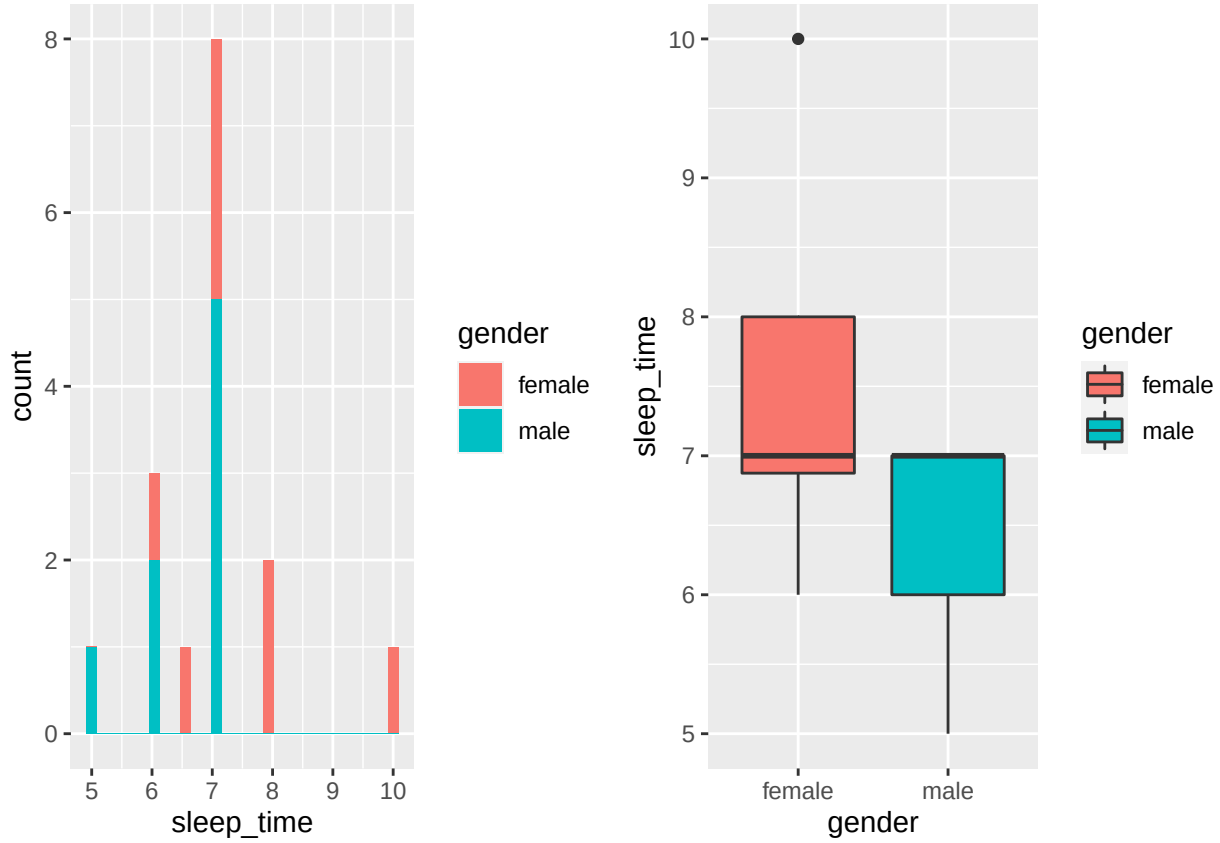
- gender: 0 for woman; 1 for man;
- age: 1 for 0-20; 2 for 21-30; 3 for 31-40; 4 for 41-50; 5 for 51-60; 6 for 61-70; 7 for 71-80;
- sleep_time: the average time people sleep at night per day
- getup/asleep: the tome people get up or fall asleep
- nap: the average time people take a nap
- phone_time: the average time people use smart phones per day
- most_act: 1 for game; 2 for work/study; 3 for social media; 4 for video;
- before_sleep/after_getup: the time people use smart phones within an hour before sleep or after grtting up

```
sleep=data.frame(read.csv("sleep.csv"))
sleep$gender=factor(sleep$gender,labels = c("female","male"))
sleep$age=factor(sleep$age)
sleep$most_act=factor(sleep$most_act)
#n1=17,n2=8
ind=sample(c(1:17),8)
#ind=8,7,3,11,14,4,13,2
sleep=sleep[c(2,3,4,7,8,11,13,14,18:25),]

female=sleep %>% filter(gender=="female")
n1=nrow(female)
male=sleep %>% filter(gender=="male")
n2=nrow(male)
n1==n2
n=n1
```

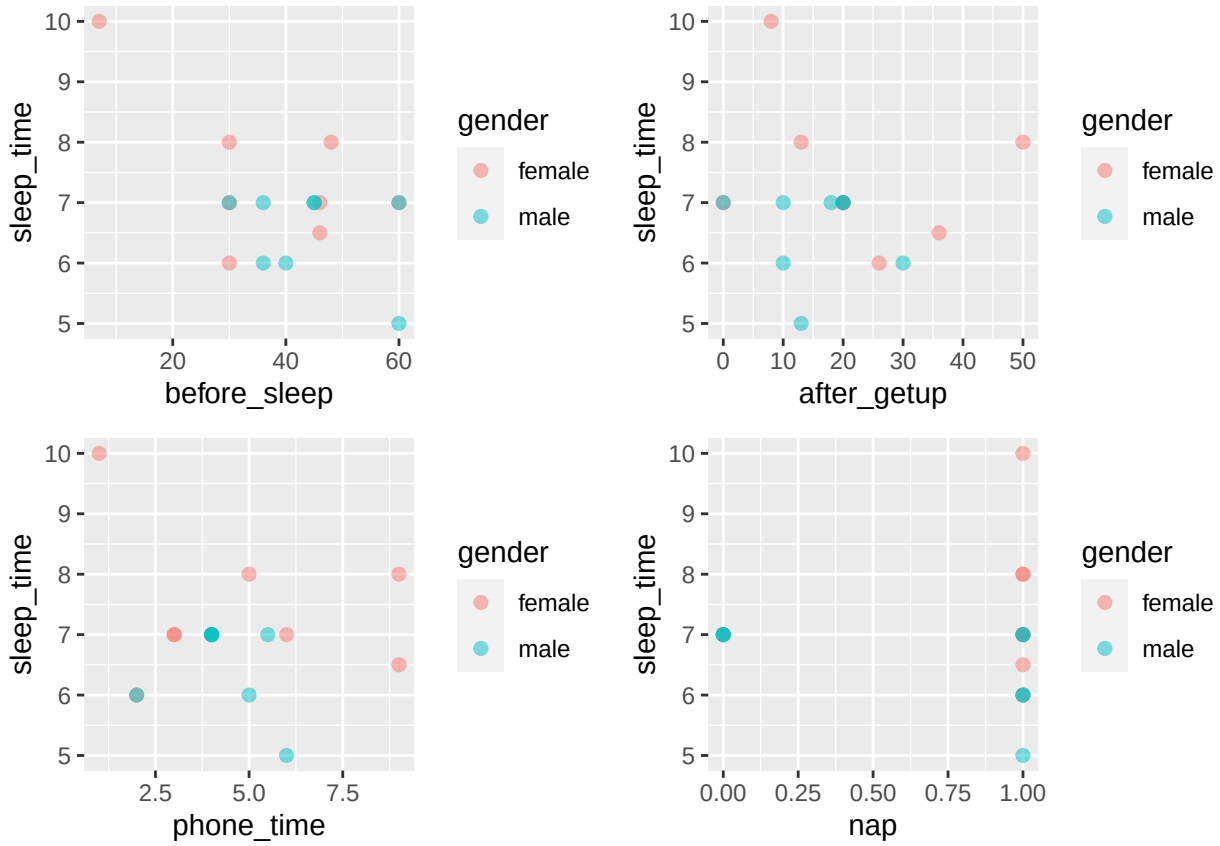
EDA (10pts)

```
p0=ggplot(sleep)+
  geom_histogram(aes(x=sleep_time,fill=gender))
# boxplot
p1=ggplot(sleep)+
  geom_boxplot(aes(x=gender,y=sleep_time,fill=gender))


# continuous variables: line plot
p2=ggplot(sleep)+
  geom_point(aes(x=before_sleep,y=sleep_time,col=gender),size=2,alpha=0.5)
```

```
p3=ggplot(sleep)+
  geom_point(aes(x=after_getup,y=sleep_time,col=gender),size=2,alpha=0.5)

p4=ggplot(sleep)+
  geom_point(aes(x=phone_time,y=sleep_time,col=gender),size=2,alpha=0.5)

p5=ggplot(sleep)+
  geom_point(aes(x=nap,y=sleep_time,col=gender),size=2,alpha=0.5)

cowplot::plot_grid(p0,p1,nrow=1)
```




```
cowplot::plot_grid(p2,p3,p4,p5,nrow=2)
```



There are some findings from the plots:

- The sleep_time of male is less than female
- most sample people sleep 7 hours per night
- there are non linear relationship between the phone related variables and sleep_time

Power Analysis (10pts)

Please perform power analysis on the project. Use 80% power, the sample size you used and infer the level of effect size you will be able to detect. Discuss whether your sample size was enough for the problem at hand. Please note that method of power analysis should match the analysis. Also, please clearly state why you should NOT use the effect size from the fitted model.

1. test the variances of two groups

```
sd(sleep$sleep_time)
```

```
## [1] 1.102554
```

```
var.test(sleep_time ~ gender, sleep, alternative = "two.sided")
```

```
##
##  F test to compare two variances
##
## data:  sleep_time by gender
## F = 2.6797, num df = 7, denom df = 7, p-value = 0.2168
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##   0.5364837 13.3847958
```

```
## sample estimates:
## ratio of variances
##           2.679688
```

The p-value is much more than 0.05, so we can say there's no significant difference between the sleep_time of male and female groups.

So, next we can do t-test to test the mean values of the two groups.

2. calculate the effect size

```
pwr.t.test(n=n, sig.level = 0.05, power = 0.8 ,alternative = "two.sided",type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 8
##               d = 1.50665
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```
#pwr.anova.test(k=2,n=n,sig.level = 0.05, power = 0.8)
```

The result infers that the effect size is 1.51, which is large effect size respectively.

I suppose the effect size is medium, d=0.5, I can detect more subtle difference between the two gender groups. That's also the reason why we should NOT use the effect size from the fitted model.

```
pwr.t.test(d=0.5,sig.level = 0.05, power = 0.8 ,alternative = "two.sided")
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76561
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

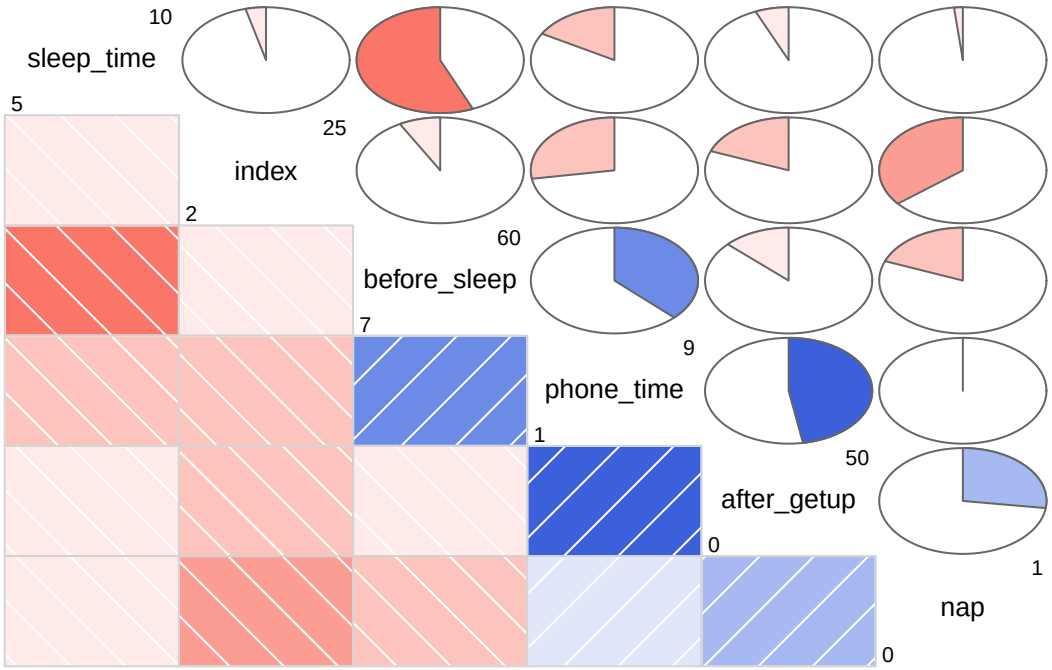
In this case, we need the number of each group is **64**.

Modeling (10pts)

Please pick a regression model that best fits your data and fit your model. Please make sure you describe why you decide to choose the model. Also, if you are using GLM, make sure you explain your choice of link function as well.

1. choose the independent variables

```
correlation=data.frame(var(sleep))
corrgram(sleep,order=T,
         lower.panel=panel.shade, upper.panel=panel.pie,
         diag.panel=panel.minmax, text.panel=panel.txt)
```



According to the correlation plot, I found that all the variables are correlated with **sleep_time**.
So, I choose the predictors by AIC in a stepwise algorithm, using `step()` function. As a result, `nap`,`age` will be the predictors.

2. build regression model

Firstly, I use linear regression model

```
# complete
fit0 = lm(sleep_time~gender+before_sleep+after_getup+phone_time+nap+age+most_act,data = sleep)
fit1 =step(fit0,direction="both")
```

```
## Start:  AIC=-0.63
## sleep_time ~ gender + before_sleep + after_getup + phone_time +
##     nap + age + most_act
##
##                Df Sum of Sq    RSS     AIC
## - nap           1    0.0061 3.8959 -2.6028
## - before_sleep  1    0.0644 3.9542 -2.3648
## - gender        1    0.2566 4.1464 -1.6057
## - after_getup   1    0.3690 4.2589 -1.1774
## <none>                      3.8898 -0.6277
## - age           2    1.1874 5.0772 -0.3651
## - phone_time    1    0.5993 4.4891 -0.3350
## - most_act      3    4.3927 8.2825  5.4650
##
## Step:  AIC=-2.6
## sleep_time ~ gender + before_sleep + after_getup + phone_time +
##     age + most_act
##
##                Df Sum of Sq    RSS     AIC
## - before_sleep  1    0.0813 3.9772 -4.2723
## - gender        1    0.3108 4.2066 -3.3749
## - after_getup   1    0.3702 4.2661 -3.1503
```

```
## <none>                      3.8959 -2.6028
## - phone_time    1    0.6120 4.5079 -2.2682
## + nap           1    0.0061 3.8898 -0.6277
## - age           2    1.7848 5.6807 -0.5683
## - most_act      3    4.6222 8.5181  3.9137
##
## Step:  AIC=-4.27
## sleep_time ~ gender + after_getup + phone_time + age + most_act
##
##                Df Sum of Sq     RSS     AIC
## <none>                       3.9772 -4.2723
## - gender        1    0.7709  4.7481 -3.4377
## + before_sleep  1    0.0813  3.8959 -2.6028
## - age           2    1.7039  5.6811 -2.5672
## + nap           1    0.0229  3.9542 -2.3648
## - after_getup   1    2.2075  6.1846  0.7916
## - phone_time    1    2.3374  6.3146  1.1244
## - most_act      3    8.9659 12.9430  8.6075
```

```
summary(fit1)
```

```
##
## Call:
## lm(formula = sleep_time ~ gender + after_getup + phone_time +
##     age + most_act, data = sleep)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97788 -0.35459  0.05754  0.34181  1.13467
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.3517     1.1678   5.439 0.000967 ***
## gendermale   -0.6488     0.5570  -1.165 0.282251
## after_getup   0.0473     0.0240   1.971 0.089347 .
## phone_time   -0.2386     0.1177  -2.028 0.082121 .
## age2          0.9055     0.7467   1.213 0.264583
## age3          2.0242     1.1700   1.730 0.127214
## most_act2     1.3537     0.6710   2.018 0.083422 .
## most_act4     0.1862     0.5194   0.359 0.730524
## most_act5    -3.1286     1.1533  -2.713 0.030083 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7538 on 7 degrees of freedom
## Multiple R-squared:  0.7819, Adjusted R-squared:  0.5326
## F-statistic: 3.137 on 8 and 7 DF,  p-value: 0.07488
```

```
R1=summary(fit1)$r.squared

# standard
sleep$sleep_time_std=(sleep$sleep_time-mean(sleep$sleep_time))/sd(sleep$sleep_time)
sleep$nap_std=(sleep$nap-mean(sleep$nap))/sd(sleep$nap)
fit2=lm(sleep_time_std~gender+age+nap_std,data=sleep)
summary(fit2)
```

```
##
## Call:
## lm(formula = sleep_time_std ~ gender + age + nap_std, data = sleep)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4651 -0.3785  0.1163  0.3721  2.1628
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1996     1.0761  -0.185    0.856
## gendermale   -0.6442     0.7040  -0.915    0.380
## age2          0.5349     0.8967   0.597    0.563
## age3          0.8209     1.2504   0.657    0.525
## nap_std      -0.0624     0.3081  -0.203    0.843
##
## Residual standard error: 1.014 on 11 degrees of freedom
## Multiple R-squared:  0.246,  Adjusted R-squared:  -0.02818
## F-statistic: 0.8972 on 4 and 11 DF,  p-value: 0.4979
```

```
R2=summary(fit2)$r.squared

# compare the R^2
print(c(R1,R2))
```

```
## [1] 0.7818859 0.2460000
```

These two regression model have the same result, so we choose the simplest one – the first model.

We can also try multilevel regression model as the following:

```
#### multilevel regression
# varying in intercept of gender, considering without group variance
fit3 = lmer(sleep_time~(1|gender)+age+nap,data = sleep)
coef(fit3)$gender
```

```
##        (Intercept)     age2     age3       nap
## female     5.81383 1.005319 1.728723 0.1861702
## male       5.81383 1.005319 1.728723 0.1861702
```

```
# varying in intercept and scope of gender, considering within group variance
fit4=lmer(sleep_time~(1+gender|gender)+age+nap,data = sleep)
coef(fit4)$gender
```

```
##           gendermale (Intercept)     age2     age3       nap
## female  0.000000e+00     5.81383 1.005319 1.728723 0.1861702
## male   -1.176256e-08     5.81383 1.005319 1.728723 0.1861702
```

```
print(c(AIC3=display(fit3)$AIC,AIC4=display(fit4)$AIC))
```

```
## lmer(formula = sleep_time ~ (1 | gender) + age + nap, data = sleep)
##             coef.est coef.se
## (Intercept) 5.81     1.02
## age2        1.01     0.87
## age3        1.73     1.04
## nap         0.19     0.66
##
## Error terms:
```

```
##  Groups   Name        Std.Dev.
##  gender   (Intercept) 0.00
##  Residual             1.11
## ---
## number of obs: 16, groups: gender, 2
## AIC = 53.8, DIC = 46.5
## deviance = 44.2
## lmer(formula = sleep_time ~ (1 + gender | gender) + age + nap,
##     data = sleep)
##             coef.est coef.se
## (Intercept) 5.81     1.02
## age2        1.01     0.87
## age3        1.73     1.04
## nap         0.19     0.66
##
## Error terms:
##  Groups   Name        Std.Dev. Corr
##  gender   (Intercept) 0.00
##           gendermale  0.00      NaN
##  Residual             1.11
## ---
## number of obs: 16, groups: gender, 2
## AIC = 57.8, DIC = 46.5
## deviance = 44.2
##     AIC3     AIC4
## 53.80373 57.80373
```

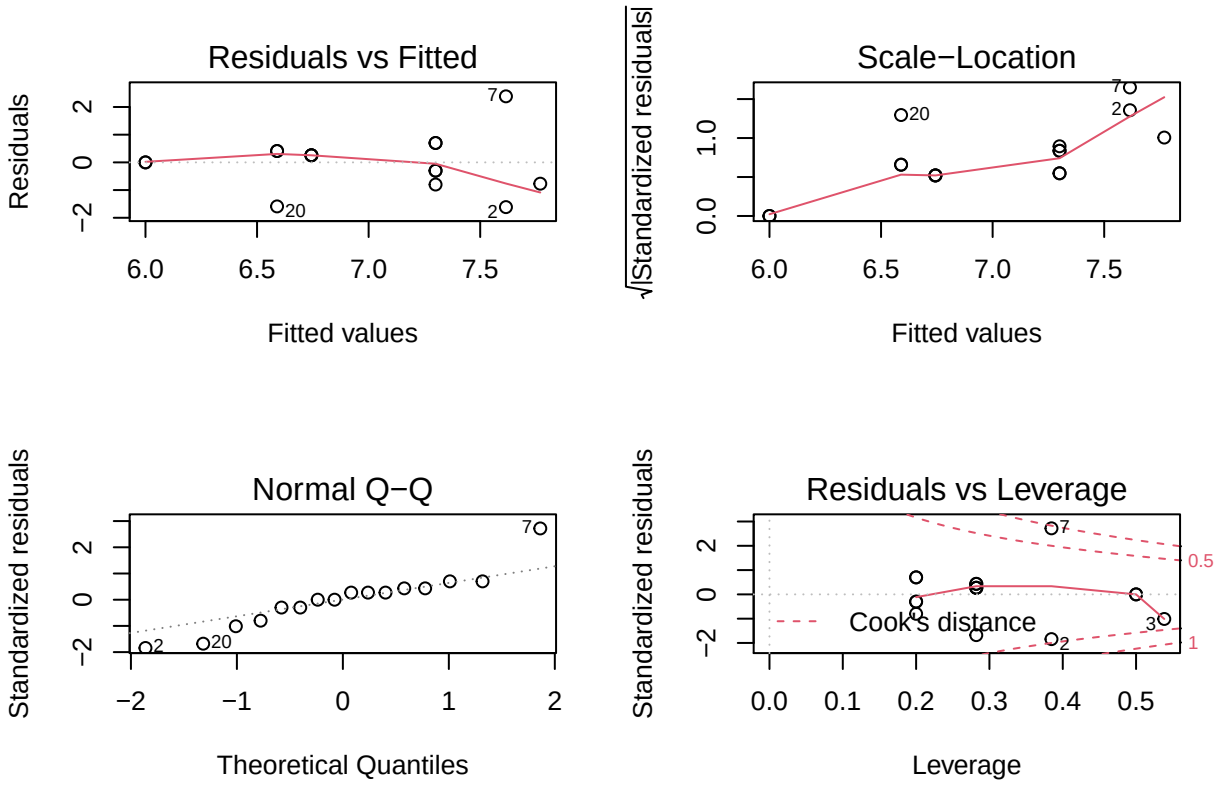
Since the $AIC4 > AIC3$, we choose the `lmer(formula = sleep_time ~ (1 | gender) + age + nap` model.

According to the result, I found that the residual-fitted value plots of model 1 and model 3 are very similar. So I choose the model 1 for further work.

Validation (10pts)

Please perform a necessary validation and argue why your choice of the model is appropriate.

```
fit1=lm(formula = sleep_time ~ gender + nap + age, data = sleep)
layout(matrix(c(1,2,3,4),2,2))
plot(fit1)
```



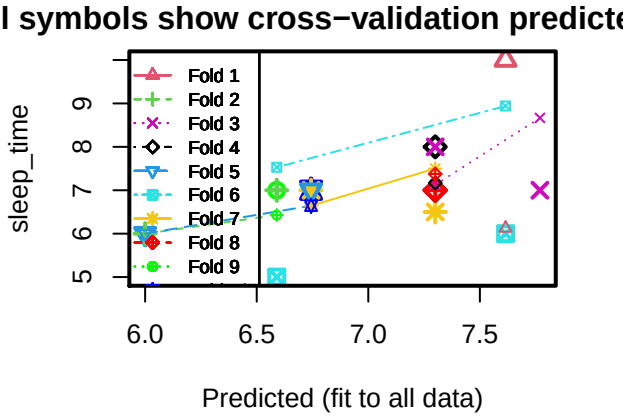

```
# 10-fold cross validation
cv.lm(fit1,data=sleep,m=10)
```

```
## Analysis of Variance Table
##
## Response: sleep_time
##           Df Sum Sq Mean Sq F value Pr(>F)
## gender     1   3.52    3.52    2.81   0.12
## nap        1   0.41    0.41    0.33   0.58
## age        2   0.56    0.28    0.22   0.80
## Residuals 11  13.75    1.25
```

```
##
## fold 1
## Observations in test set: 1
##                 7
## Predicted    7.62
## cvpred       6.12
## sleep_time  10.00
## CV residual  3.88
##
## Sum of squares = 15    Mean square = 15    n = 1
##
## fold 2
## Observations in test set: 2
##                21        25
## Predicted   6.590  6.00e+00
## cvpred      6.429  6.00e+00
## sleep_time  7.000  6.00e+00
## CV residual 0.571 -8.88e-16
```

```
##
## Sum of squares = 0.33    Mean square = 0.16    n = 2
##
## fold 3
## Observations in test set: 2
##                 3     8
## Predicted    7.77 7.300
## cvpred       8.67 7.125
## sleep_time   7.00 8.000
## CV residual -1.67 0.875
##
## Sum of squares = 3.54    Mean square = 1.77    n = 2
##
## fold 4
## Observations in test set: 2
##                 4     13
## Predicted   7.300  7.300
## cvpred      7.167  7.167
## sleep_time  8.000  7.000
## CV residual 0.833 -0.167
##
## Sum of squares = 0.72    Mean square = 0.36    n = 2
##
## fold 5
## Observations in test set: 2
##                    19    24
## Predicted    6.00e+00 6.744
## cvpred       6.00e+00 6.643
## sleep_time   6.00e+00 7.000
## CV residual -8.88e-16 0.357
##
## Sum of squares = 0.13    Mean square = 0.06    n = 2
##
## fold 6
## Observations in test set: 2
##                 2    20
## Predicted    7.62  6.59
## cvpred       8.94  7.53
## sleep_time   6.00  5.00
## CV residual -2.94 -2.53
##
## Sum of squares = 15.1    Mean square = 7.52    n = 2
##
## fold 7
## Observations in test set: 2
##                11    23
## Predicted    7.3 6.744
## cvpred       7.5 6.643
## sleep_time   6.5 7.000
## CV residual -1.0 0.357
##
## Sum of squares = 1.13    Mean square = 0.56    n = 2
##
## fold 8
```

```
## Observations in test set: 1
##                 14
## Predicted    7.300
## cvpred       7.375
## sleep_time   7.000
## CV residual -0.375
##
## Sum of squares = 0.14    Mean square = 0.14    n = 1
##
## fold 9
## Observations in test set: 1
##                 18
## Predicted   6.590
## cvpred      6.429
## sleep_time  7.000
## CV residual 0.571
##
## Sum of squares = 0.33    Mean square = 0.33    n = 1
##
## fold 10
## Observations in test set: 1
##                 22
## Predicted   6.744
## cvpred      6.643
## sleep_time  7.000
## CV residual 0.357
##
## Sum of squares = 0.13    Mean square = 0.13    n = 1
##
## Overall (Sum over all 1 folds)
##   ms
## 2.28
```



According to the Residuals vs Fitted plot and Normal Q-Q plot, the model 1 fit well when sleep_time in the range of [6, 7.3].

In addition, the cross-validated standard error of estimate (overall cross-validation residual mean of square) is 2.28.

Inference (10pts)

Based on the result so far please perform statistical inference to compare the comparison of interest.

For male:

$$sleep\ time_{male} = 6.15 - 0.154nap + 0.59age2 + 0.905age3$$

For female:

$$sleep\ time_{female} = 6.864 - 0.154nap + 0.59age2 + 0.905age3$$

```
conf_gender=confint(fit1)[1,]
conf_gender
```

```
##  2.5 % 97.5 %
##   3.47  10.26
```

The CI of gender_male (regard gender_female as the baseline) doesn't include 0, so I reject the null hypothesis H_0: The sleep_time of male equals to the sleep_time of female. We can also calculate the difference of mean values of sleep_time between two groups is 0.71. In other words, the male sleep less than female.

Discussion (10pts)

Please clearly state your conclusion and the implication of the result.

In conclusion, male's average sleep time per night is 0.71 hours less than female's.

Limitations and future opportunity. (10pts)

Please list concerns about your analysis. Also, please state how you might go about fixing the problem in your future study.

- The biggest limitation is that the assumption to use t test is the normal distribution of populations, however, I don't know the distribution of sleep_time. And I plan to use non-parametric test to revise it.
- The before_sleep and after_getup both have big correlations with sleep_time, but AIC step method just drop these variables. I feel confused about it.
- I might find that the different age periods has more impact on sleep_time than group by gender. And I plan to research difference of sleep_time between ages independently, or process the cluster analysis considering all the categorical variables.

Comments or questions

If you have any comments or questions, please write them here. - I want to know how to compare the multilevel model with linear model using R

Acknowledge

I learn from these resources:

1.Quick-R by datacamp

2.Weiping Zhang Homepage

3.*Data Analysis and Graphics with R*, Second Edition by Robert I. Kabacoff, published by Manning Publications. 178 South Hill Drive, Westampton, NJ 08060 USA. Copyright © 2015 by Manning Publications.