Midterm Exam

Jingwen Xu

11/2/2020

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.

```
# The introduction of my data
## hair_loss: When I took the survey, I set a 6 points scale to assess the severity of hair loss of res
## gender: "1" indicates female, "0" indicates male.
## age: "1" indicates 18-25 years old, "2" indicates 26-40 years old, "3" indicates more than 40 years
## insomnia: "1" indicates that respondents have insomnia, "0" indicates no insomnia.
## sleep_t: The average daily sleeping time. "1" indicates less than 4 hours, "2" indicates 4-6 hours,
## computer_t: The average daily computer facing time. "1" indicates less than 1 hour, "2" indicates 1-
## sport_t: The average daily sport time. "1" indicates less than 1 hour, "2" indicates 1-2 hours, "3"
## genetic: "1" indicates that the respondent has family hereditary hair loss, "0" indicates none.
## pregnant: "1" indicates that the respondent is in the post-pregnancy stage, "0" indicates none.
## menopause: "1" indicates that the respondent is menopause, "0" indicates none.
## chemical: "1" indicates that the respondent has used poor quality hair dye or perm, "0" indicates no
## disease: "1" indicates that the respondent has hair follicle disease, "0" indicates none.


# The comparison of interest
## Given the information of different respondents such as their gender, age, having insomnia or not and

# Load and read my data into R
```

```
hairloss <- read.csv("https://raw.githubusercontent.com/VivianXu66/Midterm_Exam/main/collection.csv",hea
head(hairloss)
```

```
##   X hair_loss gender age insomnia sleep_t computer_t sport_t genetic pregnant
## 1 1         1      1   1        1       3          2       2       0        0
## 2 2         0      0   1        0       3          3       1       0        0
## 3 3         0      1   1        0       3          1       4       0        0
## 4 4         0      0   1        1       3          5       1       0        0
## 5 5         0      0   1        0       4          3       1       0        0
## 6 6         0      0   3        0       3          4       1       0        0
##   menopause chemical disease
## 1         0        1       0
## 2         0        0       0
## 3         0        0       0
## 4         0        0       0
## 5         0        0       0
## 6         0        0       0
```

EDA (10pts)

Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

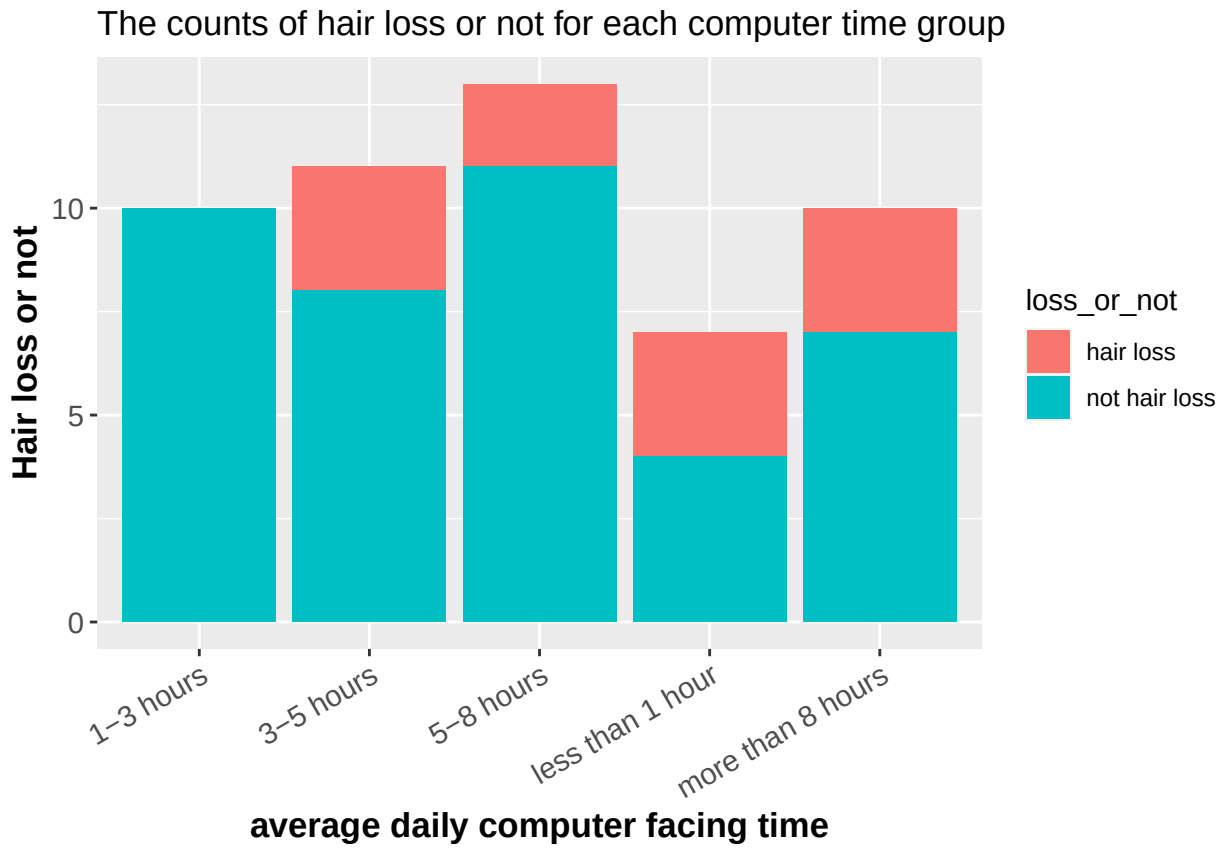
```
summary(hairloss)
```

```
##        X           hair_loss          gender            age
##  Min.   : 1.00   Min.   :0.0000   Min.   :0.0000   Min.   :1.00
##  1st Qu.:14.75   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.00
##  Median :28.50   Median :0.0000   Median :1.0000   Median :1.00
##  Mean   :28.50   Mean   :0.2679   Mean   :0.6071   Mean   :1.75
##  3rd Qu.:42.25   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:3.00
##  Max.   :56.00   Max.   :1.0000   Max.   :1.0000   Max.   :3.00
##     insomnia         sleep_t       computer_t       sport_t
##  Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :0.0000   Median :3.000   Median :3.000   Median :1.000
##  Mean   :0.2679   Mean   :2.946   Mean   :3.071   Mean   :1.518
##  3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:2.000
##  Max.   :1.0000   Max.   :4.000   Max.   :5.000   Max.   :4.000
##     genetic          pregnant   menopause          chemical
##  Min.   :0.00000   Min.   :0   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0   Median :0.00000   Median :0.00000
##  Mean   :0.03571   Mean   :0   Mean   :0.01786   Mean   :0.01786
##  3rd Qu.:0.00000   3rd Qu.:0   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :0   Max.   :1.00000   Max.   :1.00000
##     disease
##  Min.   :0.00000
##  1st Qu.:0.00000
##  Median :0.00000
##  Mean   :0.01786
##  3rd Qu.:0.00000
##  Max.   :1.00000
```

```r
# Simply organize the data
## According to common sense, the predictor "genetic", "pregnant", "menopause", "chemical" and "disease
hairloss <- subset(hairloss,hairloss$genetic=="0"&hairloss$pregnant=="0"&hairloss$menopause=="0"&hairlos
hairloss %<>% select(-c(1,9:13))
rownames(hairloss) <- c(1:51)


# There are many predictors in this data. I want to operate EDA on the predictor "computer_t" and the o
## Firstly, get the counts of hair loss or not in different computer_t groups.
bar_data <- hairloss %>% group_by(hair_loss,computer_t) %>% summarise(Count=n())
```

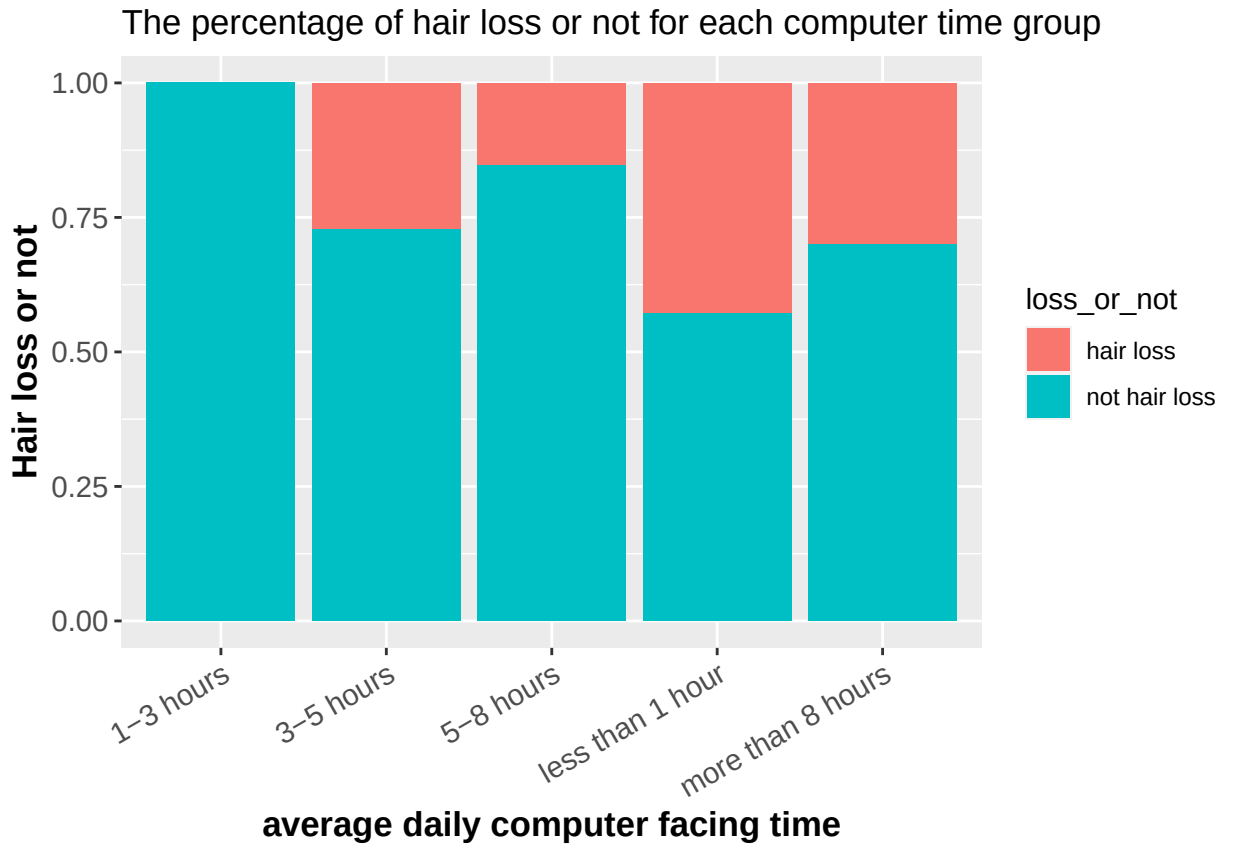
```
## `summarise()` regrouping output by 'hair_loss' (override with `.groups` argument)
```

```r
## Plot the stacked bar
computer_time <- c(rep("less than 1 hour",2),rep("1-3 hours",2),rep("3-5 hours",2),rep("5-8 hours",2),re
loss_or_not <- rep(c("hair loss","not hair loss"),5)
value <- c(3,4,0,10,3,8,2,11,3,7)
data <- data.frame(computer_time,loss_or_not,value)
ggplot(data,aes(fill=loss_or_not,y=value,x=computer_time)) +
    geom_bar(position="stack",stat="identity")+
  theme(axis.text.x=element_text(angle=30,hjust=1),
        axis.text=element_text(size=11),
        axis.title=element_text(size=13,face="bold"))+ggtitle("The counts of hair loss or not for each c
```




```r
## Plot the percentage stacked bar
ggplot(data,aes(fill=loss_or_not,y=value,x=computer_time)) +
    geom_bar(position="fill",stat="identity")+
```

```
  theme(axis.text.x=element_text(angle=30,hjust=1),
        axis.text=element_text(size=11),
        axis.title=element_text(size=13,face="bold"))+ggtitle("The percentage of hair loss or not for ea
```



Power Analysis (10pts)

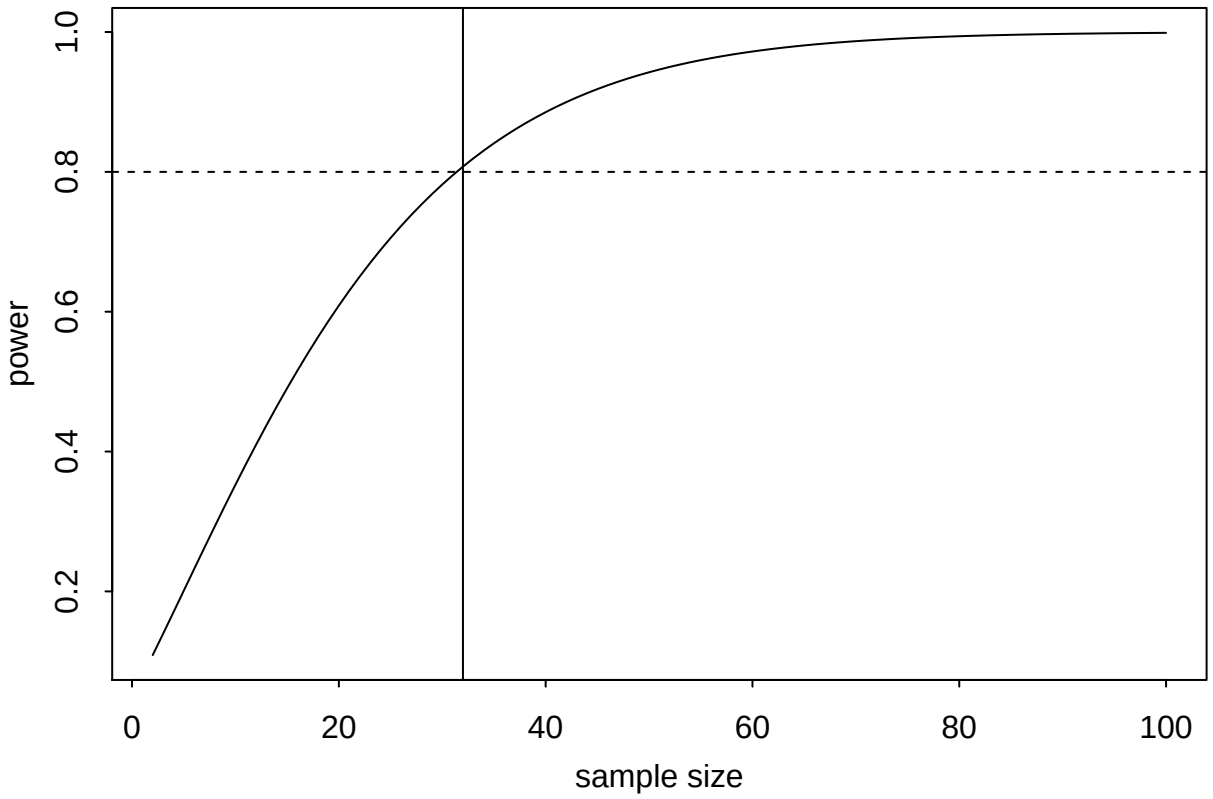
Please perform power analysis on the project. Use 80% power, the sample size you used and infer the level of effect size you will be able to detect. Discuss whether your sample size was enough for the problem at hand. Please note that method of power analysis should match the analysis. Also, please clearly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

```
# Use the function in pwr package to perform power analysis
pwr.p.test(h=NULL,n=51,sig.level=0.05,power=0.8)
```

```
##
##      proportion power calculation for binomial distribution (arcsine transformation)
##
##               h = 0.3923029
##               n = 51
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

```
## The effect size h = 0.3923029


# Cohen suggests that h values of 0.2, 0.5, and 0.8 represent small, medium, and large effect sizes res
```

```r
par(mar=c(3,3,2,1), mgp=c(2,.7,0), tck=-.01)
nlist<-2:100
powres<-sapply(nlist,function(x)pwr.p.test(n=x,h=0.5,sig.level=0.05,NULL)$power)
plot(nlist,powres,type="l",xlab="sample size",ylab="power");abline(h=0.8,lty=2); abline(v=nlist[which(po
```




```r
## From the plot, we can see that if effect size is medium and power is above 0.8, we need to ensure th


# Effect size is usually hypothesized due to the unknown true value.
```

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.

```r
fit <- stan_glmer(hair_loss~gender+insomnia+(1|age)+(1|sleep_t)+(1|computer_t)+(1|sport_t),data=hairloss
fit
```

```
## stan_glmer
##  family:       binomial [logit]
##  formula:      hair_loss ~ gender + insomnia + (1 | age) + (1 | sleep_t) + (1 |
##     computer_t) + (1 | sport_t)
##  observations: 51
## ------
##             Median MAD_SD
## (Intercept) -4.1    1.6
```

```
## gender       3.1    1.3
## insomnia     1.7    1.0
##
## Error terms:
##  Groups     Name        Std.Dev.
##  computer_t (Intercept) 1.46
##  sport_t    (Intercept) 0.83
##  sleep_t    (Intercept) 0.99
##  age        (Intercept) 1.09
## Num. levels: computer_t 5, sport_t 4, sleep_t 4, age 3
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
ranef(fit)
```

```
## $computer_t
##   (Intercept)
## 1   0.4580832
## 2  -1.1559754
## 3   0.2804530
## 4  -0.1730039
## 5   0.1093094
##
## $sport_t
##    (Intercept)
## 1  0.015331466
## 2 -0.043461565
## 3 -0.019621323
## 4  0.008789071
##
## $sleep_t
##   (Intercept)
## 1 -0.02237157
## 2  0.08290727
## 3  0.01013896
## 4 -0.08407684
##
## $age
##   (Intercept)
## 1 -0.21598645
## 2  0.20485323
## 3 -0.06763606
##
## with conditional variances for "computer_t" "sport_t" "sleep_t" "age"
```

```
## This is a multilevel logistic model. Firstly, the outcome in the data is binary so we need to fit a
```

Validation (10pts)

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

```
# Leave-one-out cross validation
logis_loo <- loo(fit)
```

```
## Warning: Found 1 observation(s) with a pareto_k > 0.7. We recommend calling 'loo' again with argumen
```

```
logis_loo
```

```
##
## Computed from 4000 by 51 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo    -26.4  5.2
## p_loo         8.4  2.3
## looic        52.8 10.4
## ------
## Monte Carlo SE of elpd_loo is NA.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     44    86.3%   1307
##  (0.5, 0.7]   (ok)        6    11.8%   1344
##    (0.7, 1]   (bad)       1     2.0%   125
##    (1, Inf)   (very bad)  0     0.0%   <NA>
## See help('pareto-k-diagnostic') for details.
```

```
logis_loo_1 <- loo(fit,k_threshold = 0.7)
```

```
## 1 problematic observation(s) found.
## Model will be refit 1 times.
```

```
##
## Fitting model 1 out of 1 (leaving out observation 22)
```

```
logis_loo_1
```

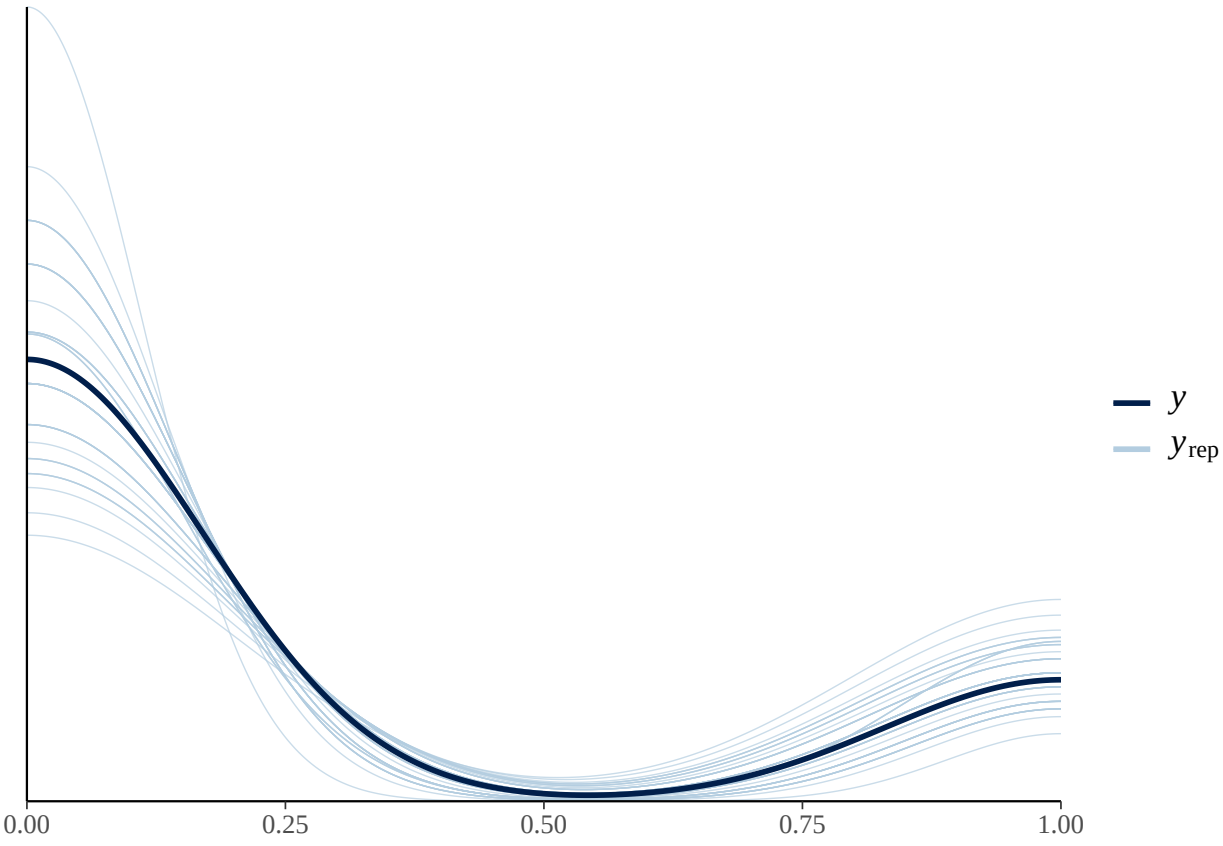
```
##
## Computed from 4000 by 51 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo    -26.5  5.2
## p_loo         8.5  2.4
## looic        52.9 10.5
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## Pareto k diagnostic values:
##                          Count Pct.    Min. n_eff
## (-Inf, 0.5]   (good)     44    88.0%   1307
##  (0.5, 0.7]   (ok)        6    12.0%   1344
##    (0.7, 1]   (bad)       0     0.0%   <NA>
##    (1, Inf)   (very bad)  0     0.0%   <NA>
##
## All Pareto k estimates are ok (k < 0.7).
## See help('pareto-k-diagnostic') for details.
```

```
## There is one unstable observation for the first loo. But after refitting the model 1 times, we can s

# To better excludes the unstable observation in LOO, I use the k-fold cross validation
kfold(fit,K=10)
```

```
## Fitting model 1 out of 10
```

```
## Fitting model 2 out of 10
## Fitting model 3 out of 10
## Fitting model 4 out of 10
## Fitting model 5 out of 10
## Fitting model 6 out of 10
## Fitting model 7 out of 10
## Fitting model 8 out of 10
## Fitting model 9 out of 10
## Fitting model 10 out of 10
##
## Based on 10-fold cross-validation
##
##            Estimate  SE
## elpd_kfold    -25.3 4.9
## p_kfold          NA  NA
## kfoldic        50.6 9.8
```

```
## The results are similar to that of LOO valiation.

# plot the posterior predictive checks
pp_check(fit)
```



```
## According to the plot, the model captures certain characteristics of the data so that they have the
```

Inference (10pts)

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

```
# Summary the inference
summary(fit)
```

```
##
## Model Info:
##  function:     stan_glmer
##  family:       binomial [logit]
##  formula:      hair_loss ~ gender + insomnia + (1 | age) + (1 | sleep_t) + (1 |
##     computer_t) + (1 | sport_t)
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 51
##  groups:       computer_t (5), sport_t (4), sleep_t (4), age (3)
##
## Estimates:
##                                              mean   sd   10%   50%   90%
## (Intercept)                                 -4.2    1.6 -6.2  -4.1  -2.2
## gender                                       3.2    1.3  1.6   3.1   5.0
## insomnia                                     1.7    1.1  0.4   1.7   3.1
## b[(Intercept) computer_t:1]                  0.7    1.0 -0.3   0.5   2.1
## b[(Intercept) computer_t:2]                 -1.6    1.7 -3.8  -1.2   0.0
## b[(Intercept) computer_t:3]                  0.4    0.9 -0.5   0.3   1.6
## b[(Intercept) computer_t:4]                 -0.3    0.9 -1.4  -0.2   0.7
## b[(Intercept) computer_t:5]                  0.2    0.9 -0.8   0.1   1.2
## b[(Intercept) sport_t:1]                     0.1    0.6 -0.6   0.0   0.7
## b[(Intercept) sport_t:2]                    -0.2    0.7 -1.0   0.0   0.4
## b[(Intercept) sport_t:3]                    -0.1    0.8 -0.9   0.0   0.5
## b[(Intercept) sport_t:4]                     0.0    0.7 -0.6   0.0   0.7
## b[(Intercept) sleep_t:1]                    -0.1    0.9 -1.0   0.0   0.6
## b[(Intercept) sleep_t:2]                     0.3    0.8 -0.4   0.1   1.2
## b[(Intercept) sleep_t:3]                     0.1    0.6 -0.6   0.0   0.8
## b[(Intercept) sleep_t:4]                    -0.4    0.9 -1.4  -0.1   0.4
## b[(Intercept) age:1]                        -0.5    0.8 -1.5  -0.2   0.2
## b[(Intercept) age:2]                         0.5    0.9 -0.3   0.2   1.6
## b[(Intercept) age:3]                        -0.2    0.8 -1.2  -0.1   0.5
## Sigma[computer_t:(Intercept),(Intercept)]  2.1    3.3  0.0   1.0   5.3
## Sigma[sport_t:(Intercept),(Intercept)]     0.7    2.0  0.0   0.2   1.7
## Sigma[sleep_t:(Intercept),(Intercept)]     1.0    2.4  0.0   0.2   2.4
## Sigma[age:(Intercept),(Intercept)]         1.2    2.3  0.0   0.4   3.1
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 0.2    0.1  0.1   0.2   0.3
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
```

```
##                                                mcse Rhat n_eff
## (Intercept)                                    0.0  1.0  2604
## gender                                         0.0  1.0  3737
## insomnia                                       0.0  1.0  3569
## b[(Intercept) computer_t:1]                    0.0  1.0  2575
## b[(Intercept) computer_t:2]                    0.0  1.0  1915
## b[(Intercept) computer_t:3]                    0.0  1.0  3635
## b[(Intercept) computer_t:4]                    0.0  1.0  3290
## b[(Intercept) computer_t:5]                    0.0  1.0  3653
## b[(Intercept) sport_t:1]                       0.0  1.0  2656
## b[(Intercept) sport_t:2]                       0.0  1.0  2572
## b[(Intercept) sport_t:3]                       0.0  1.0  3793
## b[(Intercept) sport_t:4]                       0.0  1.0  3448
## b[(Intercept) sleep_t:1]                       0.0  1.0  3411
## b[(Intercept) sleep_t:2]                       0.0  1.0  3158
## b[(Intercept) sleep_t:3]                       0.0  1.0  3093
## b[(Intercept) sleep_t:4]                       0.0  1.0  3172
## b[(Intercept) age:1]                           0.0  1.0  2410
## b[(Intercept) age:2]                           0.0  1.0  3764
## b[(Intercept) age:3]                           0.0  1.0  3185
## Sigma[computer_t:(Intercept),(Intercept)] 0.1  1.0  2186
## Sigma[sport_t:(Intercept),(Intercept)]    0.0  1.0  4038
## Sigma[sleep_t:(Intercept),(Intercept)]    0.0  1.0  3672
## Sigma[age:(Intercept),(Intercept)]        0.0  1.0  2783
## mean_PPD                                       0.0  1.0  3914
## log-posterior                                  0.1  1.0  1110
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
## We can see that except for intercept and the coefficient of gender, other coefficients are not signi

# Maybe a prediction simulation is better for inference, but I don't know how to do such a simulation f
```

Discussion (10pts)

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

1.Conclusion: (1)According to the coefficient, the people who have insomnia will be more likely to be bothered by hair loss. It's reasonable referring to the common sense. (2)According to the coefficient, the probability that women concern about hair loss is larger than men do. (3)From the EDA analysis and model coefficient, we can see that there are relationship between computer_t(sport_t, sleep_t and age) and hair loss. But we need larger sample to determine the precise positive or negative relationship.

2.Implication: In China, hair loss problem has became a popular topic especially among the young people due to their increasing life pressure. Either job or study requires a lot of time facing computer which may results in hair loss. And sedentariness can reduce metabolism. Besides, more and more people suffer from insomnia. I want to clarify the relationship between these phenomenons and hair loss to alert everyone and promote a healthy lifestyle.

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.

1.Concerns: (1)The data is not large enough with only 56 observations so that it's not persuasive with certain missing value in certain groups. For example, There is only one respondent from group 2 of computer_t

having concern about hair loss. But her hair loss is mostly related to the poor quality hair dye or perm. After removing such a disturb, there is no respondent from group 2 of computer_t having concern about hair loss. It's not reasonable and persuasive.
(2)The survey is mainly issued in my Wechat where most of the people are 18-25 or more than 40 years old so the sample is not random enough. (3)The judge of hair loss or not is based on the subjective consciousness of the respondents which may lead to measurement error. (4)The LOO cross validation may be not proper for multilevel model. There is a validation method called leave-one-group-out which will be better for multilevel model.

2.Possible fixing solutions:(1)Take the survey in more general platform which can ensure the randomness. And find more respondents to do the survey. (2)Clarify the evaluation about hair loss or not with more objective and scientific criterion. For example, I can ask the professional doctor for help to set the criterion. (3)There may be interaction among computer_t, sleep_t and sport_t, I should study if the interaction exists or not for fitting model more precisely. (4)I think I should simplify the model in further study.

Comments or questions

If you have any comments or questions, please write them here.

Question: (1)Is there a good way to deal with the overlapping of factor baseline and intercept if the predictors have more than one index variables? (2)How to properly choose the function for power analysis?