**Homework#1, BIOSTAT 620**

*Due: Tuesday, February 6, 2024 submitted in class*

*You may discuss in group but hand in an independent copy of your own solutions. ChatGPT or other LLMs are not allowed to solve the problems of this homework.*

**Link from GitHub**: <https://github.com/RuoerBei/BIOSTAT620>

**PATT I: DATA COLLECTION AND DATA PROCESSING**

**Problem 1**: Explore the your own screen activity data that you collect until the end of Friday (inclusive), January 26, 2024. This type of ‘break’ time set up by scientists in practice is often referred to as *data freeze date* during data collection. If you were unable to collect such data due to the previous setting of your mobile device or other logistic reasons, please let the instructor or GSI know immediately, some backup data would be provided to you.

1. Describe the purpose of the data collection, in which you state a scientific hypothesis of interest to justify your effort of data collection. Cite at least one reference to support your proposed hypothesis to be investigated. This hypothesis may be the one of a few possible hypotheses that you like to investigate in your first group project with your teammates.

The purpose of collecting screen activity data is to analyze patterns and impacts of screen usage on various aspects of life, such as productivity, mental health, or social behavior. Specifically, the purpose is to find the association between screen time and depression.

Hypothesis: higher screen time is associated with a higher level of depression among UMich Biostats620 students.

In my data, I would use both total screen time and social screen time to find their association with depression level. I would also distribute a questionnaire to BIOSTAT620 students to find their depression scores.

In 2017, Dr. K.C. Madhav, Dr. Shardulendra Prasad Sherchand, and Dr. Samendra Sherchan wrote a paper with the title “Association between screen time and depression among US adults.” They used the secondary data from NHANES. The participants were 3201 20-year-old or more US adults. The data came from self-reported questionaries. Although the self-reported data may bring bias, they gave the conclusion that moderate or severe depression level was associated with higher screen time.

Citation: K.C. Madhav, Shardulendra Prasad Sherchand, Samendra Sherchan. “Association between screen time and depression among US adults.” 13 March 2017.

[**https://www.sciencedirect.com/science/article/pii/S2211335517301316**](https://www.sciencedirect.com/science/article/pii/S2211335517301316)

1. Explain the role of Informed Consent Form in connection to the planned study and data collection.

The Informed Consent Form is important in research because it ensures that participants are fully aware of the study, including the purpose of the study, the possible side effects and benefits of this study, how to protect their rights and privacy, and etc. It is also important to note the voluntary nature of participation. Therefore, Informed Consent Form is essential for ethical research practice, respecting participants' autonomy and rights.

1. Describe the data collection plan, including when the data is collected, which types of variables in the data are collected, where the data is collected from, and how many data are collected before the data freeze. You may use tables to summarize your answers if necessary.

The data was collected from Ruoer Bei’s phone of personal screentime use.

The data was from 1/3/24 to 1/26/24 (the freeze date), included 24 days. Variables included Date, Total Screen Time (in hours), Social Screen Time (in hours,) Number of Pickups, and the First Pickup Time (as well as Total Screen Time (in mins), Social Screen Time (in mins) to increase the credibility of the study.)

Date: discrete quantitative, integer, date

Total Screen Time (in hours): text

Total Screen Time (in mins): continuous quantitative, numeric

Social Screen Time (in hours): text

Social Screen Time (in mins): continuous quantitative, numeric

Number of Pickups: count, numeric

the First Pickup Time: date

The data is sourced from built-in screen time tracking features in mobile devices or third-party applications designed for this purpose.

1. Create and add two new variables into your dataset; they are, “daily proportion of social screentime” (defined as the ratio of daily total social screen time over daily total screen time) and “daily duration per use” (defined as the ratio of daily total screen time over daily total of pickups).

Added in my original excel file.

A screenshot of a spreadsheet

Description automatically generated

**Problem 2**: Data visualization is one of the early steps taken to see the data at hand. Consider the variables measured in the screen activity data, including daily total screen time, daily total social screen time, and daily number of pickups as well as two new variables derived from the raw data, daily proportion of social screen time and daily duration per use.

1. Make a time series plot of each of the five variables in your data. Describe temporal patterns from these time series plots.

Below are the five plots I created. Black dots mean weekdays (Mondays to Fridays,) and red dots mean weekend days (Saturdays and Sundays.) Moreover, the winter break ends at Jan 9th. Jan 10th is the first school day.

**Time Series Plot of Total Screen Time (min)**

A graph with lines and dots

Description automatically generated

We can find a kinda strange trend in this plot: the Total Screen Time increased when I switched my lifestyle from winter break to school time. The rest of the plots are pretty random, showing no significant increase of longer Total Screen Time on weekends. We can also find that longer screen time use happened at Jan 11th, Jan 14th, and 19th, which correspond to Thursday (the second school day,) Sunday and Friday, which may show that I played harder after I have few busy days.

**Time Series Plot of Social Screen Time (min)**

A graph with lines and dots

Description automatically generated

Although there is a peak at 11th, after school days, I have longer Total Screen Time on weekends. Also, I have longer Social Screen Time during the first few days of starting this semester.

**Time Series Plot of Pickups**

A graph with lines and dots

Description automatically generated

The trend is pretty zigzag. However, we can roughly find out that after school days, I have relatively fewer pickups on weekends.

**Time Series Plot of Proportion of Social Screen Time (min)**

A graph with lines and dots

Description automatically generated

Again, the trend is pretty zigzag. The proportion of social screen time did not have strong association with weekends. Jan 11th and Jan 20th had bigger proportion of social screen time use. It seems that I played harder (on social medias) after I have few busy days. Overall, there is no apparent trend.

**Time Series Plot of Daily Duration per Use (min)**

A graph with lines and dots

Description automatically generated

Daily Duration per Use first increase then decrease, showing the trend I that have more Daily Duration per Use during the first few days of starting this semester and the following weekend. Starting from the second week of the new semester, the duration per use became more stable and decreased a lot. There are two peaks at Jan 9th and Jan 14th.

1. Make pairwise scatterplots of five variables. Describe correlation patterns from these pairwise scatterplots. Which pair of variables among the five variables has the highest correlation?

A screenshot of a graph

Description automatically generated

The pair of Social Screen Time and Total Screen Time has the highest positive correlation (0.781,) showing that social screen time is positively correlated with total screen time. In other word, I may spend a lot of time on social apps when I use my phone.

However, there are multiple pairs that have smaller but similar correlation. The pair of Duration per Use and Total Screen Time has a correlation at 0.708, showing that duration per use is positively correlated with total screen time. In other word, when I spent more time on my phones per time, cumulatively, I may have longer total screen time.

The pair of the Proportion of Social Screen Time and Social Screen Time has a correlation at 0.721, which is understandable since social screen time is the numerator of the proportion. It can further prove that I may spend a lot of time on social apps when I use my phone.

The pair of the Duration per Use and Social Screen Time has a correlation at 0.727, showing that I may spend more per time on social apps.

Moreover, the pair of Duration per Use and Pickups has the highest negative correlation (-0.819,) showing that duration per use and pickups are negatively correlated. In other words, a higher number of pickups may correspond to a shorter per time of using my phone.

1. Make an occupation time curve for each of the five time series. Explain the pattern of individual curves.

**Total Screen Time**

A graph showing a line

Description automatically generated

A graph showing a curve

Description automatically generated

From the occupation time plot for total screen time, the curve decreased slower at the edge and faster at the middle of the curve, showing that there was fewer total screen time that were shorter than 250 mins and longer than 400 mins. The similar conclusion could also be shown in the density plot above.

**Social Screen Time**

A graph showing a line

Description automatically generated with medium confidence

A graph showing a blue line

Description automatically generated

From the occupation time plot for social screen time, the curve decreased slower at the edge and faster at the middle of the curve, showing that there was fewer social screen time that were shorter than 100 mins and longer than 250 mins. The similar conclusion could also be shown in the density plot above.

**Pickups**

A graph showing the number of pickups

Description automatically generated

A graph showing a curve

Description automatically generated

From the occupation time plot for pickups, the curve decreased slower at the edge and faster at the middle of the curve, showing that there was fewer pickups that were smaller than 60 picks and bigger than 120 picks. The similar conclusion could also be shown in the density plot above.

**Proportion of Total Screen Time and Social Screen Time**

A graph showing a line

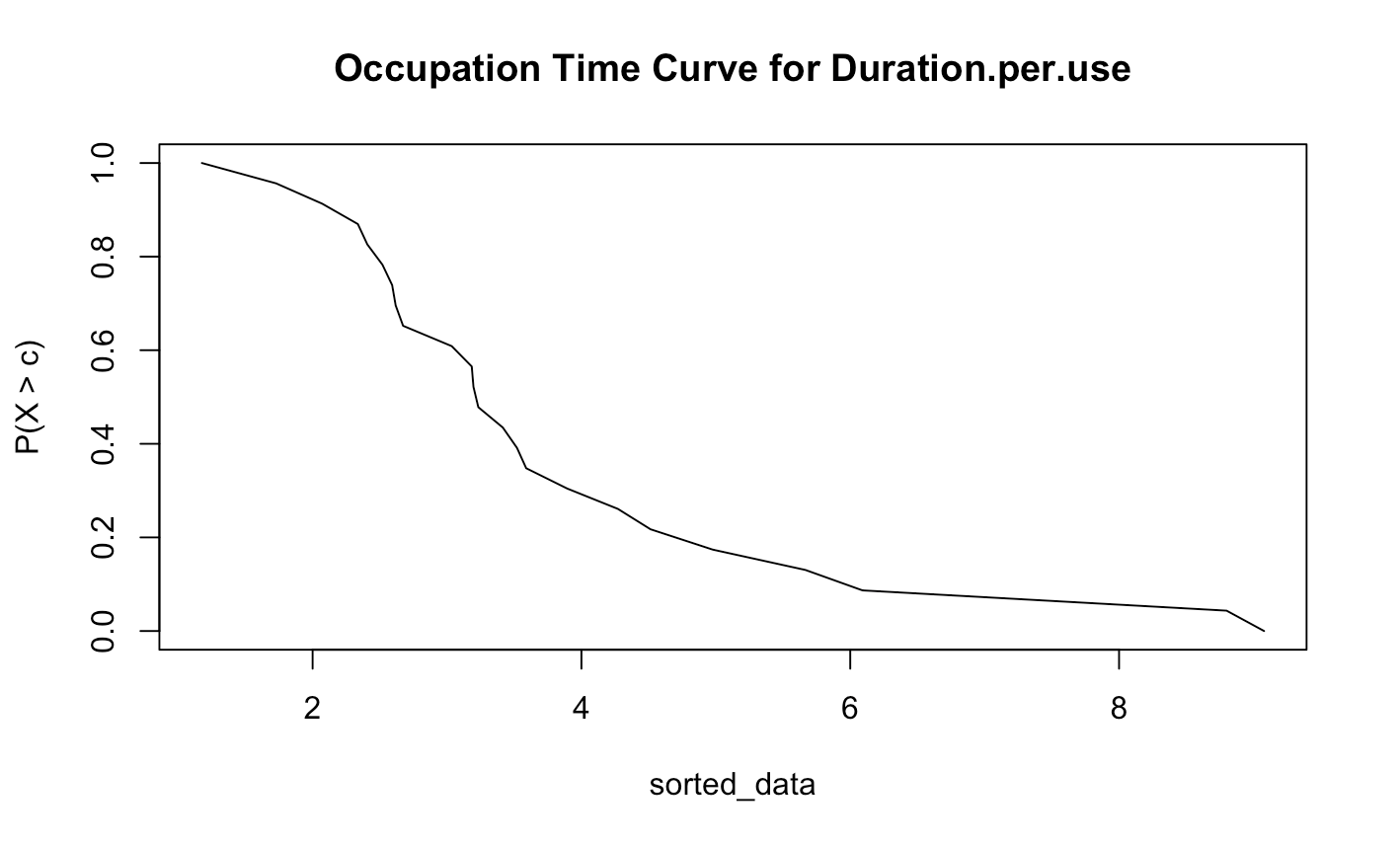
Description automatically generated with medium confidence

A graph of a graph

Description automatically generated

From the occupation time plot for Proportion of Total Screen Time and Social Screen Time, the curve decreased slower at the edge and faster at the middle of the curve, showing that the proportion that was smaller than 0.4 and bigger than 0.65 were fewer. The similar conclusion could also be shown in the density plot above.

**Daily Duration per Use**



A graph of a normal distribution

Description automatically generated

From the occupation time plot for duration per use, the curve decreased slower at the edge and faster at the middle of the curve, showing that the duration per use that was shorter than 1 and bigger than 6 were fewer. However, duration that was longer than 7.5 happened more than at 7.5. The similar conclusion could also be shown in the density plot above.

1. Use the R function acf to display the serial dependence for each of the five time series. Are there any significant autocorrelations? Explain your results. Note that in this R function, you may set plot=FALSE to yield values of the autocorrelations.

**Total Screen Time**

**A graph of a line

Description automatically generated with medium confidence**

**Social Screen Time**

**A graph with numbers and lines

Description automatically generated**

**Pickups**

**A line graph with numbers and a line

Description automatically generated**

**Proportion of Total Screen Time and Social Screen Time**

**A graph with lines and numbers

Description automatically generated**

**Daily Duration per Use**

A graph with lines and numbers

Description automatically generated

A white sheet with black numbers and letters

Description automatically generated with medium confidence

For these five auto-correlation plots, we can find that except the first one (lag at 0) would exceed the blue lines, the rest of them did not exceed the blue lines. Therefore, the autocorrelations for all these five variables at the lag were not statistically significant at 95% confidence level. Therefore, there was no enough evidence to say the true autocorrelation was a the lag for all the five variables. Also, the values in the past could not predict future values. The values from ACF could further prove the conclusion.

**Problem 3**: Explore the use of the R package circular to display the time of first pickup as a circular variable or angular variable.

1. Transform (or covert) the time of first pickup to an angle ranged from 0 to 360 degree, treating midnight as 0 degree. For example, 6AM is 90 degree and noon is 180 degree.

A screenshot of a table

Description automatically generated

1. Make a scatterplot of the first pickup data on a 24-hour clock circle. Describe basic patterns from this scatterplot in terms of personal habit of first pickup.

A circle with numbers and dots

Description automatically generated

A circular object with numbers and a white background

Description automatically generated

This circle indicates 24 hours (a 24-hour clock.) The first pickup of mine is often happened from 7:30am to 9:00am. The earliest pickup time was roughly at 5am, and latest was roughly at 12am.

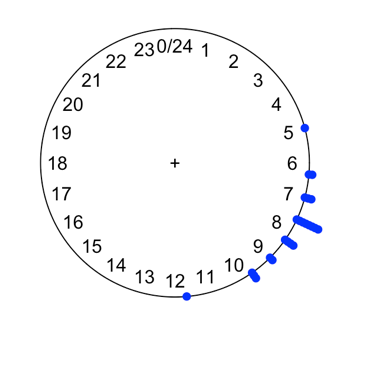
1. Make a histogram plot on the circle in that you may choose a suitable bin size to create stacking. For example, you may set a bin size at 2.5 degree, which corresponds an interval of 10 minutes. Adjust the bin size to create different forms of histogram, and explain the reason that you choose a particular value to report your final histogram plot.

A circular clock with numbers and a plus

Description automatically generated

A circular object with numbers and a plus

Description automatically generated



I set the bin size at 20 (left most one.) The middle one has the bin size at 2. The right most one has bin size at 10. We can find that when bin size is at 20, it could do a better job of showing the density. When the bin size is smaller, the graphs are harder to read (less intuitive.)

Most first Pickups happened around 7:50am. The second most Pickups time happened around 6:35am and 9:20am. They shared roughly the same number of Pickups. 5:10am and 12:05 am seemed also shared roughly the same number of Pickups.

**PART II: DATA ANALYSIS**

**Problem 4**: Consider the data of the daily number of pickups. Let *Yt* be the daily number of pickups at day *t*, and let *St* be the daily total screen time at day *t*. Then, we assume that

*Yt* ∼ Poisson(*Stλ*)*, t* = 1*,...,T,*

where *T* is the number of days for data collection, and *λ* is the expected hourly rate of pickups (or the expected number of pickups per hour). Note that here *St* needs to be recorded in unite of hour.

1. Explain why the factor *St* is needed in the Poisson distribution above.

The Poisson distribution is often used to model the number of times an event occurs within a fixed interval of time or space. In this case, Yt (the daily number of pickups) is assumed to follow a Poisson distribution. Lambda represents the average rate of these events. However, if the amount of time or space varies, the rate needs to be adjusted accordingly. Therefore, St is important since it acts to adjust expected daily total screen time at different days. Moreover, St ensures that the rate lambda is standardized to a per-hour basis regardless of how much total screen time there is each day.

1. Use the R function glm to estimate the rate parameter *λ* in which ln(*St*) is included in the model as an *offset*.

A computer code with numbers and symbols

Description automatically generated

1. Define two dummy variables: *Xt* = 1 for day *t* being a weekday and 0 for day *t* being a weekend day; and *Zt* = 1 for day *t* being January, 10 (the first day of the winter semester) or after, and 0 for day *t* before January, 10 (the winter holiday day).

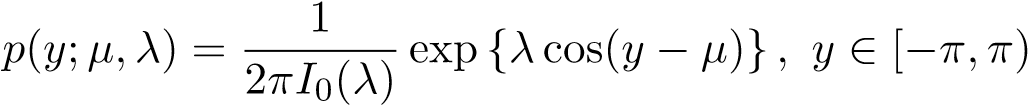
Repeat part (b) for a model ln(*λ*(*t*)) = *β*0+*β*1*Xt*+*β*2*Zt*, under which the rate parameter *λ* differs between weekdays and weekends as well as between the winter semester and the winter holiday. This model is called *log-linear model*. Cleary, this rate parameter depends on day *t*. Use the R function glm to estimate the regression coefficients and answer the following questions. (c.1) Is there data evidence for significantly different behavior of daily pickups between weekdays and weekends? Justify your answer using the significance level *α* = 0*.*05.

(c.2) Is there data evidence for a significant change on the behavior of daily pickups after the winter semester began? Justify your answer using the significance level *α* = 0*.*05.

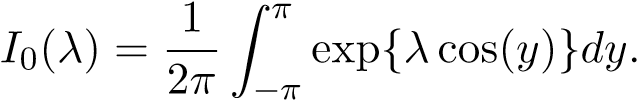
For Xt, p-value for beta1 is 0.414, which is bigger than 0.05, showing that there is no significant difference in the number of pickups between weekdays and weekends.

For Zt, p-value for beta2 is 6.76e-11, which is smaller than 0.05, showing a significant change in the number of pickups after the winter semester began.

**Problem 5**: Now analyze the first pickups data from Problem 3. The von Mises distribution is widely used to model a circular random variable *Y* . The density of this distribution takes the following form:

*,*

where *µ* ∈ [−*π,π*) is the mean direction, and *λ >* 0 is the concentration parameter. In addition, *I*0(*λ*) is the modified Bessel function of the first kind of order 0 given by



1. Use the R function mle.vonmises from the R package circular to obtain the estimates of the two model parameters *µ* and *λ* from your data of first pickups.

A math equations and numbers

Description automatically generated

1. Based on the estimated parameters from part (a), use the R function pvonmises from the R package circular to calculate the probability that your first pickup is 8:30AM or later.

A number with a number on it

Description automatically generated with medium confidence

The probability of first pick up being at 8:30am or later is 0.3318581.