Midterm 1 Review

Math 141, Week 6

For extra practice, several review problems are provided on this Review Sheet. However, they are not comprehensive, so do not limit your studying to just these problems. While the problems are intended to match the difficulty of those on the exam, the length of this review set does not represent the actual length of the exam. Solutions to these problems are posted the Exams page of the course website.

Problem 1

A recent study examines the relationship between sleep habits, alcohol use, academic performance, measures of depression and stress, and other variables in US college students. The data were obtained from a sample of 253 students who did skills tests to measure cognitive function, completed a survey that asked many questions about attitudes and habits, and kept a sleep diary to record time and quality of sleep over a two-week period. Some data is available in the `SleepStudy` data frame loaded with the code below.

```
# Load the data
library(Lock5Data)
data(SleepStudy)
```

Run this code chunk to view the documentation for this data set. **Do not** change `eval = F` to `eval = T` for this code chunk.

```
?SleepStudy
```

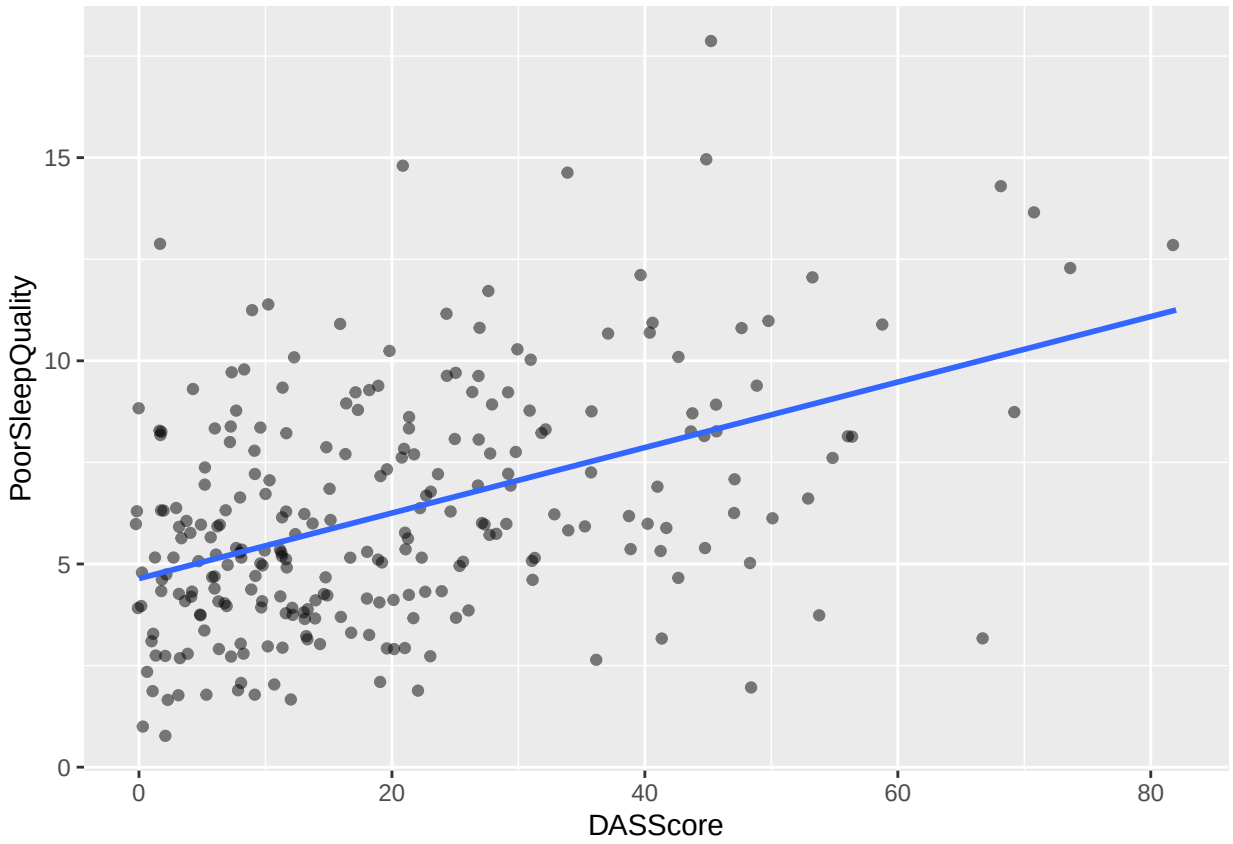
We want to determine if you can model the sleep quality of student using other variables

a. Produce a plot comparing sleep quality (response) and DAS score (explanatory). Include the least squares regression line. Discuss any trend you observe.

Generally, higher DAS scores correspond to higher values of the `PoorSleepQuality` variable. However, there are a few observations with high DASScores (>70) that may be very influential in determining the line of best fit.

```
ggplot(SleepStudy, aes(x = DASScore, y =PoorSleepQuality )) +
  geom_jitter(alpha = .5) +
  geom_smooth(method = "lm", se = F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

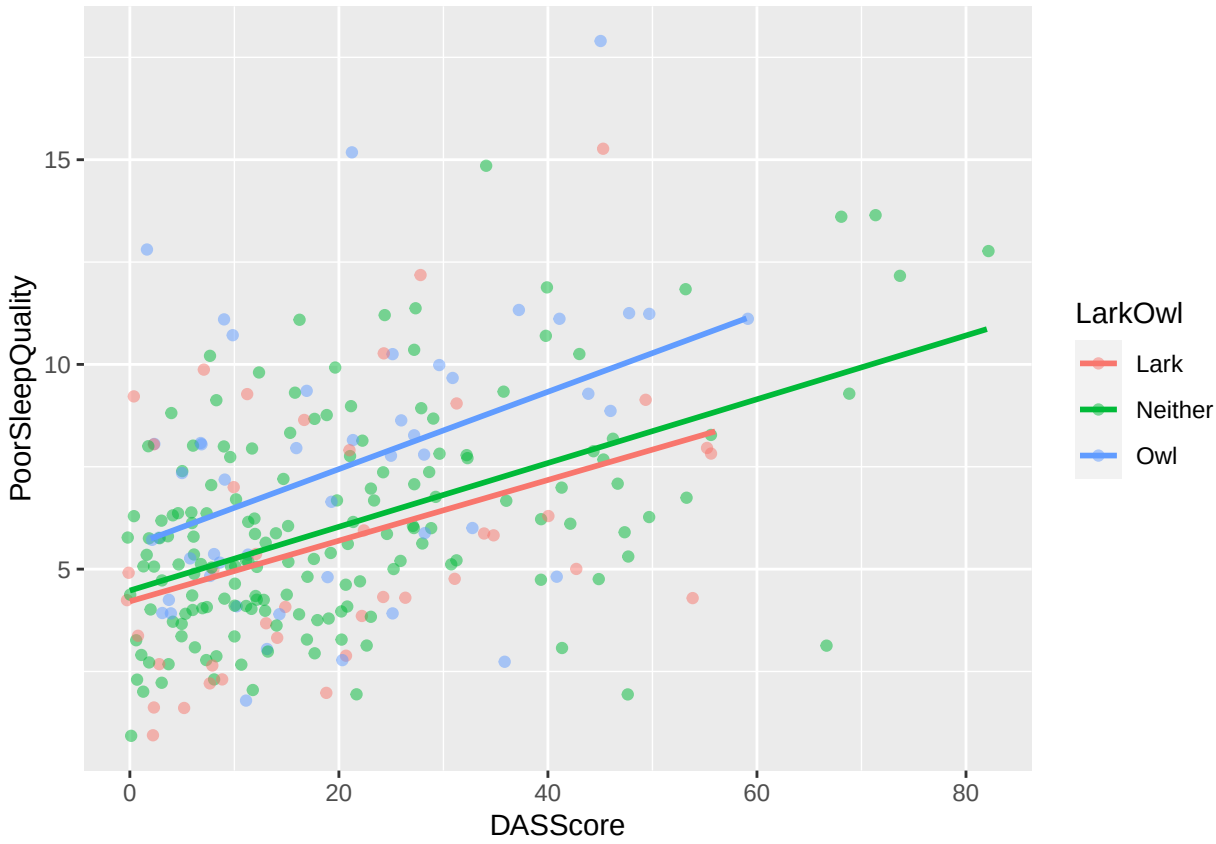


b. Recreate the plot but this time the `LarkOwl` variable, as well. What new trends are evident?

Based on the new scatterplot, Owls tend to report the highest values of `PoorSleepQuality`, compared to Larks and Neither.

```
ggplot(SleepStudy, aes(x = DASScore, y =PoorSleepQuality, color = LarkOwl )) +
  geom_jitter(alpha = .5) +
  geom_smooth(method = "lm", se = F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



---

c. Build the linear regression model for sleep quality using DASScore as the explanatory variable. Write down the formula for the least squares regression line. Interpret the slope and intercept terms in the context of this model.

```
sleep_mod <- lm(PoorSleepQuality ~ DASScore, data = SleepStudy)
get_regression_table(sleep_mod)
```

```
## # A tibble: 2 x 7
##   term      estimate std_error statistic p_value lower_ci upper_ci
##   <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept    4.64      0.257     18.0        0    4.14      5.15
## 2 DASScore     0.081     0.01       8.13       0    0.061     0.1
```

The equation for the least squares regression line is

$$\text{PoorSleepQuality} = 4.642 + 0.081 \cdot \text{DASScore}$$

The slope on this model indicates that for every 1 point increase in DASScore, there is an associated increase of 0.081 points in predicted PoorSleepQuality.

The intercept indicates that model predicts a PoorSleepQUality of 4.642 for a individual who reports a DASScore of 0.

---

d. For a student with DASScore of 50, predict the sleep quality.

Based on the linear model, the predicted sleep quality for a student with a DASScore of 50 is

```
4.642 + .081*50
```

```
## [1] 8.692
```

---

e. Now, build a linear regression model predicting sleep quality as a function of LarkOwl level. Which level is used as a baseline? What is the mean sleep quality for this level?

```
sleep_larkowl_mod <- lm(PoorSleepQuality ~ LarkOwl, data = SleepStudy)
get_regression_table(sleep_larkowl_mod)
```

```
## # A tibble: 3 x 7
##   term             estimate std_error statistic p_value lower_ci upper_ci
##   <chr>               <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept           5.71      0.447    12.8     0        4.83      6.59
## 2 LarkOwl: Neither    0.311     0.5       0.622   0.535   -0.674     1.30
## 3 LarkOwl: Owl        1.80      0.606     2.98    0.003    0.609     3.00
```

Based on the regression table, "Lark" is used as the baseline, since it does not appear as a row in the table. Therefore, the predicted mean sleep quality for Larks is the intercept of the model: 5.707

---

f. Based on the coefficients of your linear model in the previous part, which LarkOwl level has best sleep quality (recall, higher values of `PoorSleepQuality` indicate poorer sleep)? Which has the worst?

The level with the largest value of `PoorSleepQuality` is the Owl, indicating that owls have the worst sleep quality. The level with the lowest value of `PoorSleepQuality` is the Lark, indicating that Larks have the best sleep quality.

---

g. What is the predicted sleep quality for Night Owl? What is the residual for a Night Owl who reports a sleep quality of 6?

The model predicts that a Night Owl will have a sleep quality of $5.707 + 1.803 = 7.51$. If a particular night owl has a sleep quality of 6, then the residual for this individual is $6 - 7.51 = -1.51$.

Problem 2

It is well-known that lack of sleep impairs concentration and alertness, and this might be due partly to late night food consumption. A 2015 study took 44 volunteers aged 21 to 50 and gave them unlimited access to food and drink during the day, but allowed them only 4 hours of sleep per night for three consecutive nights On the fourth night, all participants again had to stay up until 4am, but this time participants were randomized into two groups; one group was only given access to water from 10pm until their bedtime at 4am while the other group still had unlimited access to food and drink for all hours. The group forced to fast from 10pm on performed significantly better on tests of reaction time and had fewer attention lapses than the group with access to late night food.

a. Is this an experiment or an observational study? Explain how you know.

This is an experiment. The researchers randomized treatments (access to only water / access to unlimited food and drink) among participants.

b. What is the population the researchers wish to study? What is the sample?

The implied population for this investigation is all adults (perhaps all adults country / region where the data was collected) The sample consists of 44 volunteers aged 21 to 50.

c. List the explanatory and response variables in this investigation, and for each, classify whether it is quantitative or categorical.

Explanatory variables:

- Access to unlimited food / water on 4th day (binary categorical with levels "Yes" and "No")

Response variables:

- Reaction time (quantitative)
- Number of attention lapses (quantitative)

d. What sampling technique was used to construct this sample? Based on the sampling technique, can conclusions from the investigation be extended to the population?

Since the investigation obtained a sample of 44 **volunteers**, this sample was constructed using *convenience sampling*. This does not represent a random sample from the population, and therefore, we cannot generalize conclusions from the study to the population.

e. Are there likely to be confounding variables in this investigation? Why or why not?

Because the treatments were randomized among subjects, there are unlikely to be confounding variables that influence the result of the experiment.

f. Based on the situation described, is it reasonable to conclude that eating late at night causes some of the typical effects of sleep deprivation (reaction time and attention lapses)? Why or why not?

Based on the investigation, because a randomized experiment method was used, we can deduce that late night eating causes some of the typical effects of sleep deprivation **among the sample of volunteers**. However, because a random sampling method was not used, we cannot generalize this conclusion to the population of all adults.

Problem 3 Use the following graphic to answer this question. https://reed-statistics.github.io/math141s21/img/m1geoms.png

a. What are the variables displayed in this graphic? For each variable, specify if it is categorical or numerical.

The variables displayed in the graphic are:

- Date (quantitative)
- Average Departure Delay (quantitative)
- Origin Airport (categorical)

---

b. What **geoms** are the variables mapped to?

The only geometric object in this plot is the *line* (or curve or spaghetti string)

---

c. What are the **aesthetic**s of each **geom**? For each **aesthetic**, give the variable that sets the values of that **aesthetic**.

Date determines the x-coordinate of the line, while Departure Delay determines the y coordinate. Origin airport determines the type of line (dashed / dotted / regular)

Problem 4

The `mtcars` data set contains information about 32 different car models from the 1970s. For each model, the data set includes records: the name of the model `name`, the fuel efficiency `mpg`, the number of engine cylinders `cyl`, the weight `wt`, the horsepower `hp`, and the transmission type `transmission` (either "manual" or "automatic").

Load the data using the following code chunk:

```
data(mtcars)
mtcars <- mtcars %>% mutate(transmission = ifelse(am == 1, "manual", "automatic")) %>% select(-am) %>% r
```

a. Use `dplyr` verbs to create a data frame consisting of: three columns recording the name of each model, the number of cylinders, and the transmission type, for exactly those cars which have a fuel efficiency between 10 and 25 miles per gallon, ordered alphabetically by model name.

```
mtcars %>%
  filter(mpg > 10 , mpg < 25) %>%
  select(name, cyl, transmission) %>%
  arrange(name)
```

```
##                   name cyl transmission
## 1          AMC Javelin   8    automatic
## 2   Cadillac Fleetwood   8    automatic
## 3           Camaro Z28   8    automatic
## 4    Chrysler Imperial   8    automatic
```

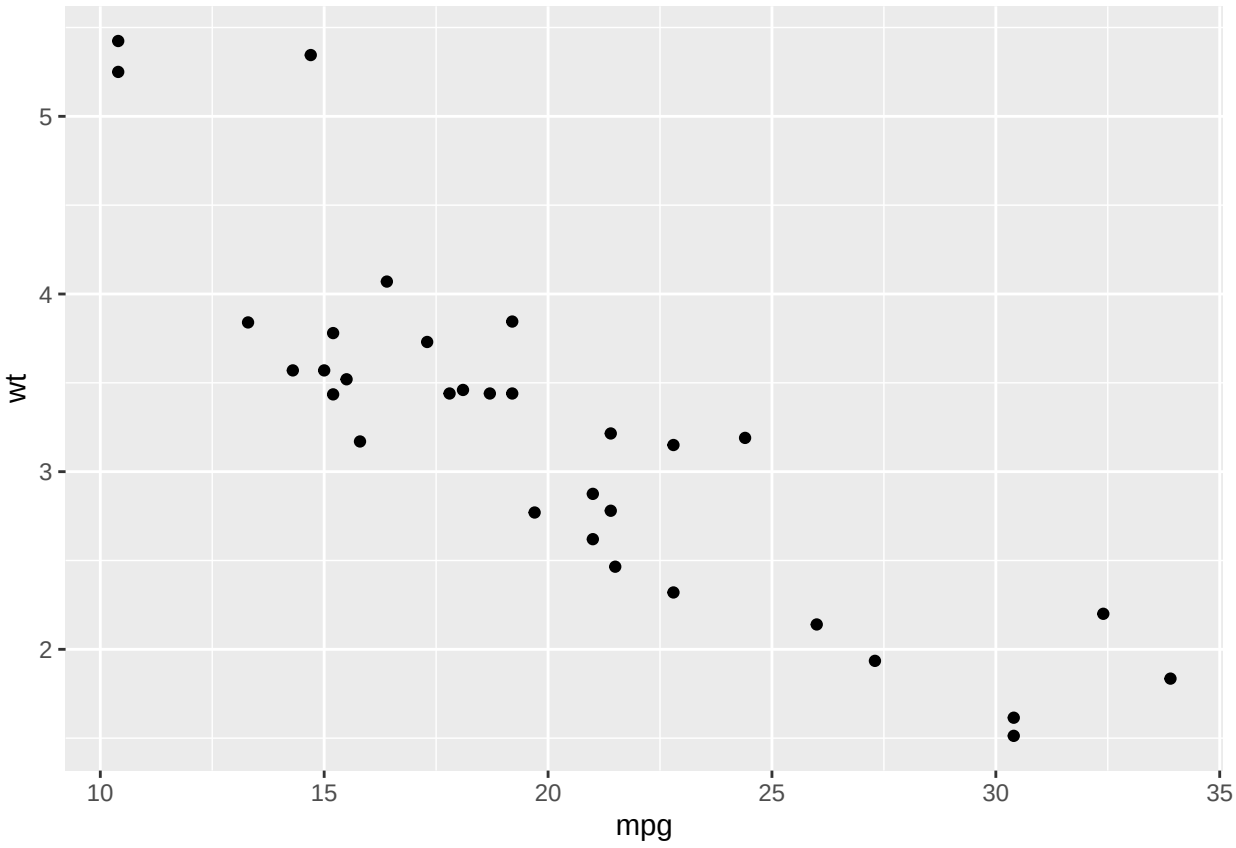
```
## 5           Datsun 710   4       manual
## 6       Dodge Challenger   8    automatic
## 7            Duster 360   8    automatic
## 8          Ferrari Dino   6       manual
## 9        Ford Pantera L   8       manual
## 10       Hornet 4 Drive   6    automatic
## 11    Hornet Sportabout   8    automatic
## 12 Lincoln Continental   8    automatic
## 13       Maserati Bora   8       manual
## 14           Mazda RX4   6       manual
## 15       Mazda RX4 Wag   6       manual
## 16            Merc 230   4    automatic
## 17           Merc 240D   4    automatic
## 18            Merc 280   6    automatic
## 19           Merc 280C   6    automatic
## 20          Merc 450SE   8    automatic
## 21          Merc 450SL   8    automatic
## 22         Merc 450SLC   8    automatic
## 23    Pontiac Firebird   8    automatic
## 24       Toyota Corona   4    automatic
## 25             Valiant   6    automatic
## 26          Volvo 142E   4       manual
```

---

b. Create a scatterplot comparing `mpg` and `wt`. Based on the graphic, do the variables appear to have a weak, moderate, or strong correlation? Moreover, is it positive or negative?

The variables mpg and wt appear to have a strong negative correlation.

```
ggplot(mtcars, aes(x = mpg, y = wt)) + geom_point()
```



---

c. Give a line-by-line explanation for what the following code does:

```
mtcars %>%
  filter(carb != 8) %>%
  mutate(hp_per_cyl = hp/cyl) %>%
  group_by(transmission) %>%
  summarize(avg = mean(hp_per_cyl ))
```

This code takes the mtcars data frame, then filters for cars that do not have 8 carborators, then creates a column called hp_per_cyl equal to horsepower divided by cylinder, then groups cars by transmission type, and then computes the average horsepower per cylinder, within each group.

---

d. Create a data frame consisting of the proportion of cars that have each type of transmission.

```
mtcars %>%
  group_by(transmission) %>%
  summarize(number = n()) %>%
  mutate(proportion = number/sum(number))
```

```
## # A tibble: 2 x 3
##   transmission number proportion
##   <chr>         <int>      <dbl>
## 1 automatic        19      0.594
```

```
## 2 manual           13      0.406
```

---

e. Identify all reasons why the following code won't run (other than the `eval = FALSE` in the chunk options). Make corrections to create side-by-side boxplots with `mpg` on the horizontal axis (labeled Miles per Gallon) and `am` on the vertical axis (labeled Transmission Type).

```
mtcars %>%
  ggplot(x = mpg, y = transmission) %>%
  geom_box() %>%
  labs(x = Miles per Gallon, y = Transmission Type)
```

---

1. inside ggplot, the aesthetic mapping needs to be including insight the `aes()` function.
2. In ggplot, layers should be added with `+` rather than connected with `%>%`,
3. The name for the boxplot geom is `geom_boxplot`
4. The labels for x and y in the `labs` layer need to be surrounded by quotations, since they are character strings.

```
mtcars %>%
  ggplot(aes(x = mpg, y = transmission)) +
  geom_boxplot() +
  labs(x = "Miles per Gallon", y = "Transmission Type")
```