**Public Health Statistics 2, 2022**
**Homework #3, Part B, Due 10/13/22**

A couple of notes with regard to submitting Homework 3 part B:

1. HW1Solutions.RmdPlease submit this as a .pdf document to the dropbox folder linked from the Homework 3 page in Courseplus.
2. While there is no required format for the submitted documents, please number you responses to each item corresponding to the numbering scheme on the assignment. In short please make it so that we can easily find your answers to each item!
3. There are two exercises to this assignment, each with several parts. Please note that the last item for each exercise asks you to post your R code for all computations involved in each part. We will not judge you on your code (there are several ways to do some of the analyses), but please annotate with comments (using the hashtag symbol to denote a comment in the code, i.e. "#") so that we can parse your code.

**Exercise 1:** Consider again the 2001 Harvard School of Public Health College Alcohol Study (CAS) was a nationally representative survey of 10,904 college students asking about lifestyle and alcohol use. For this analysis, you will be provided with data on a sample of 500 students that is similar to the data collected in the survey. For this assignment, you will examine factors associated with a student's grade point average (GPA).

The data file `casData.csv` contains information on the following variables, and can be downloaded from the online library:

- `GPA`: Student's self-reported grade point average
- `schoolyr`: Student's current school year (1=freshman, 2=sophomore, 3=junior, 4=senior, 5=5th year or beyond undergraduate student)
- `DRINK`: Student's drinking status (0=student did not drink in past 30 days, 1=student did drink in past 30 days)
- `age`: Student's age in years
- `male`: Student's sex (0=female, 1=male)

Using this data and the starter code in the accompanying document, complete the following:

**Part A: What is the *unadjusted* relationship between GPA and consumption of alcohol?**

Use R to calculate/create the following things:

1. Side-by-side boxplots showing the difference in GPA between students who drink and students who do not drink. (We will consider students with `DRINK=0` to be non-drinkers and those with `DRINK=1` to be drinkers.)

After performing a simple linear regression to relate GPA to a student's drinking status, report:

2. The estimated mean difference in GPA for students who drink compared to those who do not
3. A 95% confidence interval for this estimated difference

**Part B: What is the relationship between GPA and consumption of alcohol, when *adjusted for* a student's age?**

Use R to calculate/create the following things:

4. A scatterplot showing the relationship between GPA and a student's age
5. Side-by-side boxplots showing the difference in age between students who drink and students who do not drink.

After performing a multiple linear regression to relate GPA to both a student's drinking status and age, report:

6. The estimated mean GPA for 20-year-old students who drink
7. The estimated mean GPA for 20-year-old students who do not drink
8. The estimated mean difference in GPA for students who drink, compared to students *of the same age* who do not
9. A 95% confidence interval for this estimated difference

**Part C: What is the relationship between GPA and consumption of alcohol, when *adjusted for* a student's year in school?**

Use R to calculate/create the following things:

10. Side-by-side boxplots showing the relationship between GPA and a student's year in school. [Hint: To include school year as a categorical variable, you must use `as.factor(schoolyr)` instead of just `schoolyr` in your R code.]
11. A table showing the number and proportions of students in each school year who drink/don't drink

After performing a multiple linear regression to relate GPA to both a student's drinking status and school year, report:

12. The estimated mean GPA for freshman students who drink
13. The estimated mean GPA for senior students who drink
14. The estimated mean difference in GPA for students who drink, compared to students *of the same school year* who do not
15. A 95% confidence interval for this estimated difference

**Part D: What is the relationship between GPA and consumption of alcohol, when *adjusted for* a student's sex?**

Use R to calculate/create the following things, and please report the results for numerical answers, and paste the results for graphics:

16. Side-by-side boxplots showing the difference in GPA between male and female students
17. A table showing the number and proportions of students of each sex who drink/don't drink

After performing a multiple linear regression to relate GPA to both a student's drinking status and sex, report:

18. The estimated mean difference in GPA for students who drink, compared to students *of the same sex* who do not
19. A 95% confidence interval for this estimated difference
20. The estimated mean difference in GPA for male students, compared to female students *of the same drinking status*
21. A 95% confidence interval for this estimated difference

22. Using the results from all of the above analyses, write a short paragraph to answer the question: What is the relationship between GPA and a student's drinking status? Include estimates and confidence intervals in this short write-up.
23. Copy/paste the R code from your calculations for Exercise 1 at the end of the question in `Courier` font. Please only include code that is necessary for these calculations, rather than everything that you tried to do.

**Exercise 2:** Consider again the 1987 National Medical Expenditures Survey (NMES) collected data on health expenditures at the individual level. For this analysis, you will be provided with information on a sample of 4078 such individuals. For this assignment, you will examine factors associated with a high health expenditure, as defined to be more than $10,000 during the year.

The data file `nmesData.csv` contains information on the following variables:

- `eversmk`: Whether the individual has ever been a smoker (1=Yes, 0=No)
- `age`: Individual's age in years
- `beltuse`: How often the individual wears a seat belt (1= Always/almost always, 2=Some, 3= Rare
- `totalexp`: Self-reported total medical expenditures for 1987
- `highexp`: Whether expenditures were more than $10,000 (1=Yes, 0=No)

Using this data and possibly the code in the accompanying document, complete the following:

**Part A: What is the *unadjusted* relationship between high expenditures and smoking?**

Use R to calculate/create the following things:

24. Number and proportion of smokers with high health expenditures; number and proportion of nonsmokers with high health expenditures

After performing a simple logistic regression to relate the log-odds of a high expenditure to smoking status, report:

25. The estimated probability of a high expenditure for a smoker
26. The estimated probability of a high expenditure for a non-smoker
27. The estimated odds ratio of a high expenditure, comparing smokers to nonsmokers
28. A 95% confidence interval for this true odds ratio a high expenditure, comparing smokers to nonsmokers

**Part B: What is the relationship between high expenditures and smoking when *adjusted for* an individual's age?**

Use R to calculate/create the following things: Perform a multiple logistic regression to relate the log-odds of a high expenditure to an individual's age. Based on the results, report the following:

29. The estimated probability of a high expenditure for a 50-year-old smoker
30. The estimated probability of a high expenditure for a 60-year-old non-smoker
31. The estimated odds ratio of a high expenditure, comparing smokers to nonsmokers *of the same age*
32. A 95% confidence interval for this estimated odds ratio
33. The estimated odds ratio of a high expenditure for two groups *of the same smoking* status but whose age differs by 10 years
34. A 95% confidence interval for this estimated odds ratio

35. Using the results from all of the above analyses, write a short paragraph to answer the question: What is the relationship between high expenditures and an individual's smoking status? Include estimates and confidence intervals in this short write-up.
36. Copy/paste the code/output from your calculations for Exercise 2 at the end of the question in `Courier` font. Please only include code/output that is necessary for these calculations, rather than everything that you tried to do.