**Biostatistics 140.653**

**Third Term, 2021**

**March 1, 2021**

Quiz 2

**The purpose of this quiz is to assess your knowledge of the course materials covered during the second two weeks of class and covered in Problem Set 2.**

**Instructions:**

* **This is an open book quiz; you may consult your course notes and handouts.**
* **You should not discuss this quiz with any other student during Monday March 1st.**
* **This quiz is designed to be completed in 20-30 minutes.**
* **You may provide your solution by editing the word version of this quiz, annotating the pdf version of this quiz or writing your solution on paper and submitting a picture of your solution.**

By signing my name, I enter agree to abide by the instructions above and the Johns Hopkins University School of Public Health Academic Code:

Name (Print): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Suppose you use the Problem Set 2 NMES data for persons aged 19 to 94 to address the question of whether males and females use roughly the same quantity of medical services at each age by regressing **log total expenditures** on a non-linear function of age, sex and the interactions. Specifically, you regress:

Y = LN(totalexp +1) = loge(totalexp+1), on:

* Intercept
* age variables: *age-40, age\_sp1 = (age- 40)+, age\_sp2 = (age- 65)+*
* *female*= 1 if female, 0 if male
* the interaction of the three age variables with *female*

1. The coefficient for *female* estimates the difference in average log expenditures comparing:
   1. 40 year-old males to females (i.e. males minus females)
   2. 40 year-old females to males
   3. 65 year-old males to females
   4. 65 year-old females to males
2. To determine whether there is a difference in the average log expenditures between males and females at any age you would:
   1. Fit a second model including only the *intercept* and perform the ANOVA F-test
   2. Fit a second model excluding the 2 main effect terms of *age\_sp1* and *age\_sp2* as well as the 2 interaction terms of *female* with *age\_sp1* and *age\_sp2* and perform the ANOVA F-test
   3. Fit a second model excluding the 3 interaction terms between the age variables and f*emale* and perform the ANOVA F-test
   4. Fit a second model excluding the main effect for *female* and the 3 interaction terms between the age variables and perform the ANOVA F-test
3. Based upon your model results, a colleague asks for your best estimate for the average log expenditure for 50 year-old males. Give a formula for the estimate of the **average log expenditure** and **its standard error** using the estimated regression coefficients and their covariance matrix V. Writing R code to get the result would be as good as a more mathematical answer.
4. This same colleague realizes that interpreting the average log expenditure is difficult and asks that you provide your best estimate for the averageexpenditure for 50 year-old males in dollars.

Recall that Y is log normal if LN(Y)~N(, ).

If Y is log-normal, then E(Y) = exp(), median(Y) = exp().

Give a formula for the estimate of the average expenditure for 50 year-old males using the estimated regression coefficients and estimate of , if necessary. You DO NOT have to derive the standard error for this mean.

1. In most cases, we want to also provide a 95% CI for means of interest. Describe, in words, an approach to generate a 95% confidence interval for the average expenditure for 50 year-old males.