**Biostatistics 140.654**

**Fourth Term, 2021**

**May 3, 2021**

**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 May 3rd.**
* **This quiz is designed to be completed in 20-30 minutes.**
* **Each multiple choice question has a single best answer.**
* **There are 5 questions on this quiz; one question is a BONUS. Completing the bonus can only earn you extra points, i.e. if you choose to not answer this question, you will not lose any points.**
* **You can use calculators or R on your computer for arithmetic.**
* **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: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

The goal of the analysis is to explore predictors of having a major smoking caused disease (*MSCD*). We will consider two main predictors: whether the person ever smoked (*eversmk:* 1 if ever smoker, 0 if never smoker) and age. We used the rfImpute command to impute the missing ever smoker information and considered a non-linear function of age using a linear spline with breaks at 60 and 80 year. Specifically, I centered at 60 (*age\_c*), and created two linear spline terms with breaks at 60 and 80 years (*age\_sp1, age\_sp2*).

We fit a logistic regression model for the log odds of having a MSCD as a function of being an ever smoker and the non-linear function of age.

d1$agec = d1$lastage - 60

d1$agesp1 = ifelse(d1$agec>0,d1$agec,0)

d1$agesp2 = ifelse(d1$lastage>80,d1$lastage-80,0)

fit = glm(mscd~eversmk+agec+agesp1+agesp2,data=d1,

na.action=na.omit,family="binomial")

> summary(fit)$coefficients

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.53362870 0.074768354 -33.886378 1.057476e-251

eversmk 0.68465835 0.062073289 11.029838 2.743575e-28

agec 0.10593581 0.007982449 13.271092 3.405821e-40

agesp1 -0.05303630 0.011298114 -4.694262 2.675709e-06

agesp2 -0.04047916 0.017664672 -2.291532 2.193267e-02

> round(summary(fit)$cov.scaled,5)

(Intercept) eversmk agec agesp1 agesp2

(Intercept) 0.00559 -0.00297 0.00029 -0.00054 0.00039

eversmk -0.00297 0.00385 -0.00001 0.00003 0.00005

agec 0.00029 -0.00001 0.00006 -0.00008 0.00004

agesp1 -0.00054 0.00003 -0.00008 0.00013 -0.00009

agesp2 0.00039 0.00005 0.00004 -0.00009 0.00031

1. Using the fit of the model, we estimate that the relative odds of having a MSCD, comparing a 60 year-old ever smoker to a 60 year-old never smoker is:
   1. 0.68
   2. exp(0.68)
   3. exp(-2.53)/{1+exp(-2.53)}
   4. exp(-2.53+0.68)/{1+exp(-2.53+0.68)}
2. Using the fit of the model, we estimate that the *relative risk* of having an MSCD, comparing a 60 year-old ever smoker to a 60 year-old never smoker is:
   1. 0.68
   2. exp(0.68)
   3. exp(-2.53 + 0.68)\*{1+exp(-2.53)}/[ exp(-2.53){1+exp(-2.53 + 0.68)} ]
   4. exp(-2.53+0.68)
   5. cannot estimate the relative risk with a logistic regression
3. BONUS: Using the fit of the model, provide an estimate of and 95% confidence interval for the probability a 60 year-old ever smoker has a MSCD. Show your work.

Next, we evaluated the ability of our model to predict MSCD status. We partitioned the data into a 70:30 training and validation sample. The training and validation samples were drawn within strata of MSCD status. We refit the model above on the training sample and obtained the estimated Pr(MSCD = 1 | ever smoker, age) for each person in the validation sample. Some of the key output of this process is below:

Figure 1: Estimated probability of having a MSCD as a function of ever smoker and age, stratified by MSCD status, for individuals in the training data



Figure 2: Receiver operating characteristic, ROC, plot



1. Define the classifier: , where is a value between 0 and 1, is the estimated probability of having a MSCD obtained from the logistic regression model fit, is the indicator function evaluating to 1 if is true and 0 otherwise. Further, the sensitivity and specificity are given by and , respectively. When we set c = 0.2, the sensitivity and specificity are 0.36 and 0.85, respectively. To create a classifier with better sensitivity, i.e. higher, we would
   1. Increase c
   2. Decrease c
   3. There is no way to improve the sensitivity
   4. There is not enough information provided
2. Figure 2 displays the receiver operating characteristic curve (ROC) generated from the fit of the logistic regression model. The area under the curve (AUC) is 0.74. Propose a method for generating a 95% confidence interval for the AUC.