Biostatistics 140.654
Fourth Term, 2021
May 3, 2021

SOLUTION

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:
    a. 0.68
    b. exp(0.68)
    c. exp(-2.53)/{1+exp(-2.53)}
    d. 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:
    a. 0.68
    b. exp(0.68)
    c. exp(-2.53 + 0.68)*{1+exp(-2.53)}/[exp(-2.53){1+exp(-2.53 + 0.68)}]
    d. exp(-2.53+0.68)
    e. 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.

    See solution at the end of the document.

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

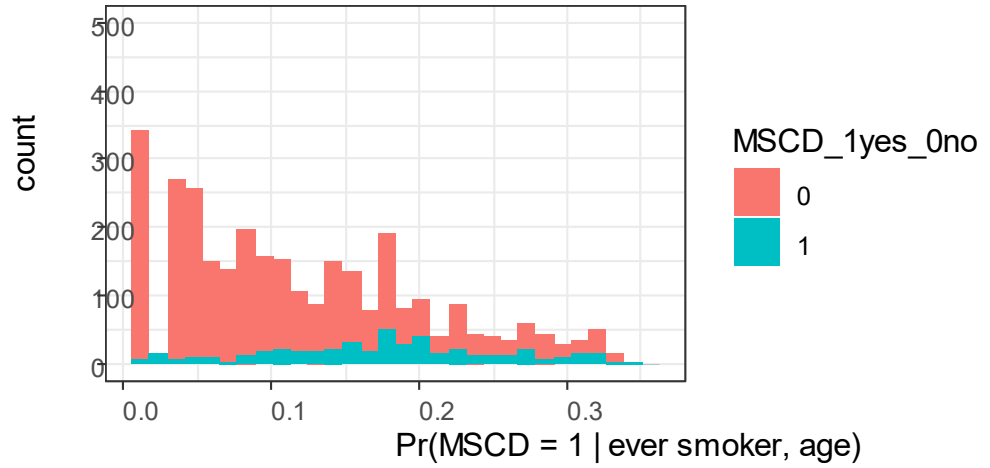
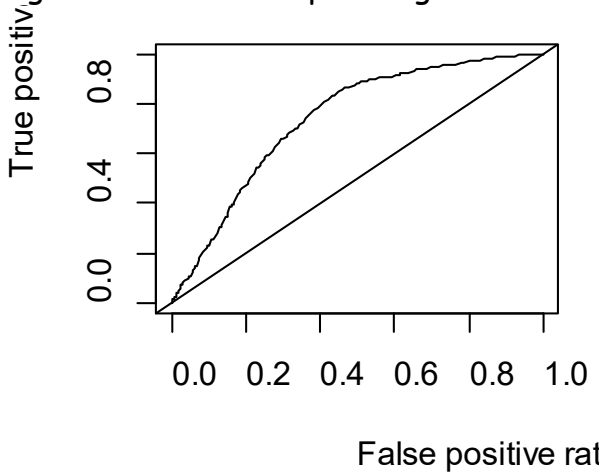


Figure 2: Receiver operating characteristic, ROC, plot



4. Define the classifier: $d(c, \hat{\mu}_i) = I(\hat{\mu}_i > c)$, where c is a value between 0 and 1, $\hat{\mu}_i$ is the estimated probability of having a MSCD obtained from the logistic regression model fit, $I(exp)$ is the indicator function evaluating to 1 if exp is true and 0 otherwise. Further, the sensitivity and specificity are given by $\Pr(d(c, \hat{\mu}_i) = 1 \mid MSCD_i = 1)$ and $\Pr(d(c, \hat{\mu}_i) = 0 \mid MSCD_i = 0)$, 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
    a. Increase c
    b. Decrease c
    c. There is no way to improve the sensitivity
    d. There is not enough information provided

5. 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.

Answer: One approach would be to apply a bootstrap procedure. To implement the bootstrap procedure, you would complete the following steps:

For b = 1 to B (a large number):

a) take a sample of size n (the number of observations in the sample) with replacement from the original data
b) Partition the data into the 70:30 training:validation sample
c) Fit the logistic regression model to the training sample
d) Get predicted Pr(MSCD|ever smoker, age) for the validation sample
e) Compute the ROC curve and save the AUC as AUC_b

Generate a 95% CI for the AUC by taking the 2.5^{th} and 97.5^{th} percentiles of AUC_b or compute the BCa confidence interval.

Two options for the Bonus question:

Option 1) Compute the Logodds, Var(Logodds) and 95% CI for the Logodds then inverse-Logit the endpoints.

$\widehat{est} = \hat{\beta}_0 + \hat{\beta}_1 = -2.53 + 0.68 = -1.85$

$Var(\widehat{est}) = Var(\hat{\beta}_0) + Var(\hat{\beta}_1) + 2\,Cov(\hat{\beta}_0, \hat{\beta}_1)$

$= .00559 + .00385 + 2(-.00297) = .0035$

$se(\widehat{est}) = .059$

95% CI for logodds: $-1.85 \pm 1.96 \times .059$

$= -1.96$ to -1.73

Inverse logit: Estimate .136

CI % CI: .123 to .151

Option 2: Apply the delta method

define $f(\beta_0, \beta_1) = \exp(\beta_0 + \beta_1) / [1 + \exp(\beta_0 + \beta_1)]$

We need $\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$ and var/cov matrix $\begin{bmatrix} .00559 & -.00297 \\ -.00297 & .00385 \end{bmatrix}$

and $\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0}$ and $\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_1}$

by applying the product rule of derivatives to $\exp(\beta_0+\beta_1)[1+\exp(\beta_0+\beta_1)]^{-1}$ we get

$$\frac{\partial f(\beta_0,\beta_1)}{\partial \beta_0} = \frac{\partial f(\beta_0,\beta_1)}{\partial \beta_1} = \frac{\exp(\beta_0+\beta_1)}{[1+\exp(\beta_0+\beta_1)]^2}$$

---

est: $f(\hat{\beta}_0,\hat{\beta}_1) = .136$

$$Var[f(\hat{\beta}_0,\hat{\beta}_1)] = \begin{bmatrix} \frac{\partial f(\hat{\beta}_0,\hat{\beta}_1)}{\partial \beta_0} \\ \frac{\partial f(\hat{\beta}_0,\hat{\beta}_1)}{\partial \beta_1} \end{bmatrix} Var/Cov(\hat{\beta}_0,\hat{\beta}_1) \begin{bmatrix} \frac{\partial f(\hat{\beta}_0,\hat{\beta}_1)}{\partial \beta_0} & \frac{\partial f(\hat{\beta}_0,\hat{\beta}_1)}{\partial \beta_1} \end{bmatrix}$$

$$\begin{bmatrix} .12 \\ .12 \end{bmatrix} \begin{bmatrix} .00559 & -.00297 \\ -.00297 & .00385 \end{bmatrix} \begin{bmatrix} .12 & .12 \end{bmatrix}$$

$$= .0000504$$

$$se(f(\hat{\beta}_0,\hat{\beta}_1) = .0071$$

95% CI: $0.136 \pm 1.96 \times .0071$

.123 to .151