Biostatistics 140.653
Third Term, 2021
March 1, 2021

Quiz 2 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 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) = $\log_e$(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:
    a. 40 year-old males to females (i.e. males minus females)
    b. 40 year-old females to males
    c. 65 year-old males to females
    d. 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:
    a. Fit a second model including only the *intercept* and perform the ANOVA F-test
    b. 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
    c. Fit a second model excluding the 3 interaction terms between the age variables and f*emale* and perform the ANOVA F-test
    d. 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 $\hat{\beta}$ and their covariance matrix V. Writing R code to get the result would be as good as a more mathematical answer.

SOLUTION: Let Y = LN(totalexp + 1) $= \beta_0 + \beta_1(age - 40) + \beta_2(age - 40)^+ + \beta_3(age - 65)^+ + \beta_4 female + \beta_5 female(age - 40) + \beta_6 female\,(age - 40)^+ + \beta_7 female(age - 65)^+ + \varepsilon, \varepsilon \sim N(0, \sigma^2)$

The average log expenditures for 50 year-old males is: $\beta_0 + \beta_1(50 - 40) + \beta_2(50 - 40) = \beta_0 + 10\beta_1 + 10\beta_2$. To estimate this average, you would plug in your estimates for $\beta_0, \beta_1, \beta_2$.

To estimate the standard error, you would need to compute the variance:

$$Var(\hat{\beta}_0 + 10\hat{\beta}_1 + 10\hat{\beta}_2) = Var(\hat{\beta}_0) + 100Var(\hat{\beta}_1) + 100Var(\hat{\beta}_2) + 20Cov(\hat{\beta}_0, \hat{\beta}_1) + 20Cov(\hat{\beta}_0, \hat{\beta}_2) + 200Cov(\hat{\beta}_1, \hat{\beta}_2)$$

Then take the square root to get the standard error.

In matrix notation, you would define A = (1,10,10,0,0,0,0,0) and compute A %*% fit$coefficients to obtain the estimate of the average log expenditure for 50 year-old males and the variance is A %*% fit$vcov %*% t(A). The standard error would the square root of the variance.

4. This same colleague realizes that interpreting the average log expenditure is difficult and asks that you provide your best estimate for the average expenditure for 50 year-old males in dollars.

Recall that Y is log normal if LN(Y)~N(μ, σ^2).

If Y is log-normal, then E(Y) = exp($\mu + \sigma^2/2$), median(Y) = exp(μ).

Give a formula for the estimate of the average expenditure for 50 year-old males using the estimated regression coefficients and estimate of σ^2, if necessary. You DO NOT have to derive the standard error for this mean.

SOLUTION: The get the average expenditure for 50 year-old males, you would take the estimated average log expenditures: $\hat{\mu} = \hat{\beta}_0 + 10\hat{\beta}_1 + 10\hat{\beta}_2$ and $\hat{\sigma}^2$, the estimated residual variance and compute: $\exp(\hat{\mu} + \hat{\sigma}^2/2) - 1$, note the "minus 1" is included because of the original transformation of the data is LN(expenditure + 1) (which makes it possible to take the natural log of the data for someone with no expenditures).

5. 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.

SOLUTION: There are two choices here:

a) Apply the delta method to $exp\ (\hat{\mu} + \hat{\sigma}^2/2)$ - 1, where you would assume $\hat{\sigma}^2$ is fixed, to estimate $Var(exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right) - 1)$, take the square root of this variance estimate and compute the 95% CI.

b) Apply the bootstrap procedure by taking repeated samples of the data with replacement, fit the model described in Question 1, compute and save $\hat{\boldsymbol{\mu}}$. Take the values of $\hat{\boldsymbol{\mu}}$ apply the percentile method of BCa method to generate the 95% CI.

The second approach has the advantage of not relying on the normality assumption of the residuals and the constant variance assumption.