HW8

Due:3/6/2019

Instructions. In this homework, we will continue to work on analyzing the association between *Insurance* and *Stage at diagnosis* with four covariates *Age*, *Sex*, *Race*, and *Poverty*. Please submit your homework by uploading the .RMD file or the HTML NB file to Canvas under the HW8 assignment. Please also store the dataset that you download from Canvas in the same folder as the .RMD file for this homework so that you will not have to change the import code (and neither will we when we grade it).

Run the below code to import the data and open the libraries needed for the homework. You may have to install some packages for the libraries if you have not already done so during the class demo.

```
library(readr) #for read txt file
library(knitr) #for creating nicer tables
library(tidyverse) # for various packages
library(nnet) #Multinomial logistic regresison
library(MASS) #Ordinal logistic regression
Breast_SEER_Class6 <- read_delim("Breast_SEER_Class6.txt", "\t", escape_double = FALSE, trim_ws = TRUE)
str(Breast_SEER_Class6)
```

Run the below code to make new variables for *marital status* categorizing marital status into single and married and converting the *poverty* variable to a numeric variable.

```
names(Breast_SEER_Class6)<-c("Age", "Race", "Sex", "Diagnosis_year", "Stage", "First",
                             "PatientID", "Insur", "Marital", "Poverty")
Breast_SEER_Class6 <- Breast_SEER_Class6 %>%
   mutate(Male_cat=factor(Sex),
          First_cat=factor(First),
          Age_num=as.numeric(gsub("([0-9]+).*$", "\\1",Age)), #make age as numeric variables
          #Non-Black as reference
          Race_cat=case_when(Race=="White"
                             |Race=="Other (American Indian/AK Native,
                             Asian/Pacific Islander)" ~ 0,
                              Race=="Black" ~ 1),
         Race_cat=factor(Race_cat, 0:1, c("Non-Black", "Black")),
          #Stage I as reference
          Stage_cat=case_when(Stage == "IA" | Stage == "IB" ~ 0,
                              Stage == "IIA" | Stage == "IIB" ~ 1,
                              Stage == "IIIA" |Stage == "IIIA"
                              |Stage == "IIIC" |Stage =="IIINOS" ~ 2,
                              Stage =="IV" ~ 3),
          Stage_cat=factor(Stage_cat, 0:3, c("StageI", "StageII", "StageIII", "StageIV")),
          #Uninsured as reference
          Insur_cat = case_when(Insur == "Uninsured" ~ 0,
                                Insur == "Any Medicaid"  ~ 1,
                                Insur == "Insured"|Insur == "Insured/No specifics" ~ 2),
          Insur_cat=factor(Insur_cat, 0:2, c("Uninsured", "Medicaid", "Insured")),
```

```
        Poverty_num=as.numeric(Poverty))%>%
    na.omit() %>%
    filter(First_cat=="Yes")
```

1. Re-level (using the *relevel* function) the dependent variable to stage I as the reference category for modeling.

2. Execute a mutlinomial regression model to examine the effect of insurance category on stage at diagnosis. Include the covariates (Race_cat, Age_num, Poverty_num, and Male_cat) in your model to adjust for the association between insurance category and stage at diagnosis.

3. Perform hypothesis tests for all non-reference levels of the *insurance* variable, explain what you find at each level compared to the reference level.

4. Get odds ratios and 95% confidence intervals for associations between insurance status and stage at diagnosis, interpret the results. Hint: you should have a total of 6 ORs.

5. Execute an ordinal logistic regression with the *insurance* variable as the independent variable and the same covariates.

6. Perform hypothesis tests for all non-reference levels of the *insurance* variable, explain what you find for each level compared to the reference level.

7. Get odds ratios and 95% confidence intervals for associations between insurance status and stage at diagnosis, interpret the results.

8. What general conculsions can you draw from both models in terms of the association betwwen insurance and stage among breast cancer patients?