Assignment 2

Biostatistics 1: Introduction to Biostatistics, 5BD000

November 2024

The Data Step

Package

Begin by loading all the necessary packages you need at top.

```
pacman::p_load(tidyverse,rms,haven,mgcv,epitools,logistf,nlpred,geepack,
               skimr,pROC,tableone,emmeans,glmtoolbox,CalibrationCurves,mice, dcurves)
```

Read in the data and define variables

In the package epitools we have access to the dataset wcgs or the Western Collaborative Group Study data. It is a prospective cohort study, that recruited middle-aged men (ages 39 to 59) who were employees of 10 California companies and collected data on 3154 individuals during the years 1960-1961. These subjects were primarily selected to study the relationship between behavior pattern and the risk of coronary hearth disease (CHD). A number of other risk factors were also measured to provide the best possible assessment of the CHD risk associated with behavior type. Additional variables collected include age, height, weight, systolic blood pressure, diastolic blood pressure, cholesterol, smoking, and corneal arcus

Here you can find information on the dataset and variables: https://rdrr.io/cran/epitools/man/wcgs.html

Now read in the original data and then define derived variables we need or the in the format we need to use them.

```
# read in the data
data(wcgs)
# create factors from var ditpat0 with levels B and A in stead of 0 and 1
wcgs$dibpat0f<-factor(wcgs$dibpat0,levels=0:1,label=c("B","A"))
# create categories for age
wcgs$agegroup <- cut(wcgs$age0,breaks=c(39,45,55,60),include.lowest = T,right = FALSE)
# binary variable for smoker or not
wcgs$smoker <- ifelse(wcgs$ncigs0>0,1,0)
wcgs$smokerf <- factor(wcgs$smoker,levels=c(0,1),labels=c("No","Yes"))
# convert height from inches to cm
wcgs$heightcm <- wcgs$height0*2.54
wcgs$weightkg <- wcgs$weight0*0.45359237
wcgs$bmi <- wcgs$weightkg / (wcgs$heightcm/100)^2

wcgs$bmicat <-  cut(wcgs$bmi,breaks=c(0,25,30,40),include.lowest = T,right = FALSE)

wcgs$cholmmol <- wcgs$chol0/39
```

```
wcgs$sbp10 <- wcgs$sbp0/10

# Create categories of sbp (systolic blood pressure). Make sure to have the lowest and highest numbers
wcgs$sbpcat <- cut(wcgs$sbp0,breaks=c(0,140,240),include.lowest = T,right = FALSE)

# For use when tabulating the data you can create labels
wcgs$chd69f <- factor(wcgs$chd69,levels=c(0,1),labels=c("No","Yes"))

wcgs$cholmmol <- ifelse(wcgs$cholmmol<15,wcgs$cholmmol,NA)
```

The Background

Here we study the outcome coronary heart disease (chd69). Your task is to find a prediction model for the risk of CHD. We will consider the following covariates/ predictors which are previously known risk factors for heart disease: behaviour type A/B, age, systolic bp, cholesterol, bmi, smoking, and arcus.

Decide on the variables to use

In this analysis we'll consider the following variables.

chd69 age systolic bp cholesterol bmi smoking arcus behaviour type

```
d <- wcgs %>% select(id,agegroup,age0,cholmmol,sbp10,bmi,smokerf,arcus0,dibpat0f,chd69)
```

Create the analysis set only including the variables we want to use.

```
dc  <- d %>% drop_na()
```

Create a complete case version Compare them

```
nrow(d)
```

```
## [1] 3154
```

```
nrow(dc)
```

```
## [1] 3139
```

Extra excercise: Impute the missing data. Since there is only a small fraction missing here, we will use a method called predictive mean matching and use one dataset (m=1). This is not included in the lectures.

```r
set.seed(154550)
imp <- mice(d, m=1, maxit=0)
predM<-imp$predictorMatrix
# Leave out the ID column (the first column)
predM[, 1] <- 0
meth<-imp$method
dimp <- mice(d, method= "pmm" ,m=1,predictorMatrix = predM, maxit=15, seed=71332, print=FALSE)
di <- complete(dimp,1)
```

Here you can see how the missing values for cholesterol were replaced

```r
di[is.na(d$cholmmol),"cholmmol"]
```

```
##  [1] 8.410256 4.256410 6.615385 7.333333 5.256410 7.076923 5.615385 6.794872
##  [9] 5.692308 4.487179 7.846154 5.846154 4.410256
```

1. Table 1

Create table 1. Describe all available variables in the data. Show both, the original data and the imputed data. Hint: use package 'tableone' and one function from that package to create the table

We will continue with the imputed data we have now called di.

2. Overall risk or overall rate

a. What is the outcome we are interested in?

b. What are the known risk factors for our outcome of interest?

c. How many persons are included?

d. What is the overall risk or rate and prevalence of the disease in our cohort?

3. Building the model and choosing predictors

a. Use the available data to create the optimal prediction model for our outcome, choose the appropriate model and the predictors that will improve your predictions. Please explain the reasons for choosing the model type you chose and why the predictors were chosen.

b. Are there any predictors that should be included as interaction terms with other variables or as categorical variables? If so, please explain the reasoning for this in the model?

c. Calculate the predicted risk of the outcome for every person in the dataset according to the model you have chosen and add it to your dataset.

Diagnostics

4. Discrimination

a. AUC. How good is the model you have built to discriminate between cases and non-cases? Please plot the ROC curve and calculate the AUC of the ROC curve including 95% confidence intervals.

b. Please plot the ROC curve and find the threshold that maximizes the sum of the sensitivity and specificity. Please report the sensitivity and specificity at that threshold.

c. AUC adjusted for optimism.To adjust for optimism in the predictions, we can use the bootstrapping method using 200 repitions. Please calculate the adjusted AUC with 95% confidence intervals and compare it to the unadjusted AUC. Is there a difference in the unadjusted and the adjusted AUC?

hint: you can use the *validate* function from package *rms* to estimate the adjusted value for auc due to optimism via the bootstrap method.

d. Cross validation. Use 10-fold cross-validation for a logistic model to obtain the adjusted AUC and compare to the unadjusted and adjusted AUC.

5. Calibration

a. Calibration Curve. To evaluate the performance of the prediction model we have computed the AUC from ROC analysis, now please plot the calibration curve and report the slope and the intercept of the calibration curve. Use the model you chose before.

b. Please estimate the goodness of fit by the method of Hosmer and Lemeshow. Interpret the test results.

c. Create a prediction model only using variable agegroup as a predictor and estimate the discrimination.

d. Please compare the discrimination to the model you used before with a statistical test and interpret the results.

e. Plot both ROC curves in one figure.

6. Decision Curve Analysis

a. Plot the decision curve for the model you created and compare its net benefit to the default strategies provided in the function.

b. Is the model you have built clinically useful or clinically harmful? At what clinical thresholds is it beneficial?

c. Differences in NB between models. Plot the model you created in 5.c and add it to the figure you have plotted of the net benefit in 6.a. Make comparisons between the two curves.

7. Discussion

Please discuss if the model you have created is clinically useful and what the next possible steps would be for you as a researcher.