DATA 303-DATA473 Assignment 4

Due: 11:59 PM Sunday 4 June 2023

Intructions

- Prepare your assignment using Rmarkdown
- Submit your solutions in two files: an Rmarkdown file named `assignment4.Rmd` and the PDF file named `assignment4.pdf` that results from knitting the Rmd file.
- The YAML header of your Rmarkdown file must contain your name and ID number in the author field, and should have the output format set to `pdf_document`. For example:

```
---
title: "DATA 303 Assignment 4"
author: "Ryan Admiraal, 12345678"
date: "4 June 2023"
output: pdf_document
---
```

- While you are developing your code you may find it easiest to have the output set to `html_document` but change it to `pdf_document` when you submit.
  - A common error that occurs in producing a PDF is when unicode characters (e.g., β, α) are included in your text. The error usually looks something like "! Package inputenc Error: Unicode char \u8: not set up for use with LaTeX." or something similar. A potential workaround is provided at https://bookdown.org/yihui/rmarkdown-cookbook/latex-unicode.html. Alternatively, if you perform a Google search for the unicode character in question (e.g., "unicode \u8"), you can more readily identify the unicode character in question, find it in your Rmarkdown file, and replace it with a more appropriate character for knitting to PDF (e.g., β can be produced by typing "β").
- In your submission, embed any executable R code in code chunks, and make sure both the R code and the output is displayed correctly when you knit the document.
- If there are any R code errors, then the Rmarkdown file will not knit, and no output will be created at all. If you cannot get your code to work but want to show your attempted code, then put `error = TRUE` in the header of the R code chunk that is failing.

```
```{r, error = TRUE}
your imperfect R code
```
```

- Where appropriate, make sure you include your comments in the output within the Rmarkdown document.
- **You will receive an email confirming your submission. Check the email to be sure it shows that both the Rmd file and the PDF file have been submitted.**

Background and Data

Heart disease is the annual leading cause of death worldwide, accounting for more than 25% of deaths in 2016 (World Health Organization 2018). It is also a significant economic burden for the healthcare system with Nichols et al. (2010) estimating that heart disease and other cardiovascular diseases cost an average of roughly USD $19,000 per patient, according to a study in the United States over the period of 2000-2005. Early detection of heart disease (along with many other diseases) is important in terms of reducing both mortality and costs to the healthcare system.

We will examine data on 4,240 participants in the Framingham Heart Study (Boston University and the National Heart, Lung, & Blood Institute 2020), an ongoing study that began in 1948 and has been instrumental in the identification of a number of risk factors for heart disease and other cardiovascular diseases. The data are available in the file `Framingham Heart Study.xlsx`, which can be read into R using the code below but with the path changed to point to the location of the file on your computer. A full list of variables contained in the dataset and descriptions of these variables is also provided, both here and in the Excel file.

```
# Load the "readxl" package to read in data from an Excel file.
library(readxl)
# Read in the heart disease dataset.
hd <- read_xlsx("~/Documents/Dropbox/Courses/DATA303/Data/Framingham Heart Study.xlsx",
sheet = "Data", na = "NA")
```

Table 1: Variables and their descriptions for data contained in the file `Framingham Heart Study.xlsx`.

Variable	Description
`SEX`	Sex of the individual (0 = "Female", 1 = "Male").
`AGE`	Age (in years) of the individual at the time of the health exam.
`EDUC`	Highest level of education of the individual (1 = "Some high school", 2 = "High school or Graduate Equivalency Diploma", 3 = "Some university or vocational school", 4 = "University").
`SMOKER`	Indicator of whether or not the individual is a current smoker (0 = "No", 1 = "Yes").
`CIG`	Average number of cigarettes that the individual smokes each day.
`BP_MED`	Indicator of whether or not the individual is on blood pressure medication (0 = "No", 1 = "Yes").
`STROKE`	Indicator of whether or not the individual previously had a stroke (0 = "No", 1 = "Yes").
`HYPER`	Indicator of whether or not the individual was hypertensive (0 = "No", 1 = "Yes").
`DIAB`	Indicator of whether or not the individual is diabetic (0 = "No", 1 = "Yes").
`CHOL`	Total cholesterol level (in mg/dL).
`SBP`	Systolic blood pressure (in mmHg).
`DBP`	Diastolic blood pressure (in mmHg).
`BMI`	Body mass index.
`HR`	Resting heart rate (in beats per minute)
`GLUC`	Glucose level (in mg/dL)
`HD_RISK`	Indicator of whether the individual has 10-year risk of future coronary heart disease (0 = "No", 1 = "Yes")

Our focus will be on 10-year risk of coronary heart disease (CHD). Ten-year risk of CHD is a predicted risk (*i.e.*, a probability ranging between 0 and 1) of developing CHD within the next 10 years. Although this is not an observed outcome but rather an estimated value, 10-year risk of CHD is a well-established measure in the medical community. We will consider a binary version of this variable which indicates whether or not a person would be considered as at risk of developing CHD within the next 10 years.

Assignment Questions

1. **Missing data and variable recode: (10 marks)**

   Although our objective will be to consider inferential and predictive models for 10-year risk of CHD, we will first ensure that we understand aspects of the underlying data as well as create a new variable that may prove useful in producing comparisons of 10-year risk of CHD for medically-meaningful blood pressure ranges. (In practice, we would want to examine each relevant variable to identify extreme observations and be sure that there are not any erroneous values. As this dataset has already been cleaned, we will not do so for this assignment.)

   a. **(2 marks)** First, perform an analysis of the level of missing data for each variable. For only those variables for which there are missing data, produce a table of the form shown below, where `VARIABLE_i` is the name of the variable with missing data, n_i is the count for number of missing observations for that variable, and p_i is the proportion (to 5dp) of missing observations for that variable. Which variable has the highest level of missing data?

   Table 2: Frequency and proportion of missing values for variables with missing data.

Variable	`VARIABLE_1`	`VARIABLE_2`	...	`VARIABLE_k`
Frequency (n)	n_1	n_2	...	n_k
Proportion (p)	p_1	p_2	...	p_k

   b. **(3 marks)** Create a new data frame called `hd.complete`, which only keeps people/observations that have no missing data. In total, what proportion (to 5dp) of people have been removed from the original dataset to produce this final data frame?

   c. **(3 marks)** Add a variable to the data frame `hd.complete` called `SBP_CAT`, which converts systolic blood pressure (`SBP`) from a numeric variable to a categorical variable according to the blood pressure ranges specified by Madell and Cherney (2018). (See references listed at the end of the assignment.) For the purposes of coding `SBP_CAT`, you can assume that the values for each blood pressure category go to just below that of the next category, as our dataset does not consist of blood pressures that are rounded to the nearest whole number. This means that, for instance, the systolic blood pressure range of 120 – 129 should in fact be interpreted as 120 – $<$ 130. This should produce five levels (*i.e.*, blood pressure ranges) for `SBP_CAT`. (Note that the final level corresponds to systolic blood pressure above 180 mmHg.) Produce a table for `SBP_CAT` which shows how many observations fall into each blood pressure range.

   d. **(2 marks)** Explain when we would expect that using the categorical variable `SBP_CAT` rather than the numeric variable `SBP` would lead to a better fit for a regression model (whether logistic regression, linear regression, or Poisson regression).

2. **Inferential analysis: (25 marks)**

   Now we will focus on 10-year risk of CHD and look at the role that blood pressure may play in whether or not someone is considered to be at risk of developing CHD within the next 10 years.

   a. **(3 marks)** We will first consider a logistic regression model of 10-year risk of CHD (`HD_RISK`) on systolic blood pressure (`SBP`) and diastolic blood pressure (`DBP`). Previous research suggests that the following variables are potential confounders for the true relationship between blood pressure and 10-year risk of CHD and should also be included in the logistic regression model:

   - sex of the individual (`SEX`)
   - age of the individual (`AGE`)
   - highest level of education of the individual (`EDUC`)
   - average number of cigarettes smoked per day (`CIG`)

- total cholesterol level (`CHOL`)
- body mass index (`BMI`)
- glucose level (`GLUC`)

For this logistic regression model, calculate the variance inflation factors for predictors (to 3dp) to determine whether or not there is evidence of significant multicollinearity among the predictors in the model. If so, comment on which predictor(s) should be removed, and use this model for subsequent parts of this question.

b. **(3 marks)** Using your model from part (a), produce a table of logistic regression model output and write out the estimated logistic regression equation using the form

$$\log\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \quad = \quad \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \cdots + \widehat{\beta}_k X_k,$$

where you clearly define the variables X_1, X_2, ..., X_k and replace $\widehat{\beta}_0$, $\widehat{\beta}_1$, ..., $\widehat{\beta}_k$ with their estimated values (to 4dp).

c. **(6 marks)** Carry out Wald tests for the coefficients for

- systolic blood pressure and
- diastolic blood pressure.

For each coefficient, clearly state

i. the hypotheses you are testing,
ii. the value of the test statistic,
iii. the p-value, and
iv. your conclusion in terms of whether the "effect" of the predictor on the response is statistically significant.

d. **(3 marks)** For any significant Wald tests in part (c), provide a precise interpetation of what the estimated coefficient suggests about the "effect" of the predictor on the response, and calculate a corresponding 95% confidence interval (to 3dp) for the estimated "effect".

e. **(4 marks)** A 2015 study by Wu et al. (2015) found that

> "cardiovascular and expanded-cardiovascular mortality risks were lowest when systolic blood pressures were 120 to 129 mm Hg, and increased significantly when systolic blood pressures (SBPs) were $\geq$ 160 mm Hg. . . ."

Although Wu et al. (2015) considered different ranges of systolic blood pressures ($<$ 120, 120-–129, 130-–139, 140-–149, 150—159, $\geq$ 160 mmHg) than Madell and Cherney (2018), we will use those specified by Madell and Cherney (2018) in investigating whether ranges of blood pressures may differ in terms of associated 10-year risk of CHD.

Fit the same model as before, but replace `SBP` with `SBP_CAT`.

i. Produce a table of logistic regression model output for this model.
ii. Based strictly on p-values, comment on what conclusions you would make for Wald tests based on coefficients for `SBP_CAT`. (Note that you do not need to state hypotheses or values for test statistics. You simply need to use the p-values to explain what these results mean about comparisons of systolic blood pressure ranges.)
iii. Do your results agree with the findings of Wu et al. (2015)?

f. **(3 marks)** Does the model that uses `SBP_CAT` (*i.e.*, the model fit in part (e)) provide a better fit than the model that uses `SBP` (*i.e.*, the model from part (a))?

g. **(3 marks)** Finally, for the best model of the two you fit (in parts (a) and (e)), perform a Hosmer-Lemeshow test for $g = 10$, 20, and 30 groups, and comment on what these suggest about the goodness-of-fit of this model to the 10-year risk of CHD data.

3. **Statistical learning: (15 marks)**

   Now we perform an exploratory analysis to try to identify the best set of predictors in predicting 10-year risk of CHD. Consider as predictors all variables other than the new variable that you constructed in Question 1 (`SBP_CAT`).

   a. **(4 marks)** Find the optimal models identified by forward and backward selection algorithms. Report the predictors included in these optimal models. If these models are different, highlight how they differ, and explain why forward and backward selection algorithms may not arrive at the same optimal model.

   b. **(5 marks)** Find the optimal models identified by best subset selection using AIC and BIC as selection criteria. Report the predictors included in these optimal models. If these models are different, highlight how they differ, and explain why the criteria of AIC and BIC may lead to different "best" models. If these models differ from those identified as "best" by forward and backward selection, explain why that may be the case.

   c. **(6 marks)** Although it would be most appropriate to consider all possible combinations of the 15 predictor variables for a cross-validation routine to select a model based on maximising the accuracy or maximising area under the receiver operating characteristic curve (AUC), it is not feasible to do so on home computers in a reasonable amount of time. Consequently, use the predictors identified by best subset selection according to the criterion of minimising AIC from part (b). (If unable to perform the required subset selection in part (b), make note of that here and use the predictors in the optimal model identified by backward selection in part (a).) For this set of predictors, use 20 repetitions of 10-fold cross-validation to identify the optimal model(s) identified according to the criteria of

      i. maximising accuracy and
      ii. maximising AUC.

      If the optimal model(s) identified according to these criteria are different, highlight how they differ, and explain why the criteria of maximising accuracy and maximising AUC may lead to different "best" models. If these models differ from those identified as "best" in parts (a) and (b), explain why this may be the case.

Assignment total: 50 marks

References

Boston University and the National Heart, Lung, & Blood Institute. 2020. "The Framingham Heart Study." https://framinghamheartstudy.org/.

Madell, R., and K. Cherney. 2018. "Blood Pressure Readings Explained." *Healthline*. https://www.healthline.com/health/high-blood-pressure-hypertension/blood-pressure-reading-explained.

Nichols, G. A., T. J. Bell, K. L. Pedula, and M. O'Keeffe-Rosetti. 2010. "Medical Care Costs Among Patients with Established Cardiovascular Disease." *The American Journal of Managed Care* 16 (3): e86–93.

World Health Organization. 2018. "The Top 10 Causes of Death." https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

Wu, C.-Y., H.-Y. Hu, Y.-J. Chou, N. Huang, Y.-C. Chou, and C.-P. Li. 2015. "High Blood Pressure and All-Cause and Cardiovascular Disease Mortalities in Community-Dwelling Older Adults." *Medicine* 94 (47): e2160. https://doi.org/10.1097/MD.0000000000002160.