DPH112 Biostatistics 2
Coursework 2

Elmer V Villanueva
Elmer.Villanueva@xjtlu.edu.cn

1 Background

You will make use of the Excel file called *HINTS 4 Cycle 4 Data.csv*. The file contains data from the Health Information National Trends Survey (HINTS) conducted by the US National Cancer Institute. The survey collects nationally representative data on the American public's use of cancer-related information. The survey has been conducted nine times since 2003. The data you will be analysing comes from the 2014 iteration. Codebooks and methodology papers may be found on the HINTS website, which you may access as you wish.

You will be exploring the relationship between a female respondent's knowledge about the existence of a cervical cancer vaccine and her satisfaction with the quality of their health care in the past year, as well as specific covariates. In particular, you will be assessing the following variables:

- `heardhpvvaccine2` - the main dependent variable. Respondents were asked the following question: " A vaccine to prevent HPV infection is available and is called the HPV shot, cervical cancer vaccine, GARDASIL, or Cervarix. Before today, have you ever heard of the cervical cancer vaccine or HPV shot?"
- `qualitycare` - the main independent variable of interest. Respondents were asked, "Overall, how would you rate the quality of health care you received in the past 12 months?"
- `whenpaptest` - the length of time since the respondent's last Pap test.
- `age` - the respondent's age
- `speakenglish` - how well the respondent speaks English. You must dichotomise this variable as follows: a "confident" speaker has `speakenglish==1` or `speakenglish==2`, while a "non-confident" speaker has `speakenglish==3` or `speakenglish==4`.

You must perform a *logistic regression* and report your findings. The data analysis must be conducted in R. You must clean the data set, ensuring that only

valid data are analysed. You must perform the analysis and produce regression equations and suitable graphical summaries. You must also test the assumptions of the model. Finally, you must present your R commands in an RMD file. In a written report, you must describe your methods and findings.

2 Instructions

Prepare a written report of no more than 250 words, at least one and no more than five figures and at least one and no more than three tables. The report must show that you are able to

1. describe the methods you've employed (10 marks);
2. summarise the results of the regression analysis (30 marks);
3. summarise the results of the diagnostic tests (10 marks);
4. interpret the findings in the context of the research question (10 marks); and
5. use technical English appropriately (10 marks)

The R file must show that you are able to

1. import data from Excel (5 points);
2. identify and exclude invalid data from analysis (5 points);
3. perform a logistic regression (5 points);
4. produce a graphical summary of the data (5 points); and
5. perform tests of the validity of the model (10 points)

3 Submission

You must submit *two* files.

- File 1 is the research report in PDF format. The file must be named $<$ StudentID $>$ *.pdf*. For example, 123456789.pdf.
- File 2 is the RMD file with your R commands. The file must be named $<$ StudentID $>$ *.Rmd*. For example, 123456789.Rmd.

You must submit the files via ICE. All penalties for late or incomplete submissions apply.

4 Marking Rubric

Criterion	Level 1	Level 2	Level 3	Level 4
Report: Methods	Not done (0 marks)	Major mistakes or omissions (3 marks)	Minor mistakes or omissions (6 marks)	No mistakes or omissions (10 marks)
Report: Results - Regression equation	Not done (0 marks)	Major mistakes or omissions (3 marks)	Minor mistakes or omissions (6 marks)	No mistakes or omissions (10 marks)
Report: Results - Confidence intervals and p-values	Not done (0 marks)	Major mistakes or omissions (3 marks)	Minor mistakes or omissions (6 marks)	No mistakes or omissions (10 marks)
Report: Results - Effect of covariates	Not done (0 marks)	Major mistakes or omissions (3 marks)	Minor mistakes or omissions (6 marks)	No mistakes or omissions (10 marks)
Report: Diagnostic tests	Not done (0 marks)	Major mistakes or omissions (3 marks)	Minor mistakes or omissions (6 marks)	No mistakes or omissions (10 marks)
Report: Interpretation	Not done (0 marks)	Major mistakes or omissions (3 marks)	Minor mistakes or omissions (6 marks)	No mistakes or omissions (10 marks)
Report: Technical English	Major mistakes (0 marks)	Minor mistakes (2 marks)	Appropriate (10 marks)	
R: Data import	Not done (0 marks)	Done (5 marks)		
R: Invalid data	Not done (0 marks)	Done (5 marks)		
R: Regression	Not done (0 marks)	Inappropriate (2 marks)	Appropriate (5 marks)	
R: Graphs	Not done (0 marks)	Inappropriate (2 marks)	Appropriate (5 marks)	
R: Model validity	Not done (0 marks)	Inappropriate (4 marks)	Appropriate (10 marks)	