Generalized Linear Models (2462): Assignment 2023

Instructions for Submission

You must hand in your assignment by **12 noon** on **Wednesday 8th February 2023**. You should submit your assignment electronically via the submission point on the GLM page on Moodle.

The report should be **no more than four pages long** including tables and figures. Reports longer than four pages will be penalized by one grade. The font size should not be smaller than Times New Roman 11 point (you do not have to use Times New Roman), with margins set as Normal.

The report should be accompanied by a suitably commented Stata "Do" file that produces all of the results that you refer to in your report. Please do not submit an R file.

Background

The data come from a cross-sectional study conducted to investigate the aetiology of carotid plaque in a particular region of a European country. Carotid plaque is a risk factor for coronary heart disease measured using carotid ultrasonography. The measurement technique first simply classifies plaque as present or absent, although refinements can also allow calculation of an integer valued plaque count. In this study only a randomly selected 30% of participants have plaque counts.

As well as carotid plaque, participants in the study reported numerous demographic and socio-economic factors. Potential risk factors for coronary heart disease were also measured. Here you are provided with data on age, sex, education (a proxy measure of social class), smoking and waist-hip ratio as well as plaque counts.

The focus in this assignment is on the causal effects of smoking and waist-hip ratio on plaque. Previous studies have suggested that the effect of waist-hip ratio may differ in males and females, so this is one issue that should be explored.

Data

The file **assign2023.dta** on Moodle holds the following variables on individual participants.

Variable	*Description and coding*
id	Subject identifier
age_cat	Age category (1 = 40-44, 2 = 45-49, 3 = 50-54, 4 = 55-59, 5 = 60-64, 6 = 65-69 years)
sex	Sex (0 = female, 1 = male).
education	Education (0 = Secondary education or less, 1 = University graduate)
smoking	Are you a current smoker? (1 = Never smoked, 2 = No, ex-smoker 3 = Yes, a regular smoker)
whr	Waist-hip ratio
plaque	Presence/absence of carotid plaque (0 = absent, 1 = present)
plaque_count	Carotid plaque count (only available in a subset)

Aims

The aims of your analysis are as follows. The tasks you should carry out to achieve these aims are detailed in the next section.

- To quantify the effects of smoking on the presence of plaque, by fitting a statistical model that makes appropriate adjustment for relevant confounding variables.
- To estimate the prevalence of plaque if no one in the study had ever smoked, and compare this with the observed prevalence.
- To quantify the effect of waist-hip ratio on the presence of plaque, by fitting a statistical model that makes appropriate adjustment for relevant confounding variables.
- To explore the possibility that the effects of waist-hip ratio on the presence of plaque may differ in males and females.
- To estimate the prevalence of plaque if World Health Organisation (WHO) guidelines on obesity were followed. The WHO classifies men with a waist-hip ratio above 0.9, and women with a waist-hip ratio above 0.85, as obese.
- To quantify the effect of smoking on plaque count, by fitting a statistical model that makes appropriate adjustment for relevant confounding variables.
- To quantify the effect of waist-hip ratio on plaque count, by fitting a statistical model that makes appropriate adjustment for relevant confounding variables.
- To explore the possibility that the effects of waist-hip ratio on plaque count may differ in males and females.

Statistical Analysis

Use Stata to help carry out the following tasks. You should bear in mind that for all of these analyses there is no single definitively correct approach. However, you should try to ensure that your approach always matches the aims.

All of the analysis aims can be addressed using techniques taught in the GLM and other Medical Statistics MSc. modules. Although you can use other techniques should you wish to, you are discouraged from doing so. Devoting substantial time and space in the report to alternatives could be a distraction from the main aims.

1) Preliminaries

Carry out any preliminary descriptive analysis of the data that you judge necessary.

2) Relating smoking to the presence of plaque (binary)

a) As explained in lecture notes, if all predictor variables are categorical there are advantages in analysing data as grouped binary, so first convert the data to this format. The following series of Stata commands can be used to do this.

```
gen plaquecopy=plaque
collapse (count) plaquecopy (sum) plaque, by(sex age_cat smoking education)
rename plaquecopy n
```

b) Using the grouped data created in 2a), fit a logistic regression model relating the presence of plaque to smoking status adjusting for age, sex and education (since these are potential confounders) in an appropriate fashion. Note that you should not adjust for waist-hip ratio, since it could plausibly be on a causal path between smoking and plaque. Note also that, here and elsewhere, you do not need to use variable selection techniques in order to determine which variables should be included in the model. Assess the fit of the model using an appropriate statistical test. Interpret your findings.

c) Reverting to the individual participant data, fit the same model as in 2b) and use this to compute estimates of the causal effect of smoking status on plaque expressed as marginal odds ratios comparing i) ex-smokers and ii) regular smokers with never smokers. Also estimate what the marginal prevalence of plaque would be if no one in the study had ever smoked. Compare this with the observed prevalence. Report and comment on your findings.

3) Relating waist-hip ratio to the presence of plaque (binary)

a) Add waist-hip ratio as an additional predictor variable to your model in 2c) in order to estimate the causal effect of waist-hip ratio on the presence of plaque controlling for potential confounding by age, sex, education and smoking. Since a unit difference in waist-hip ratio is large any results that you report for waist-hip ratio should relate to 0.1 unit differences. Assess the fit of the model using any statistical techniques that you judge appropriate, making any modifications to the model if you judge these necessary in order to achieve the aim of the analysis. Interpret your findings.

b) Modify your model in 3a) to allow the effect of waist-hip ratio to vary by gender. Interpret your findings.

c) Based on your findings from 3a) and 3b), use an appropriate model of your choice to estimate what the prevalence of plaque would be if all males with a waist-hip ratio above 0.90 reduced their waist-hip ratio to 0.90 and all females with a waist-hip ratio above 0.85 reduced their waist-hip ratio to 0.85.

4) Relating smoking and waist-hip ratio to plaque count

a) Use an appropriate statistical technique to model the relationship between plaque count and smoking status adjusting for age, sex and education. Interpret your findings. For the purposes of this assignment you need not explore the fit of your model, but you should justify your choice of model.

b) Use an appropriate statistical technique to model the relationship between plaque count and waist-hip ratio adjusting for smoking status, age, sex and education. Interpret your findings. As above, it is not necessary for you to explore the fit of this model.

c) Modify the model in 4b) to allow the effect of waist-hip ratio to vary by gender. Interpret your findings. As above, it is not necessary for you to explore the fit of this model.

Report Specification

a) The report should be in five sections, one relating to each of the above four parts and the final one a 'Discussion and Conclusions' section. Each of sections 2, 3 & 4 should start by describing the statistical methods used in that section, then present the key findings using text, tables and figures as you judge necessary. In the final section you should summarise your findings, consider the implications of the results, and discuss any limitations of this study and your analysis. There is no need to include any review of any literature.

b) The report should be accompanied by a commented Stata "Do" file that produces all of the results in your report. "Do" files that do not run, or which do not produce results given in the report, will be penalised.

c) The following table gives the percentage of the total marks that will be allocated to each part of the submission.

Section	%
1) Preliminaries	8
2) Relating smoking to the presence of plaque (binary)	20
3) Relating waist-hip ratio to the presence of plaque (binary)	20
4) Relating smoking and waist-hip ratio to plaque count	30
5) Discussion and Conclusions	14
6) Stata "Do" file	8
Total	100

d) You do not need to describe in detail all of the statistical procedures that you carry out. Rather you should decide which results you consider most important in addressing the aims of the assignment, and present these clearly in tables and/or figures, with accompanying text. Other results and analyses can be referred to in less detail, although you should state clearly what you have done.

e) There is no need to define or explain the general meaning of statistical terms such as '*p*-value'.

f) Do not include any non-graphical Stata output, *i.e.* results copied directly from the Stata screen.