16.8 Practical 16

Dataset required: `nhanesglm.dta`

Introduction

In this practical we will again use the NHANES data introduced in previous sessions, but here we will analyse a categorised version of systolic blood pressure (SBP) as the outcome variable, using an ordinal logistic regression model.

For the purposes of this session, we will ignore diastolic blood pressure and categorise blood pressure according to SBP as follows:

Category	SBP
Normal	< 120
Prehypertension	$120 - 139$
High (stage 1)	$140 - 159$
High (stage 2)	≥ 160

Given the above classification system, it is of interest to model how the probability of falling into the four categories depends on age and gender.

Aims

- Interpret coefficients from ordinal logistic regression
- Understand the differences between multinomial and ordinal regression
- Know how to assess the fit of an ordinal regression model

Analysis

1 Use the `egen` command (or otherwise) to create a new variable for SBP, categorised as above. Tabulate the new variable to see its distribution.

What is the mean age in each blood pressure group?

2 We will fit an ordinal logistic regression model to the categorised SBP variable with age as the only covariate. But to help the interpretation of the intercept estimate, we will first centre the age variable at 50 years.

```
gen cen_age = ageyrs - 50

ologit sbpcat cen_age
```

Discuss: Carefully describe the interpretation of each of the estimated parameters.

3 (a) From your fitted model, calculate the estimated probability of having SBP in the highest (High (stage 2)) category for an individual aged 70 years.

(b) From your fitted model, calculate the estimated probability of having SBP in the second highest category (High (stage 1)) for an individual aged 70 years.

(c) Use the `margins` command in Stata to obtain these estimates.

4 Use the `predict` command to generate predicted probabilities of subjects being in each of the four SBP categories, and plot the probabilities against age. Hint: in order to see the lines sort the data by age before plotting the probabilities.

5 To assess whether the proportional odds assumption is reasonable here, fit the corresponding `mlogit` model. Compare the log-likelihoods, AIC values, and fitted probabilities from the two models.

6 As a final check of the model fit, we can compare the coefficients to those from a series of logistic regression models with different cut-offs. First create three new variables, to dichotomise the SBP variable in three different ways:

(a) Normal vs All other categories

(b) Normal or Prehypertension vs Hypertension (Stage 1 or 2)

(c) All other categories vs Stage 2 Hypertension

Fit a standard logistic regression model for each of these new variables as the outcome, and centred age as the only covariate. Compare the age coefficient from each model to the equivalent from the ordinal logistic regression model.

Discuss: Considering all of the above, which model do you prefer for these data?

7 Fit an ordinal logistic regression model and perform an appropriate test to assess if there is evidence that gender is associated with SBP category, after adjusting for age. What is the estimated effect of gender?

8 Use one or more of the approaches in parts 5 and 6 to assess whether the ordinal logistic regression model with age and gender provides a good fit to these data.

Discuss: Which model do you now prefer for these data?

9 (optional) Prove that the latent variable model defined in Section 16.5 of the notes implies the ordinal logistic regression model given by equation (16.1).