9.8 Practical 9

Dataset required: `rubber.dta`

Introduction

The purpose of this exercise is to fit and interpret some Poisson regression models for data from a cohort study of men from the rubber manufacturing industry.

In one rubber factory isolated conditions were maintained for the manufacturing process, while in a second workers were exposed to more dirt and fumes. The aim is to investigate differences in death rates between men who worked in the two factories, making appropriate allowances for age.

The dataset (rubber.dta) consists of four variables for groups defined by factory and age category:

Variable	Description
`agegrp`	Age group: 1=50-59, 2=60-69, 3=70-79, 4=80-89
`factory`	Factory (1 or 2)
`deaths`	Number of deaths
`pyrs`	Number of man-years of exposure

Aims

The aim of this session is to address the following questions about these data:

1 Is there a difference between deaths rates in the two factories?

2 Is death rate associated with age?

3 What is the best way to model age (categorical or continuous)?

Analysis

1 Calculate the total number of deaths and total person years of observation. What is the overall rate of death amongst these men? What is the log death rate?

2 Calculate the death rate and log death rate in each factory.

3 In Stata, generate a variable for the log death rates in each row, and plot these against age group, labelled according to factory (hint: use the option `mlabel([varname])` with the `twoway` command). What does this plot suggest?

Discuss: What are your initial conclusions from these descriptions of the data?

4 Write down algebraically a model that can be used to investigate the relationship between death rate and age group (treated as a categorical variable, with age group 50-59 taken as the baseline category).

Discuss: Take a moment to check with your colleagues that you have specified the models in the same way.

5 Using Stata:

(a) obtain the maximum likelihood estimates of the parameters in your model. How strong is the evidence for an age effect?

(b) calculate the estimated rate ratios comparing age groups:

i. 60-69 with 50-59

ii. 80-89 with 70-79

Calculate 95% CI's for these estimates, using `lincom` (or by re-parameterising the model) for ii.

6 Now include both age group and factory effects in an additive linear predictor. What is the evidence for a difference in death rates between the factories (adjusted for age group)? Calculate the estimate (and 95% CI) of the rate ratio comparing the two factories, adjusted for age group.

7 Now fit a model that includes the interaction between age group and factory.

(a) What term can be used to describe this model (Hint: look at its deviance)?

(b) What is the evidence for an interaction? How do you interpret this for these data?

(c) Write down algebraically the linear predictor for this model, and use this to help you calculate the estimated rates of death in

i. factory 1, age 70-79

ii. factory 2, age 50-59

iii. factory 2, age 60-69

(d) Check your answers against your original data

Discuss: Take a moment to check with your colleagues that you have specified the models in the same way.

8 Now consider the age group variable as continuous. Fit a model with age group and factory as covariates and consider its deviance.

Discuss: What do you conclude about the overall fit of the model with continuous age and factory as additive effects? How might the fit be improved?

9 Now consider your analysis as a whole.

Discuss: What are your epidemiological conclusions from this analysis? Working together with one or more colleagues (in your Breakout Room if online), write a short paragraph to summarise your findings concerning the comparison of the two factories. If online, one of you should post your group's paragraph in the Zoom chat.