## 14.9 Practical 14

Datasets required: `vitE.dta` and `infertility.dta`

### Introduction

There are two parts to this session.

In the first part we re-visit the vitamin E dataset from the previous practical and apply the conditional logistic regression methods we learned in lecture 13.

In the second part we use a new dataset and investigate the association between infertility in women and number of previous spontaneous and induced abortions.

### Aims

- Learn how to fit a conditional logistic regression model to matched data (in Stata).
- Learn how to fit a standard logistic regression model to matched data by including the matching variables (in Stata).
- Understand the different assumptions underpinning the two models described above

### Part A. Vitamin E dataset

Reload the vitamin E dataset and generate a binary variable to indicate high vitamin E levels (above 12mg/dl).

As a reminder, in the previous session we found that the estimated conditional odds ratio for the association between the binary vitamin E variable and a person's cancer status (case or control) was 0.76, indicative of a protective effect of high vitamin E levels.

1 We will first demonstrate an invalid method of analysis which might seem superficially attractive.

Fit a logistic regression model with case as the dependent variable, and the binary vitamin E indicator variable and the set variable as categorical explanatory variables.

```
logistic case h_vitE i.set
```

Confirm that this does not give the same (correct) estimate of the odds ratio as we calculated in the previous session (0.76).

This situation contrasts with the linear regression we used in the previous practical session, where adjusting for the matched set produced a valid estimate of the mean difference in vitamin E levels between cases and controls. This type of analysis is not valid when the dependent variable is binary.

2 Fit a conditional logistic regression analysis as below.

```
clogit case h_vitE, group(set)
```

3 We could also use the original, continuous vitamin E variable as an explanatory variable. However, as the dependent variable is still binary, we must again use conditional logistic regression.

```
clogit case vitE, group(set)
```

4 Suppose now that we wish to explore whether the effect of dichotomised vitamin E on the conditional odds of cancer varies with observation time (time from blood collection to diagnosis of cancer in the case). Carry out an appropriate analysis to explore this.

**Discuss: What do you conclude from this analysis?**

## Part B. Infertility dataset

The data for Part B come from a matched case-control study of the association between infertility in women and previous spontaneous or induced abortion. This is an example of a matched case-control study with two controls per case. The dataset is called infertility_data.dta. The main research question of interest is whether previous abortions (induced or spontaneous) affect risk of infertility.

Further details of the study can be found in the original paper by Tzonou et al, which is available online: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1059707`

Each case (a woman diagnosed as infertile) was matched to two controls on age, parity and education level. The variables shown in the data are as follows:

| Variable | Description |
|---|---|
| `case` | case/control indicator (0=control, 1=case) |
| `stratum` | index for matched set, numbered 1–271 |
| `spontaneous` | Number of prior spontaneous abortions [0 (zero), 1 (one), 2 (2 or more)] |
| `induced` | Number of prior induced abortions [0 (zero), 1 (one), 2 (2 or more)] |
| `age` | age in years |
| `parity` | number of children |
| `education` | years of education [1 (0-5 years), 2 (6-11 years), 3 (12+ years)] |

5 Open the dataset and familiarise yourself with the data and its structure.

(a) Do rows contain information on individual people or matched sets?

(b) How many cases, controls, and matched sets are included in the dataset?

(c) Are all of the sets the same size?

6 We will examine the effect of spontaneous abortion on infertility using conditional logistic regression. Use `clogit` to perform a matched analysis using the variable `spontaneous` as a categorical explanatory variable.

What do you conclude?

7 As discussed in the lecture, an alternative analysis for a matched case-control study with ‘well-defined’ matching variables is to perform an unmatched analysis with regression adjustment for the matching variables. Fit the following three models using `glm` or `logit`.

(a) A model that only includes the matching variables (age, parity and education level),

(b) A model that only includes the exposures of interest, the categorised count of spontaneous abortions, as a categorical explanatory variable,

(c) A model that includes the exposure of interest and the matching variables.

**Discuss: Compare the estimated coefficients for the matching variables in the models in parts (a) and (c). What do you conclude?**

**Also compare the estimated coefficients relating to the categorised count of spontaneous abortions in (b) and (c) with those from the matched analysis using conditional logistic regression. Which approaches give valid estimates of the odds ratios?**