13.11 Practical 13

Dataset required: `vit_E.dta`

Introduction

The aim of this practical is to consolidate the ideas of matching discussed in the lecture. We will use data from a matched case-control study of the association between vitamin E measured in the blood and occurrence of cancer. The dataset is called vit_E.dta

High levels of vitamin E are thought by some to be protective against cancer. This hypothesis was investigated by measuring Vitamin E in stored blood samples from 271 men in a large cohort study initially aged 35–64 years who subsequently developed cancer. These values were compared with those from control men who had not, at that time, developed cancer. One control for each case was selected randomly from within the cohort study, subject to matching for age (within 5 years), duration of storage of the blood sample (within 3 months) and smoking status.

Note that this dataset does not include the matching variables. It does include 'observation time', which is the time from blood collection to diagnosis of cancer in the cases. In this session we will analyse the data assuming no confounding.

Variable	Description
`set`	Case-control unique identifier taking values from 1 to 271
`case`	0: control 1: case
`vitE`	Vitamin E (mg/dl)
`time`	Observation time: 1: up to 1 year 2: 1-3 years 3: over 3 years

Aims

- Understand how to appropriately conduct hypothesis tests on data from a matched case-control study.
- Use appropriate statistical tests to analyse matched studies with continuous dependent variables.

Analysis

1 Start by looking at the form of the data. It will help to sort the data so that the data from the same set (block) are shown on consecutive rows. Use the code below to list the first five sets, with the control appearing first within each pair.

```
sort set case
list in 1/10, sep(2)
```

2 (a) Summarize the vitamin E measurements for cases and controls. What is the difference between the means in the two groups?

Hint: use the `bysort` prefix to `summarize` to get the summaries within cases and controls automatically.

(b) Use suitable plots to display vitamin E levels in the cases and controls.

For questions 3-5 we will analyse these data treating vitamin E as the dependent variable, and cancer as the independent variable. In reality the vitamin E reading is best thought of as an exposure since blood was taken and stored before the onset of cancer.

3 We will first analyse these data using linear regression.

(a) We first demonstrate an invalid approach, to demonstrate the importance of allowing for block in any analysis. Fit a linear regression model with just the case-control variable as an explanatory variable.

```
regress vitE i.case
```

What is the coefficient for the case-control variable? What is the reported standard error?

(b) Now add the set variable, as a categorical predictor in the regression model.

```
regress vitE i.case i.set
```

What is the coefficient for the case-control variable? What is the reported standard error?

Discuss: Compare the estimates of the case-control coefficient and its standard error. Is there evidence that people who developed cancer had different vitamin E levels than people who did not develop cancer?

4 In order to make the data suitable for analysis with a paired t-test (and other analyses), we need to reconfigure it so that the data from each pair is contained in one row. This can be done using the reshape command in Stata, which is an extremely useful command, but can be fiddly to use in practice.

Look at your data in the browser, then use the following command to reshape it.

```
reshape wide vitE, i(set) j(case)
```

Now look at the data again in the browser to see how it is now arranged.

5 With the data in wide format, carry out a paired t-test of the null hypothesis that the mean blood vitamin E levels are the same in cases and controls.

Discuss: How do the estimated mean difference and its standard error compare to those from the linear regression models above?

For the remainder of this session we will dichotomise vitamin E into a binary variable and perform analysis to estimate the odds ratio of association. Here we can use the reversible nature of an odds ratios so that we can interpret vitamin E as the exposure and cancer as the outcome or vise-versa.

6 (a) Define a "high" vitamin E status as a level above 12mg/dl and complete the table below. The table records counts of sets, that is pairs of people, one of whom developed cancer and the other who did not.

		Controls		
	Vitamin E status	Low	High	Total
Cases	Low			
	High			
	Total			

[Hint: create two new variables for vitamin E status in cases and controls respectively, which take the value 1 for individuals with high vitamin E and the value 0 for individuals with low vitamin E. In the Stata do file these variables are h_vitE_control and h_vitE_case.]

(b) Using the completed table, compute the test statistic for McNemar's test. What is the null hypothesis?

(c) Use the `bitesti` command to perform a test of this null hypothesis.

(d) Again using the completed table, calculate the estimated odds ratio by hand.

(e) Use the `cii` command to calculate an exact 95% confidence interval for this estimate.

7 Use the `mcc` command to analyse this matched case-control study using dichotomised vitamin E as the exposure.

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