30

A STRUCTURED APPROACH TO DATA ANALYSIS

Objectives

After reading this chapter, you should be able to:

1. Conduct a detailed analysis of a complex dataset arising from an epidemiologic study with a minimum of wasted time and a maximum probability of avoiding serious errors in the analysis.

2. Congratulate yourself on getting through all of the material in this text, provided you didn't skip directly to this final 'substantive' chapter.

30.1 Introduction

When starting into the analysis of a complex dataset, it is very helpful to have a structured approach in mind. In this chapter, we provide one template which, we trust, will be applicable in most situations. Others with experience in data analysis might have different approaches and we would not suggest that what is presented below is either the 'only' approach, or necessarily the 'best' one—as with models, every approach is imperfect but some are useful. However, for individuals getting started in epidemiology, the following will serve as a template which can be used to guide their initial efforts in data analysis.

For most, there is a strong tendency to want to jump straight into the sophisticated analysis which will provide the ultimate answer for your study. This rarely works out, but in order to satisfy your curiosity ... go ahead and try it anyway. Just don't waste more than an hour on it and ignore whatever results you get, as they will inevitably be wrong. Having thus satisfied that primal urge to take the short cut to the end, you can proceed with a structured approach to the analysis.

We will work through the process in a logical sequence, starting with the handling of data-collection sheets and ending with keeping track of results. However, bear in mind that data analysis is an iterative process which often requires that you back up several steps as you gain more insight into your data.

However, before you start any work with your data, it is essential that you construct a plausible **causal diagram** of the problem you are about to investigate. This will help identify which variables are important outcomes and predictors, which ones are potential confounders and which might be intervening variables between your main predictors and outcomes. Keep this causal diagram in mind throughout the entire data-analysis process. **Note** With large datasets, it will not be possible to include all predictors as separate entities. This can be handled by including blocks of variables (*eg* demographic characteristics) in the diagram instead of listing each variable.

30.2 Data-collection sheets

It is important to establish a permanent storage system for all original data-collection sheets (survey forms, data-collection forms *etc*) that makes it easy to retrieve individual sheets if they are needed during the analysis. If individuals (or groups) in the study have identification numbers, this makes a convenient way to store (and later retrieve) individual files. Some things to consider when dealing with the file follow.

- Do not remove originals from this file. If you need to take a specific sheet for use at another location, make a photocopy of it.
- Never ship the original to another location without first making copies of all forms. (You don't want to lose your whole study because the post office or courier loses your package.)
- Set up a system for recording the insertion of data-collection sheets into the file so that you know how many remain to be collected before further work begins.
- Once all of the forms have been collected—before you do anything else—scan through all sheets to get an impression for their completeness. If there are omissions in the data-collection sheet (*ie* forgetting to complete the last page of a questionnaire), returning to the data source to complete these data will more likely be successful if it is done soon

after data were initially collected rather than weeks or months later (after data analysis has begun).

30.3 Data coding

Some issues related to data coding have already been discussed in Chapter 3, in particular, the advisability of having a space to allow for coding directly on the data-collection sheet. Some other issues to consider when coding your data are as follows.

- As noted in Chapter 3, assign a specific number to all missing values. Be sure that this specific number is not a legitimate value for any of your responses. Some statistical packages allow for multiple types of missing values so you may want to start off with different codes for different types (*eg* -999=no data provided, -998=invalid entry, -997=not applicable).
- If you have 'open' questions, scan the responses and develop a list of needed codes before starting coding.
- Maintain a master list of all codes assigned.
- Use numeric codes. In general, avoid the use of string variables except for rare instances where you need to capture some textual information (*eg* a comment field).
- Only code one piece of information in a single variable. Never make compound codes! For example, if you have recorded both the sex and race of participants in a study, it might be tempting to code them as 1=male, Caucasian, 2=female, Caucasian, 3=male, Hispanic *etc*. Do not do this. Create separate variables for sex and race.
- For all types of data, note any obvious outlier responses (*eg* an individual's body mass index reported as 127) and correct them on the data sheet.
- Use a different coloured pen so your coding notations can clearly be differentiated from anything previously recorded on the data-collection sheets.

30.4 Data entry

Some of the issues to consider when entering your data into a computer file are as follows.

- Double-data entry, followed by comparison of the 2 files to detect any inconsistencies, is preferable to single-data entry.
- Spreadsheets are a convenient tool for initial data entry, but these must be used with extreme caution; because it is possible to sort individual columns, it is possible to destroy your entire dataset with one inappropriate 'sort' command.
- Custom data-entry software programs provide a greater margin of safety and allow you to do more data verification at the time of entry. One such program in the public domain is EpiData (http://www.epidata.dk/).
- If you expect large quantities of multilevel data (*eg* multiple blood pressure measurements on many individuals in many centres), using hierarchical database software can make data entry and retrieval more efficient. Alternatively, you can set up separate files for data at each level (*eg* a centre file, an individual file *etc* and merge the files after data entry).
- As soon as the data-entry process has been completed, save the original data files in a safe location. In large, expensive trials it might be best to have a copy of all originals stored in another location.
- If the data-entry program which you use does not have the ability to save your data in the

format of the statistical package that you are going to use, there are a number of commercially available software programs geared specifically to convert data from one format to another.

- If you use a general purpose program (*eg* spreadsheet) to enter your data, as soon as the data are entered, convert them to files usable by the statistical program that you are going to use for the analysis. Do all of the analyses in that statistical program (*ie* don't start doing basic statistics in the spreadsheet). You are going to need the statistical program eventually, and it will be a lot easier to keep track of all of your analyses if they are all done there. This will also simplify the process of tracking modifications to the data.

30.5 Keeping track of files

It is important that you have a system for keeping track of all your files. Some suggestions that will help you do this are:

- Assign a logical name with a 2-digit numerical suffix (*eg* brazil01). Having a 2-digit suffix allows you to have 99 versions which still sort correctly when listed alphabetically.
- When data manipulations are carried out, save the file with a new name (*ie* the next available number). Do not change data and then overwrite the file.
- Keep a simple log of files created (Table 30.1) with some very brief information about the contents of the file (*eg* number of observations and variables).

Table 30.1 Example of data of files created in blood pressure study

File name	Date created	Description	# Obs.	# Vars.
bp01.odc	27/09/07	original blood pressure study data; spreadsheet; 1 record per measurement	1092	8
bp01.dta	28/09/07	original file; Stata format	1092	8
bp02.dta	30/09/07	45 records with missing values dropped	1047	8
etc...				

30.6 Keeping track of variables

You are often faced with keeping track of a bewildering array of variables in a dataset from an epidemiologic study. We are not advocating studies with huge numbers of predictors (in fact, we discourage such studies), but even a relatively focused study can give rise to a large number of variables once transformed variables, and/or recoded categorical variables, have been created. To help keep track of these variables, we recommend the following.

- Use short (but informative) names for variables and have all related variables start with the same name. For example, the following might be a logical set of variable names for information relating to age.
- age = the original data (in years)
- age_ct = age after centring by subtraction of the mean
- age_ctsq = quadratic term (age_ct squared)
- age_c2 = age categorised into 2 categories (young vs old)

- age_c3 = age categorised into 3 categories.
- Long names can often be shortened, but kept recognisable, by removing vowels (*eg* wtr_cstrn as a short form for water cistern).
- In some cases, adding a single letter prefix might help keep groups of variables together. For example, a series of bacteriological results might be named b_ecoli, b_cjejuni, b_salm *etc.*
- If the statistics program you use is 'case sensitive' (*ie* differentiates between 'd' and 'D'), use ONLY lower-case letters.
- At some point you will want to prepare a master list of all variables with some very basic information. It should be possible to have the statistical program prepare this listing (or one similar to it).

30.7 Program mode versus interactive processing

Some statistical programs can be used in an interactive mode where individual functions are carried out by either selecting items from menus or typing in a command. While very useful for exploring your data and trying out analyses, this interactive mode should not be used for any of the 'real' processing and/or analysis of your data because it is very difficult to keep a clear record of steps taken when using programs in this manner. Consequently, it is difficult, or impossible, to reconstruct the analyses you have completed.

The alternative is to use the program in 'program mode' in which you compile the commands necessary to carry out a series of processing steps or analyses into a program and then run the program. These program files can be saved (again, a logical naming convention is required) and used to reconstruct any analyses you have carried out. Nearly all of the programs used in the analyses presented in the examples in this text were carried out using these types of program.

Several things to keep in mind when writing these program files are as follows.

- Name the files logically so you can easily find the correct file if you want to rerun your analyses (*eg* bw5k_dataprep, bw5k_uncond, bw5k_logit for a series of programs for preparing and analysing the birth-weight dataset).
- Structure the program to make it easy to follow.
  - Many programs start with a common block of text that might specify the working directory on your computer, open a log file *etc*.
  - Use sequential indents to keep blocks of commands together (see example below).
- Document the file thoroughly. All programs will allow you to add comments to these program files and these should be used to document:
  - what the program does, and
  - in some cases, record key results in the program file.

Example 30.1 shows a small portion of the program file used to carry out the analyses in Chapter 14.

Example 30.1 Sample of a program file for analyses
data = none

The following is a portion of the program file (-do- file in Stata) used to carry out the analyses in Chapter 14.

```
* MER
* ch14_all.do
* Linear regression - all calculations
*
   version 12
   clear all
   set more off
   set memory 20m
   cd c:\mer\ch14\data            /* this line is computer dependent */
   capture log close
   log using mer14_all, replace text

*******************************************************************************
* Disclaimer
*******************************************************************************
/*
The analyses shown in this program are carried out for providing examples for the text.
   They do not, in any way, represent a complete or appropriate analysis of the dataset
*/

*******************************************************************************
* Open the data
*******************************************************************************

use c:\mer\data\bw5k.dta, clear

*******************************************************************************
* Ex 14.1 - 14.2 Simple linear regression models - birthweight as outcome
*         - gestation length as continuous predictor (limited to <=40 weeks)
*         - multiple births as dichotomous predictor
*******************************************************************************
   * continuous predictor
      summ bwt, d
      regress bwt gest
      * generate prediction SE (of mean and individual)
      predict SE_mean, stdp
      predict SE_ind, stdf
      * br gest SE_mean SE_ind
   * prediction intervals
      twoway (scatter bwt gest, msize(small)) ///
         (lfitci bwt gest, ciplot(rline) blcolor(black) blpattern(dash)) ///
         (lfitci bwt gest, stdf ciplot(rline) blcolor(black) blwidth(medium) ///
            blpattern(shortdash_dot_dot)), ///
         legend(off) scheme(s1mono) ytitle("birth weight (gm)", m(r+2)) ///
         ylabel(, angle(horizontal) format(%4.0f)) ///
         xtitle("length of gestation (weeks)", m(t+2)) ///
         plotregion(style(none))
   graph save "figs/fig14_predict.gph", replace
   graph export "figs/fig14_predict.emf", replace
```

30.8 DATA EDITING

Before beginning any analyses, it is very helpful to spend some time editing your data. The most important components of this process are labelling variables and values within variables, formatting variables and correctly coding missing values.

- All variables should have a label attached to them which more fully describes the contents of the variable. While variable names are often quite short (*eg* <8 or <16

characters), labels can be much longer. **Note** With some computer programs, the labels are stored in a separate file.

- Categorical variables (hopefully they are all numeric) should have meaningful labels attached to each of the categories. For example, sex could be coded as 0 or 1, but should have labels for 'male' and 'female' attached to those values.
- The number(s) that was assigned to all missing values needs to be converted into the code(s) used by your statistics program for missing values.
- Some programs will allow you to attach 'notes' directly to the dataset (or to individual variables within the dataset). These explanatory notes can be invaluable in documenting the contents of files.

30.9 Data verification

Before you start any analyses, you must verify that your data are correct. This can be combined with the following 2 processes (processing your outcome and predictor variables) because both involve going through all of your variables, one-by-one.

If you have a very small dataset, you might want to print the entire dataset (make sure it aligns all values for one variable in one column) and review it for obvious errors. However, this is rarely feasible for datasets from epidemiologic studies.

- For continuous variables:
  - determine the number of valid observations and the number of missing values
  - check the maximum and minimum values (or the 5 smallest and 5 largest) to make sure they are reasonable (if they are not, find the error, correct it and repeat the process)
  - prepare a histogram of the data to get an idea of the distribution and see if it looks reasonable.
- For categorical variables:
  - determine the number of valid observations and the number of missing values
  - obtain a frequency distribution to see if the counts in each category look reasonable (and to make sure there are no unexpected categories).

30.10 Data processing—outcome variable(s)

While you are going through the data verification process, you can also start processing your outcome variable(s). To do this you will need to review the stated goals of the study to determine the format(s) of the outcome variable(s) which best suits the goal(s) of the study. For example, you might have conducted a clinical trial of a new vaccine for the control of norovirus gastroenteritis in nursing homes and have recorded daily incidence of gastroenteritis for the duration of the study. From this single mortality variable, you could compute the mean daily incidence rate during outbreak periods, the cumulative mortality over the study, the peak incidence rate observed, whether or not the nursing home met some set of criteria for having an 'outbreak' of norovirus, or the length of time to the onset of an outbreak. Which you choose to analyse will depend on the goals of the study. Once you have identified the appropriate outcome variable(s), consider the following.

- If the outcome is categorical, is the distribution of outcomes across categories acceptable? For example, you might have planned to carry out a multinomial regression

of a 3-category outcome, but if there are very few observations in 1 of the 3 categories, you might want to recode it to a 2-category variable.

- If the outcome is continuous, does it have the characteristics necessary to support the analysis planned?
- If linear regression is planned, is it distributed approximately normally? If not, explore transformations which might normalise the distribution. **Note** It is the normality of the residuals which is ultimately important, but if the original variable is far from normal, and there are no strong predictors, it is unlikely that the residuals will be normally distributed.
- If it is a rate (or count) and Poisson regression is planned, are the mean and variance of the distribution approximately equal? If not, consider negative binomial regression or alternative analytic approaches. (As above, the assumption of equality of the mean and variance applies to the residuals, but should be approximately true in the original data, unless there are one or more strong predictors, if this is to be the case.)
- If it is time-to-event data, what proportion of the observations are censored? You might also want to generate a simple graph of the empirical hazard function to get an idea what shape it has.

30.11 Data processing—predictor variables

It is important to go through all predictor variables in your dataset to determine how they will be handled. Some issues to consider include the following.

- Are there many missing values? If there are, you might have to abandon plans to use that predictor, or conduct 2 analyses, one on the subset in which the predictor is present and one on the full dataset (by ignoring the predictor).
- What is the distribution of the predictor?
- If it is continuous, is there a reasonable representation over the whole range of values? If not, it might be necessary to categorise the variable (see comments about evaluating the relationship between predictors and outcome in section 26.13).
- If it is categorical, are all categories reasonably well represented? If not, you might have to combine categories.

30.12 Data processing—multilevel data

If your data are multilevel (*eg* blood pressure measurements within individuals within centres), it is necessary to evaluate the hierarchical structure of the data.

- What is the average (and range) number of observations at one level in each higher level unit? For example, what is the mean, minimum and maximum number of blood pressure measurements per individual in the dataset? Similarly, what are those values for the number of people per centre?
- Are individuals uniquely identified within a hierarchical level? It is often useful to create one unique identifier for each observation in the dataset. This will help identify specific points when evaluating outliers, influential observations *etc*. This can be done either by creating a variable that consists of a combination of the group and individual identifiers, or simply assigning a unique sequential number to each unit in the dataset.

30.13 Unconditional associations

Before proceeding with any multivariable analyses, it is important to evaluate unconditional associations within the data.

- Associations between pairs of variables can be evaluated using the following techniques.
  - Two continuous variables—correlation coefficient, scatterplot, simple linear regression
  - One continuous and one categorical variable—one-way ANOVA, simple linear or logistic regression
  - Two categorical variables—cross-tabulation and χ^2. Cross-tabulations are particularly useful for identifying unexpected observations (*eg* cases of cervical cancer in males).
- Associations between predictors and the outcome variable(s) need to be evaluated to:
  - Determine if there is any association at all, as it might be possible to ignore predictors with virtually no association with the outcome at this stage (see Chapter 15)
  - Determine the functional form (*eg* is it linear?) of the relationship between any continuous predictor and an outcome (discussed in Chapter 15)
  - Get a simple picture of the strength and direction of the association between predictors and outcome, to aid in the interpretation of results of the complex statistical models you will subsequently build.
- Associations between pairs of predictors need to be evaluated to determine if there is a potential for collinearity problems (highly correlated predictors).
- Special attention needs to be paid to potential confounding variables. Evaluate the associations between these variables and the key predictors of interest and the outcome. This will provide some insight into whether or not there is any evidence of confounding in your data (*ie* particularly if there is a strong association with both the key predictor and the outcome).

30.14 Keeping track of your analyses

You are now ready to proceed with the more substantial analysis of your data. However, before starting, it is wise to set up a system for keeping track of your results. Here are a few points to keep in mind to facilitate this process.

- Carry out your analyses in substantial 'blocks.' For example, if computing descriptive statistics, do so for all variables, not just 1 or 2. (Eventually you will need descriptive statistics for all of them, so you might as well keep them together.)
- Most statistical packages allow you to keep a 'log' file which records all of the results from a set of analyses. Give these log files the same name as the program file (except with a different extension).
- As facilities for electronic storage, retrieval and searching of files get better, there is less need for printouts of results. However, if you feel these will help you keep your results organised, 3-ring binders (2 or 4 rings in Europe) are a very convenient way to store printouts of all analytical work. Label and date all printouts and describe briefly what each contains on the first page of the printout. This will simplify finding results later.

The steps described above are essential if you have very large datasets and/or complex analyses which take a long time to process. With smaller datasets, you may find it more convenient to simply store the data file(s) and the program file(s) used in the analyses. In this case, you can

recreate the results by re-running the programs whenever you want to refer back to your results.

Following the steps outlined above will not guarantee that you obtain the best possible results from your analyses. However, the process will minimise the number of mistakes and lost time that affect all researchers that are just starting to develop experience with data analysis (and some of us who have been doing it for years). As you gain experience, you might choose to modify some of the items described above as you identify more efficient ways to conduct your analyses.

Good luck!