Exercise 2: Inspecting Data

Launch Stata, open a new do-file and save as *Stata_Exercise2.do*. Add appropriate comments at the beginning of the do-file. Remember to keep saving the do-file as you go along. Run through these exercises referring to **chapter 3** in the module notes.

Exercise 2.1

- Add commands to your do file to:
  - Change the working directory to the *Exercise 2* folder and load the dataset *bl_demog.dta*.
  - Obtain a summary of the dataset in memory using the `describe` command.
    - How many variables and observations are there?
    - How many variables have value labels and variable labels?
- Use the `browse` command to:
  - View all the variables
    - Make sure you understand the structure of the dataset e.g. how many rows per patient? Are there any obvious issues with any of the variables?
  - View just the four variables *ptid*, *wt*, *ht* and *wc*
  - Repeat having sorted the data in ascending order of the variable *wt*
  - Repeat having sorted the data in descending order of the variable *wc*
  - Repeat limiting to patients where *wt* is greater than 130kg.

- Use the `list` command to obtain a listing of:
  - The variables *ptid, age, wt, sbp* and *dbp* for patients whose *sbp* is less than 90 mmHg.
    - How many patients have an *sbp*<90 mmHg?
  - The variables *ptid, age, wt, sbp* and *dbp* for patients whose *sbp* is greater than 180 mmHg.
    - What do you notice about the values that are listed?
    - How many patients have an *sbp*>180 mmHg?
  - The variable *ptid* and any variable ending in *bp* in the first 10 rows of the dataset
  - Repeat for the last 10 rows of the dataset

- Use the command `codebook` to obtain summaries of the variables:
  - *ptid*, *birthdt*, *age*, *agegroup*, *race*, *smkstat*, *wt, lvef, diab*
    - Are there any duplicate patient ids?
    - What type (i.e. string, numeric, continuous, categorical) of variable is each of the above?

- Use the `summarize` command to find the mean and standard deviation, median and interquartile range and the range of values for *age* and *wt*.

- Use the `tabulate` command to obtain:
  - One-way tables of (i) *agegroup*, (ii) *sex* and (iii) *smkstat* (also try the `tab1` command)
  - Two-way tables of (i) *agegroup* and *hfdiag* and (ii) *agegroup* and *diab*. Are the totals the same for each table? Add the option *missing*.

- Use the `histogram` command to inspect the distributions of:

  - *sbp*, *wc*, *hrate, egfr* and *lvef*
  - Are there any issues with any of these variables?
- Use the `twoway scatter` command to get a scatter plot of:
  - (i) *sbp* and *dbp* and (ii) *wt* and *wc* (use hollow circles for the marker symbols).

Exercise 2.2

- Load the dataset *bl_labsall.dta*
  - Use `describe`, `browse` and `codebook` to familiarise yourself with the data
    - How many variables and observations are there? What are the variables?
    - What is the structure of the dataset?
    - Are there any duplicated patient ids?

- Use `summarize` and `histogram` to inspect the distributions of:
  - *creat hb, pot, sodium and totbil*
    - What issues are there?
    - How many missing values are there for each variable?
    - What is the mean for each of the variables?
    - Produce histograms omitting the problem values.

Exercise 2.3

- Load the dataset *vitals_long.dta*
  - Use `describe`, `browse` and `codebook` to familiarise yourself with the data
    - How many variables and observations are there?
    - What are the variables?
    - What is the structure of the dataset?
    - Produce one-way tables and a two-way table of *visit* and *param*.

- Use `summarize` inspect the distributions of:
  - *value*
  - *value* for each category of *param*
    - Are there any suspicious values? What are they?

- Produce a histogram of *value* for each category of *param* (omitting the problem values, if any).
- Produce a histogram of *value* for *sbp* at each *visit* (omitting the problem values, if any).

- Use `duplicates report` to investigate whether there are any duplicate values:
  - by patient id,
  - by patient id AND visit,
  - by patient id AND visit AND param,
    - Are there any issues?
    - Use `duplicates tag` to help investigate the duplicate values:
    - What is the nature of the problem? How would you deal with this?