Data analysis with R

Lecture 5

More on preliminary data analysis

Jouni Junnila

Lattice graphics

- Lattice plots allow the use of layout on the page to reflect aspects of the data structure.
- R has a special package for these kind of graphs, called *lattice*.
- With this package we can construct very informative trellis style graphical displays.
- This quality can be consider a speciality of R. Other statistical software are far behind from R, when talking about lattice graphs.

Lattice graphics; example

- Let's consider the dataset *jobs*.
  - The number of workers in the Canadian labor force broken down by region for the 24-month period from January, 1995 to December, 1996
- The dataset is arranged so that each region is in it's own column. For drawing our graph, we need to first manipulate our data so that we'll have all the values in the same column.
- We also use a logarithmic scale, to represent our data in a most informative way.

Example continues

```
> library(lattice)
> library(DAAG)
> Jobs <- stack(jobs, select=1:6)
> Jobs$Year <- rep(jobs[,7], 6)
> names(Jobs) <- c("Number", "Province","Year")
> xyplot(Number~Year | Province, data=Jobs, scales=list(y=list(log=2, relation="sliced")), type="l", layout=c(2,3))
```

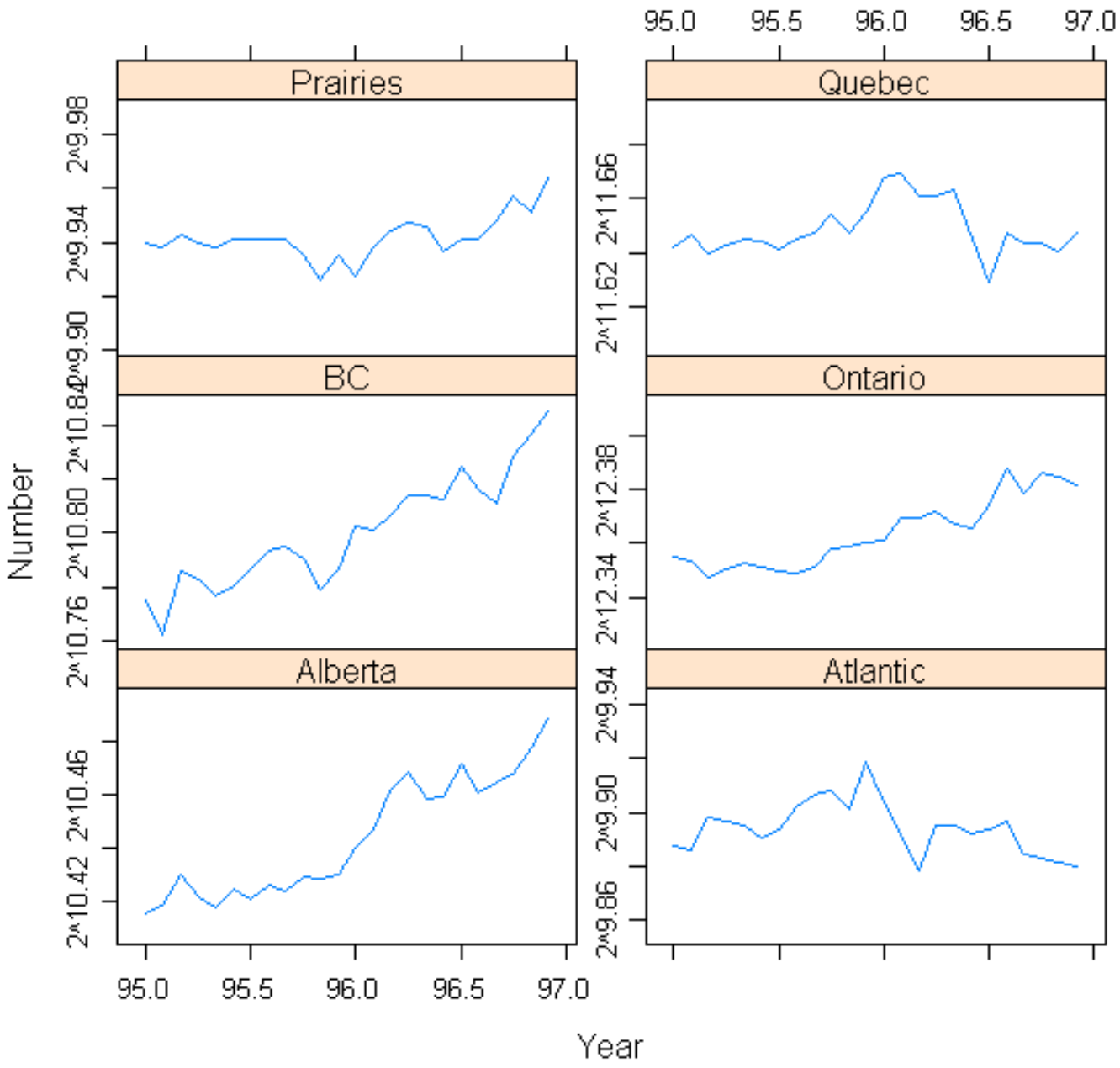
Necessary libraries

Stacking the data

Repeating the date, for each region

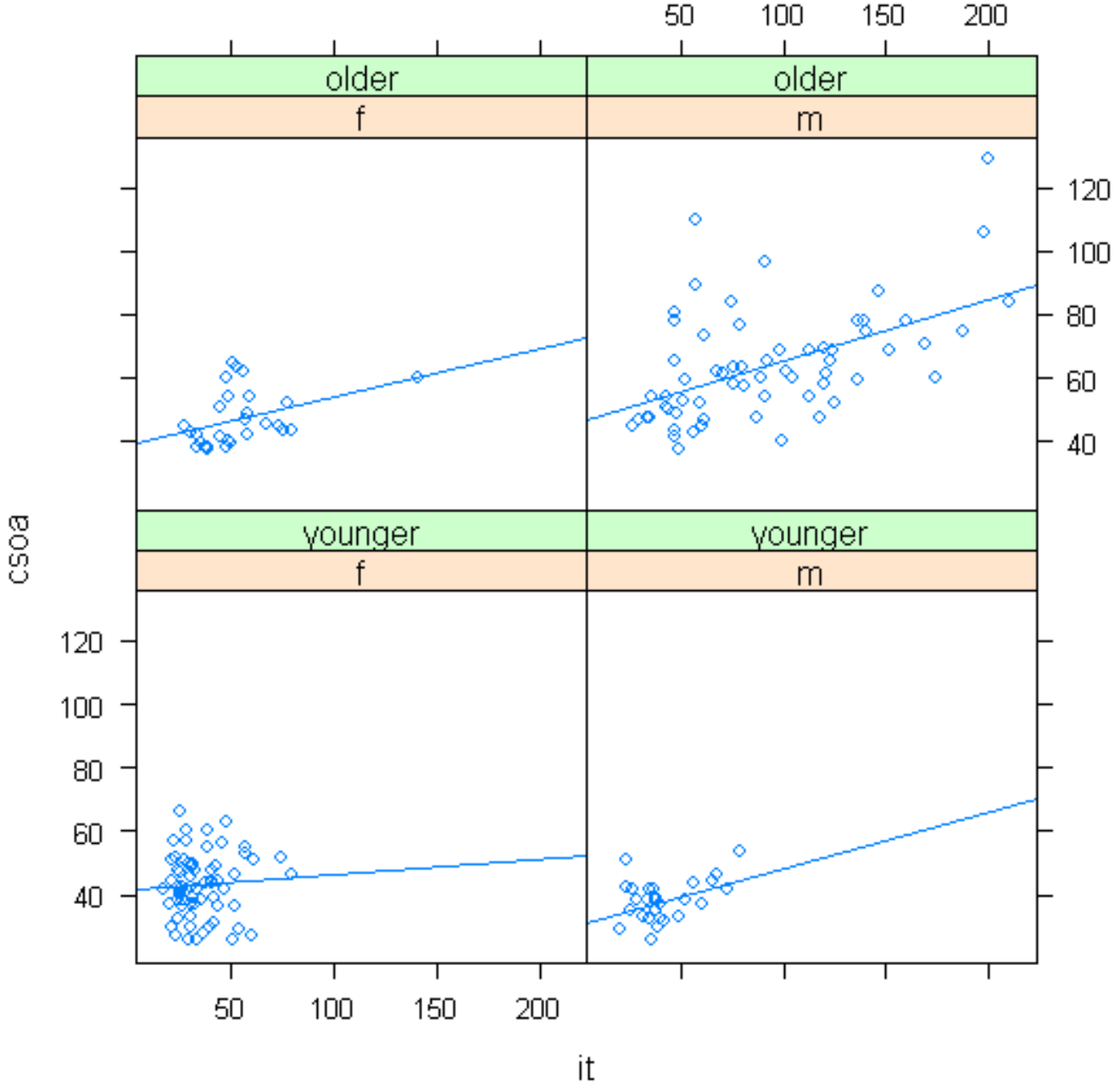
Naming the data-frame

Plotting six lineplots with Sliced logarithmic scale


Prairies
Quebec
BC
Ontario
Alberta
Atlantic
Number
Year
95.0
95.5
96.0
96.5
97.0
2^9.98
2^9.94
2^9.90
2^11.66
2^11.62
2^10.84
2^10.80
2^10.76
2^12.38
2^12.34
2^10.46
2^10.42
2^9.94
2^9.90
2^9.86

Example 2

- Let's try a bit simpler example.
- This time we want to compare two numeric variables *csoa* (time to recognize) and *it* (inspection time), separately in two different age groups and in two different genders.
  - We want four scatter plots laid in a same window and regression added on. The code is given below
  - `xyplot(csoa~it | sex*agegp, data=tinting)`


older
f
older
m
younger
f
younger
m
csoa
it
50
100
150
200
40
60
80
100
120

What to look for in plots

- Analyst should look for several different things when looking at plots.
- Next we will grasp five main issues.
    - Outliers
    - Asymmetry of the distribution
    - Changes in variability
    - Clustering
    - Non-linearity

Outliers

- We are always interested in noting points, that seem to be isolated from the main body of the data.
- Such points could corrupt any model that we fit.
- Boxplots highlight outliers in one dimension.
  - Outliers can also be detected using normal probability plots (handled later on the course).
- Scatter plots may highlight outliers in two dimensions.
- It is also important to try to determine is the outlier because of an error or is it a genuine value.

Asymmetry of the distribution

- Most asymmetric distributions can be characterized as either positively or negatively skewed.
- Longer tail on the right means positively skewed and vice versa.
- Logarithmic transformations usually make these kind of distributions more symmetric.
- Severe skewness is a more serious problem, when considering the validity of the results, compared to other kinds of non-normality.

Changes in variability

- Boxplots and histograms give us an impression of the amount of variability in the data.
- Side by side box plots are useful when comparing variability across different samples or treatment groups.
- We must keep in mind, that many statistical models depend on the assumption that variability is constant across treatment groups.
- If variability is constant on logarithmic scale, then the relative variation on the original scale is constant

Clustering

- Clusters in scatter plots may suggest structure in the data that may or may not have been anticipated.
- When moving on to formal analysis, these clusters must be kept in mind.
- Analyst must think: do the clusters correspond to different values of some relevant variable.

Non-linearity

- We should not fit a linear model to a data where relationships are clearly non-linear.
- Often transforming variables will help us making the relationships closer to linear.
- If transformations don’t help, there are numerous amount of non-linear methods that could be applied.
- Non-linear methods will not be handled during this course.

Statistical analysis strategies

- Analyses ask questions of data.
- The questions can be simple:
  - Does giving this medicine to the patient make his blood pressure lower. If so, how much lower.
- Or we may ask about relatioships of variables
  - What is the relationship between increasing dose and the blood pressure of the patient.

Planning the formal analysis

- If existing data is available, investigator can use it to determine the form of analysis that is appropriate for the main body of data.
- If possible, it's always desirable to plan the analysis in advance.
  - This reduces the chance of biasing the results of the analysis in a direction closest to the analyst's preference.
- Even so, graphical checks should still precede the formal analysis.

Preliminary examination

- If no previous data is available and it is not altogether clear what to expect, careful preliminary examination of the data is important. Specially the analyst should look for following qualities:
  - ✓ Outliers
  - ✓ Unexpected patterns within groups
  - ✓ Between group differences in the scatter of the data
  - ✓ Clusters in the data
  - ✓ Unanticipated time trends (eg. With order of data collection)

Data validation

- In all studies, it is necessary to check for obvious data errors or inconsistencies.
  - However, the analyst should not change the data if he/she cannot be absolutely sure about the correction.
- The level of validation varies a lot, depending on data collection technique, resources, etc.
- The less planning has been done before the data collection, the less validation is usually possible.

Changing the analysis plan

- Eventhough it's good to plan your analysis in advance, the plan should allow you to some limited changes based on exploratory analysis.
- An acceptable change to the analysis could be a data transformation to make the data closer to normal distribution
- On the other hand if there are a potentially large amount of comparisons that could be made, data-based selection of one or two comparison should be avoided.

Data manipulation

- One quite an important and time consuming issue data analyst must consider can be called data manipulation.
- Manipulation here means that we have to do some kind of data derivation to the original dataset, to make it fit the function(s) we want to use.
  - Examples of this kind of manipulation
    - Putting all values to a same column and generating factor variables.
    - Generating ID-variables

Data derivation

- Some analyses require some calculations before the actual analysis (eg. survival analysis)
- Also, we might want to class some original variable in one or more different ways.
  - Eg. From the original *Age*-variable we might want to create two age-group variables (one dichotomous and one with more classes)
- It's a good practise to do all the data derivation tasks in advance and somehow check that your derivation has been correct, before going into the analysis.

Taking the next step

- After we have plotted our data enough, looked for things mentioned before, cleaned/validated the data as much as possible and done necessary derivations it is time to move on to formal data analysis.
- However, moving to formal analysis does not mean we cannot do exploring also later on. So going back is possible and often even desirable.
- Formal analysis usually starts with descriptive statistics and moves on to modelling.