The study of relationship[1]

Important concepts

We need graphical and numerical tools, and we need to understand how they work. Graphical tools are often more useful and powerful than numerical tools, as they show (should show) more information without summarizing in a more or less arbitrary way. Graphical tools are *always* important as diagnostic tool.

Plots

Plots are the workhorses of exploration; the PLOT command is a central command in the EDA software. It has two sides: (1) PLOT as a normal EDA commands with a number of basic options and (2) PI (Plot_Inspect), a special module used to concentrate on a single relationship offering many commands for detailed examination. A typical analysis sequence might be...

```
>PLOT 1,20               ! A simple Plot
>PLOT 21,20 CASID        ! A second PLOT using casids as markers
>PLOT 20,22 CASID        ! A third PLOT
>PI                      ! Let's have a closer look at this one !!!!
```

The PLOT command[2] offers options to change the marker types (CASID, NUMBER, GVAR)[3], plot size(XUNITS=,YUNITS)[4], scaling and limiting options (NOFAR, NOOUTLIERS, LIMIT=(xlow,xhigh,ylow,yhigh)) and some others. Note that these options are available within PI as commands.

PI Plot_Inspect is a module within EDA with its own commands, i.e. when entering PI (the prompt changes to *Pi:* and the status line shows different information) the ordinary EDA commands are no longer available. At the end of this handout you will find the syntax chart for the PI command.[5]

PI is used for several purposes (1) detailed examination of observations and their position (2) analysis of straight line relationships and residuals and (3) hunting for transformations making a relationship more linear. In this section only the first is considered.

Below you will find commented examples, meant to stimulate your exploration of the possibilites of the PLOT/PI command (this list is not necessarily a meaningful sequence of commands to perform, but examples of the various possibilities of the PLOT command.[6]

```
>PLOT 20,21           ! Normal plot with two variables
pi:TYPE CASID FULL    ! set marker type and ask for full casids (4 letters)
pi:SIZE XUNIT=72      ! make the x-axis 72 units wide
pi:WINDOW OFENCES     ! set the viewing window to the outer fences
pi:WINDOW IFENCES     ! set it to the inner fences
pi:WINDOW HINGES      ! set it to the hinges
pi:WINDOW ALL         ! back to see all cases
pi:WINDOW AROUND=UK   ! make a window around the UK
pi:WINDOW ALL
pi:WINDOW CORNER=(UK,BRAS)   ! window with corners 'UK' and 'BRAS'
pi:WINDOW CORNER=BOLI ! Corners are 'BOLI' and nearest window corner
pi:DCASE              ! display all cases in the window
```

---

1. E. Horber, 13.12.98 : REL1.mss

2. Here we are concentrating on the study of bivariate relationships. The PLOT command is also used to plot a single variable against sequence (often time) and to plot one variable against several other variables.

3. The default is DOTS, but the SET PLOT TYPE command lets you change the default value

4. The SET PLOT SIZE command lets you preset a default size.

5. The normal syntax concepts applies, but the syntax is somewhat simplied. Commands need only 2 letters, variable lists do no apply (as you are looking at the same pair of variables all the time) and most options can also be shortened to 2 letters.

6. Please note that "pi:" in front of each PI command, is the prompt you will see on your screen in PI-mode; note also that the status line will show information meaningful in this context.

```
pi:P                  ! redisplay the plot
pi:WI ALL
pi:IDENTIFY NEIGB=UK  ! show the neighbours of the UK
pi:TRACE              ! add trace lines (cross-median)
pi:IDENTIFY Q=1       ! show observations in quadrant 1
pi:ADDSCALE           ! add a row/col scale
pi:IDENTIFY COL=3     ! show observations in column 3
pi:ID COL=3 ROW=9     ! show observations in col3 and row 9
pi:                   ! quit PI
>BOXPLOT ..           ! continue the road....
```

Straight lines

Relationships between two variables are often described and summarized by straight lines crossing the cloud of points. A line is expressed as

```
Y  = b X + a
```

where *b* is the slope (changes observed in Y, when X changes by one unit) and *a* the intercept (value of Y, when X=0).

Note that here we will deal with lines summarizing the relationship between two variables, i.e. fit a line to a cloud; sometimes however we draw other useful (reference) lines as landmarks making the graphic more readable, for instance a diagonal line of intercept 0 and slope of 1.

This linear summary of a relationship is of course never perfect for a real dataset, but hopefully adequate, i.e. it summarizes an essential aspect of the relationship.

```
Y =     b X + a       +       residual

Y =     summarized/   +       what is left to be
        "explained"           "explained" otherwise
```

There are a number of methods of computing such a line. In exploration we use a resistant line; the most common way of computing a line is by a method called *least squares* which, like all methods based on means, has the disadvantage of being heavily influenced by the presence of outliers or other unusual features.

Residuals

Almost all the greatest discoveries in astronomy have resulted from the consideration of what we have elsewhere termed Residual Phenomena, of a quantitative or numerical kind, that is to say, of such portions of the numerical or quantitative result of observation as remain outstanding and unaccounted for all that would result from the strict application of known principles.[7]

As the summary catches only a part of the variation of Y, it is important to examine what is left, i.e. it is important to look at the residuals, in order to

- Judge whether the summary is adequate (is the fit a good summary? Does it explain enough of Y?)
- Diagnose linearity, groups, outliers or other unusual features ... you did not discover before...
- Look for guidance for the next step (other variables??) in the analysis

Residual analysis and diagnosis is mostly graphical, but there are some numerical summaries, like measures of goodness of fit (sometimes badness of fit) and we can use any numerical summary to describe the residuals, e.g. any measure of spread is useful to describe the average size of the residuals.

Lines describe a relationship as "Y is dependent on X" (i.e. we distinguish a dependent and an independent variable).

Correlation

---

7. Sir John F. W. Herschel; Bart, K.H. in: Outlines of Astronomy, Lea and Blanchard Philadelphia, 1849, p. 548

Correlation is a related concept, however we do not distinguish a dependent and an independent variable; we just want to know how strong the relationship between X and Y is. All correlaton coefficents attempt to measure the *strength* of a relationship (correlation, association) and typically produce a value in the range -1 to +1 or 0 to 1, where 0 is no correlation, and 1 perfect correlation. The sign distinguishes positive and negative correlations.

In exploration we use *resistant* correlations whenever possible. The most common coefficient used (for quantitative data) in "ordinary" statistics is the Pearsonian correlation coefficient, a measure of linear correlation based on the mean, i.e. disqualified for data exhibiting unusual features especially outliers.

Caveat

When using numerical summaries it is important to understand (e.g. in the case of correlation):

- ***When to use?*** Is the level of measurement adequate? Many tools explained in this course are suited for quantitative data (interval scaled) only, but NOT for nominal or ordinal data!
- What information do I need? Is a single summary *really* enough ***(Very unlikely!)***?
- ***What aspect do they try to capture?*** Namely are they linear or something else. Using a linear summary means that we assume (assumption!) that we focus on the linear aspect of the relationship. Of course looking at the residuals of a linear summary means that we are focusing on what is NOT linear!
- ***What can go wrong?*** Difficult situations, outliers? groups? Problems: non-linearity? many assumptions? Are they applicable to all situations??

DO NOT FORGET

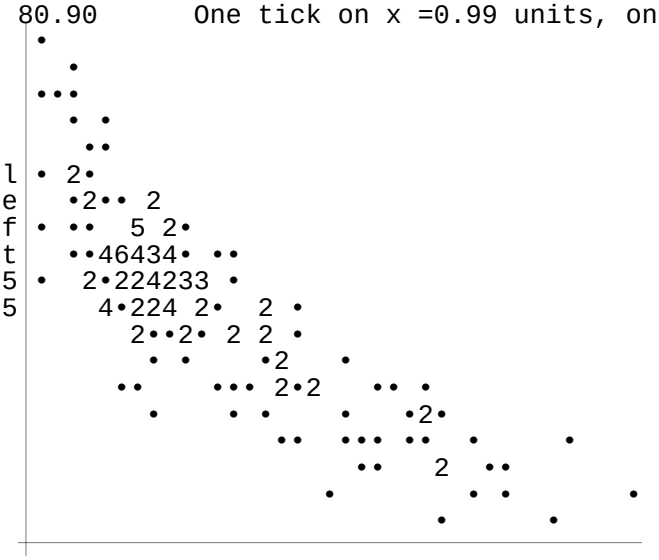
- All summaries, especially numerical summaries, are often[8] wrong.
- Plot, plot and look.... plot and look again.... (iteration)
- Let your substantive problem guide you, rather than be guided by purely technical considerations.

Some examples

```
>PLOT 41 48

>LINE 41 48 TRASH RESIDUAL
```

8. too often with real data!

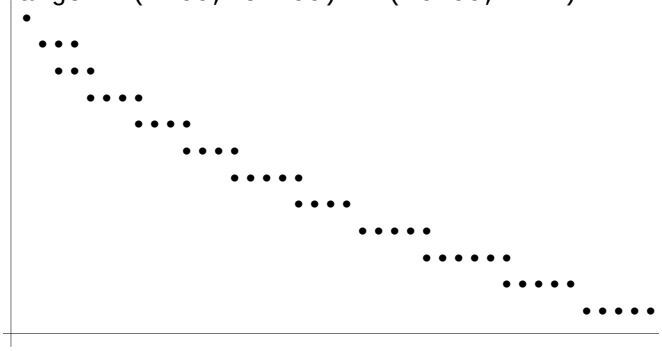
```
 157 towns
Plot of class12( 41) with left55( 48)
>>> All towns shown.
 80.90      One tick on x =0.99 units, on y =3.21
```



```
 7.70   class12                      45.40
 20.00
```

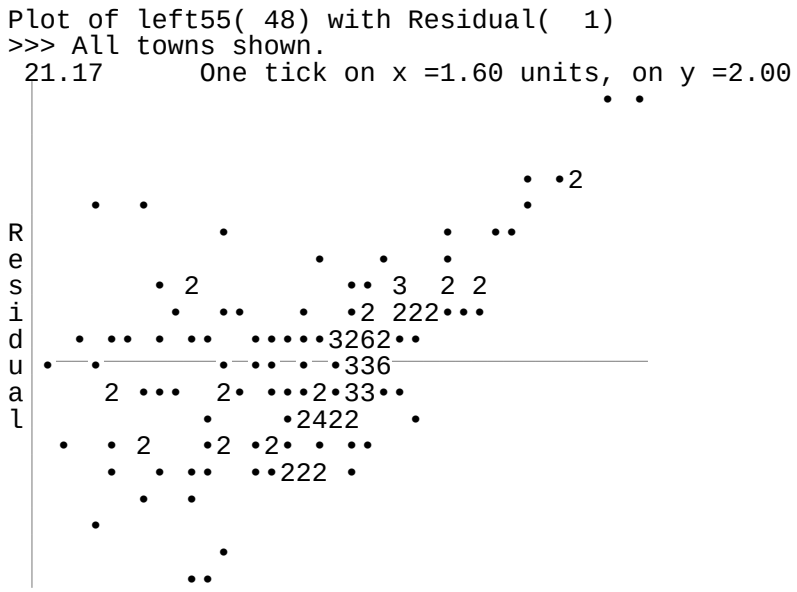
The TRASH option asks for the TRASH curve, whereas the RESIDUAL option will create a new variable in the WA containing the residuals of the resistant line.[9]

```
independent  : class12(41) % occ/retired males social class I,II
dependent  : left55(48) % votes for Labour, other left-wing

Resistant line (Velleman&Hoaglin): left55 = -1.25 * class12 + 70.25
Half-slope-ratio 1.39 (HSR1 0.72)
>>> Linear fit may be inappropriate
Fit = 0.52 Resistant correlation = -0.79

Residuals: Sharpness (overall,-1,-2):     4.74,     4.61,     4.51
TRASH curve
Range X=(1.00,157.00) Y=( 0.00,4.74)
```



```
Residual copied into var#  1

 >PLOT 48,1
```

9. Note that RESIDUALS copies the residuals into the next free location in the WA. If you prefer you may write RESIDUALS=10, i.e. store the residuals into variable 10, overwriting a variable if it exists. The FIT option is similar, but copies the fit from the resistant line.



PI (continued)

The LINE command[10] is available from within PI and additional commands let you analyse the result in great detail.

The ALINE command will show the line within the plot (adding two tick marks at the edges of the plot).[11]

Whenever a LINE command has been issued, residuals and the fit are available for further analysis. To tell PI what to plot you will use the PLOT command[12] The general form of the plot command is

```
PLOT x,y
```

where x and y are any combination of the following elements:[13] (1) X (or I for independent), (2) Y (or D for Dependent) (3) R for residuals (available only after a LINE command) (4) F (or YHAT or E) for the fitted values and (5) S (for sequence).

```
pi:LINE
pi:P S R     ! Plot sequence and residuals
pi:P X Y     ! back to the original X,Y plot
po:P X R     ! X against residuals.
```

If you have made a selection e.g. using a WINDOW command, the LINE command will operate on the selected observations.

Command syntax

PLOT

```
PLOT v1,[v2] [XUNITS=val][YUNITS=val] [FULL | SYMWID=n] [BIG*]
           [DOTS | CASID | NUMBER | GVAR{=v#} | THREE=v# {“altsym”}]
           [NO{X|Y}| BOX | [BA{X|Y}] [FRAME]
           [GLOBAL | PERCENT | LIMIT[=(xmin,xmax,ymin,ymax)]
                   | NOFAROUT | NOUTLIER
 => Use PI to inspect plot
```

---

10. The syntax is the same, with the exception of the variable list, ignored within plot inspect.

11. Note that ALINE does not need a LINE command to work as it is used to add any reference line you might want to add. If no LINE command has been issued, ALINE defaults to a reference line of y=x, i.e. a slope of 1.

12. This is not the EDA PLOT command, but the PLOT command available within PI with a different syntax.

13. This list is incomplete, as further elements will be added later (re-expressions).

```
PLOT v1,v2,v3,[v4,v5,v6] [XUNITS=val] [YUNITS=val]
           [LETTERS|“symbols”] [BIG*] [LIMIT{=(xmin,xmax,ymin,ymax)}]
           [DENSITY {“symbols”}][BIG*] [LIMIT{=(xmin,xmax,ymin,ymax)}]
PLOT v1,v2  SCAT        same params

  *) Print file must be open
Symbol type and plot size defaults as set by SET PLOT
```

PI: plot inspect

```
PI (Plot_inspect) commands

<return> or Quit                                    Leave PI
?/HElp; INfo/STatus                                 for help; status information
ADdscale                                            Add a scale (for identify)
ALine [A=v][B=v][“s”]                               Add a line (ticks)
AXis | [NO[X|Y] [FRAME] | BOX | BA{X|Y}]            Change Axis
DCASE [case]                                        Display casids (wildcards)
DIagnostic SUMDIFF                                  Diagnostic
DRaw                                                Show current plot
IDentify | [ALL] |  [COLUMN=col] [ROW=row]         Identify
         | QUAD= 1|2|3|4                             observations
         | [ROW] [COLUMN] WithCase=cas#
         | Neighbours_of=cas# [PROX=(xpos,ypos)]
LINE, LOWESS                                        Line and Lowess commands
MArk=c# [“s”]                                       Mark specified case
P [<vspec>] [<vspec>]                               Plot, see <vspec>
POint [X=v] [Y=v] [“s”]Mark specified coodinates
PRINT                                               write current plot to PF

REverse                                             Reverse plot (x <-> y)
SAve <vspec>                                        Save specified vspec to WA.
SEt SILENT | PRINT | SYM=(s,e)                      Set switches.
SIze [XUNITS=x] [YUnits=Y] [FULL | SYMBWID=n]       Change plot/symbol size
SLimit [GLOBAL] |NOUT |NOFAR |PERCENT               Set plotting
       | LIMIT[=(x1,x2,y1,y2)]                       limits
TRace  | [MEDIAN*] | [X=val] [Y=val] [POSIT]       Add a
       | THROUGH=cas#                                 crosstrace
TYpe <symb> [FULL | SYMWID=c] [POS=cpos]            Select plot symbol type
WIndow | [ALL*] | HINGES | IFENCES | OFENCES        Plot specified
       | LAST | Q=1|2|3|4 | Around=c# [PROX=(x,ywindow
       | WINDOW=(x1,x2,y1,y2))
       | Corners=(cas#,cas#) | CORNER=cas#
X/Y <nvar>                                          Set X or Y to a new variable
XTransform,YTransform <t>                           Transform X or Y

Symbols used:
<nvar> VAR=var# | SEQUENCE | C1dim=d# | K2dim=d# | RESID | FIT
<symb> [DOTS] | CASID | NUMBER | GVAR{=v#} | THIRD=v# <c.opt> [NOSYM]
<vspec> X | Y | XT | YT | Resid | Fit | Seq | X<t> | Y<t>
        Synonyms: D (for X) I (Y) E (FIT)
<t>     UP | DO | RS | RE | RR | LO | SR | RA | SQ | CU | ?
<c.opt> BINS | [FRAC] | EXACT | READ [“symbols”
        DISTRIBUTIONAL [SIMPLE] [“symbols”]
        REFERENCE{=value} [“Symbols”] [FUZZ=val]
        MARK|=val | IF>val |IF=val| IF<val | IF~val
        [“symbols”] [FUZZ=val]

>> Use PRINT with all commands to write a result to the PF.
```

LINE

```
Computes a straight line through a cloud of points

LINE x,y <method> [RESIDUAL{=var#}][FITTED{=var#}] [TRASH]
```

```
<method>::= | [RLINE] [TRACE][STEPS=maxsteps]
                       [SHORT][Tol=tolerance]
            | TUKEY [STEPS=n]
            | LSQ
            | LSQ2
            | LSQORTHOGONAL
If the y-variable is omitted, the variable set by SET YVAR is used.
```

REGRESS

```
Perform Multiple biweight regression
REGRESS vlist [Y=yvar#]*)
              [NOCONSTANT] [CUTOFF=val] [EPSI=v]
              [MAXITERATIONS=num] [TRACE {FULL}]
              [TRASH]
              [RESIDUALS{=v#}] [FITTED{=v#}]
              [BCOEFFS{=v#}]

You may use SET YVAR=var#.
*) Obsolete form REGRESS vdep,vlist <options>
```