FIXREG - Update 2.6.2001

Christian Hennig - Universit"at Hamburg

2.6.2001

This is a short documentation for some features of the FIXREG-software, which are changed compared to the manual. The software is now consistent with my forthcoming publications and nearly consistent to the theoretical basis

Christian Hennig (2000): "Regression Fixed Point Clusters: Motivation, Consistency and Simulations", Preprint 2000-02, Fachbereich Mathematik - SPST.

The manual and the paper Hennig (2000) are avialable from the same website.

1 Iteration specifications

Almost all changes concern the iteration specifications (Section 4.4 in the manual). They now look as follows:

```
Settings for cluster search:
 (*) indicates test functions for curious users; possibly unstable or useless.
       a) 853 runs with random start (choose a for direct change),
       b) Change of iteration number through parameters
           minimum size 30 of FPC to be found with probability 0.950
           maximum iteration number 200000
       c) Scale factor c determined by approximation formula: yes
       d) Scale factor c = 10.070
       f) (*) Number of data points for search start: 3
       g) Minimum size of clusters: 23
       h) Output of standardized values (median and MAD): no
       i) Minimum times to be found for relevant FPC: 3
       j) (*) Do you want to enter points with decreased search probability: no
       k) Suppress standardization: no
       l) Iteration step limit: 150
       m) Similarity cutoff for FPC-groups: 0.8500
       n) Point numbers output in fix-file: no
       q) Start of calculations
Enter a letter, if you want to change the default settings
```

In detail (menu items that are not mentioned here are explained in the manual):

a) The number of iteration runs is now automatically defined by a theoretically founded formula, see Hennig (2000), Section 7.3. The number can be modified directly by choosing "a", or according to the theory by typing "b".

b) The rationale for the new formula is that an FPC of size $n/5$ should be found with an estimated probability of 0.95. These settings can be changed, and a new iteration run number is calculated. The maximum number of iterations can be changed as well. The number is limited, because for large p the automatically defined number can be very large.

c) The scale factor c is now by default determined automatically by a formula, which guarantees a very low probability to find more than one FPC at homogeneous data (determined by the simulations in Hennig, 2000). That is, if more than one FPC is found, significant non-homogeneous structure can be expected. If "c" is changed, c can be specified via the "standard level", see "e".

d) Type "d" to change c directly. For small n, a smaller c may be useful for exploratory purposes, and for very large n, one may want to have a minimum of c larger than 3 (from the formula).

e) Change of c via "standard level", as in the manual. (Only available if "c" is changed.)

g) The default is now determined on grounds of an estimated probability of at least 0.5 to recover an FPC of this size.

i) "Relevant" FPCs should be stable and occur more than once during the search. The default is 3. A larger value makes the analysis more stable, a smaller value makes computation faster. Note that the automatical choice of c is based on the value 3, as well as the automatic choice of the number of iteration runs. The latter can be adapted by choosing "b" after changing "i".

m) To make the interpretation of the output easier, too similar FPCs are grouped together by single linkage clusters of similarity ≥ 0.85, see Hennig (2000), Section 7.2. If many FPCs are grouped together to few groups (especially, if many FPCs are joined with the FPC of almost all points, see below), a value of 0.9 may be worth a try.

n) By default, the numbers of points belonging to an FPC are no longer visible in the main output file (see manual, Section 5.1). Choose "n" to get them back.

Some features of the old version, which were not of much use (and not fully documented in the manual), are discarded.

2 Output

The output now ends with the groups of FPCs. Because there are often too many FPCs, the interpretation should concentrate on the "representative FPCs" of the Single Linkage groups, based on a similarity measure between the FPCs, defined as formula (7.1) in Hennig (2000). For these groups, the members are written to the main output file, as well as the numbers of the representative FPCs, which can be assessed as explained in the manual (Sections 5.1, 5.2). The cluster intersection table (called "similarity table" in the manual), as well as a table of the values of the new similarity measure are given above the groups.

I discarded the old "cluster quality" measure in favour of a better measure called "expectation ratio", which is the number of times that an FPC was found divided by the expected number, based on the estimated probability of the finding of an FPC dependent on its size. The expectation ratio is given for all FPC groups. The representative FPCs for each group are chosen as having maximal expectation ratio. This is modified compared to Hennig, 2000, where the simple number of findings was proposed to choose the representative FPCs.

3 Hint

At least under UNIX, the inputs to `fixreg` may be specified in a file, e.g. `belcalls.in`, and `fixreg` can be started by

```
fixreg < belcalls.in
```

This is useful if you perform more than one analysis with the same dataset, and input errors do no longer force you to enter again everything. The input file must contain all inputs. Note that these sometimes depend on earlier inputs, so that there is no unique format for such files. Here is the example `belcalls.in` for the "calls from Belgium-data", see manual, Section 4.2:

```
2
belcalls.dat
n
n
3
calls
1
year
2
*
y
belcallsnew
d
80
q
```

An empty line before the first number is important. The file must end with "q" to start the calculations. "2" is the input option (other options need more lines to specify n and p), `calls` and `year` are variable names, the following numbers are positions. `belcallsnew` is the root name of the output files. `d 80` changes the scale constant to 80. If you are a little familiar with the software and the manual, you will understand the other lines and possibilities.