Chapter g10 – Smoothing in Statistics

1. Scope of the Chapter

This chapter is concerned with methods for smoothing data. At present this chapter contains one function for smoothing a sequence of observations or a time series.

2. Smoothing Methods

Many of the methods used in statistics involve fitting a model, the form of which is determined up to a small number of parameters. For example, a distribution model like the gamma distribution, a linear regression model or an autoregression model in time series. In these cases the fitting involves the estimation of the small number of parameters from the data. In modelling data with parametric models there are two important stages in addition to the estimation of the parameters, these are: the identification of a suitable model, for example the selection of a gamma distribution rather than a Weibull distribution, and the checking to see if the fitted model adequately fits the data. While parametric models can be fairly flexible they will not adequately fit all data sets especially if the number of parameters is to be kept small.

Alternative models based on smoothing can be used. These models will not be written explicitly in terms of parameters. They are sufficiently flexible for a much wider range of situations than parametric models. The main requirement for such a model to be suitable is that the underlying models would be expected to be smooth, so excluding those situations where, for example, a step function would be expected.

These smoothing methods can be used in variety of ways, for example:

(1) Producing smoothed plots to aid understanding.

(2) Identification of a suitable parametric model from the shape of the smoothed data.

(3) Eliminating complex effects that are not of direct interest so that attention can be focused on the effects of interest.

3. Smoothers for Time Series

If the data consists of a sequence of n observations recorded at equally spaced intervals, usually a time series, several robust smoothers are available. The fitted curve is intended to be robust to any outlying observations in the sequence, hence the techniques employed primarily make use of medians rather than means. The ideas are come from the field of exploratory data analysis (EDA), see Tukey (1977) and Velleman and Hoaglin (1981). The smoothers are based on the use of running medians to summarize overlapping segments. These provide a simple but flexible curve.

In EDA terminology, the fitted curve and the residuals are called the smooth and the rough respectively, so that:

Data = Smooth + Rough.

Using the notation of Tukey, one of the smoothers commonly used is, 4253H, twice. This consists of a running median of 4, then 2, then 5, then 3. This is then followed by what is known as Hanning. Hanning is a running weighted mean, the weights being 1/4, 1/2 and 1/4. The result of this smoothing is then 'reroughed'. This involves computing residuals from the computed fit, applying the same smoother to the residuals and adding the result to the smooth of the first pass.

4. References

Tukey J W (1977) *Exploratory Data Analysis.* Addison-Wesley.

Velleman P F and Hoaglin D C (1981) *Applications, Basics and Computing of Exploratory Data Analysis* Duxbury Press, Boston, Massachusetts.

5. Available Functions

Computes a smoothed series using running median smoothers `g10cac`