Appendix chapter 5
Explanation of the techniques in “Searching for discontinuity”

Marijn van Dijk

Part 1: Illustration of the peak criterion (linear model)

First, select a column and paste the time and the observed data, for instance in columns A and B, from row 10 downwards (in the illustration the final value of the data is in B64; see Screenshot 1, Figure 5.6). We now put a linear regression in column C, by computing \((\text{slope} \times A10) + \text{intercept}\) (the slope and intercept have to be computed first; in the screenshot these are in E1 and E2, respectively) and copying it downwards. We now compute the difference between the data and the model by inserting \(= \text{ABS}(C10-B10)\) in D10 and copying it downwards. In column E, we divide column D by column C (residuals divided by the model) in order to obtain the relative variability.

The second step is to define some statistical characteristic of the observed data set. We do this because we simulate data that are based on a normalized distribution of the data. We therefore insert a small block in columns F and G, computing the sum of the values of the linear model (C) and the sum of the residuals (D). The resulting values are plotted in cells G9 and G10. We now calculate the total relative variability by dividing G10 by G9 in cell G11 (we have called this cell “var_prop_lisa”).

We now set up our simulation in row H. In this case we have chosen a permutation technique that draws values from a normalized distribution of the data presented in column B using the formula \(= \text{dNormalInt}(C10;C10\times \text{var_prop_lisa})\). It should be noted that other randomization formulas are also possible as long as the values are randomly selected from the observed data. We now compute the difference between the linear model and the simulated data by adding \(= \text{ABS}(C10-H10)\) in I10 and copying it downwards. Finally we calculate the relative variability by dividing I by C.

The third step is to define the criterion values where we compare the observed data with the simulated data. In our illustration we have chosen J4 and K4 for this. J4 contains the criterion value of the observed data and is defined as the maximal relative peak in the data \(= \text{MAX}(E10:E57)\); K4 does the same thing for the simulated data by computing \(= \text{MAX}(J10:J57)\).

The final step is to test the observed criterion against the simulated criterion. This can be done using Poptools. Choose “Simulation Tools” > “Monte Carlo Analysis”. The dependent range contains the simulated criterion (K4) and the test
value is the observed criterion value (J4). Set the number of simulations at 1000 and select a cell for the test output. In the illustration, this is done in cell L20.

The output (P16 and below) can be interpreted by looking at the number of times the simulated data reached a value that exceeds the observed criterion of a value higher than that. In the illustration, this value can be found in cell P21. This value has to be divided by the number of simulations (in cell Q21) to obtain a p value. In the illustration, the resulting p value is <.001 since in 0 out of 1000 simulations a comparably high or higher relative variance was found in the simulated data. This means that the 0 hypothesis (that the peak is produced by a linear model) can be rejected.

In the Excel file “Demo chapter 5_1 peaks” you will find the complete model.

Part 2: Illustration of the membership criterion (linear model)

First, again plot the data and time in columns A and B (in Screenshot 2, Figure 5.7, they are presented in A2–A49 and B2–B49). We now go to columns G–I, where we reshuffle the data in such a way that we obtain a sample from the original pool of data in J2–J20. In the illustration we use three consecutive array formulas, and add and subtract the values to obtain new values that in principle could have stemmed from the observed data pool. Note that this is just one of the ways to simulate the data. We now calculate how often the observed values occur in the

Figure 5.6 Screenshot 1: Peak criterion applied to the data for Lisa.
simulated data (by means of a simple frequencies formula) and put this in columns E (values) and F (frequencies in the present resampling).

Secondly, we resample the data 10,000 times (see cell M41) and put the output in columns L and M. In column N we can transform this into a percentage of all iterations. We now put a threshold at .025 (implying that 95% of the set is defined as the initial set); in the illustration this value is exceeded at the value of 17.

We now start assigning membership values to the observed values (based on the relative frequencies in columns L and M) in column C by writing =LOOKUP(B2; $E$2:$E$38;$O$2:$O$38). We then plot the data and the membership values in a graph and note when the membership suddenly drops to a lower value. In the illustration, this happens at time point 0.

In the Excel file “Demo chapter 5_2 membership” you will find the complete model.