# Case Study – ROI Estimation of Speaker Programs

## Problem Statement:

Bizer, a pharmaceutical company, has been conducting Speaker Programs for increasing the awareness of its Hepatitis – C drug, HCD, among physicians since the past 2 years in 5 states of USA. The states are California, Florida, Illinois, Michigan and New York. They wish to measure the impact of the programs and calculate its ROI.

## Data Description:

Two datasets have been provided –

* Attendee Data
* Universe Data

### Attendee Data –

The data has 6 columns, Attendee ID, Attendee Class, Attendee State, Month, Sales and SP flag, Attendee Class represent the class of Attendees (i.e. Physicians, Nurses etc.), we are supposed to measure the impact only for class “Physicians”, SP flag is 1 for the month in which the physician attended the program and 0 otherwise. The remaining columns are self – explanatory. The data contains Physicians who attended the program at some point in the past year.

### Universe Data –

Universe data is similar to Attendee Data except for lacking the SP flag; it contains Physicians who attended the program at some point in the past year as well as those who didn’t.

## Test & Control Matching Approach:

Test/Control matching is one of the basic ways of ROI estimation; it essentially works on the concept of comparing a pair of Subjects who were behaviorally identical before an event and then measures any difference in their behavior after one of them has been a part of an event while the other hasn’t. Any such difference can then be attributed to the event.

The subject who has been a part of the event is known as the Test, the one who hasn’t been part of the event is known as Control, the pair is known as test – control pair.

In our case, a physician who has attended any Speaker program is the test and Speaker Program is our event, the challenge is to find identical control physicians from the Universe Data who did not attend any Speaker Program and then compare them to our tests.

## Finding the identical Physicians:

Since our data is limited to sales and Geography, we will be using that for finding out the matched control physicians. Our criterions are:

* Both the test and Control should belong to the same State.
* The Euclidean distance between the sales of the 3 months preceding the Speaker Program month should minimum, i.e. , If the speaker program happened in month 4 for test A.
  + - Assume sales of A in month 1, 2 and 3 be A1, A2, A3 respectively.
    - Let the sales of the control in corresponding months be C1, C2, C3 respectively.
    - Find the control which minimizes the square root of W1(A1 – C1)2 + W2(A2 – C2)2 + W3(A3 – C3)2
    - Here Ws are weights; they are used to increase/decrease the importance of a particular term in the distance. For instance in our example, month immediately before SP might have a higher value than the one 3 months before the SP.

## Instructions:

1. You might be required to clean the data before processing, have a good look at it and clarify any doubts with your instructors.
2. Only Attendees with Class “Physicians” need to be matched to controls.
3. The matching must happen to a control which has not attended a Speaker Program at any point of time, so you will need to subset the Universe data by removing the Attendee IDs from it.
4. In case a physician has attended the program more than once, use the first occurrence and ignore the second one.
5. W1 = 1, W2 = 2, W3 = 4.
6. ROI needs to calculated by the difference of difference method, i.e. (Sum(Sales of tests in 3 months post SP – Sales of Control in those 3 months) – Sum(Sales of tests in 3 months pre SP – Sales of Control in those 3 months)) / Total Cost of the Programs.
7. Sales can be converted to Dollars by the factor 1 Unit = $5K, the total SP cost is $500K.

## Expected Results:

* An R Code file which has the following properties –
  + Can be run only with the help of the input data set present in CSV format in the working directory, no other CSVs / code files should be required.
  + Clearly writes in CSV format the matched test - control pairs with their IDs as well as pre and post 3 month sales along with the calculated Euclidean distance. The name of the CSV should be Matched\_pairs\_<Your name and emp ID>
  + ROI number should be calculated at the end of the code and printed for reference.