SAS Macro %PiecewiseTransfCuts

Explicación del método de selección de los cortes óptimos para armar variables piecewise

El presente documento tiene como objetivo explicar el proceso seguido por la macro SAS ***%PiecewiseTransfCuts*** desarrollada por el autor, a los fines de elegir los cortes óptimos a la hora de armar una variable piecewise.

**ÍNDICE**

1 Resumen 2

2 Supuestos 2

3 Detalles 2

3.1 Descripción general del método 2

3.2 Elección de los valores de corte a intentar 2

3.3 Algoritmo 2

3.3.1 Regresiones a 2 tramos 2

3.3.2 Regresiones a 3 tramos 3

3.3.3 Cálculo del R2 a maximizar 3

3.3.4 Elección del conjunto de cortes óptimo 3

3.4 Aplicación del algoritmo a una variable target binaria 3

3.5 Limitaciones 4

3.6 Ejemplo 4

# Resumen

El método implementado en la macro se basa en elegir los cortes de la variable input analizada de manera de maximizar el R2 de la regresión piecewise, respecto a una variable target continua.

# Supuestos

Se parte de los siguientes supuestos:

* La variable target es continua
* La variable input analizada (i.e. a convertir en variable piecewise) es continua.

# Detalles

## Descripción general del método

El método es empírico, en el sentido de que se considera un conjunto de "conjuntos de cortes" posibles, se calcula el R2 de la regresión piecewise para cada "conjunto de cortes" y se elige el "conjunto de cortes" que arroja el R2 más grande.

En la implementación del método, cada conjunto de cortes considerado tiene **a lo sumo 2 cortes**. Es decir que un conjunto de "conjuntos de cortes" posibles podría ser algo así:

{ {5}, {7}, {10}, {13}, {5, 10}, {5, 13}, {7, 13}, {10, 13} }

donde los primeros cuatro conjuntos corresponden a una regresión piecewise con 2 tramos y los últimos cuatro conjuntos corresponden a una regresión piecewise con 3 tramos.

## Elección de los valores de corte a intentar

Los valores de corte son elegidos a partir de los valores tomados por la variable input continua, según los siguientes parámetros recibidos por la macro:

* ncuts: el número de valores distintos a considerar como valores de corte (default: 15)
* maxnpieces: el máximo número de tramos a considerar (default: 3 --y como máximo puede ser 3)
* minpropcases: la mínima proporción de puntos que debe haber en cada tramo considerado al definir los cortes (default: 0.10 o 10%)

En base al rango de valores de la variable input y al valor del parámetro ncuts, se definen los valores de corte distintos a considerar como una grilla regular cuyo mínimo y máximo valores quedan determinados por el parámetro minpropcases. Por ej. si ncuts = 4, estos valores pueden ser {5, 7, 10, 13}.

## Algoritmo

### Regresiones a 2 tramos

Para cada uno de los valores de corte a intentar se arman las regresiones piecewise de 2 tramos y para cada una de ellas se calcula el R2 obtenido.

### Regresiones a 3 tramos

Si maxnpieces = 3, se consideran combinaciones de a pares de estos valores para armar las regresiones piecewise de 3 tramos, que podrían quedar determinadas por el siguiente conjunto de pares de cortes:

{ {5, 10}, {5, 13}, {7, 13}, {10, 13} }

cuyas combinaciones pueden estar limitadas por el valor de minpropcases (por esta restricción es que por ej. no aparece el par de cortes {5, 7}).

### Cálculo del R2 a maximizar

El valor del R2 de cada regresión piecewise se calcula como el R2 de cualquier modelo, a saber:



donde:

*RSSM*: Residual Sum of Squares del modelo (en este caso la regresión piecewise)

*RSS0*: Residual Sum of Squares del modelo nulo (es decir sin ninguna variable input o regresora)

Estas cantidades por definición se calculan como:





donde:

*n*: número de puntos u observaciones válidas utilizadas para el análisis.

*yi*: valor de la variable target para la observación *i*.

*y’i*: valor estimado de la variable target por el modelo para la observación *i*.

: valor promedio de la variable target (i.e. estimación de *yi* por el modelo nulo).

### Elección del conjunto de cortes óptimo

Sobre todos los conjuntos de cortes considerados, se elige como conjunto de cortes óptimo aquel para el cual el R2 es el máximo de todos los obtenidos.

## Aplicación del algoritmo a una variable target binaria

Para poder aplicar el algoritmo a una variable target binaria es necesario primero crear una variable target continua. Para esto, se propone el siguiente paso de preparación de datos previo:

1. Agrupar en bins de igual tamaño la variable input continua. Cada bin debería tener una cantidad apreciable de puntos, por ej. mínimo 100 puntos (o menos si es que la cantidad de casos totales del problema es muy chica. En todo caso el número de puntos en cada bin, no debería ser inferior a 30)
2. Para cada bin calcular:

* , el promedio de los valores de la variable input continua *x* que caen dentro del bin.
* logit(*pi*), donde *pi* es la probabilidad de ocurrencia del evento de interés de la variable target binaria en el bin *i* y .

1. Utilizar como variables input y target para las regresiones piecewise a:

* 
* *yi* = logit(*pi*)

## Limitaciones

Actualmente el algoritmo implementado en la macro ***%PiecewiseTransfCuts*** presenta las siguientes limitaciones:

1. La elección de los cortes *no* depende de la distribución de la variable input.

Esto podría ser problemático en caso de que haya una región de valores de la variable input con alta densidad de puntos, ya que podría ocurrir que nunca caiga un corte dentro de este conjunto de alta densidad.

1. Las regresiones piecewise efectuadas no permiten incorporar una variable de peso.

En el caso en que los valores utilizados en el análisis provienen de un proceso de agrupación o binning de la variable input, o si la variable continua toma valores discretos que ya presentan una agrupación natural (ej. la edad), el algoritmo actualmente no tiene en cuenta la cantidad de casos de cada grupo o bin al efectuar las regresiones.

Está previsto mejorar la macro en ambos sentidos, y generar una nueva versión donde:

1. la elección de los cortes tenga en cuenta la distribución de valores de la variable analizada,
2. las regresiones piecewise utilicen una variable de peso (weight) provista por el usuario.

Por ahora un paliativo para la limitación (1) es aumentar el valor de ncuts para así considerar más puntos de cortes posibles. La limitación (2) puede evitarse en algunos casos considerando bins de igual tamaño, pero esto no siempre es posible (por ej. cuando la variable input toma valores que se repiten una gran cantidad de veces).

## Ejemplo

A continuación se muestra un ejemplo de aplicación de la macro ***%PiecewiseTransfCuts*** sobre el siguiente conjunto de 68 puntos de entrada (*x*, *y*), donde *x* corresponde a datos de *edad* y “*y*” corresponde al logit de un evento de interés.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
| y | 47.3% | 41.9% | 38.8% | 35.8% | 33.3% | 30.7% | 29.4% | 27.2% | 30.5% | 29.5% | 28.8% | 26.8% |
| # casos | 5770 | 6230 | 6588 | 5851 | 4627 | 4178 | 3735 | 3686 | 4387 | 4245 | 3980 | 3749 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 |
| y | 28.5% | 28.4% | 26.4% | 25.4% | 24.8% | 24.6% | 25.5% | 24.1% | 23.7% | 23.1% | 23.0% | 22.4% |
| # casos | 4294 | 4702 | 4656 | 4686 | 4724 | 4977 | 5266 | 5441 | 5542 | 5679 | 5556 | 5053 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 |
| y | 21.0% | 20.6% | 20.3% | 20.0% | 18.9% | 19.3% | 18.1% | 18.9% | 18.8% | 17.8% | 15.8% | 16.5% |
| # casos | 4938 | 5081 | 5126 | 4747 | 4934 | 4707 | 4381 | 4682 | 4764 | 4670 | 4790 | 4660 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 | 64 | 65 |
| y | 15.3% | 15.9% | 16.1% | 16.4% | 16.2% | 14.5% | 14.1% | 14.4% | 13.5% | 13.1% | 13.8% | 13.1% |
| # casos | 4497 | 4281 | 4179 | 4151 | 4097 | 3979 | 3699 | 3577 | 3370 | 3185 | 3134 | 2947 |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 | 74 | 75 | 76 | 77 |
| y | 13.9% | 12.5% | 13.5% | 12.6% | 13.4% | 13.9% | 13.1% | 13.6% | 12.2% | 11.8% | 11.1% | 12.3% |
| # casos | 2904 | 2751 | 2488 | 2511 | 2296 | 2157 | 1987 | 1883 | 1746 | 1650 | 1573 | 1298 |

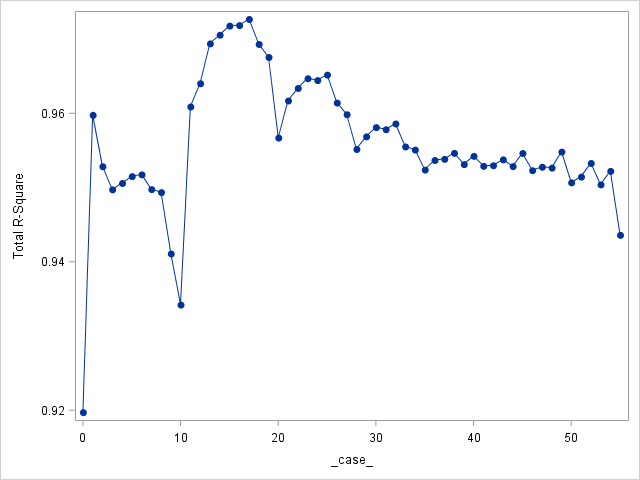
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 78 | 79 | 80 | 81 | 82 | 83 | 84 | 85 |
| y | 12.7% | 13.6% | 11.2% | 10.1% | 13.3% | 12.5% | 7.3% | 10.7% |
| # casos | 1211 | 1019 | 893 | 755 | 618 | 424 | 218 | 121 |

La macro se ejecuta con los siguientes valores de los parámetros discutidos anteriormente:

* ncuts = 10
* maxnpieces = 3
* minpropcases = 0.10 (es decir como mínimo debe haber 7 puntos en cada tramo)

los cuales generan 56 regresiones piecewise distintas analizadas, una con ningún corte, 10 con un solo corte, y 45 con dos cortes.

Los valores R2 obtenidos para las 56 regresiones así obtenidas se muestran en la siguiente gráfica:



R2 del modelo

case

donde el eje horizontal (*case*) indexa el conjunto de cortes considerados al armar la variable piecewise en cada regresión, comenzando a la izquierda por el valor 0 que representa la “variable original sin tramar”, los 10 puntos que siguen corresponden a conjuntos de corte que definen 2 tramos, y los 45 puntos restantes corresponden a pares de corte que definen 3 tramos.

Los valores de corte considerados son los siguientes 10 valores:

{23, 29, 35, 41, 47, 53, 59, 65, 71, 77}

Y luego tomados de a pares para las regresiones con 3 tramos, lo que da un total de:

regresiones con 3 tramos.

Siendo el número de combinaciones posibles de tomar de a 2 los 10 valores del conjunto arriba indicado.

En este caso, el conjunto de cortes óptimo es el correspondiente al *case* = 17 con un **R2 = 0.9727**, es decir 6% superior al R2(*case*=0) = 0.9197 correspondiente a no tramar la variable *x*.

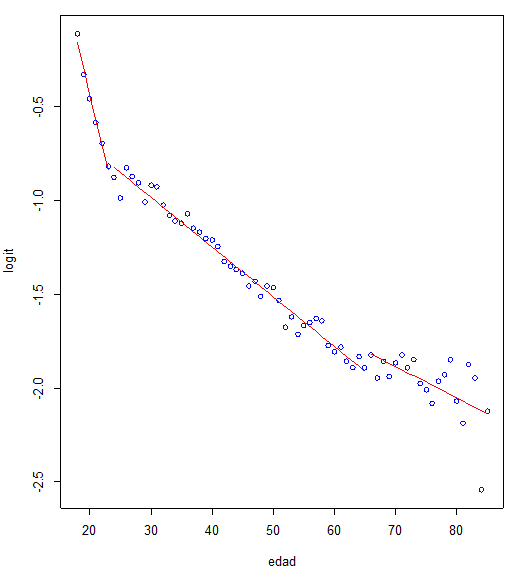
Los cortes correspondientes son: *x* = 23 y *x* = 65, es decir los tramos elegidos para la variable *x* son:

Tramo 1: *x* < 23

Tramo 2: 23 <= *x* < 65

Tramo 3: 65 <= *x*

La regresión piecewise obtenida para estos cortes se muestra en la siguiente gráfica:



x

y