

Hồi quy tuyến tính với phần mềm RapidMiner



Nguyễn Huỳnh Luận

Hồi quy là gì?

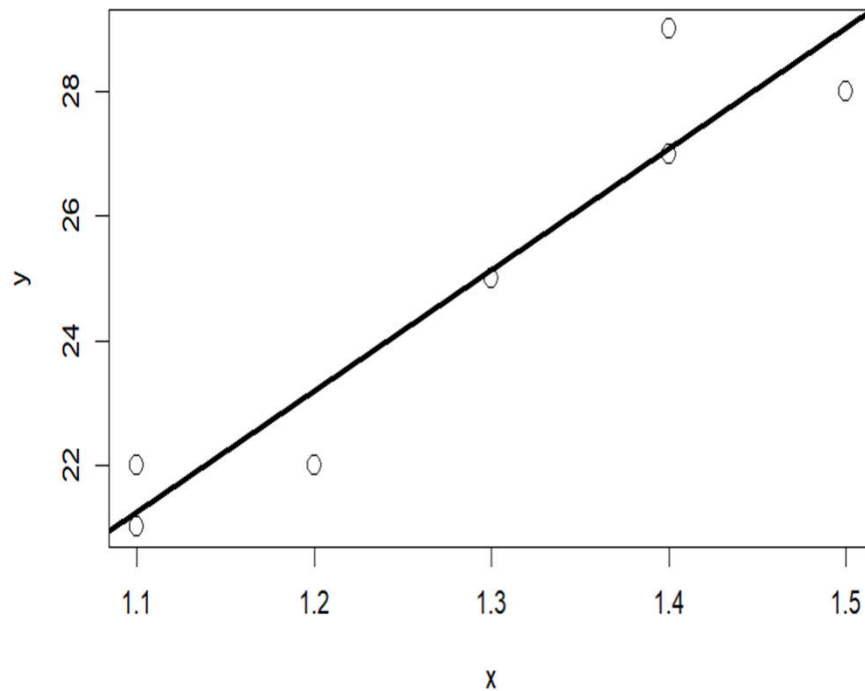
x	y
1,1	11
1,2	12
1,1	11
1,4	14
1,4	14
1,5	15
1,3	13

$$y = 10x$$

x	y
1,1	22
1,2	22
1,1	21
1,4	27
1,4	29
1,5	28
1,3	25

$$\hat{y} = 19,4x - 0.1$$

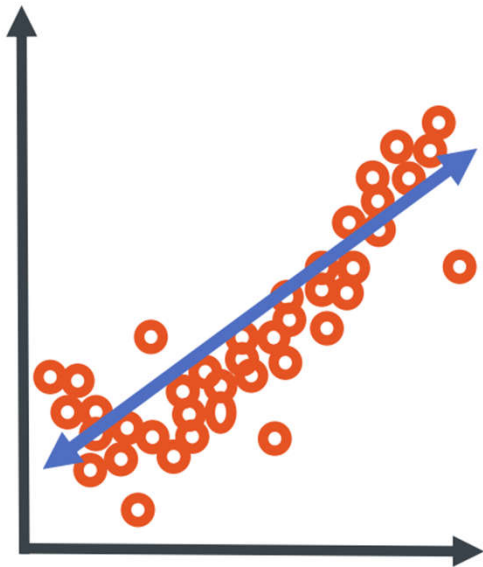
Đường hồi quy



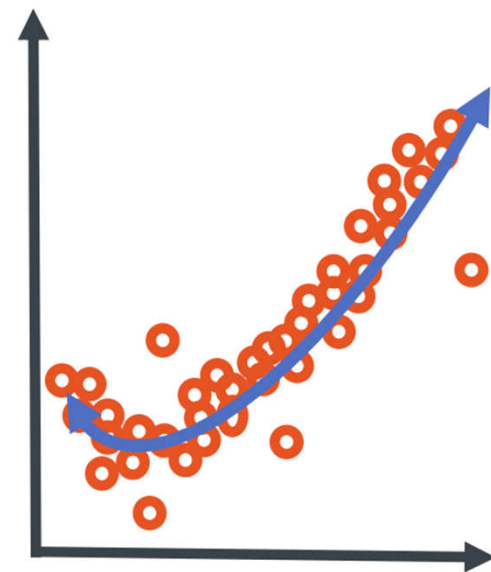
<i>x</i>	<i>y</i>
1,1	22
1,2	22
1,1	21
1,4	27
1,4	29
1,5	28
1,3	25

$$\hat{y} = 19,4x - 0.1$$

Các loại hồi quy



Hồi quy tuyến tính



Hồi quy phi tuyến

Xây dựng mô hình hồi quy tuyến tính

Mô hình hồi quy tuyến tính đơn:

$$\hat{y} = ax + b$$

- y : là biến phụ thuộc
- x : là biến độc lập
- a : là hệ số hồi quy
- b : là hệ số tự do

Ví dụ: Khi ta hồi quy lượng calo bị đốt cháy khi tập thể dục theo thời gian tập luyện thì

- Lượng calo bị đốt cháy là biến phụ thuộc
- Thời gian tập luyện là biến độc lập

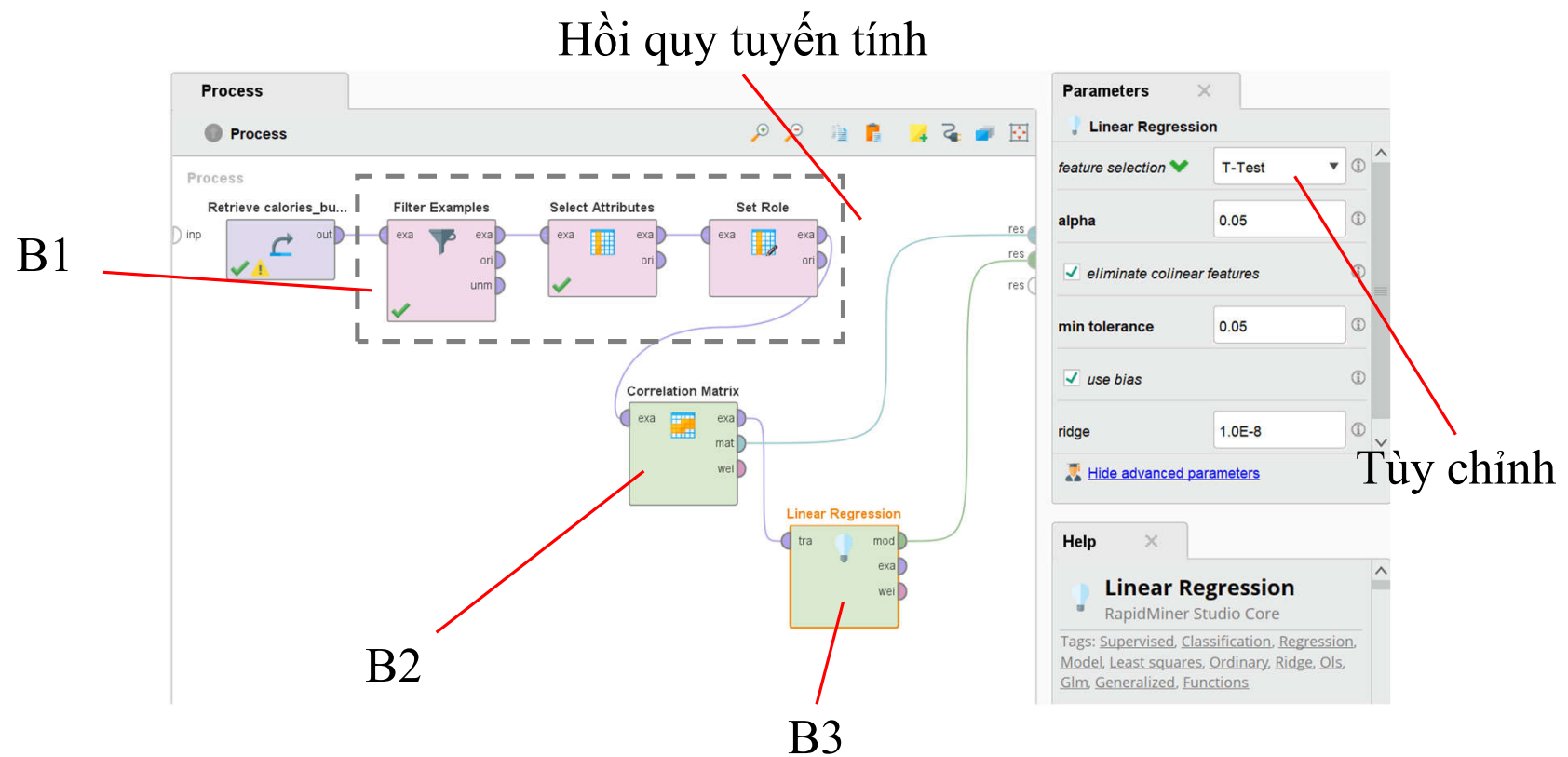
Xây dựng mô hình hồi quy tuyến tính

B1. Tiền xử lý
dữ liệu

B2. Kiểm tra
tương quan
tuyến tính

B3. Xây dựng
mô hình

Các bước thực hiện trên RapidMiner



Độc kết quả

Attribut...	Duration	Calorie...
Duration	1	0.956
Calories...	0.956	1



Tương quan
tuyến tính thuận
rất mạnh



Xây dựng mô hình hồi
quy tuyến tính

Lưu ý: Nếu tương quan tuyến tính giữa hai biến yếu hoặc không có thì ta không cần xây dựng mô hình hồi quy tuyến tính

Đọc kết quả

LinearRegression (Linear Regression) X Correlation Matrix (Correlation Matrix)							
Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Duration	7.156	0.022	0.956	1	323.908	0	****
(Intercept)	-21.610	0.389	?	?	-55.560	0	****

Các hệ số hồi quy

Kiểm định t:

- $p\text{-value} < 0.05$: các hệ số có ý nghĩa
- Ngược lại: các hệ số không có ý nghĩa

Mô hình xây dựng được: $\text{Calories_burnt} = 7.156 * \text{Duration} - 21.610$

Một người tập luyện trong 30 phút thì lượng calo bị đốt cháy được dự đoán là: $7.156 * 30 - 21.610 = 193.07$