ジャックナイフ法を用いた外れ値の検出

2014ss098:吉岡裕輔

指導教員：小藤俊幸

1　はじめに

データ分析を行う際, 多くの場合生データをそのまま分析することはできない.『分析をする際には, データの不正または異常な値を考慮する必要があります. 文献 [1]』このような理由から, データ解析前には元のデータを加工する必要がある.

今回, 生データをそのまま解析し, 相関係数, 散布図を求めた結果とデータを対数変換し, ジャックナイフ法により, 外れ値の検出を行った後の分析結果の比較を行う.『対数をとったデータは, 全体の性質が一部のデータに依存しなくなり, 全体を把握するのに望ましい.(文献 [2])』

2　ジャックナイフ法

ジャックナイフ法とは，データの重複を許さず, 元データから 1 つずつ除いて相関係数を求め，その変化を見ることで，相関係数の安定性を調べることができる手法であり，はずれ値の検出に役にたつ手法である．このとき，あるデータを除いたときの相関係数の値が，他のデータを除いたときに比べて，大きく異なっている場合，そのデータがはずれ値である可能性が高いといえる．(文献 [3],[4])

3　解析データ

No.	speicies	Brain_weight	Body_weight
1	Mountain_Beaver	1.35	8.1
2	Cow	465	423
3	Grey_Wolf	36.33	119.5
4	Goat	27.66	115
5	Guinea_Pig	1.04	5.5
6	Diplodocus	11700	50
7	Asian_Elephant	2547	4603
8	Donkey	187	419
9	Horse	521	655
10	Potar_Monkey	310	115
11	Cat	3.3	25.6
12	Giraffe	529	680
13	Gorilla	207	406
14	Human	62	1320
15	African_Elephant	6654	5712
16	Triceratops	9400	70
17	Rhesus_Monkey	6.8	179
18	Kangaroo	35	56
19	Hamster	0.12	1
20	Mouse	0.023	0.4
21	Rabbit	2.5	12.1
22	Sheep	55.5	175
23	Jaguar	100	157
24	Chimpanzee	52.16	440
25	Brachiosaurus	87000	154.5
26	Rat	0.28	1.9
27	Mole	1.222	3
28	Pig	192	180

表 1　Average body and brain weights for animals

表 1 は 28 種類の動物 (species) に関する，体重 (Body_weight)[kg] と脳の重さ (Brain_weight)[g] の平均値のデータである.(文献 [5]) まず最初に, このデータを加工せず, 体重と脳の重さの相関の強さを知るため, 以下のようなデータ解析を行う.

1　相関係数を求める.

2　散布図を利用して, 視覚的に理解する.

その後, データを対数変換し, 外れ値をジャックナイフ法を用いて, 取り除いて, 散布図を描き, 相関係数を再度求め, 結果を比較する.

4　データ分析

(1.1)　>脳データ<-read.table(”animal.txt”,header=TRUE)

(1.2)　<plot(脳データ$Body_weight, 脳データ$Brain_weight)

(1.3)　<cor(脳データ$Body_weight, 脳データ$Brain_weight)

```
[1]-0.0053
```

相関係数は，[1]-0.0053 となった．このときの相関の強さは，相関がほとんどないといえる．しかし, 元のデータの散布図 (図 1) を見てみると, 左下にデータが密集しているのがわかる. しかし, 右下と左上に一部のデータの値が他より大きい動物データがある. この散布図から, データのばらつきが大きいことがわかるため, データの対数変換を行う,

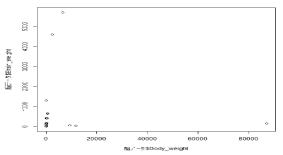

図 1　元のデータの散布図

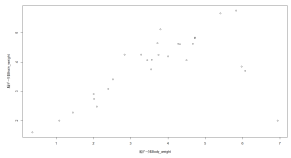

図 2　対数変換後の散布図

No.	speicies	Brain_weight	Body_weight
1	Mountain_Beaver	2	2.908
2	Cow	4.667	4.626
3	Grey_Wolf	3.56	4.077
4	Goat	3.442	4.061
5	Guinea_Pig	2.017	2.74
6	Diplodocus	6.068	3.699
7	Asian_Elephant	5.406	5.663
8	Donkey	4.272	4.622
9	Horse	4.717	4.816
10	Potar_Monkey	4.491	4.061
11	Cat	2.519	3.408
12	Giraffe	4.723	4.833
13	Gorilla	4.316	4.609
14	Human	3.792	5.121
15	African_Elephant	5.823	5.757
16	Triceratops	5.973	3.845
17	Rhesus_Monkey	2.833	4.253
18	Kangaroo	3.544	3.748
19	Hamster	1.079	2
20	Mouse	0.362	1.602
21	Rabbit	2.398	3.083
22	Sheep	3.744	4.243
23	Jaguar	4	4.196
24	Chimpanzee	3.717	4.243
25	Brachiosaurus	1.447	2.279
26	Rat	2.087	2.477
27	Mole	3.279	4.255
28	Pig	6.94	2

表 2　対数変換後のデータ (底 10)

元のデータの散布図 (図 1) と対数変換後の散布図 (図 2) を比較すると, 対数変換後の方が, データのばらつきが少ないことを視認することができる.

元データに他よりも大きい値のデータが含まれるとき, データ解析の結果が大きい一部のデータに依存してしまう. 対数をとることにより, データ全体を把握することができる.

(2.1)　>対数変換<-read.table(”animallog1.txt”,header=TRUE)

(2.2)　<plot(対数変換$Body_weight, 対数変換$Brain_weight)

(2.3)　<cor(対数変換$Body_weight, 対数変換$Brain_weight)

```
[1]0.7676
```

対数変換後の相関係数は,[1]0.7676 となり, この結果から動物の頭と身体の重さには弱い相関があるといえる.

5　外れ値の検出

対数変換後にデータ解析を行った結果, 生データの解析結果に比べて, 散布図はデータのばらつきが小さくなり, 相関係数も弱い相関を示すことが確かめられた.
しかし,1 を見てみると, 一見相関関係が強く直線状にデータがあるように見えるが, 一部の大きい値や小さい値があると視認できる.
これらの外れ値をジャックナイフ法により除外する.

(3.1)　データを一つ削除し，27 種類の動物の相関係数を順に求める．このとき，No.25,Brachiosaurus を削除したとき，相関係数は [1]0.8172 となり，強い相関を示す．他の動物を削除したときに比べ，相関係数が大きく異なるため，これは外れ値といえる．

(3.2)　次に 1 つデータを削除し，26 種類の相関係数を求める．このとき，No.6,Diploducus を削除したときの相関係数は [1]0.8687. これは, 他の動物を削除したときの相関係数に比べ，値が乖離しているので外れ値とみなすことができる．

(3.3)　続いて,25 種類の動物の相関係数を求める. このとき,No.16,Triceratops を削除したときの相関係数は [1]0,9503. これも他の動物のデータを削除した時の相関係数に比べ, 値が乖離しているので, 外れ値と判断することができる.

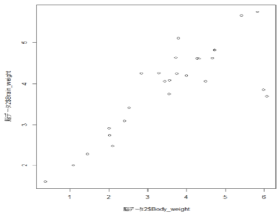

図 3　**(3.1)** 結果

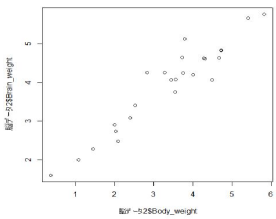

図 4　**(3.3)** 結果

ここまで，3 種類の動物のデータを外れ値として削除した．このときの，相関係数は [1]0.95032. 非常に強い正の相関持つといえる.

(3.4)　最後に,human を取り除いたとき, 相関係数は,[1]0.9626 となる. この後はどの動物データを削除しても, 相関係数に大きな乖離がみられない. 元のデータを対数変換した後,No.25,Brachiosaurus,No.6,Diploducus, No.16,Triceratops,No.14,Human. これら四種類の動物のデータを順に削除したあとの散布図 (3.3) の結果 (図 4) は極端に大きい値や小さい値の外れ値がなく, 視覚的にも相関係数が示す強い相関関係を確認することができる.

最後に, 対数変換を行う前の元のデータにおいて No.25,No.6,No.16,No.14 のデータを削除して, 相関係数を求めると,[1]0.9447 となる.
(3.4) 結果 (図 5) から視覚的にも明らかである.

以上より, 元のデータを加工せずに, 解析すると, 相関係数は [1]-0.005 となり, 無相関であったが, データを対数変換し, ジャックナイフ法により外れ値を検出し, 再度相関係数を求めると [1]0.9447 となり, 動物の頭と身体の重さの間には, 正の強い相関があるといえる.

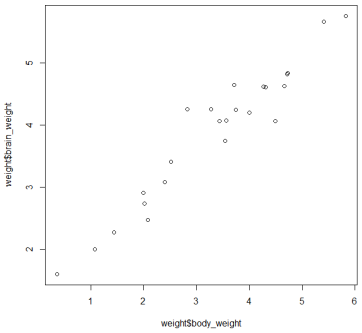

図 5　(3.4) 結果

6　考察

データの外れ値の影響を考慮せず，相関係数の値から解釈すると，相関はほとんどないという結果となった．しかし，元のデータの散布図 (図 1) を見てみると，体重や脳の重さが他と比べ明らかに大きな動物がデータの中に含まれていると気づくことができる．このとき，データのばらつきが大きく, 相関の有無がわからないとき, データを対数変換することが有用である.
今回の場合では, 元のデータの散布図 (図 1) と対数変換後の散布図 (図 2) を比べると, 対数変換後は, 相関関係が強く, 直線状にデータがあるように見える.
このようにデータを対数変換し, 加工した後, ジャックナイフ法により, 外れ値を検出する手順が有効である.

7　おわりに

外れ値の存在を考慮しないデータ解析と外れ値を検出したデータ解析の結果を比較することによって，相関係数の値が大きく変わることがわかった．これより，外れ値の影響は無視することができないため，データ分析を行う際は外れ値の存在について注意を払う必要があるといえる.

参考文献

[1] 越水直人:『データサイエンティスト養成読本』. 株式会社技術評論社, 2017

[2] 牧允皓:『データサイエンティスト養成読本』. 株式会社技術評論社,2017

[3] 岩沢宏和：『世界を変えた確率と統計のからくり』. SB Creative 株式会社, 2014

[4] 松原望: 『統計学１００のキーワード』. 弘文堂, 2014

[5] 山田剛史・村澤武俊・村井潤一郎:『R によるやさしい統計学』. オーム社,2008