計数工学実験 (統計手法) レポート

Rによる多変量データ解析
— グラフィカルモデリング —

学籍番号 XX-XXXXXX
氏名: 佐藤 瞭

概 要

アメリカ合衆国の州ごとの人口やペット (犬・猫) のデータについて，基本統計量や相関行列などを求めた．また，3 つ以上の変数の関連を視覚的に捉える為に，共分散選択アルゴリズムを用いたグラフィカルモデリングとして，データに対応した無向独立グラフを導出した．得られた結果に基づいて，データの背景を考察した．

1 データ解析の目的

ペットを飼っている人は，ペットとして飼う動物を選択するときに，どのような要因に影響されているか知りたい，という問題がある．具体的には，犬を飼っている人の割合が高い地域には，人口や面積にどのような特徴があるか知りたい，といった例が挙げられる．今回は, アメリカ合衆国の州ごとに，犬や猫を飼っている人の割合と人口・面積・人口密度の関連を, フリーの統計解析ソフト「R」を用いて解析する．

2 つの変数の関係については，相関を求めることでその傾向を知ることができる．しかし, 3 つ以上の変数の傾向をとらえるには，それぞれの相関を見るだけ不十分である．例えば，人口密度の値を固定したときに, 犬を飼っている人の割合と猫を飼っている人の割合の相関がどのようになるかはわからない．3 つ以上の変数の「絡み」をとらえるために，今回は共分散選択アルゴリズムによるグラフィカルモデリングを用いる．これにより，それぞれの変数の「絡み」を無向独立グラフに表し，視覚的にとらえることができる．

基本統計量や相関行列, グラフィカルモデリングで得られた結果をもとに, それぞれから読み取れることの違いに留意しながら, 各変数の関連を探る．

2 データ解析

アメリカ合衆国で，ペットとして，淡水魚に次いで最も多く飼われている [1] 動物である犬と猫のデータと，州ごとの人口・人口密度・面積のデータを解析する．

犬と猫のデータは data.world[2] から取得した．人口などのデータは World Population Review[3] から取得した．2 つのテーブルを地名で結合し，アラスカとハワイを除く 48 の州，及びワシントン D.C. のデータを得た．データの変数は 7 個, データ数は 49 行である．

変数名，変数の説明，表記の簡便さのための変数名の略記を表 1 に記す．

2.1 基本統計量

データに関して，R の関数 `summary` を利用して得た要約統計量を表 2 にまとめた．

表 1: 変数名の略記，変数名，変数の説明

略記	変数名	説明
S1	Density	人口密度 (人 / 平方マイル)
S2	Pop_2019	人口 (人，2019 年)
S3	Area	面積 (平方マイル)
S4	Dog.Owners.percentage	犬を飼っている人の割合 (%)
S5	Mean.Dogs.per.household	犬を飼っている世帯での犬の平均頭数
S6	Cat.Owners.percentage	猫を飼っている人の割合 (%)
S7	Mean.Cats.per.household	猫を飼っている世帯での猫の平均頭数

表 2: 要約統計量

	S1	S2	S3	S4	S5	S6	S7
Min.	6	573720	61	13.1	1.10	11.6	1.70
1st Qu.	55	1932549	35826	32.9	1.40	29.0	1.90
Median	108	4682509	53625	36.6	1.60	31.3	2.00
Mean	438	6650367	60303	37.0	1.59	31.6	2.04
3rd Qu.	230	7530552	79627	42.5	1.70	33.8	2.20
Max.	11535	39776830	261232	47.9	2.10	49.5	2.60

2.2 相関行列と対散布図

成績データの相関行列を表 3 に，偏相関行列を表 4 に示す．相関行列と偏相関行列を区別するために，ここでは偏相関行列の対角成分は — で表すという慣習にしたがう．また，ともに対称行列であるため，右上成分は示していない．

データの対散布図を図 1 に示す．

表 3: データから直接計算した相関行列.

	S1	S2	S3	S4	S5	S6	S7
S1	1.0000						
S2	-0.0884	1.0000					
S3	-0.2612	0.4419	1.0000				
S4	-0.5962	-0.0638	0.3986	1.0000			
S5	-0.4252	0.0478	0.3911	0.7919	1.0000		
S6	-0.5458	-0.2379	-0.0243	0.3767	0.1639	1.0000	
S7	-0.1478	0.0154	0.1290	0.4171	0.4534	-0.1064	1.0000

表 4: データから直接計算した偏相関行列.

	S1	S2	S3	S4	S5	S6	S7
S1	———						
S2	-0.2634	———					
S3	0.0050	0.4641	———				
S4	-0.3346	-0.2666	0.2865	———			
S5	0.0273	0.0788	0.0617	0.652	———		
S6	-0.4594	-0.2232	-0.1094	0.211	-0.1141	———	
S7	-0.0087	0.0169	-0.1270	0.221	0.1706	-0.261	———

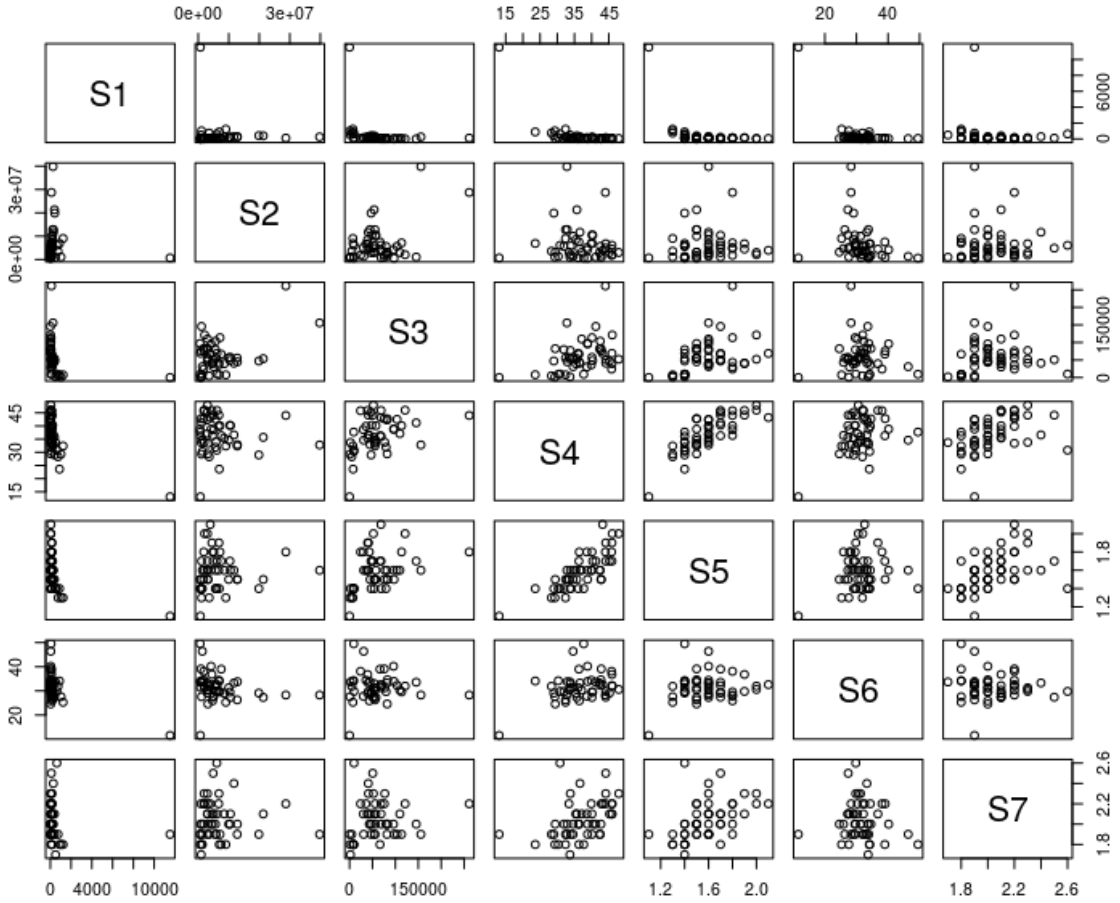

図 1: データの対散布図.

相関行列，偏相関行列や対散布図からデータに関して分かることとして，たとえば以下のようなことが挙げられる．

- 犬を飼っている人の割合と，犬を飼っている世帯での犬の平均頭数には，やや強い正の相関がある．
- 相関行列をみると，面積と犬を飼っている世帯での犬の平均頭数には正の相関があるように見えるが，偏相関係数をみるとほとんど相関がないので，他の変数が絡んでいることが予想される．
- 面積と犬を飼っている人の割合には弱い正の相関があるが，面積と猫を飼っている人の割合にはほぼ相関がない．

2.3 共分散選択

成績データに関して, 共分散選択を行なったところ, 28 個の偏相関係数のうち 11 個をゼロとおいたモデルが選択された．選択されたモデルでの相関行列を表 5 に，偏相関行列の推定値を表 6 に示す．ゼロとおいた偏相関係数には下線をつけてある，また，除去される辺の順番と AIC の変化を表 7 に示す．さらに, 対応する無向独立グラフを図 2 に示す.

表 5: 選択されたモデルでの相関行列の推定値.

	S1	S2	S3	S4	S5	S6	S7
S1	1.0000						
S2	-0.0884	1.0000					
S3	-0.2970	0.4419	1.0000				
S4	-0.5962	-0.0638	0.3986	1.0000			
S5	-0.4540	-0.0395	0.3208	0.7919	1.0000		
S6	-0.5458	-0.2379	0.0238	0.3160	0.2045	1.0000	
S7	-0.1117	0.0299	0.1542	0.3270	0.4534	-0.1064	1.0000

表 6: 選択されたモデルでの偏相関行列の推定値.

	S1	S2	S3	S4	S5	S6	S7
S1	——						
S2	-0.271	——					
S3	0	0.482	——				
S4	-0.363	-0.230	0.284	——			
S5	0	0	0	0.656	——		
S6	-0.516	-0.300	0	0	0	——	
S7	0	0	0	0	0.329	-0.184	——

表 7: 除去される辺の順番と AIC の変化.

反復回数	1	2	3	4	5	6
除去される辺	(S3, S1)	(S7, S1)	(S7, S2)	(S5, S1)	(S5, S3)	(S5, S2)
AIC	-2.00	-3.99	-5.98	-7.94	-9.76	-11.02
AIC の変化	——	-1.99	-1.99	-1.96	-1.82	-1.26

反復回数	7	8	9	10	11
除去される辺	(S7, S3)	(S6, S3)	(S6, S5)	(S6, S4)	(S7, S4)
AIC	-12.35	-13.81	-14.17	-15.14	-15.67
AIC の変化	-1.33	-1.46	-0.36	-0.97	-0.53

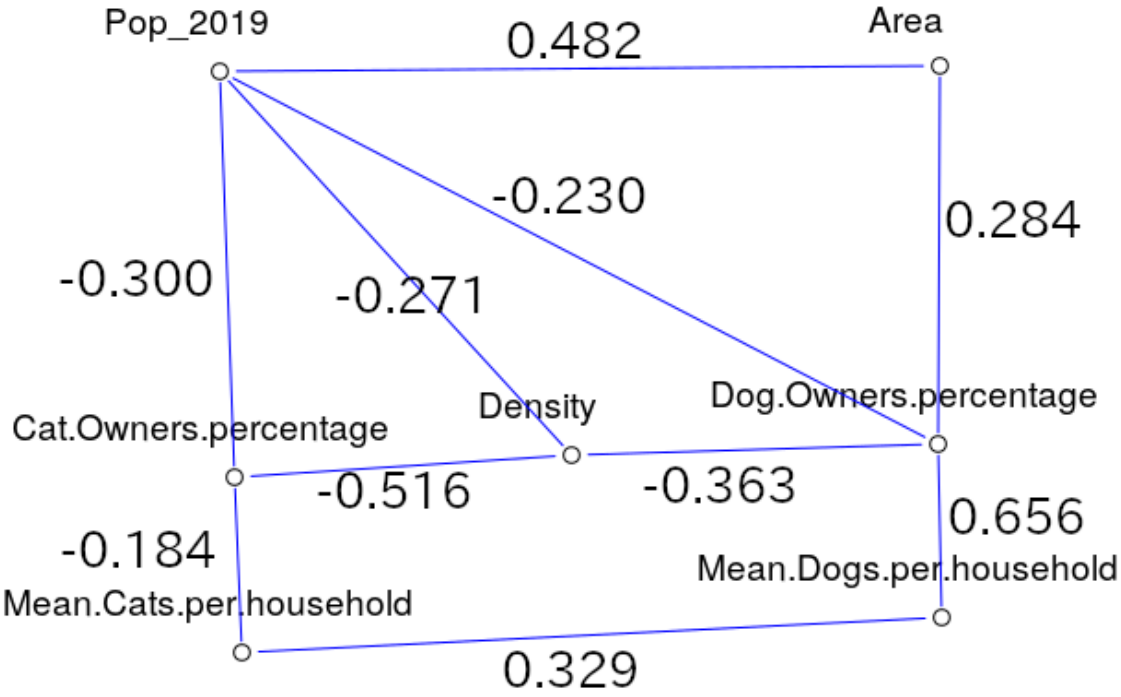

図 2: 成績データの無向独立グラフ (辺の上の数字は偏相関係数である) .

3　考察

まずは，グラフィカルモデリングに対応した無向独立グラフ（図 2）からわかることを考察する．任意の 2 頂点について，2 つを結ぶときに必ず通らなければならない点は存在しないため，分離定理を適用できるような頂点集合の組は存在しないことがわかる．そのため，この無向独立グラフからは，条件付き独立な変数の組は見つからない．そのため，人口・人口密度・面積といった情報のうち，犬や猫のうちどちらか一方の所有率や所有数に影響する変数はないことがわかる．

ここで，猫に関する変数と，犬に関する変数の，グラフ上での違いを 2 点に絞って考える．

1. まず，(猫の所有率，面積) の辺は除去されているのに対して，(犬の所有率，面積) の辺は除去されていない点が挙げられる．実際に犬や猫を飼う場面を思い浮かべてみると，犬を飼う場合は，ドッグランのような屋外の開けた場所で遊ばせる場面が想定でき，ある程度広い土地がある地域のほうが犬を遊ばせやすいことが考えられる．実際，(犬の所有率，面積) の偏相関係数は正の値になっていることから，土地が広い地域では犬を飼うことに対して積極的であるのではないか，という仮説を立てることができる．一方で，猫を飼う場合は，家の中で飼ってあまり外に出さないようなケースが犬よりも多い (日本の調査 [4] ではこのような傾向がみられる) ために，犬の場合ほど広い土地を必要としない，という説明が考えられる．

2. つぎに，(猫の所有率，猫を飼っている世帯での猫の平均頭数) には偏相関がほぼないのに加えて偏相関係数が負になっているのに対して，(犬の所有率，犬を飼っている世帯での犬の平均頭数) にはやや強い正の偏相関がある点が挙げられる．犬を飼っている世帯が多い地域では犬を飼っている頭数が多くなる傾向にあるが，猫ではこのような傾向がみられない背景として，飼い主同士の接触の機会の多寡が影響している可能性が考えられる．犬を飼っている場合は散歩などで外に出る機会がある．その際に，犬の所有率が高い地域であれば他の飼い主と遭遇する確率も高くなる．一方，猫を飼っている場合は，家の中で飼っていてあまり外に連れ出さない場合が犬よりも多く [4]，飼い主同士の接触の機会は，犬の場合と比べて少ないと予想される．仮に，飼い主同士の接触が多いと，飼い主はより多くのペットを飼いたくなる，という傾向があるとすれば，今回発見した傾向の説明はつく．飼い主同士の接触の多寡とペットを飼っている頭数の関係を新たに調べることで，この仮説は検証できるだろう．

他の特徴的な点として，(猫を飼っている世帯での猫の平均頭数，犬を飼っている世帯での犬の平均頭数) の辺がグラフに残っており，この辺が残っているために，この変数対が条件付独立でなくなっている点が挙げられる．偏相関係数をみると 0.329 となっており，弱い正の偏相関があることが確認できる．この背景として，動物の種類に限らず，周囲にペットの所有者が多いと自分も飼いたくなる，といった同調効果の可能性が挙げられるほか，ペットの所有率が高い地域はペットを飼い始める障壁が低い環境 (たとえば，予防接種や去勢の費用を自治体が負担してくれる，といった金銭面の環境) にある，といった仮説が考えられる．

つぎに，AIC の変化について考察する．まず AIC の定義は，

$$\mathrm{AIC} = -2 \times (\text{最大対数尤度}) - 2 \times (\text{制約式の個数}) + (\text{定数})$$

であった．つまり，AIC はモデルのデータの当てはまりが良いほど，また，グラフの辺が少ないほど大きくなる．また，グラフの辺を 1 つ減らすたびに，$-2 \times$ (制約式の個数) の項は -2 ずつ変化していく．そのため，AIC の変化量に 2 を足した値が，制約を増やすことによって犠牲になった，モデルの当てはまりの良さであると解釈できる．このことを踏まえた上で，表 7 の AIC の変化を見てみる．反復回数が大きくなるにつれて，AIC の変化量は概ね小さくなっている．すなわち，反復回数が大きくなるにつれて，制約を増やすために，モデルの当てはまりの良さをより大

きく犠牲にしていることが確認できる．他の除去と比べて AIC の変化の大きさが大きい例として，2 回目で除去される辺 (S7，S1) が挙げられる．この除去による AIC の変化は-1.99 であり，この辺はモデルから取り除いても当てはまりの良さはほとんど失われないため，制約を増やすために早い反復回数で除去されたと解釈できる．この辺に対応する変数対は，猫を飼っている世帯での猫の平均頭数と人口密度である．実際，最終的に得た無向独立グラフでは，両者は猫の所有率という変数を間に挟んで結ばれており，猫を飼っている世帯での猫の平均頭数と人口密度の相関は別の変数によるものであったといえる．

参考文献

[1] Number of pets in the United States in 2017/2018, by species (in millions), https://www.statista.com/statistics/198095/pets-in-the-united-states-by-type-in-2008/, 2019/04/22 閲覧.

[2] Cat vs. Dog Popularity in U.S.，https://data.world/datanerd/cat-vs-dog-popularity-in-u-s, 2019/04/22 取得.

[3] US States by Density 2019，http://worldpopulationreview.com/states/state-densities/, 2019/04/22 取得.

[4] 平成 30 年　全国犬猫飼育実態調査「主要指標　サマリー」，ペットフード協会，https://petfood.or.jp/data/chart2018/3.pdf，p.21，2019/04/25 閲覧.