TRADの進歩

データサイエンスコンソーシアム，慶應義塾大学

柴田　里程

TextilePlot, R and DandD

- データサイエンスの理論と実践の間の架け橋
  - データの視覚表示 (GUI)
  - データの総体的な理解を助け，コミュニケーションを円滑にする
  - R による計算 (CUI)
- データ
  - 楽しく
  - 創造的
  - 魅力的
  - 身近に

見えてきた統計手法の限界（例）

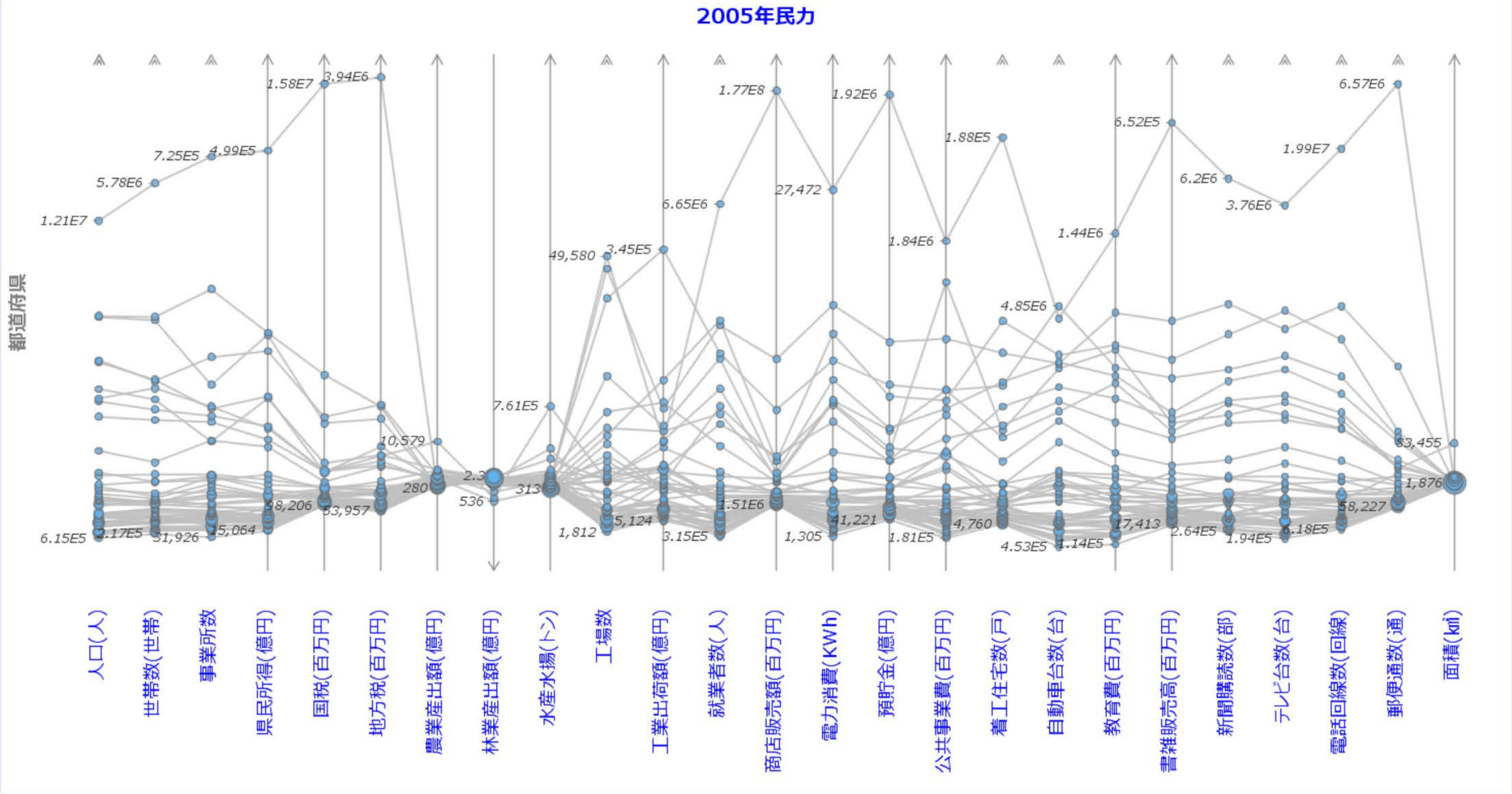
- 主成分分析
  - 2階から目薬
  - 主成分の意味不明

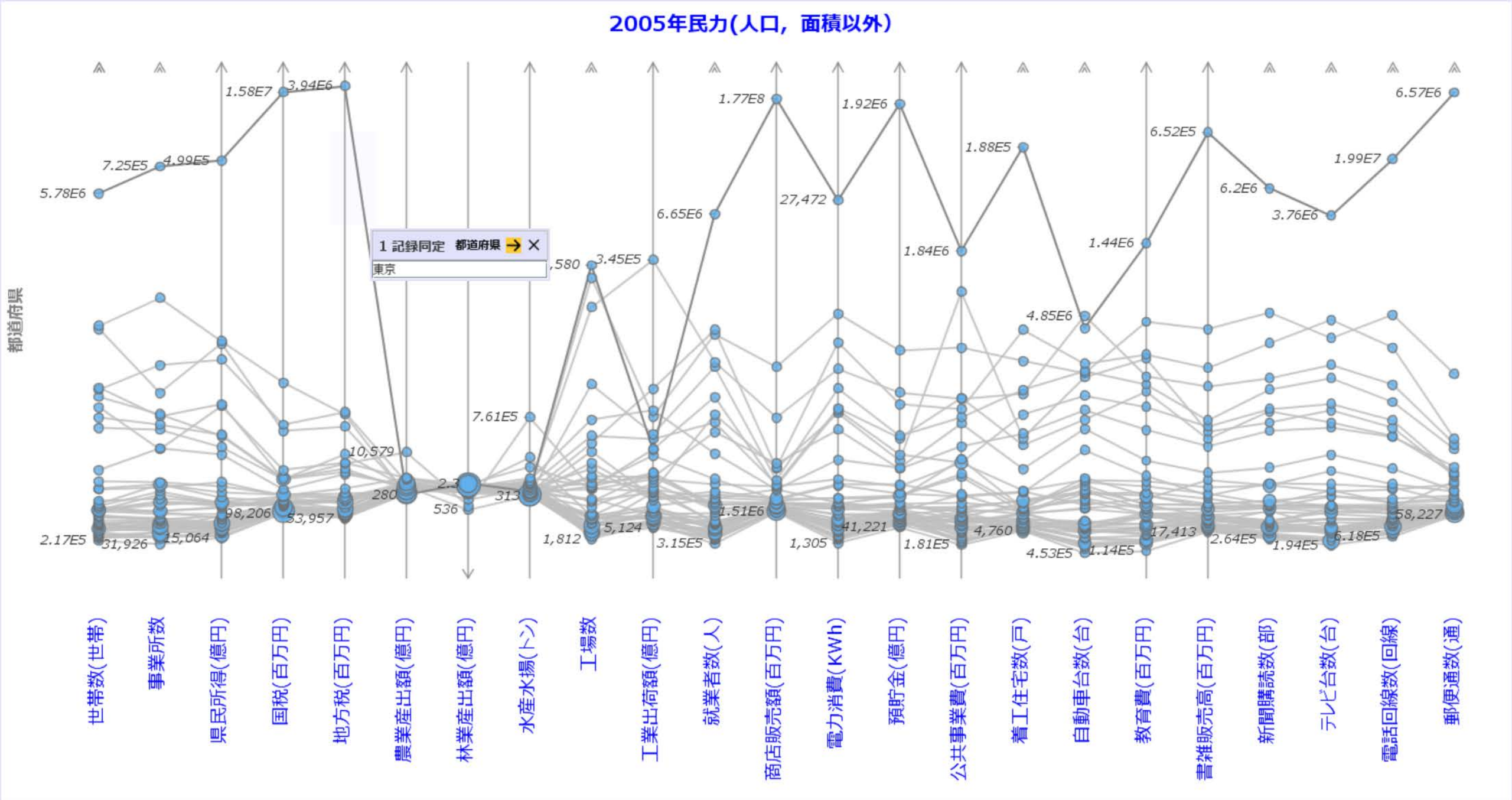
- ロジットモデル
  - ロジット変換の有効性？
  - 本当の姿を見落としている

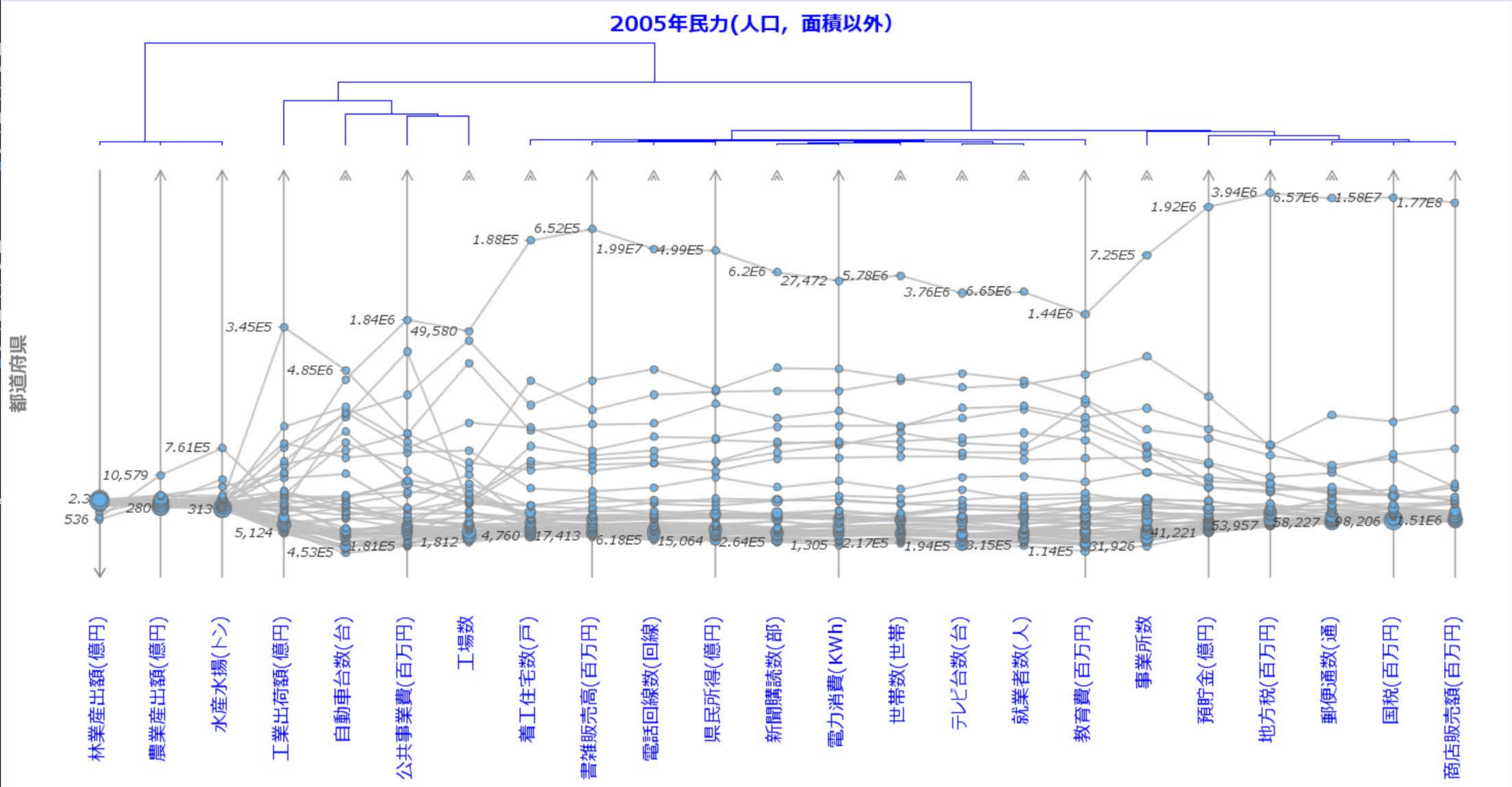
民力データ

主成分分析

	人口	世帯数	事業所数	県民所得	国税	地方税	農業産出額	林業産出額	水産水揚	工場数	工業出荷額	就業者数	商店販売額	電力消費	預貯金	公共事業費
北海道	5650573	2522295	270504	145293	1379098	546638	10579	463.7	761331	10668	51199	2796200	2.02E+07	11256	336487	1611254
青森	1479358	551806	74341	32498	242730	131442	2402	88.2	107727	3188	10997	731000	3693933	2671	70174	324170
岩手	1405060	488354	72456	34152	224269	123476	2587	193.4	108752	4126	20209	715500	3525821	2631	76438	339411
宮城	2350026	856527	115297	61092	735121	245372	1870	76.1	373037	5782	32677	1169800	1.09E+07	4396	125693	391923
秋田	1173722	410308	65300	27294	186373	97512	2208	123.2	5599	4130	11731	571000	2714120	2152	56790	328282
山形	1225990	387732	70523	29851	202270	107894	2349	62.5	4116	5971	26287	631400	2968623	2327	67083	363756
福島	2116210	716505	109652	56554	390824	210846	2640	149.7	51131	8494	49706	1052500	4898557	3922	113222	466686
茨城	2991804	1039865	135383	86778	746384	320761	4194	79.7	173448	11160	96742	1555900	6574412	5721	190256	450254
栃木	2006717	701919	103835	61184	425157	225878	2786	118.3	11282	10953	74427	1049600	5646460	3931	128826	382702
群馬	2022780	719576	109637	59127	464797	209336	2210	131.3	3843	12898	69710	1063500	5362438	4086	147553	333528
埼玉	6980889	2660152	266775	186180	1355243	633293	2004	28	14422	28809	124108	3711500	1.50E+07	13131	445037	903250
千葉	6001032	2348339	206793	184026	1704591	555876	4319	23.7	281435	10820	105257	3130600	1.23E+07	11438	360828	646053
東京	1.21E+07	5776805	724769	498507	1.58E+07	4E+06	280	7.4	45242	49580	104363	6653800	1.77E+08	27472	1915951	1841386
神奈川	8600109	3602950	309441	264122	3160545	880703	751	8.3	36807	18475	181354	4433600	2.00E+07	16612	584015	948689
新潟	2455996	810483	142123	66880	530773	242825	3281	328.5	73953	14068	40669	1261400	7330619	4694	172060	664364
富山	1118661	367754	64734	33114	318907	116521	835	24.4	17797	5741	32461	593600	3305066	2465	97496	278946
石川	1175071	417164	72638	33792	317154	125988	684	28.5	29363	8829	21648	632900	4280880	2652	92795	278602
福井	824824	260744	52855	23962	179464	98482	597	24.9	9535	6217	15560	446100	2315651	1915	71145	194788
山梨	882678	319146	52789	22807	187752	93218	823	16.3	3979	5417	21364	465000	1928163	1882	64533	245716
長野	2200896	777553	128969	60262	488468	224361	2425	536	2384	12478	53691	1189800	6464420	4768	180124	582792
岐阜	2106917	701408	122425	58868	443046	209845	1239	125.6	575	17084	44852	1110100	5234017	4349	179226	563226
静岡	3773140	1347330	207923	121941	985478	452095	2582	145.6	151944	22593	153746	2042800	1.13E+07	7685	304204	867838
愛知	7027499	2634915	360358	243703	3401779	1E+06	3259	42.9	93271	42281	344743	3808400	4.15E+07	14596	637306	1007471
三重	1857773	672654	93292	55070	576682	208908	1266	96.1	128672	8372	75093	971300	3828670	3906	143158	415260


2005年民力
都道府県
人口(人)
世帯数(世帯)
事業所数
県民所得(億円)
国税(百万円)
地方税(百万円)
農業産出額(億円)
林業産出額(億円)
水産水揚(トン)
工場数
工業出荷額(億円)
就業者数(人)
商店販売額(百万円)
電力消費(KWh)
預貯金(億円)
公共事業費(百万円)
着工住宅数(戸)
自動車台数(台)
教育費(百万円)
書雑販売高(百万円)
新聞購読数(部)
テレビ台数(台)
電話回線数(回線)
郵便通数(通)
面積(㎢)
6.15E5
1.21E7
5.78E6
7.25E5
4.99E5
1.58E7
3.94E6
2.17E5
31,926
45,064
98,206
53,957
10,579
280
2.3
536
7.61E5
313
49,580
3.45E5
1,812
5,124
6.65E6
3.15E5
1.77E8
1.51E6
27,472
1,305
1.92E6
41,221
1.84E6
1.81E5
1.88E5
4,760
4.85E6
4.53E5
1.44E6
1.14E5
6.52E5
17,413
6.2E6
2.64E5
3.76E6
1.94E5
1.99E7
6.18E5
6.57E6
58,227
83,455
1,876


2005年民力(人口，面積以外)
都道府県
1 記録同定 都道府県
東京
世帯数(世帯)
事業所数
県民所得(億円)
国税(百万円)
地方税(百万円)
農業産出額(億円)
林業産出額(億円)
水産水揚(トン)
工場数
工業出荷額(億円)
就業者数(人)
商店販売額(百万円)
電力消費(KWh)
預貯金(億円)
公共事業費(百万円)
着工住宅数(戸)
自動車台数(台)
教育費(百万円)
書雑販売高(百万円)
新聞購読数(部)
テレビ台数(台)
電話回線数(回線)
郵便通数(通)

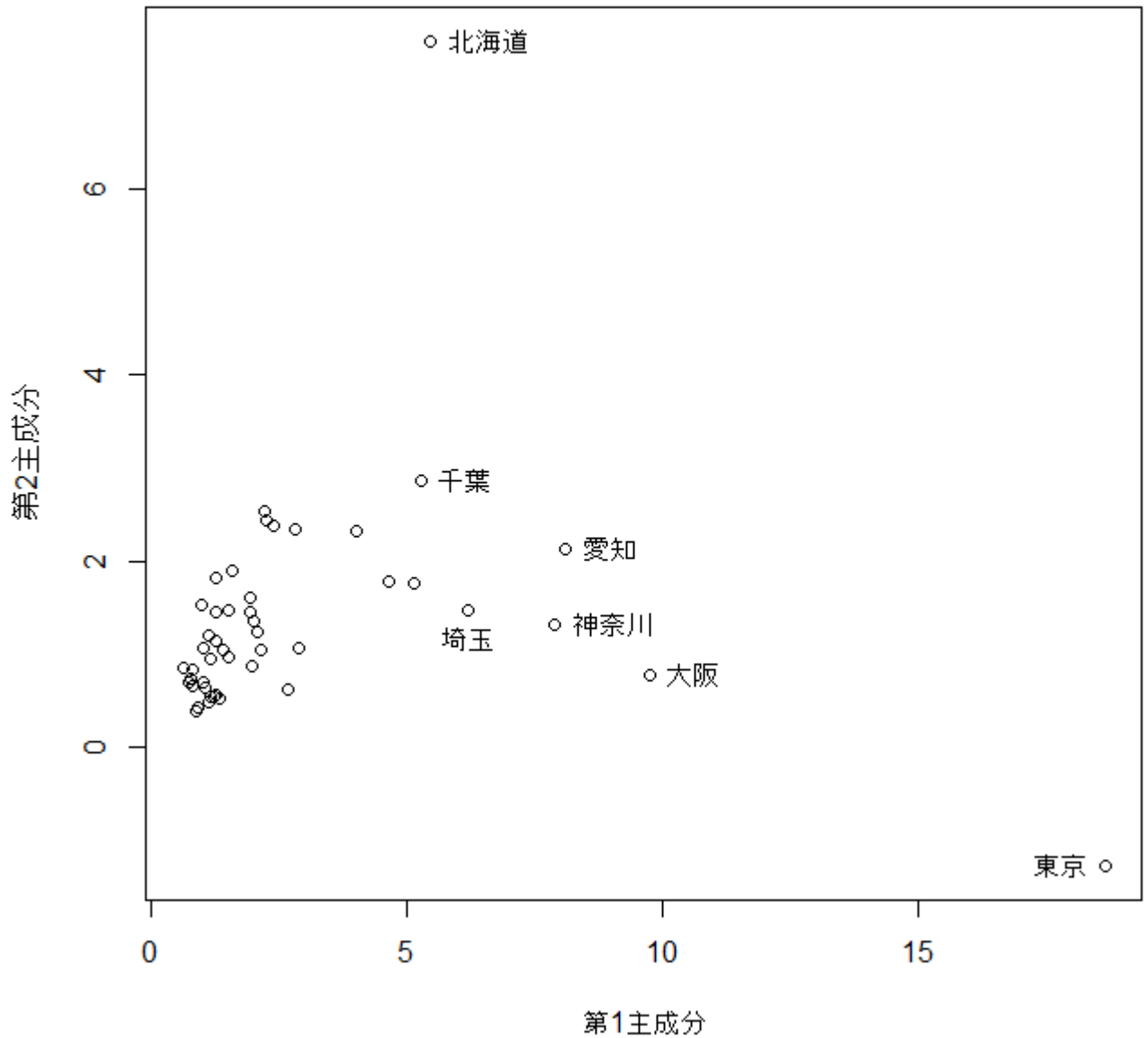

2005年民力(人口，面積以外)
都道府県
林業産出額(億円)
農業産出額(億円)
水産水揚(トン)
工業出荷額(億円)
自動車台数(台)
公共事業費(百万円)
工場数
着工住宅数(戸)
書雑販売高(百万円)
電話回線数(回線)
県民所得(億円)
新聞購読数(部)
電力消費(KWh)
世帯数(世帯)
テレビ台数(台)
就業者数(人)
教育費(百万円)
事業所数
預貯金(億円)
地方税(百万円)
郵便通数(通)
国税(百万円)
商店販売額(百万円)
2.3
536
10,579
280
7.61E5
313
3.45E5
5,124
4.85E6
4.53E5
1.84E6
1.81E5
49,580
1,812
1.88E5
4,760
6.52E5
17,413
1.99E7
6.18E5
4.99E5
15,064
6.2E6
2.64E5
27,472
1,305
5.78E6
2.17E5
3.76E6
1.94E5
6.65E6
3.15E5
1.44E6
1.14E5
7.25E5
31,926
1.92E6
41,221
3.94E6
53,957
6.57E6
58,227
1.58E7
98,206
1.77E8
1.51E6

主成分

```
[[1]]$寄与率
[1] 0.7807595

[[1]]$主成分
      鳥取       高知       島根       佐賀       徳島       福井       山梨       宮崎     和歌山       秋田       香川       沖縄
0.6148082  0.7386977  0.7586514  0.7884993  0.7936668  0.8583321  0.9144802  0.9827039  1.0125829  1.0236479  1.0537918  1.1072056
      大分       山形       富山       奈良       岩手       青森       石川       長崎       滋賀       愛媛       山口       熊本
1.1219095  1.1397622  1.1615443  1.2359736  1.2435153  1.2615485  1.2649266  1.2710743  1.3371064  1.4129386  1.4934078  1.5203431
    鹿児島       三重       福島       岡山       栃木       群馬       岐阜       長野       宮城       新潟       京都       茨城
1.5759422  1.9163512  1.9197410  1.9804242  2.0117736  2.0622448  2.1375787  2.2037674  2.2555092  2.3816837  2.6693761  2.8204758
      広島       静岡       福岡       兵庫       千葉     北海道       埼玉     神奈川       愛知       大阪       東京
2.8983628  4.0232546  4.6346197  5.1423200  5.2658840  5.4726629  6.2076668  7.8899678  8.1061165  9.7458839 18.6664942

[[1]]$変量重み
  foresty.out      agri.out   fishery.out  industry.out           car   public.work     education       factory            TV
-0.02243990    0.01272416    0.04173623    0.14833352    0.15736979    0.16587157    0.17942387    0.20031048    0.20536288
    employee        Epower       company     household     newspaper   pref.income        phones    book.sales       housing
  0.21088505    0.21284453    0.21315032    0.22170870    0.22318191    0.23340019    0.23561613    0.24566659    0.24607859
     deposit     local.tax merchan.sales          mail       gov.tax
  0.25195435    0.25891594    0.26576539    0.26583356    0.26820185
```


北海道
千葉
愛知
神奈川
埼玉
大阪
東京
第2主成分
第1主成分
0
2
4
6
0
5
10
15

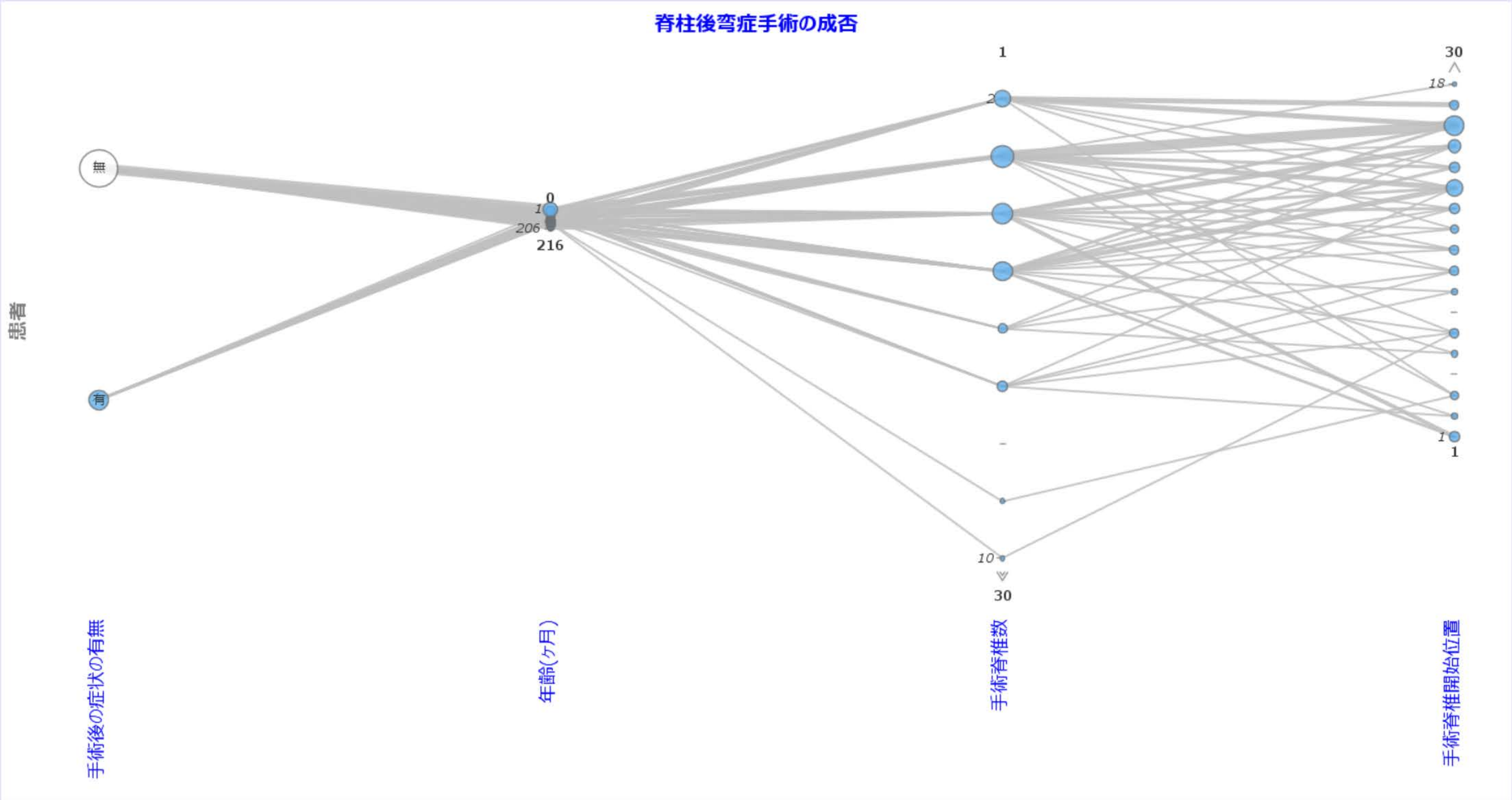
主成分分析の限界

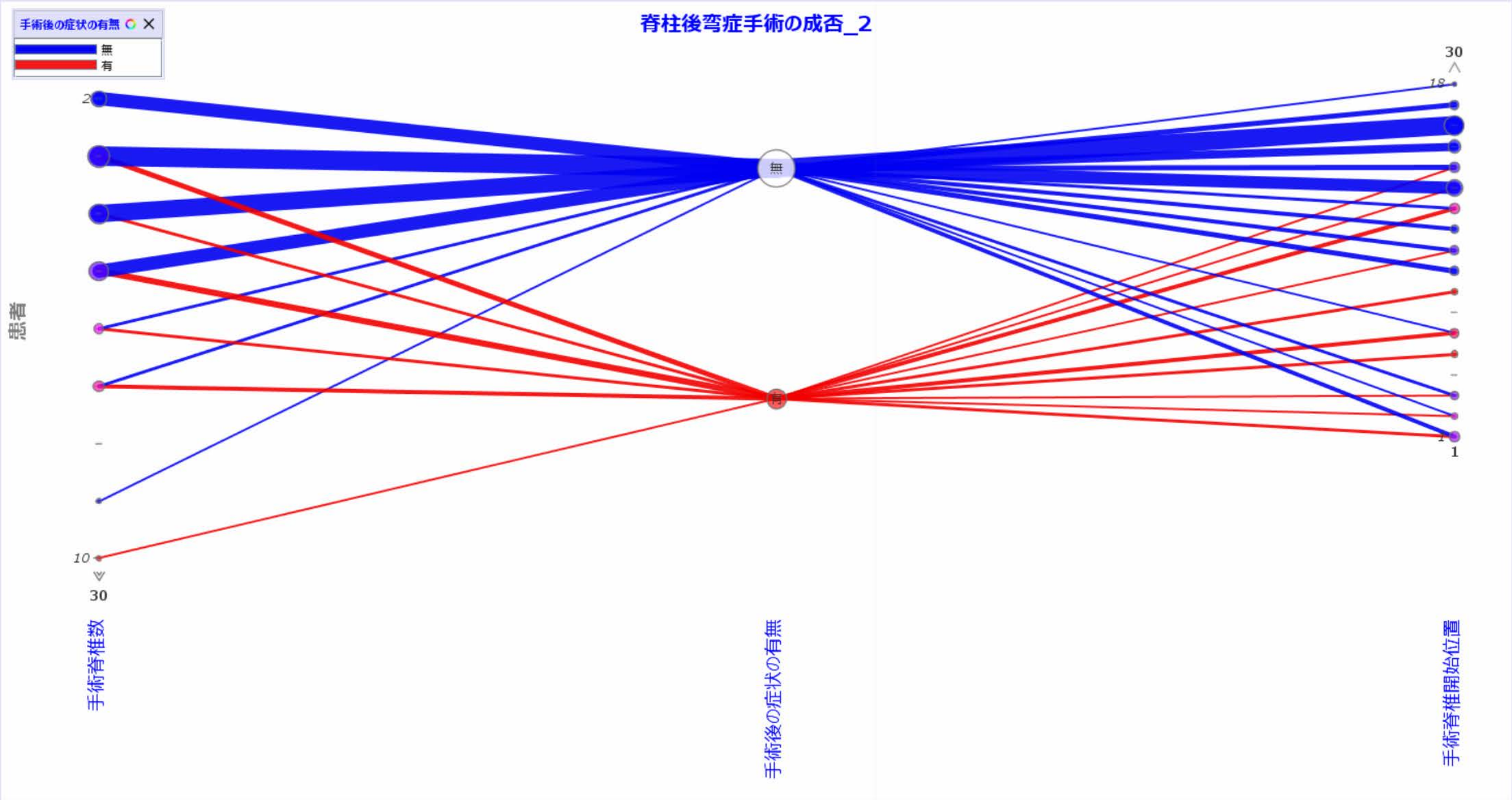
- 計算機資源が限られていた時代の遺産
- 意味を考えない線形代数の応用
- 非数値データが混じっているときには適用不可
- 行き止まり，両手を縛られたような状況に陥る

- 視覚表示からわかることのほうが，ずっと豊富で直接的

脊柱後弯症データ

ロジットモデル


脊柱後弯症手術の成否
患者
無
有
手術後の症状の有無
0
1
206
216
年齢(ヶ月)
1
2
10
30
手術脊椎数
30
>
18
1
1
手術脊椎開始位置


脊柱後弯症手術の成否_2
手術後の症状の有無
無
有
患者
2
10
30
手術脊椎数
無
手術後の症状の有無
30
18
1
手術脊椎開始位置

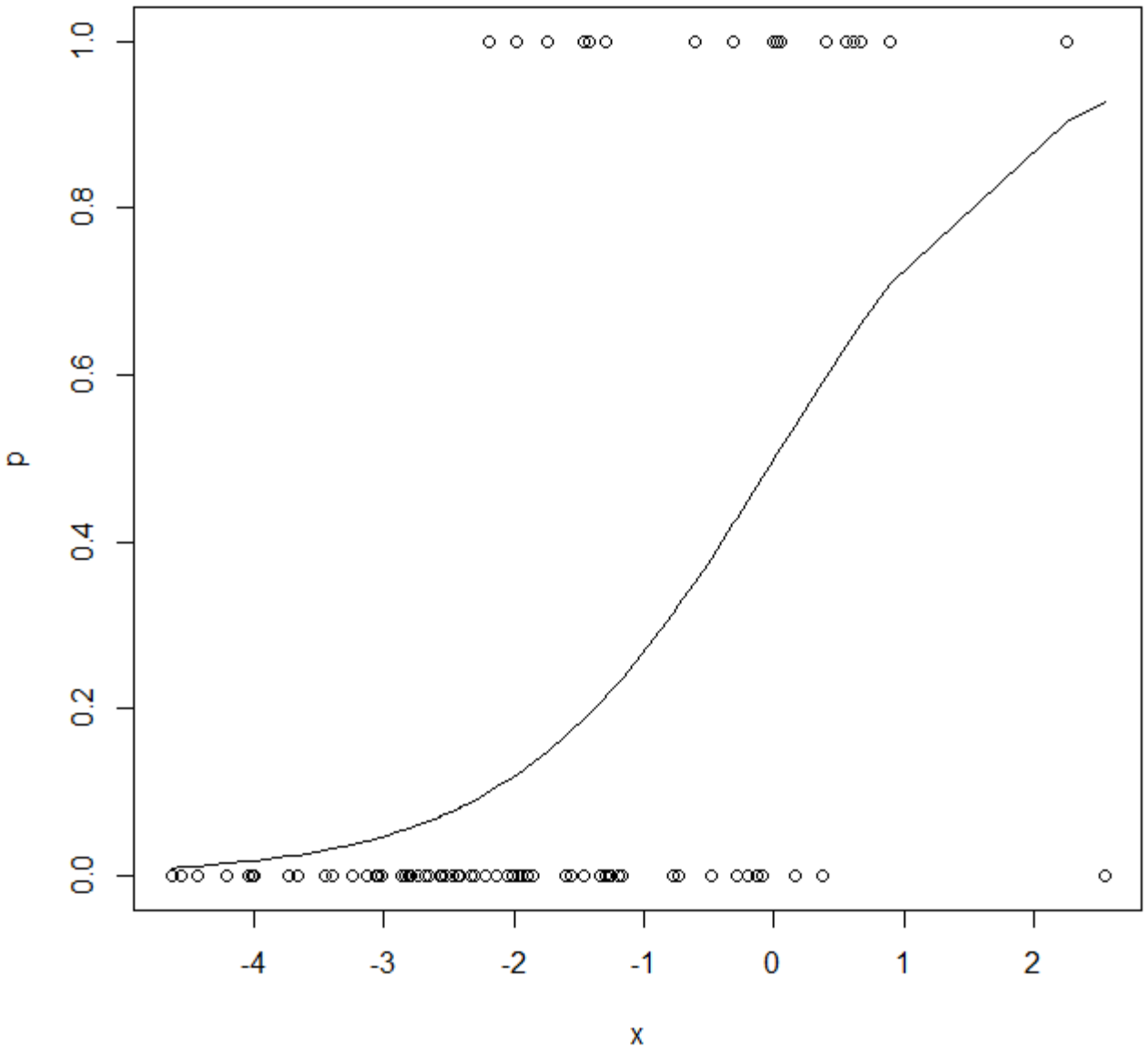
```
Call:  glm(formula = Kyphosis ~ Age + Start + Number, family = binomial, data = kyphosis)

Coefficients:
(Intercept)          Age        Start       Number
   -1.82525      0.01067     -0.21150      0.38558

Degrees of Freedom: 79 Total (i.e. Null);  76 Residual
Null Deviance:      82.76
Residual Deviance: 60.76        AIC: 68.76

Analysis of Deviance Table
       Df Deviance Resid. Df Resid. Dev
NULL                      79     82.760
Age     1   1.2627        78     81.497
Start   1  17.3477        77     64.150
Number  1   3.3878        76     60.762
```

$$p = \frac{e^x}{1+e^x} : \text{症状の残る確率},\ x = -1.8525 + 0.01067 * Age - 0.21150 * Start + 0.38558 * Number$$

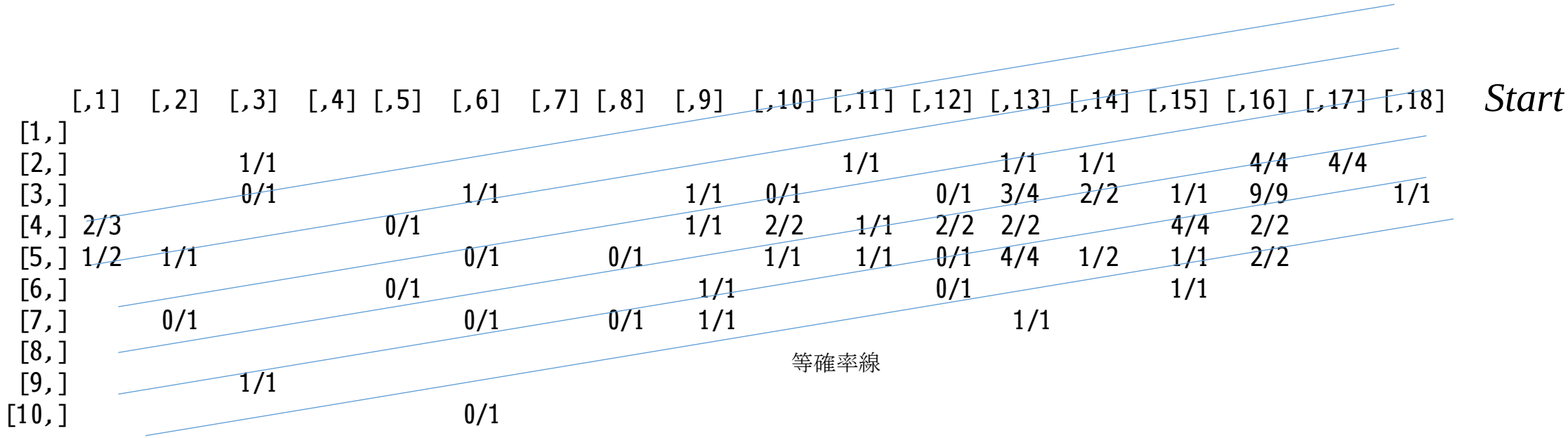

p
1.0
0.8
0.6
0.4
0.2
0.0
-4
-3
-2
-1
0
1
2
x

疑問

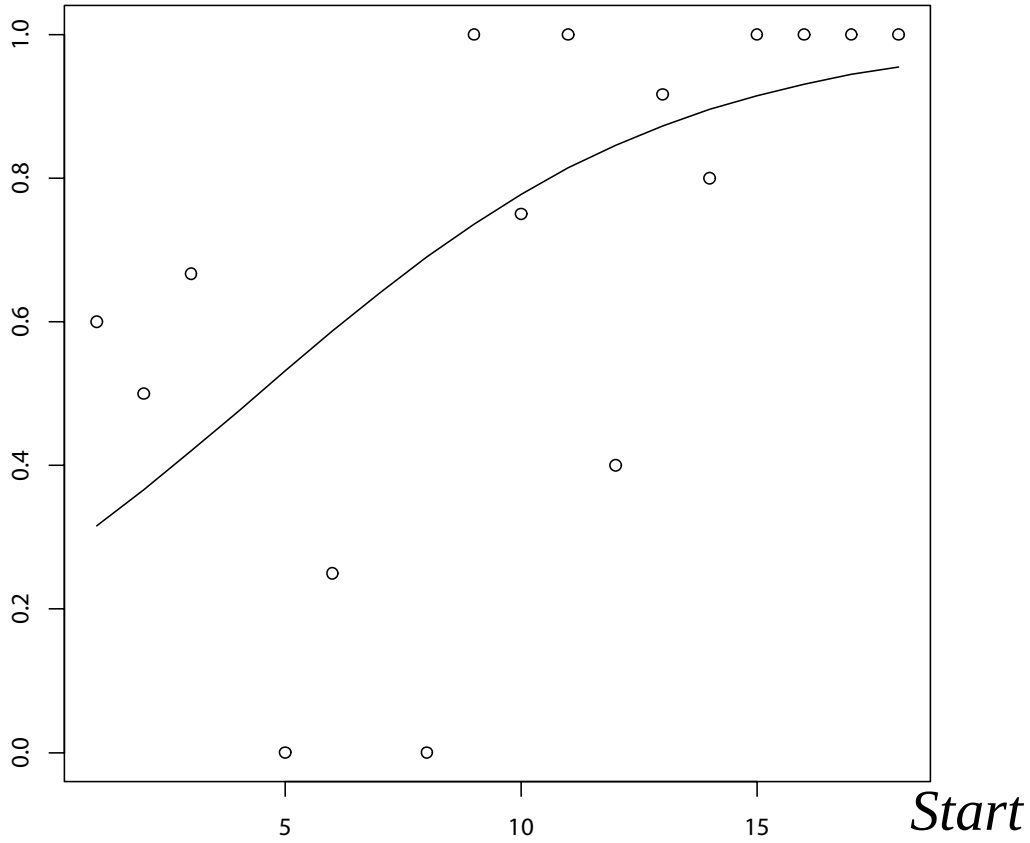
$x = -1.8525 + 0.01067 * Age - 0.21150 * Start + 0.38558 * Number$ の意味?

$p = \dfrac{e^x}{1+e^x}$ の意味？

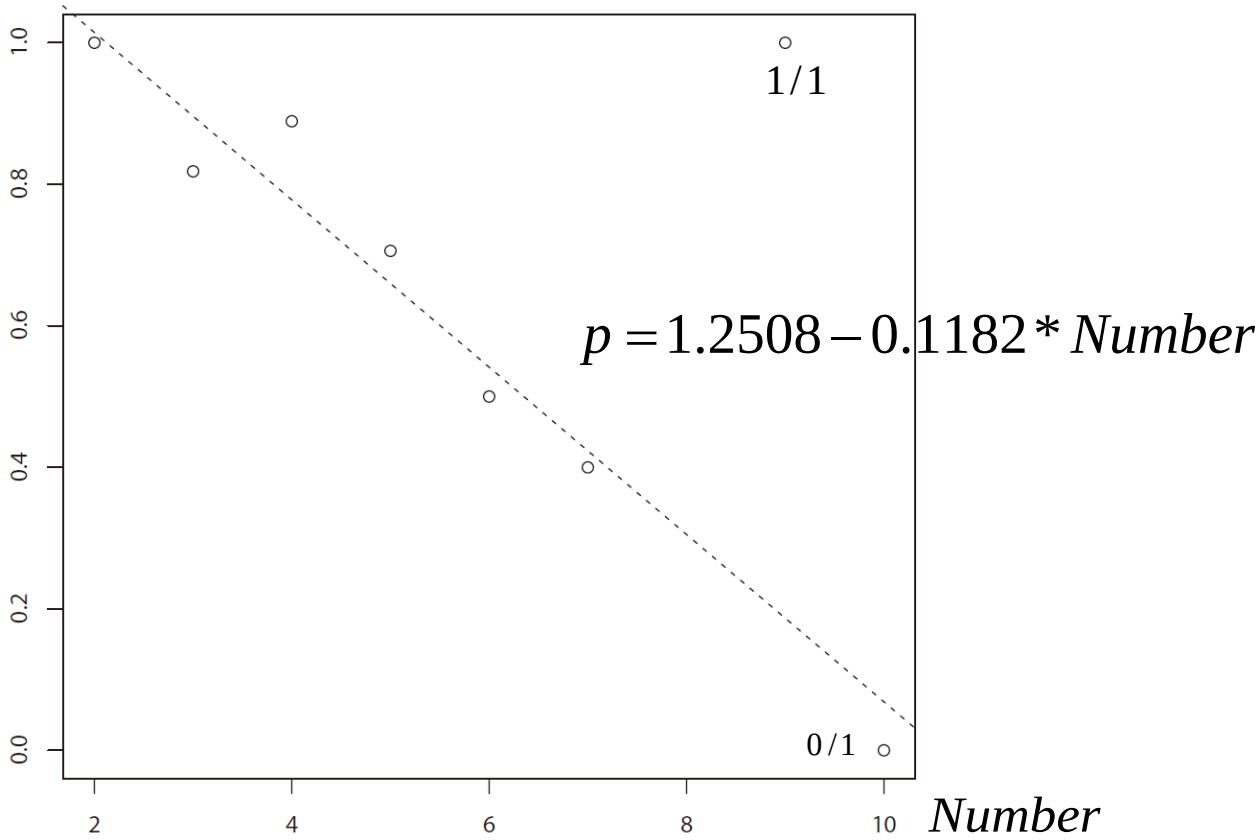
なぜ *Start* のほうが *Number* より寄与度が高いのか？


[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13] [,14] [,15] [,16] [,17] [,18]
Start
[1,]
[2,] 1/1 1/1 1/1 1/1 4/4 4/4
[3,] 0/1 1/1 1/1 0/1 0/1 3/4 2/2 1/1 9/9 1/1
[4,] 2/3 0/1 1/1 2/2 1/1 2/2 2/2 4/4 2/2
[5,] 1/2 1/1 0/1 0/1 1/1 1/1 0/1 4/4 1/2 1/1 2/2
[6,] 0/1 1/1 0/1 1/1
[7,] 0/1 0/1 0/1 1/1 1/1
[8,]
[9,] 1/1
[10,] 0/1
等確率線
Number

手術開始脊柱番号 に対する ロジット成功確率，成功率



手術脊柱数に対する成功率



ロジットモデルの限界

- かなり恣意的なモデル
- モデルは両刃の剣
  - 単純化できる
  - 特有の見方に縛られる
  - 他が見えなくなる
- まず視覚表示で大枠は押さえてから使う必要
- 疑問をもつ
  - 本当？
  - なぜ？