

JUSE-StatWorks/V5 の機能紹介 -多変量解析を中心として- 2011 年日本プロ野球におけるルール変更の影響の検証

株式会社日本科学技術研修所
数理事業部
中澤登

ここでは、2009 年から 2011 年の 3 年間の日本プロ野球レギュラーシーズンのデータについて、JUSE-StatWorks/V5 を用いて解析を行う。

1. 問題の背景

2011 年の日本プロ野球において、前年までとは異なる以下のようなルール変更があった。また、これらのルール変更によって、下記のような影響があると一般的に言われている。

① 12 球団すべてが同一のボールを使用する「統一球」の導入。

【導入の主旨】

統一球が導入された目的の一つとして国際化がある。

2006 年に続き 2009 年に開催された野球の世界大会(ワールドベースボールクラシック(WBC))において日本は優勝したが、代表選手たちは『国際球』を扱う際の違和感を訴えていた。

それまでは、日本プロ野球の公式戦は球団ごとに違うメーカーのボールを使っており、WBC で使用されたボールと日本野球機構(NPB)公認のボールを比較してみると、一般的に、日本のボールの方が飛びやすく、国際球の方が縫い目も高く、滑りやすいと言われていた。

国際試合の度に代表選手が「ボールの違和感」を訴えるようではプレーにも影響しかねない。これを解消すべく、NPB が使用するボールを反発力(飛距離)や触ったときの感触が『国際球』に近いものに統一することになり、2011 年より導入された。

【ルール変更による影響】

一般的に打球は飛ばなくなり、投手(守備側)に有利になると言われている。

② セ・パ両リーグの審判員の統合。

【導入の主旨】

円滑な運営や経費節減などを目的に 2011 年シーズンから、それまでは別々に運営されていた各リーグの審判員業務を統合して活動することになった。セ・パ両リーグで異なっていた立ち位置や細かな動作なども統一された。

【ルール変更による影響】

ストライクゾーンはもともとセリーグの方が広いとされており、今回の審判員統合でパリーグのストライクゾーンもセリーグに合わせて広くなったと言われている。ストライクゾーンが広がると投手(守備側)に有利になると考えられる。

③ 試合時間 3 時間 30 分ルールの導入。

【導入の主旨】

東日本大震災の影響により、節電対策として 2011 年シーズンは、試合開始時刻から 3 時間 30 分経過した場合、新しい延長回には入らないというルールが設けられた。

【ルール変更による影響】

勝率 = 勝数 / (勝数 + 敗数) で算出され、勝率により順位が決定される。全チームそう試合数は 144 試合と決まっており、引分の場合再試合は行われない。引分数が多いということは、(勝数 + 敗数) が少なくなるため、勝ち越している(勝数 > 敗数)のチームは、勝数と敗数の差(勝数 - 敗数)の値が同数のチームがある場合、(勝数 + 敗数)の総数が少ない方が勝率は高くなる。つまり、引分数が多い(試合消化数が多い)方が有利になる。そのため、上位チームは引分数が多い方が、最終的により上位になりやすいので、2011 年シーズン中は、引分の数が、最終的な順位に影響するのではないかと言われた。

注) 勝数 < 敗数の場合は、逆の傾向が出る。

2. 解析の目的

2009 年, 2010 年, 2011 年の日本プロ野球レギュラーシーズンの実データを用いて, 基本解析, 重回帰分析・数量化 I 類, 主成分分析, クラスター分析などを行い, 前述のルール変更による影響の有無を検証する. また, リーグ別, チーム別に影響の傾向が異なるのか検証する.

3. データ

データは, 日本野球機構(NPB)のオフィシャルサイトに公開されているデータより, 33 変数×36 サンプルを収集した.

◆変数(33 変数)

表 1. 解析に用いる変数

	基本データ	勝敗データ	打撃系データ		投手系データ		
			プラス要因	マイナス要因	プラス要因	マイナス要因	
変数	年 順位 チーム リーグ 試合数	勝数 敗数 引分数 勝率	得点 安打 本塁打 盗塁 犠打 四球 死球	打率 出塁率 長打率	三振 併殺打	防御率 セーブ ホールド 完投 完封勝 奪三振	被安打 被本塁打 与四球 与死球 失点 自責点

◆サンプル

12 球団×3 年分(2009 年～2011 年) = 36 サンプル.

4. 基本解析

多変量連関図

全変数を用いて, 多変量連関図を描き, 基本情報を確認する.
なお, 試合数はすべて 144(試合)なので, 解析からは除外する.

統計的に特徴的な傾向が表れている変数, 組み合わせを手早く確認するために, 「注目グラフ表示」機能を用いる. 以下の条件(表 2)に該当する変数, 組み合わせの各グラフについて, 着色表示される. (図 1)

表 2. 注目グラフ表示(着色表示)の条件

外れ値の有無	:外れ値があるヒストグラム, 散布図 (初期設定:2σ 外のサンプルがある場合)
正規分布に従っているかどうか	:正規分布に従っているとはいえないヒストグラム
相関関係の強さ	:相関関係が強い散布図 (初期設定:相関係数の絶対値が 0.6 以上)
関連性の有無(独立であるかどうか)	:独立であるとはいえない帯グラフまたは立体棒グラフ
層間でばらつきに違いがあるかどうか	:層間でばらつきに違いがある層別ヒストグラム (初期設定:有意水準 5%)
層間で平均に違いがあるかどうか	:層間で平均に違いがある層別ヒストグラム (初期設定:有意水準 5%)

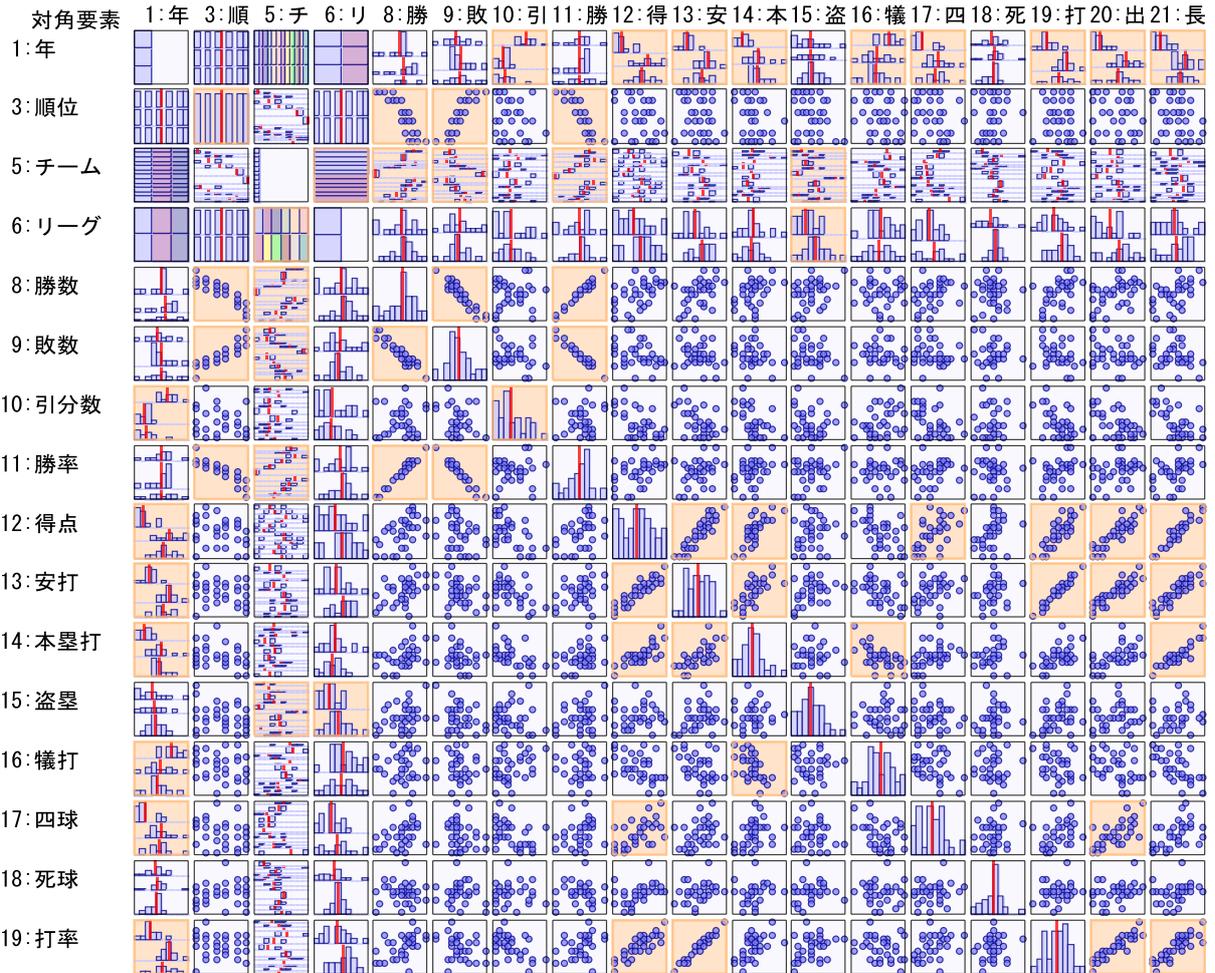


図 1. 多変量連関図(強調表示)

勝数と勝率に正の相関がある, 得点と本塁打数, 安打数, 打率に正の相関がある, など当然と思われる事象は除外して, 多変量連関図より, 以下のような特徴的な傾向が掴める.

- ・ 「年」によって, 平均値に違いがある変数が多くある.
- ・ 「リーグ」によって, 「盗塁数」, 「セーブ数」, 「完投数」に違いがある.
- ・ 「敗数」と「引分数」は, 「防御率」, 「失点」, 「自責点」といった投手成績の変数との相関がある.
- ・ 「敗数」は, 「防御率」, 「被安打数」, 「失点」, 「自責点」と正の相関があり, 「セーブ数」は負の相関がある.
- ・ 「引分数」は, 「防御率」, 「被本塁打数」, 「失点」, 「自責点」と負の相関がある.
- ・ 「勝数」は, 打撃成績とも投手成績とも高い相関があるものはない.
- ・ 「本塁打数」と「犠打数」に負の相関がある.
- ・ 「四球数」と「得点」に正の相関がある.

など

「年」による傾向がありそうなので, 詳しく確認する.
多変量連関図より, 「年」の行を切り抜き表示する.

「年」の行を切り抜き表示した連関図

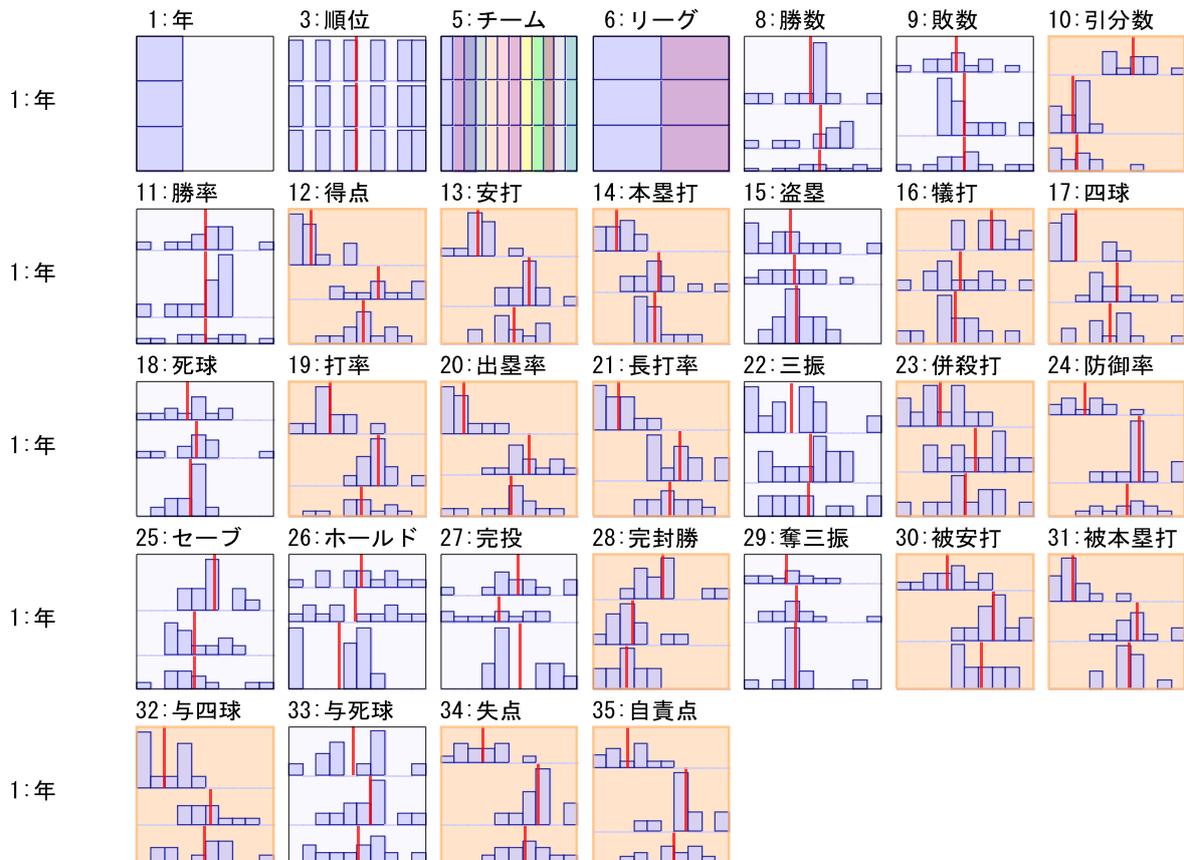


図 2. 「年」の行を切り抜いた連関図

「年」の行を切り抜き表示した連関図より、以下のような特徴的な傾向が掴める。

- 「年」によって「引分数」に違いがあり、2009年、2010年に比べて2011年の引分数は多い。
- 「得点」、「安打数」、「本塁打数」、「打率」、「出塁率」、「長打率」、など、打撃系の変数について、2009年、2010年に比べて2011年は低い。
- 「四球数」が、2009年、2010年に比べて2011年は少ない。
- 「併殺打数」が、2009年、2010年に比べて2011年は少ない。
- 「防御率」、「被安打数」、「被本塁打数」、「与死球数」、「失点」、「自責点」、など、投手系の変数について、2009年、2010年に比べて2011年は低い。
- 「完封勝数」が、2009年、2010年に比べて2011年は多い。

2009年、2010年に比べて2011年は投手に有利な結果となっており、問題の背景で述べたルール変更の影響を裏付ける結果となっている。

次に、2011 年の大きな特徴である「引分数」について詳しく確認する。
多変量連関図より、「引分数」の行を切り抜き表示する。

「引分数」の行を切り抜き表示した連関図

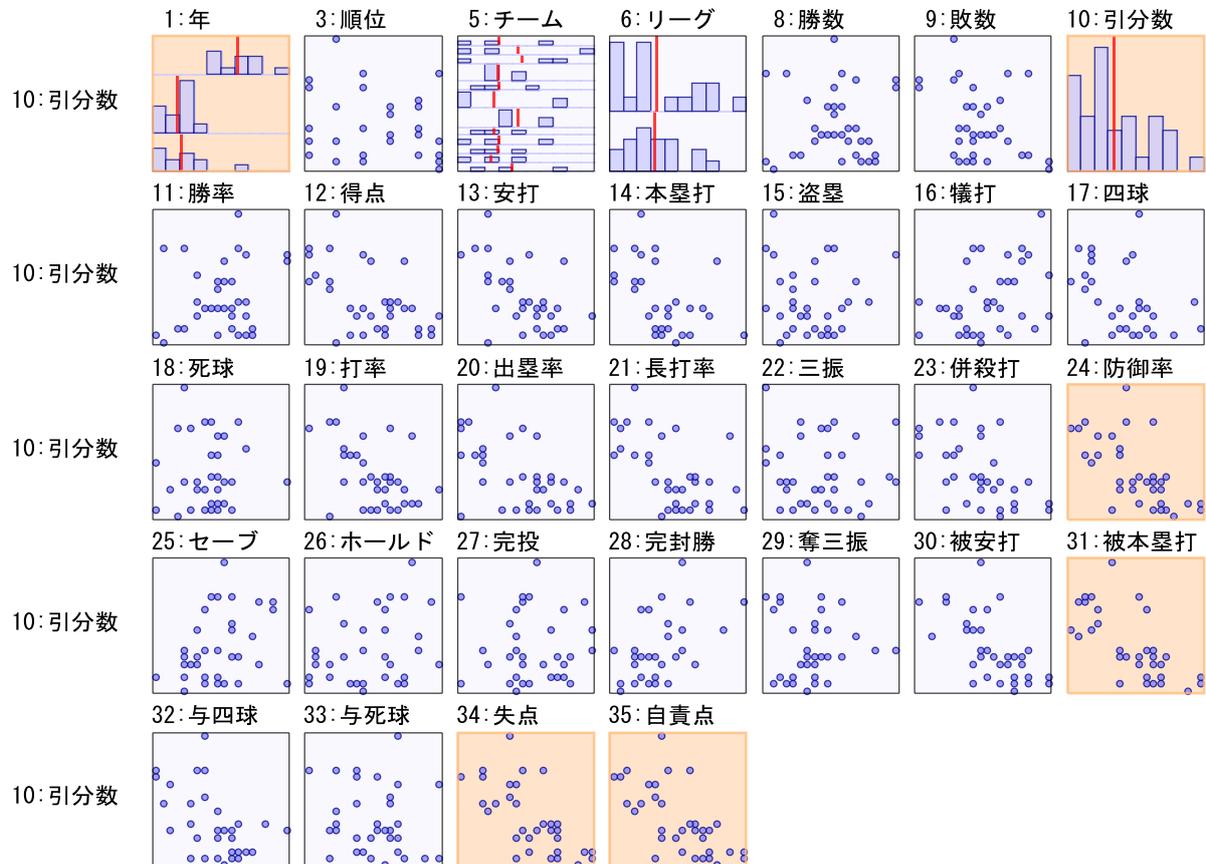


図 3. 「引分数」の行を切り抜いた連関図

「引分数」の行を切り抜き表示した連関図より、以下のような特徴的な傾向が掴める。

- 2009 年、2010 年に比べて 2011 年の「引分数」は多い。
- 「引分数」と「防御率」、「被本塁打数」、「失点」、「自責点」、など、投手系の変数について、負の相関がある。

2009 年、2010 年に比べて 2011 年は引分数が多く、3 時間 30 分ルールの影響は明らかである。

また、引分数が、最終的な順位に影響するかどうかについては、このデータからは、「引分数」と「順位」に特徴的な関係は見いだせない。

「引分数」と「防御率」、「被本塁打数」、「失点」、「自責点」など、投手系の変数について、負の相関があることから、投手成績の良いチーム(投手力のあるチーム)が引分になっている場合が多いと言える。

「引分が多い」=「負けない」ということで強いイメージを与える可能性がある。

「引分数」についてさらに詳しく確認する。
多変量連関図において、マーキングの機能を用いて「年」で層別表示した後、「引分数」の行を切り抜き表示する。

「年」でマーキング(層別表示)後、「引分数」の行を切り抜き表示した連関図

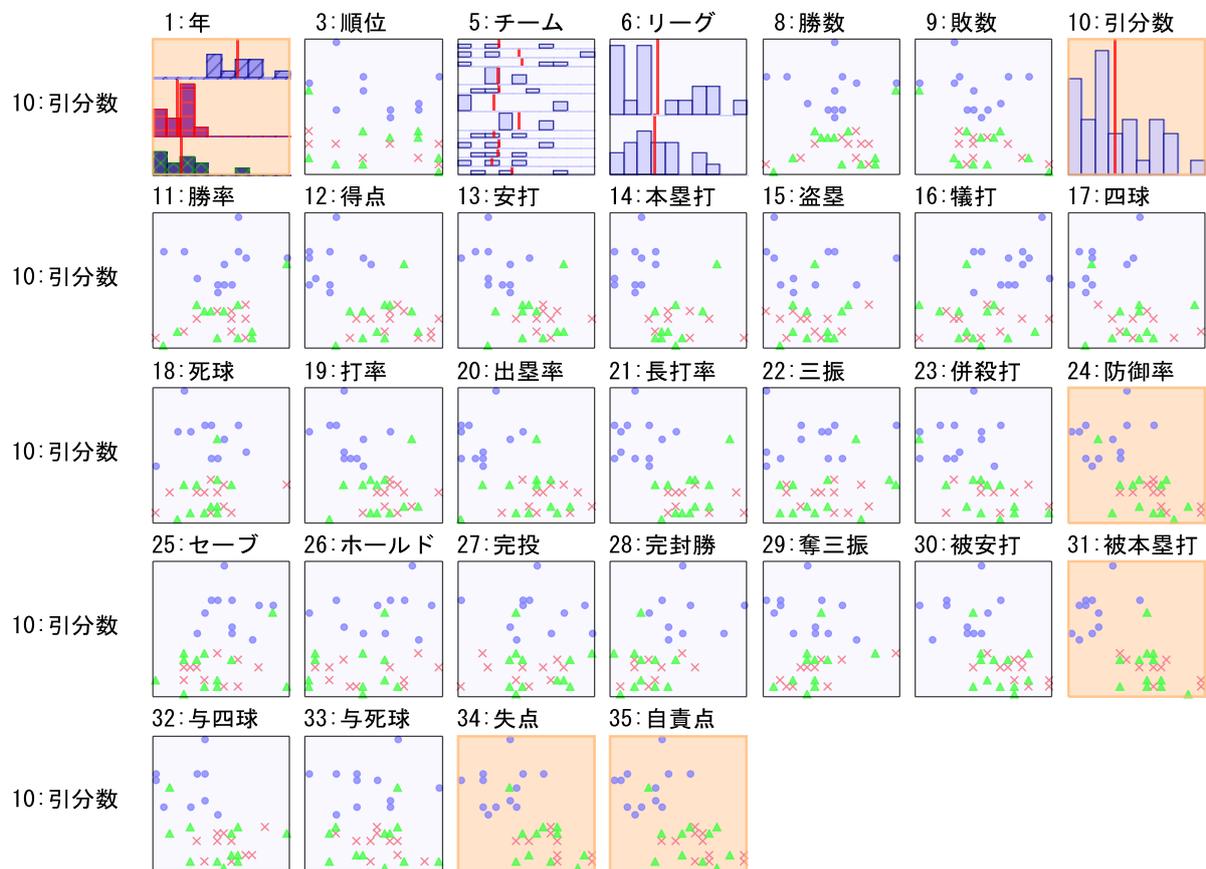


図 4. マーキング後、「引分数」の行を切り抜いた連関図

マーキング(層別表示)後の「引分数」の行を切り抜き表示した連関図より、「引分数」と「防御率」、「被本塁打数」、「失点」、「自責点」などの変数との負の相関は、「年」で層別すると擬相関であることがわかる。

その他、「年」で層別した情報を踏まえて、「引分数」については特に目立った傾向を見出すことはできない。

次に、「年」でマーキングした多変量連関図より、「勝率」の行を切り抜き、「勝率」との関係が強い変数はないか検討する。

「年」でマーキング(層別表示)後、「勝率」の行を切り抜き表示した連関図

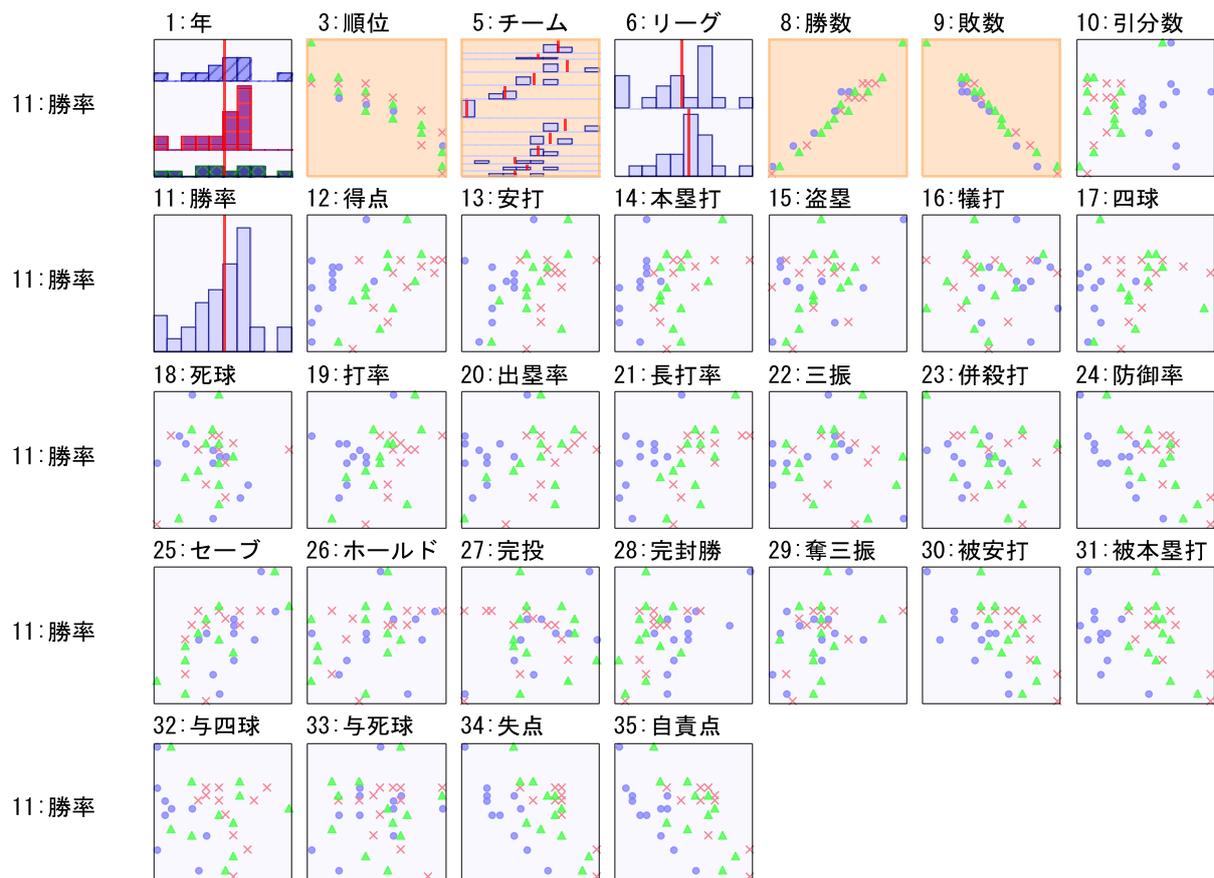


図 5. マーキング後、「勝率」の行を切り抜いた連関図

「勝率」の行を切り抜き表示した連関図より、勝率と引分数は関係が見られない。

勝率について、チームごとに平均値に傾向が見られることから、過去 3 年間の勝率が高いチームと低いチーム(強いチームと弱いチーム)があることがわかる。

「勝率」の連関図を「年」でマーキング(層別)表示することより、打撃系の変数では、得点、安打数、本塁打数などの変数が勝率と正の相関があるように見え、投手系の変数では、防御率、被安打、失点、自責点などの変数が勝率と負の相関があるように見える。

相関係数行列で確認すると、各年共に投手系の変数が勝率と相関があることがわかる。

特に 2011 年は投手系の変数の相関が高くなり、打撃系の変数の相関が低くなっている傾向がある。

基本統計量		順序統計量		相関係数行列	
質的変数: 年		カテゴリ名:			
No	変数名	勝率 2011	勝率 2010	勝率 2009	
3	順位	-0.909++	-0.832++	-0.947++	
8	勝数	0.992++	0.998++	0.993++	
9	敗数	-0.993++	-0.998++	-0.997++	
10	引分数	0.054	0.296	0.487	
11	勝率	1.000	1.000	1.000	
12	得点	0.576	0.621+	0.663+	
13	安打	0.420	0.465	0.604+	
14	本塁打	0.527	0.454	0.425	
15	盗塁	0.492	0.172	0.554	
16	犠打	0.059	-0.027	0.216	
17	四球	0.318	0.293	0.047	
18	死球	-0.495	0.300	0.365	
19	打率	0.453	0.540	0.606+	
20	出塁率	0.596	0.597	0.509	
21	長打率	0.697+	0.591	0.638+	
22	三振	-0.168	0.024	-0.040	
23	併殺打	-0.473	-0.091	-0.358	
24	防御率	-0.817++	-0.777+	-0.786+	
25	セーブ	0.553	0.494	0.743+	
26	ホールド	0.236	0.521	0.388	
27	完投	0.626+	0.127	-0.135	
28	完封勝	0.687+	0.618+	0.329	
29	奪三振	0.464	0.437	0.173	
30	被安打	-0.778+	-0.813++	-0.709+	
31	被本塁打	-0.596	-0.705+	-0.525	
32	与四球	-0.280	-0.273	-0.197	
33	与死球	0.384	-0.243	0.004	
34	失点	-0.846++	-0.773+	-0.836++	
35	自責点	-0.810++	-0.764+	-0.755+	

図 6. 勝率と各変数の相関係数(年ごとに層別)

層別散布図の拡大表示

連関図から任意のグラフを指定し拡大表示することができる。

ここでは、勝率と防御率の層別散布図を拡大表示してみる。

拡大した層別散布図に対して、層ごとに確率楕円の表示を行うことができる。

「年」で層別することにより、勝率と防御率に負の相関があることがわかる。さらに、ここでも 2009 年、2010 年に比べ 2011 年の防御率が低いことがわかる。

層別散布図は、層ごとの散布図に表示方法を切り替えて描くこともできる。

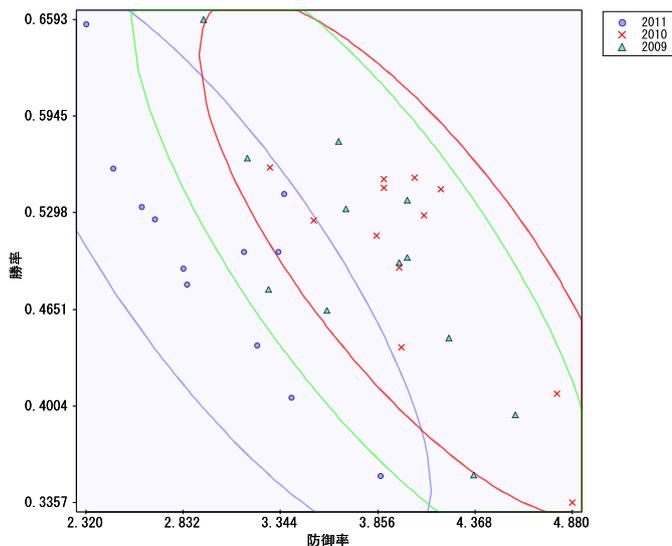


図 7. 層別散布図(年ごとに層別)

X軸: 防御率 Y軸: 勝率

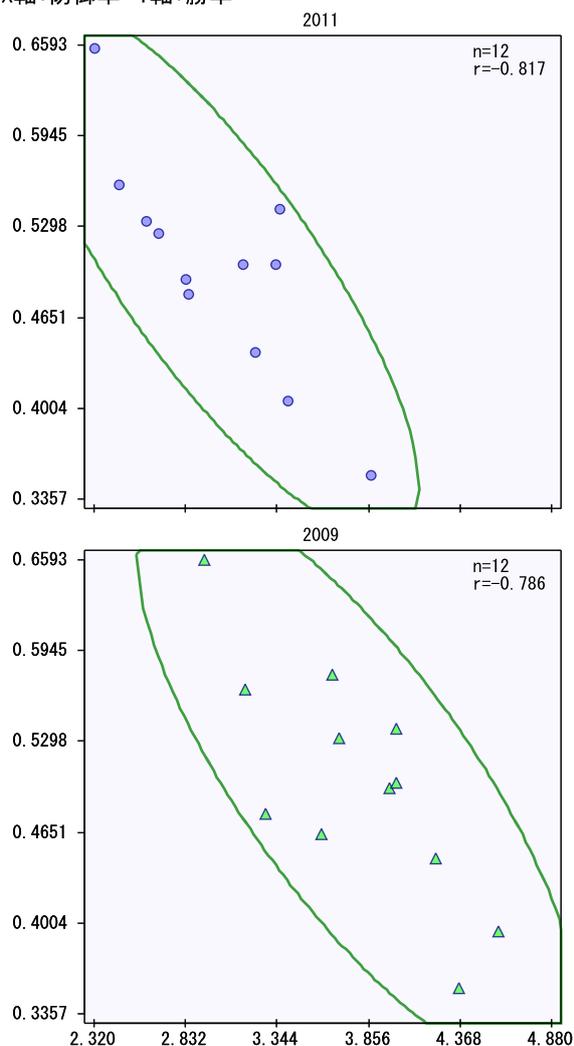


図 8. 層ごとの散布図

5. 重回帰分析・数量化 I 類

次に「勝率」を目的変数とした重回帰分析・数量化 I 類による解析を行う。

この時、順位、勝数、敗数の各変数は、勝率への寄与が当たり前なので、予め説明変数の候補から外す。

変数選択

変数選択画面において、「強調表示」機能を利用する。

「強調表示」機能は、その時点で、説明変数として取り入れられていない変数のうち、分散比の値が最も大きい変数について着色表示してくれる機能で、次に説明変数として取り込む変数の目安になる。

変数選択	選択履歴	SE変化(%)	偏回帰'DoT-一覧	偏回帰'DoT	偏回帰残差一覧	
目的変数名	重相関係数	寄与率R ²	R* ²	R** ²		
勝率	0.000	0.000	0.000	0.000		
	残差自由度	残差標準偏差				
	35	0.075				
vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	1613.3610	0.000	0.500		
1	年	0.0002	1.000			
5	チーム	6.8017	0.000			
6	リーグ	0.3744	0.545			
10	引分数	0.8351	0.367 +			
12	得点	6.1025	0.019 +			
13	安打	4.7643	0.036 +			
14	本塁打	3.9340	0.055 +			
15	盗塁	6.2567	0.017 +			
16	犠打	0.1787	0.675 +			
17	四球	1.0117	0.322 +			
18	死球	0.0213	0.865 +			
19	打率	6.2065	0.018 +			
20	出塁率	5.6321	0.023 +			
21	長打率	6.2161	0.018 +			
22	三振	0.1677	0.685 -			
23	併殺打	2.7670	0.105 -			
24	防壁率	16.0044	0.000 +			
25	セーブ	17.5173	0.000 +			
26	ホールド	5.2363	0.028 +			
27	完投	1.6588	0.206 +			
28	完封勝	9.2007	0.007 +			
29	奪三振	4.8752	0.034 +			
30	被安打	16.1402	0.000 -			
31	被本塁打	4.6173	0.039 -			
32	与四球	1.2057	0.280 -			
33	与死球	0.1830	0.672 +			
34	失点	15.0762	0.000 -			
35	自責点	13.7281	0.001 -			

図 9. 変数選択 初期状態

「セーブ数」→「打率」→「防壁率」→「得点」の順で、説明変数として取り込むことができる。

この時点で、自由度二重調整寄与率(R**²)は 0.851、残差標準偏差は 0.027、残差自由度は 31 である。

ここまで取り込まれた変数は、他の変数に比べ一般的な知見からも勝率に対して重要な変数だと思われる。

また、各変数の偏回帰係数の符号についても違和感はない。

ここで、「打率」の分散比の値が 2 を下回っているので、「打率」を説明変数から外す。

変数選択	選択履歴	SE変化(%)	偏回帰'DoT-一覧	偏回帰'DoT	偏回帰残差一覧	
目的変数名	重相関係数	寄与率R ²	R* ²	R** ²		
勝率	0.939	0.981	0.985	0.851		
	残差自由度	残差標準偏差				
	31	0.027				
vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	-1.1230	0.297	0.172		
1	年	0.1444	0.866			
5	チーム	1.5579	0.187			
6	リーグ	0.9884	0.328			
10	引分数	0.9966	0.326 +			
12	得点	17.9224	0.000	0.001	0.608	0.186
13	安打	3.4138	0.075 -			
14	本塁打	0.0105	0.919 -			
15	盗塁	1.9760	0.170 +			
16	犠打	3.7091	0.064 +			
17	四球	2.1922	0.149 +			
18	死球	0.0444	0.835 -			
19	打率	1.8567	0.183	1.026	0.186	0.207
20	出塁率	2.8445	0.102 +			
21	長打率	0.0357	0.851 +			
22	三振	0.6176	0.438 -			
23	併殺打	0.0535	0.819 -			
24	防壁率	63.8644	0.000	-0.087	-0.740	0.449
25	セーブ	9.5324	0.004	0.003	0.261	0.538
26	ホールド	1.2084	0.280 +			
27	完投	2.3642	0.136 +			
28	完封勝	2.2464	0.144 +			
29	奪三振	0.3134	0.580 +			
30	被安打	0.3512	0.558 -			
31	被本塁打	0.2939	0.592 +			
32	与四球	0.6149	0.439 +			
33	与死球	2.4846	0.125 +			
34	失点	2.4246	0.130 -			
35	自責点	2.9301	0.097 +			

図 10. 変数選択 第①段階

その後、「犠打数」→「完投数」→「長打率」の順で、説明変数として取り込むことができる。

ここで、「得点」と「長打率」のトレランスの値が 0.1 より小さくなったので着色表示される。

この 2 つの変数のうちどちらかを説明変数から外すことも検討できるが、ここでは外さずに、変数選択を続行する。

変数選択	確認率%	残差の分布	残差の連関	予測		
目的変数名	重相関係数	寄与率R ²	R* ²	R** ²		
勝率	0.957	0.915	0.898	0.881		
	残差自由度	残差標準偏差				
	29	0.024				
vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	0.0000	0.996	-0.001		
1	年	0.0352	0.965			
5	チーム	1.7322	0.145			
6	リーグ	0.2848	0.598			
10	引分数	2.0604	0.162 +			
12	得点	4.9715	0.034	0.000	0.435	0.077
13	安打	0.0000	0.998 -			
14	本塁打	0.1601	0.692 +			
15	盗塁	2.5290	0.123 +			
16	犠打	10.5478	0.003	0.001	0.220	0.633
17	四球	0.3386	0.565 +			
18	死球	0.0388	0.845 -			
19	打率	0.0295	0.865 +			
20	出塁率	0.5545	0.463 +			
21	長打率	4.2293	0.049	0.857	0.410	0.073
22	三振	1.3375	0.257 -			
23	併殺打	0.0982	0.756 -			
24	防御率	39.9950	0.000	-0.072	-0.608	0.315
25	セーブ	14.2781	0.001	0.003	0.319	0.408
26	ホールド	0.6002	0.445 +			
27	完投	6.2592	0.018	0.003	0.176	0.580
28	完封勝	1.5714	0.220 +			
29	奪三振	1.9594	0.173 +			
30	被安打	0.5314	0.472 -			
31	被本塁打	0.6100	0.441 -			
32	与四球	0.5033	0.484 +			
33	与死球	1.1443	0.294 +			
34	失点	1.2788	0.268 -			
35	自責点	1.2883	0.266 +			

図 11. 変数選択 第②段階

その後、「盗塁数」→「引分数」→「三振数」→「四球数」と取り入れたところで、「得点」の分散比の値が 2 を下回ったので、「得点」を説明変数から外す。

変数選択	確認率%	残差の分布	残差の連関	予測		
目的変数名	重相関係数	寄与率R ²	R* ²	R** ²		
勝率	0.970	0.942	0.918	0.896		
	残差自由度	残差標準偏差				
	25	0.021				
vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	0.4882	0.491	-0.088		
1	年	0.3019	0.742			
5	チーム	0.8605	0.593			
6	リーグ	0.3927	0.537			
10	引分数	4.6877	0.040	0.003	0.183	0.409
12	得点	0.4960	0.498	0.000	0.145	0.055
13	安打	0.0062	0.938 +			
14	本塁打	0.5070	0.483 +			
15	盗塁	5.0782	0.033	0.000	0.134	0.655
16	犠打	9.4684	0.005	0.001	0.193	0.584
17	四球	2.2484	0.146	0.000	0.124	0.343
18	死球	0.5788	0.454 -			
19	打率	0.0140	0.907 +			
20	出塁率	0.0001	0.991 -			
21	長打率	11.3708	0.002	1.391	0.666	0.060
22	三振	3.0892	0.091	-0.000	-0.103	0.680
23	併殺打	0.3986	0.534 +			
24	防御率	19.3392	0.000	-0.059	-0.500	0.180
25	セーブ	13.7980	0.001	0.003	0.322	0.310
26	ホールド	1.6242	0.215 +			
27	完投	4.0527	0.055	0.002	0.149	0.426
28	完封勝	0.9038	0.351 +			
29	奪三振	0.9741	0.334 +			
30	被安打	0.0557	0.815 -			
31	被本塁打	0.9464	0.340 -			
32	与四球	0.5562	0.463 +			
33	与死球	1.3002	0.265 +			
34	失点	0.5308	0.473 -			
35	自責点	0.4573	0.505 +			

図 12. 変数選択 第③段階

この時点で、分散比の値が 2 を超える変数はないので、重回帰モデルを確定させる。

vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	1.5220	0.228	-0.133		
10	引分数	5.1123	0.032	0.003	0.168	0.413
15	盗塁	7.1861	0.013	0.000	0.149	0.742
16	犠打	10.5928	0.003	0.001	0.199	0.608
17	四球	4.3647	0.047	0.000	0.151	0.439
21	長打率	107.2245	0.000	1.658	0.794	0.389
22	三振	3.8153	0.062	-0.000	-0.111	0.707
24	防御率	19.3707	0.000	-0.058	-0.495	0.181
25	セーブ	13.8267	0.001	0.003	0.314	0.315
27	完投	4.3647	0.047	0.002	0.153	0.428

図 13. 変数選択 第④段階

vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	1.5220	0.228	-0.133		
10	引分数	5.1123	0.032	0.003	0.168	0.413
15	盗塁	7.1861	0.013	0.000	0.149	0.742
16	犠打	10.5928	0.003	0.001	0.199	0.608
17	四球	4.3647	0.047	0.000	0.151	0.439
21	長打率	107.2245	0.000	1.658	0.794	0.389
22	三振	3.8153	0.062	-0.000	-0.111	0.707
24	防御率	19.3707	0.000	-0.058	-0.495	0.181
25	セーブ	13.8267	0.001	0.003	0.314	0.315
27	完投	4.3647	0.047	0.002	0.153	0.428

回帰式	定数項	引分数	盗塁	犠打	四球	長打率	三振	防御率	セーブ	完投
Y=	-0.133	+0.003 X10	+0.000 X15	+0.001 X16	+0.000 X17	+1.658 X21	-0.000 X22	-0.058 X24	+0.003 X25	+0.002 X27

図 14. 確定モデル

この時、自由度二重調整寄与率(R^{*2})は 0.901、残差標準偏差は 0.021、残差自由度は 26 である。ここで説明変数として取り込まれた各変数の偏回帰係数の符号について、特に違和感はない。

残差の検討

確定した重回帰モデルについて、残差の検討を行う。

残差の分布

まず、テコ比と残差t値の散布図により、重回帰モデルから外れの大きいサンプル、重回帰モデルに大きく影響を与えているサンプルがないか確認する。

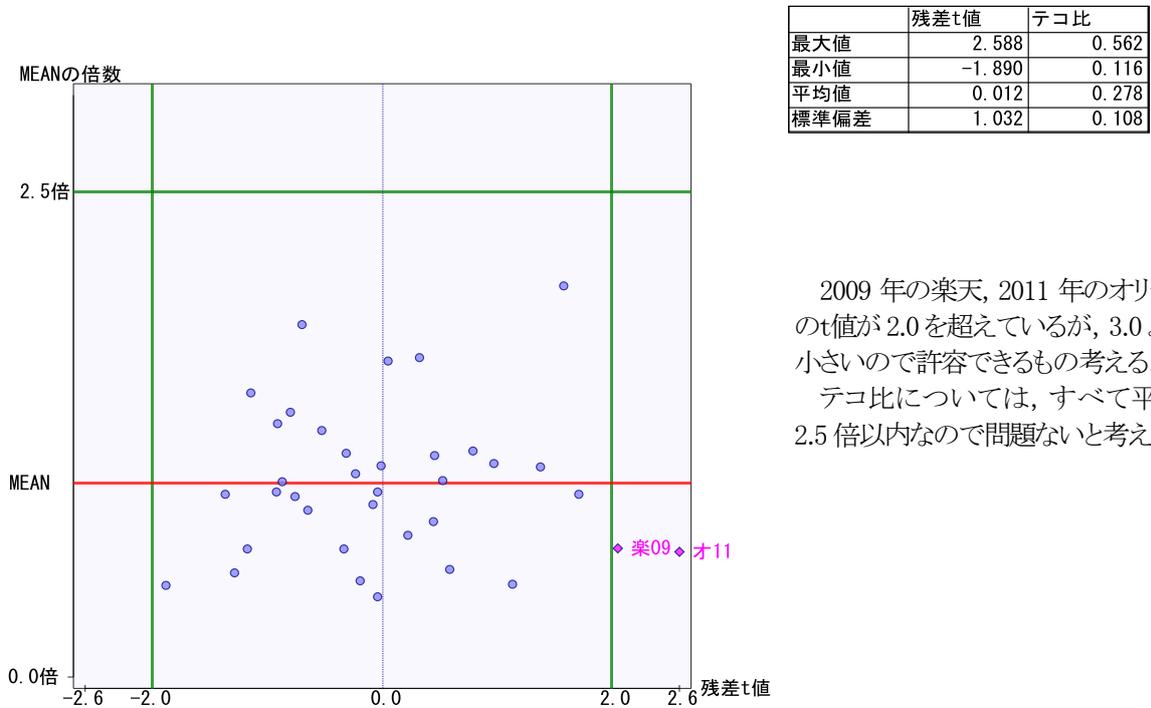


図 15. テコ比と残差t値の散布図

次に、残差のヒストグラムを確認する。

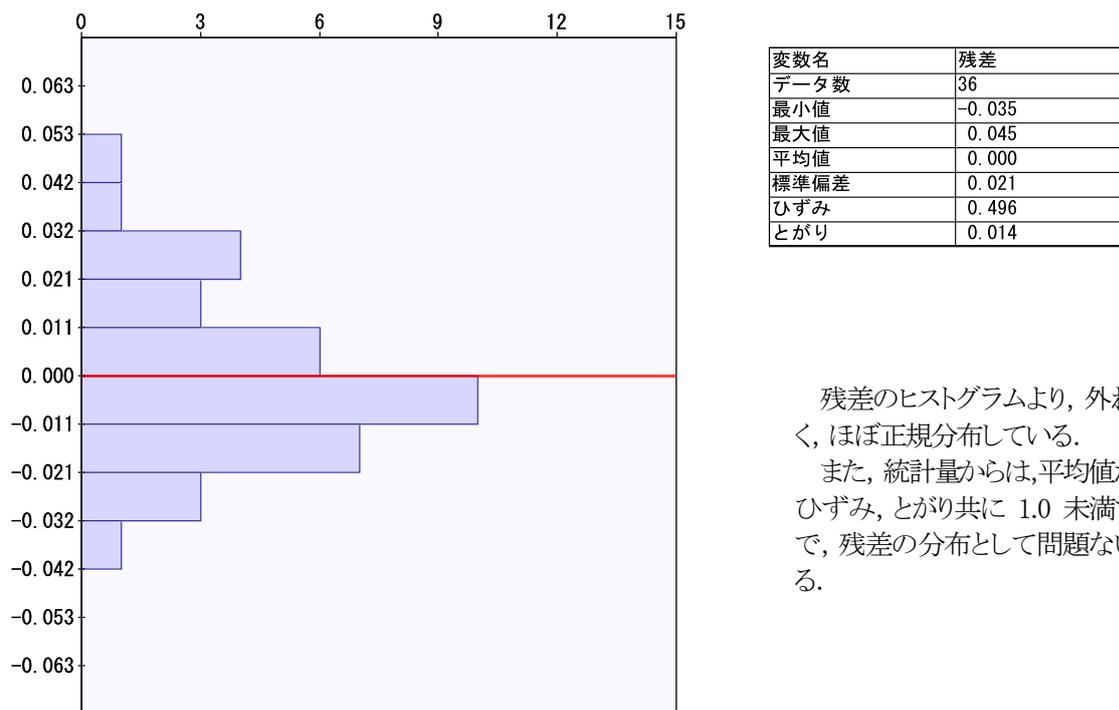


図 16. 残差のヒストグラム

残差の連関

次に、残差と各説明変数との連関図を確認する。

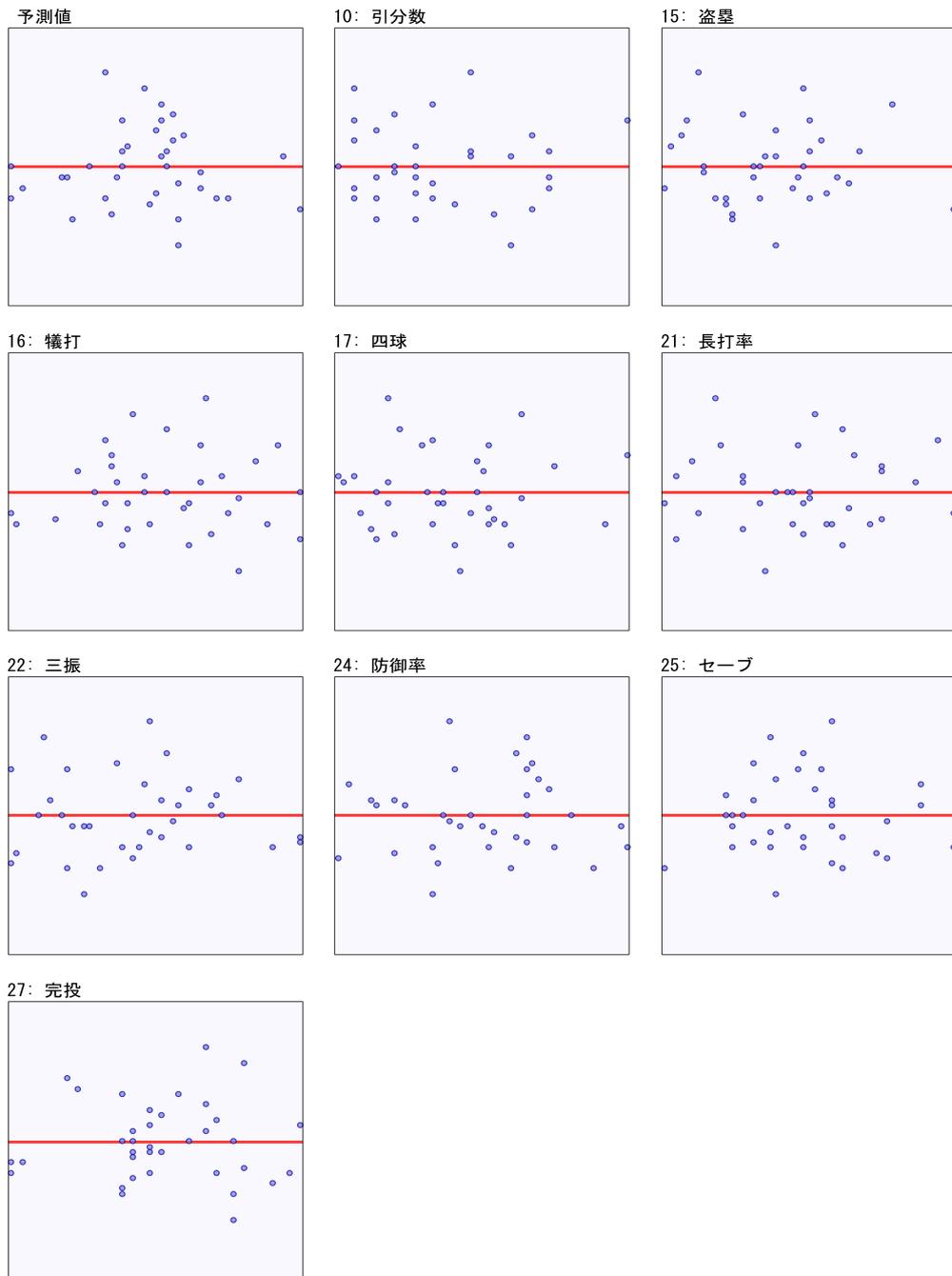


図 17. 残差と各変数の連関図

各変数共に特徴的な傾向はなく、残差のばらつき方に問題はない。
「勝利」を予測するための重回帰式の 1 つとして採用できると考えられる。

	定数項	引分数	盗塁	犠打	四球	長打率	三振	防御率	セーブ	完投
回帰式	Y= -0.133	+0.003 X10	+0.000 X15	+0.001 X16	+0.000 X17	+1.658 X21	-0.000 X22	-0.058 X24	+0.003 X25	+0.002 X27

図 18. 重回帰式(確定モデル)

別の重回帰モデルの検討

できあがった重回帰式とは別の重回帰モデルを検討してみる。

偏回帰プロット

既にできあがった重回帰モデルより、偏回帰プロットを描いてみる。

縦軸: 勝率 | i | の残差
横軸: $X | i |$ の残差

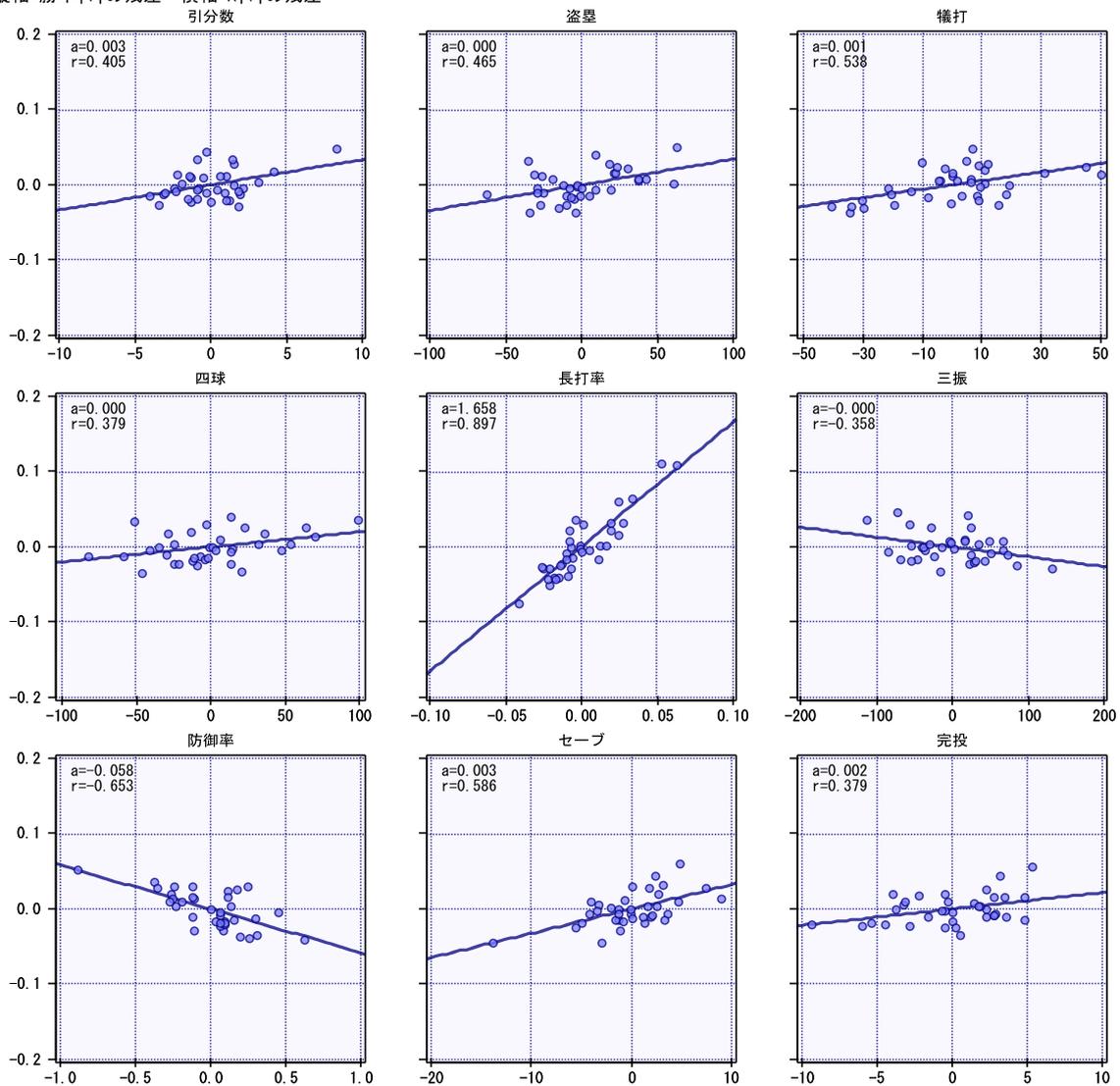


図 19. 偏回帰プロット一覧

各説明変数の中でも、長打率と防御率の相関係数(r)が大きいことがわかる。
そこで、長打率と防御率のみを説明変数として用いた重回帰モデルを検討してみる。

シンプルな重回帰モデル

変数選択画面において長打率と防御率のみを説明変数として取り込み重回帰式を作る。

vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	18.3364	0.000	0.262		
21	長打率	85.1238	0.000	1.615	0.773	0.819
24	防御率	113.8743	0.000	-0.105	-0.894	0.819

図 20. 変数選択(長打率と防御率のみ)

説明変数 2 つのシンプルな重回帰モデルができあがり、この時、自由度二重調整寄与率(R^{*2})は 0.788、残差標準偏差は 0.034、残差自由度は 33 である。

また、各説明変数の偏回帰係数の符号についても特に違和感はない。

残差の検討を行うと、テコ比と残差t値の散布図(図 22)より、2009 年の楽天、2011 年のヤクルトのt値が 2.0 を超えているが、3.0 より小さいので許容できるもの考える。

テコ比については、すべて平均の 2.5 倍以内なので問題ないと考える。

残差のヒストグラム(図 23)からは、正規分布から大きく外れているとは言えず、残差と各変数の連関図(図 24)からも特に大きな傾向は見られないので、残差のばらつきは問題ないと言える。

目的変数名	重相関係数	寄与率 R^2	R^{*2}	R^{**2}
勝率	0.900	0.810	0.799	0.788
	残差自由度	残差標準偏差		
	33	0.034		

vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	18.3364	0.000	0.262		
21	長打率	85.1238	0.000	1.615	0.773	0.819
24	防御率	113.8743	0.000	-0.105	-0.894	0.819

回帰式	定数項	長打率	防御率
Y=	0.262	+1.615 X21	-0.105 X24

図 21. 確定モデル(長打率と防御率のみ)

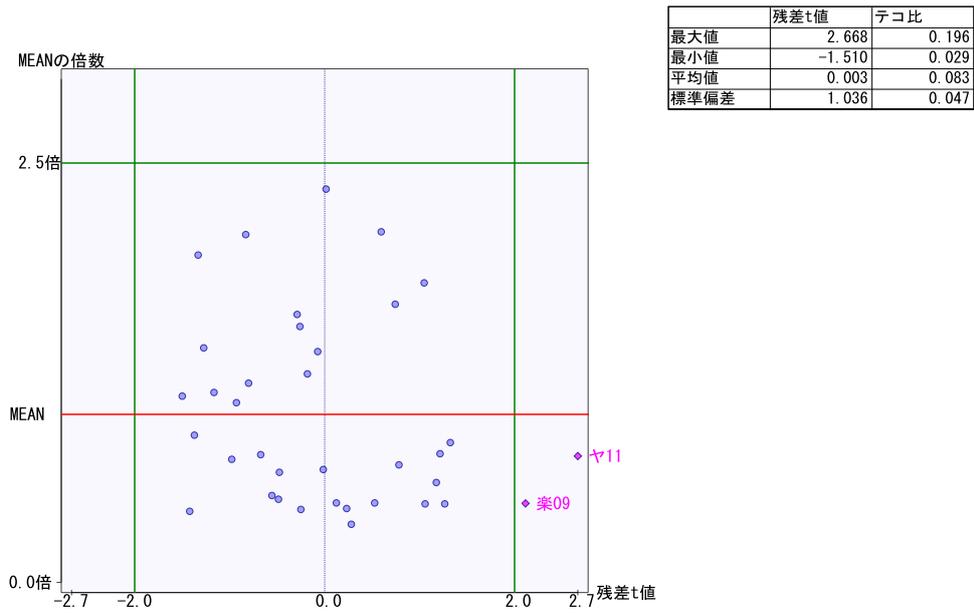


図 22. テコ比と残差t値の散布図

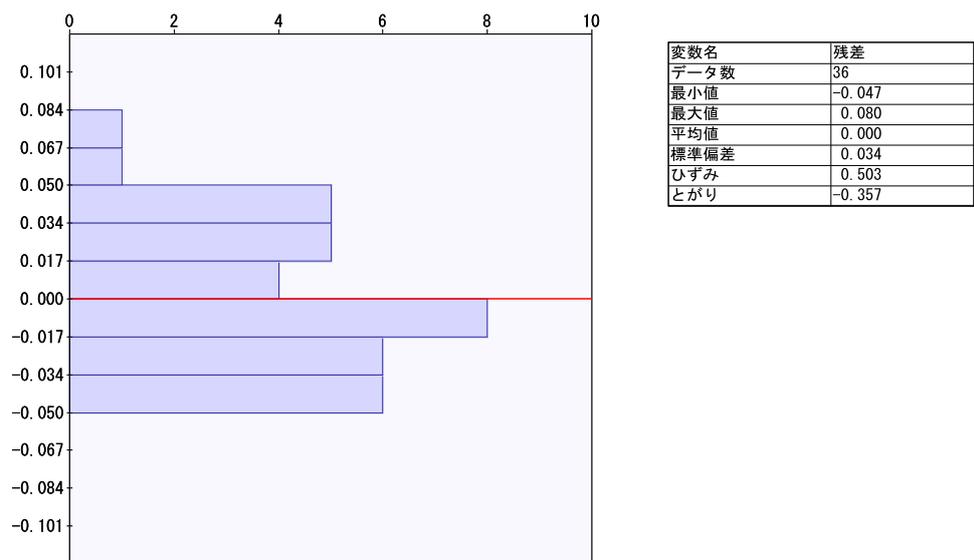


図 23. 残差のヒストグラム

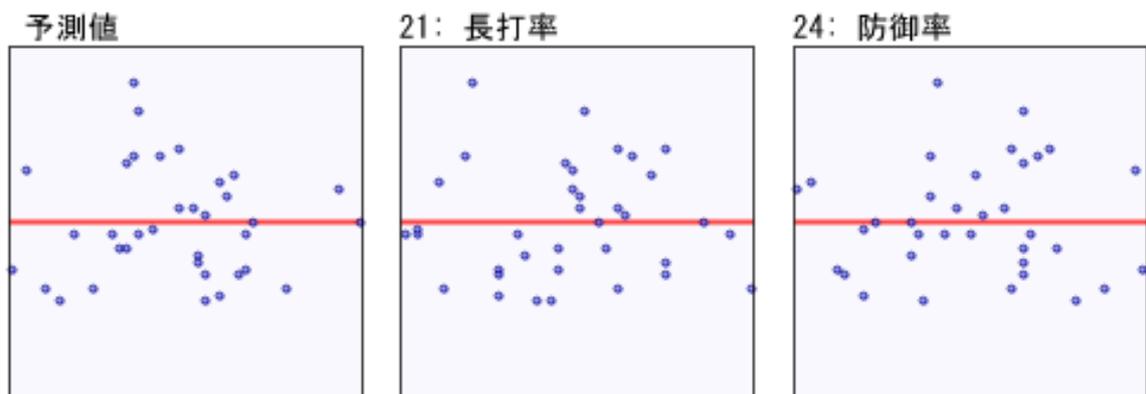


図 24. 残差と各変数の連関図

質的変数を取り込んだ重回帰・数量化 I 類モデル その①

予備解析において、特徴的な傾向が表れていた「年」を説明変数として取り込んだ重回帰・数量化 I 類モデルを検討してみる。

目的変数名	重相関係数	寄与率R ²	R ²	R** ²		
勝率	0.970	0.941	0.917	0.894		
	残差自由度	残差標準偏差				
	25	0.022				
vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	23.0122	0.000	0.404		
1	年	5.5205	0.010			
	2011			0.000		
	2010			0.064		
	2009			0.066		
5	チーム	-999.0000	0.000	共線性有		
6	リーグ	7.8343	0.010			
	セ			0.000		
	バ			0.029		
10	引分数	0.6629	0.007	0.005	0.271	0.281
12	得点	39.7290	0.000	0.001	0.601	0.262
19	安打	0.3233	0.575	*		
14	本塁打	0.9450	0.341	*		
15	盗塁	0.6590	0.425	*		
16	犠打	0.6297	0.435	*		
17	四球	0.0785	0.782	-		
18	死球	0.9640	0.552	*		
18	打率	0.8315	0.371	*		
20	出塁率	0.8897	0.414	*		
21	長打率	2.4097	0.120	*		
22	三振	7.8956	0.009	-0.000	-0.164	0.697
23	併殺打	0.0144	0.906	*		
24	防壁率	0.0186	0.893	-		
25	セーブ	12.1799	0.002	0.003	0.309	0.304
26	ホールド	2.7763	0.108	0.001	0.116	0.490
27	完投	0.5577	0.462	*		
20	完封勝	0.4925	0.499	*		
29	奪三振	0.1833	0.664	*		
30	被安打	0.8588	0.425	-		
31	被本塁打	1.8983	0.170	*		
32	与四球	0.1046	0.749	*		
33	与死球	8.2419	0.008	0.002	0.161	0.759
34	失点	42.0468	0.000	-0.001	-0.716	0.195
35	自責点	0.0001	0.992	-		

図 25. 「年」を取り込んだ変数選択

まず、「年」を説明変数として取り込み「年」は取り込んだまま外さないものとして変数選択を行った結果、説明変数として、「年」、「リーグ」、「引分数」、「得点」、「三振数」、「セーブ数」、「ホールド数」、「与死球」、「失点」を取り込んだ重回帰・数量化 I 類モデルができた。

この時の自由度二重調整寄与率(R**²)は 0894, 残差標準偏差は 0.022, 残差自由度は 25 である。

また、各説明変数の偏回帰係数の符号については、与死球数が正(+))になっており、これは一般的な知見と合致しない。

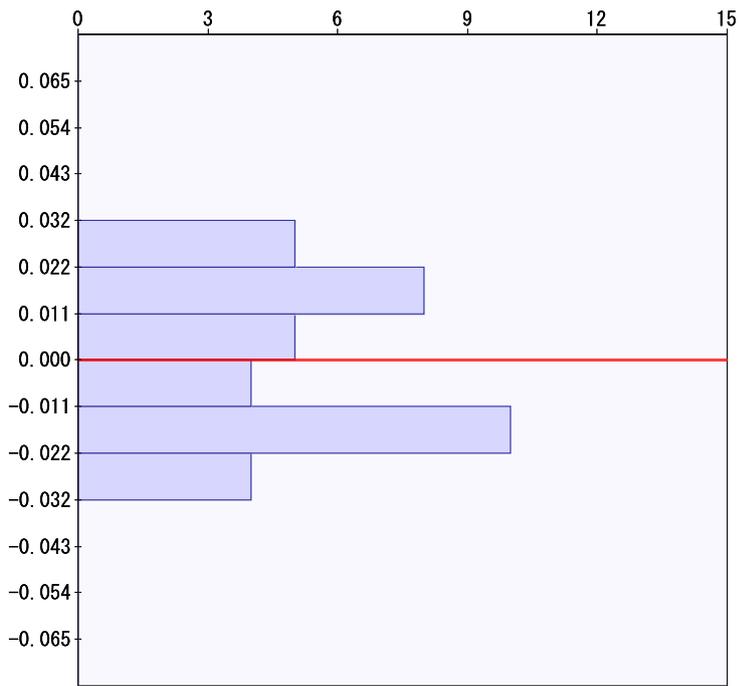
残差の検討を行うと、残差のヒストグラム(図 27)において、とがりの値が-1.417 となっている。

また、正規確率プロット(図 28)における正規性の検定(Shapiro-Wilk)において、P 値は 0.010 と有意となり、検定方法を Anserson-Darling に切り替えても P 値は 0.021 と有意となるので、正規分布とは言えない。残差の正規性が保たれていない。

目的変数名	重相関係数	寄与率R ²	R ²	R** ²		
勝率	0.970	0.941	0.917	0.894		
	残差自由度	残差標準偏差				
	25	0.022				
vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	23.0122	0.000	0.404		
1	年	5.5205	0.010			
	2011			0.000		
	2010			0.064		
	2009			0.066		
6	リーグ	7.8343	0.010			
	セ			0.000		
	バ			0.029		
10	引分数	0.6629	0.007	0.005	0.271	0.281
12	得点	39.7290	0.000	0.001	0.601	0.262
22	三振	7.8956	0.009	-0.000	-0.164	0.697
25	セーブ	12.1799	0.002	0.003	0.309	0.304
26	ホールド	2.7763	0.108	0.001	0.116	0.490
33	与死球	8.2419	0.008	0.002	0.161	0.759
34	失点	42.0468	0.000	-0.001	-0.716	0.195

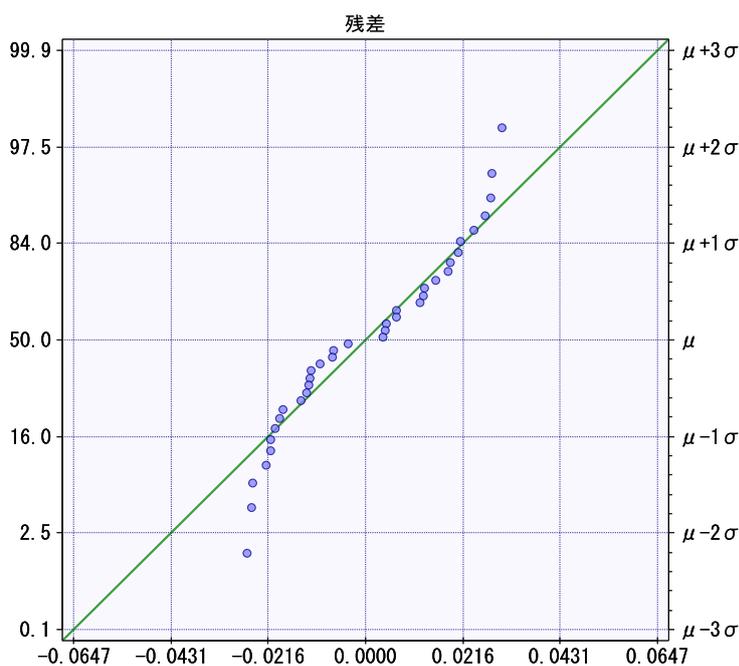
回帰式	定数項	年	リーグ	引分数	得点	三振	セーブ	ホールド	与死球	失点
	Y= 0.404									
		2011	+							
		2010	セ							
		2009	バ							

図 26. 「年」を取り込んだ重回帰・数量化 I 類モデル(確定モデル)



変数名	残差
データ数	36
最小値	-0.026
最大値	0.030
平均値	0.000
標準偏差	0.022
ひずみ	0.138
とがり	-1.417

図 27. 残差のヒストグラム



データ数	36
最小値	-0.026
最大値	0.030
平均値	0.0000
標準偏差	0.02157
ひずみ	0.138
とがり	-1.417
正規性検定	Shapiro-Wilk
W	0.914
P値	0.010

図 28. 正規確率プロット

以上より、今回得られているデータにおいて、説明変数に「年」を取り込んだ重回帰・数量化 I 類モデルを想定するよりは、年ごとに層別して、それぞれの年で重回帰モデルを検討した方が、当てはまりの良い重回帰モデルが得られる可能性がある。

質的変数を取り込んだ重回帰・数量化 I 類モデル その②

質的変数の「リーグ」を説明変数として取り込んだ重回帰・数量化 I 類モデルを検証してみる。

目的変数名	重相関係数	寄与率R ²	R* ²	R** ²		
勝率	0.969	0.939	0.915	0.891		
	残差自由度	残差標準偏差				
	25	0.022				
vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	0.8468	0.366	-0.103		
1	年	1.6091	0.222			
5	チーム	-989.0000	0.000	共線性有		
6	リーグ	4.5473	0.043			
	セ			0.000		
	バ			0.025		
10	引分数	4.8814	0.037	0.003	0.170	0.414
12	得点	2.0402	0.186	+		
13	安打	0.0086	0.936	+		
14	本塁打	0.0001	0.991	+		
15	盗塁	2.0499	0.185	+		
16	犠打	4.2535	0.050	0.000	0.143	0.508
17	四球	1.3666	0.254	+		
18	死球	0.0318	0.860	+		
19	打率	0.0068	0.935	+		
20	出塁率	1.2933	0.267	+		
21	長打率	127.9096	0.000	1.758	0.841	0.441
22	三振	10.0623	0.004	-0.000	-0.207	0.573
23	併殺打	0.3577	0.555	+		
24	防御率	15.5835	0.001	-0.047	-0.401	0.237
25	セーブ	18.4850	0.000	0.004	0.376	0.320
26	ホールド	3.2285	0.084	0.001	0.130	0.466
27	完投	5.6557	0.025	0.003	0.187	0.394
28	完封勝	0.1713	0.883	+		
29	奪三振	0.2439	0.626	+		
30	被安打	0.0143	0.906	-		
31	被本塁打	0.0165	0.899	+		
32	与四球	0.8931	0.370	+		
33	与死球	1.9698	0.173	0.001	0.077	0.803
34	失点	0.2115	0.850	-		
35	自責点	3.0273	0.095	+		

図 29. 「リーグ」を取り込んだ変数選択

まず、「リーグ」を説明変数として取り込み「リーグ」は取り込んだまま外さないものとして変数選択を行った結果、説明変数として、「リーグ」、「引分数」、「犠打数」、「長打率」、「三振数」、「防御率」、「セーブ数」、「ホールド数」、「完投数」、「与死球」を取り込んだ重回帰・数量化 I 類モデルができた。

この時の自由度二重調整寄与率(R**²)は 0891, 残差標準偏差は 0.022, 残差自由度は 25 である。

また、各説明変数の偏回帰係数の符号については、与死球数が正(+)になっており、これは一般的な知見と合致しない。

残差の検討を行うと、テコ比と残差t値の散布図(図 31)において、2011 年西武, 2011 阪神の残差t値が-2.0 を下回っているが、-3.0 よりは大きいので許容できるもの考える。

残差のヒストグラム(図 32)からは、正規分布から大きく外れているとは言えず、残差と各変数の連関図(図 33)からも特に大きな傾向は見られないので、残差のばらつきは問題ないと言える。

目的変数名	重相関係数	寄与率R ²	R* ²	R** ²		
勝率	0.969	0.939	0.915	0.891		
	残差自由度	残差標準偏差				
	25	0.022				
vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	0.8468	0.366	-0.103		
6	リーグ	4.5473	0.043			
	セ			0.000		
	バ			0.025		
10	引分数	4.8814	0.037	0.003	0.170	0.414
16	犠打	4.2535	0.050	0.000	0.143	0.508
21	長打率	127.9096	0.000	1.758	0.841	0.441
22	三振	10.0623	0.004	-0.000	-0.207	0.573
24	防御率	15.5835	0.001	-0.047	-0.401	0.237
25	セーブ	18.4850	0.000	0.004	0.376	0.320
26	ホールド	3.2285	0.084	0.001	0.130	0.466
27	完投	5.6557	0.025	0.003	0.187	0.394
33	与死球	1.9698	0.173	0.001	0.077	0.803

	定数項	リーグ	引分数	犠打	長打率	三振	防御率	セーブ	ホールド	完投	与死球	
回帰式	Y= -0.103		+()	+0.003 X10	+0.000 X16	+1.758 X21	-0.000 X22	-0.047 X24	+0.004 X25	+0.001 X26	+0.003 X27	+0.001 X33
		セ	+0.000									
		バ	+0.025									

図 30. 「リーグ」を取り込んだ重回帰・数量化 I 類モデル(確定モデル)

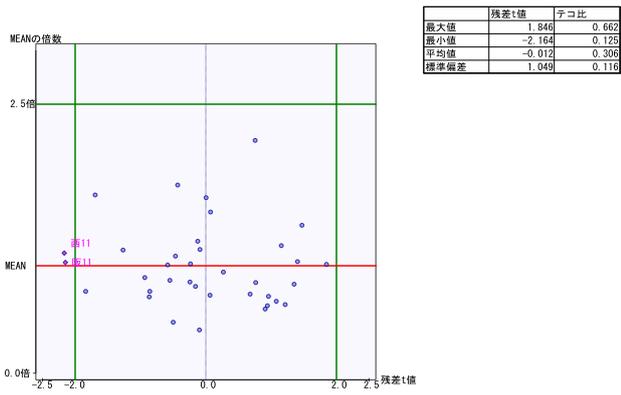


図 31. テコ比と残差t値の散布図

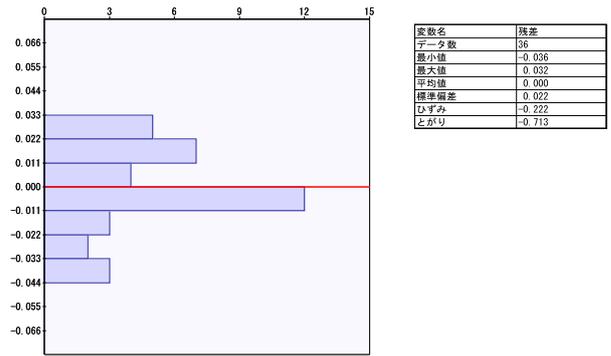


図 32. 残差のヒストグラム

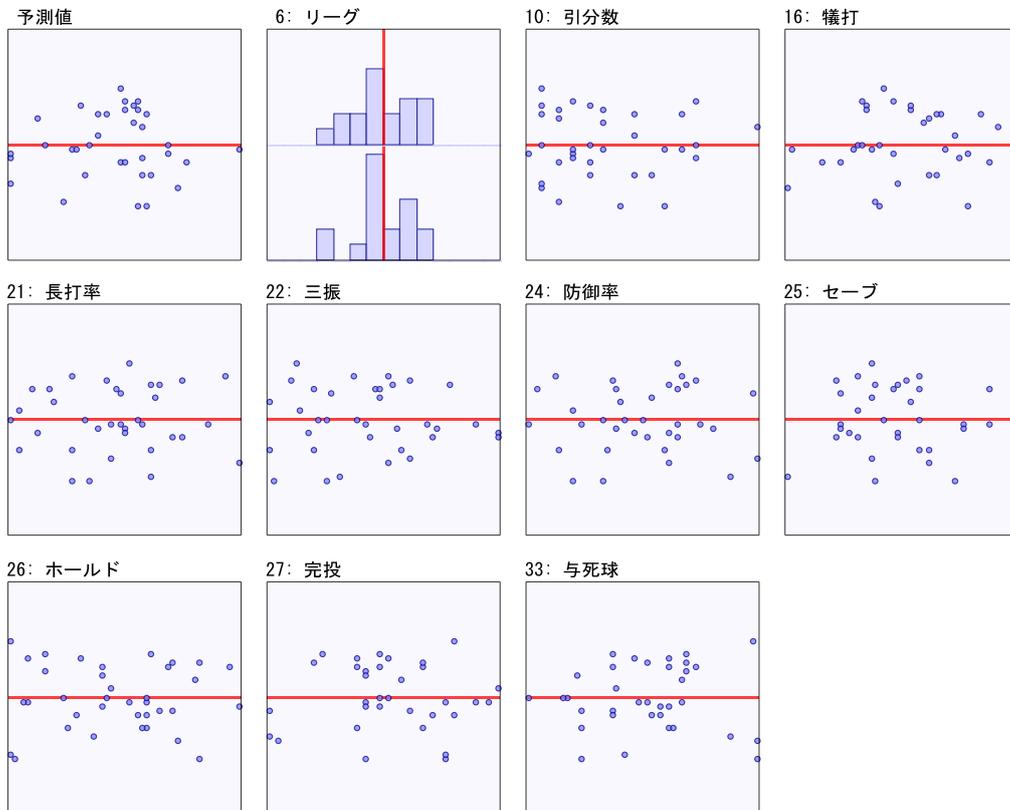


図 33. 残差と各変数の連関図

6. 主成分分析

次に主成分分析を行う。
全ての変数を用いて主成分分析を行う。

固有値

固有値を確認すると、主成分 1 から主成分 7 までの固有値が 1 を超えており、累積寄与率より、主成分 1~7 までで全体の 85.8% の情報が要約されていることがわかる。

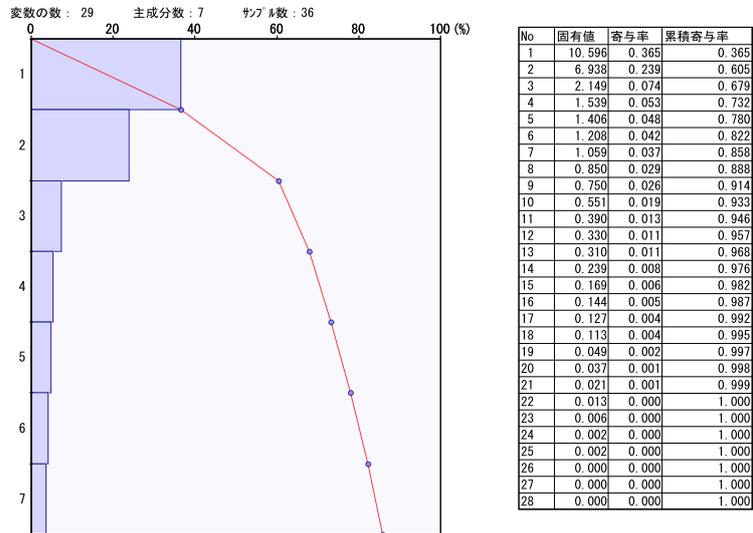


図 34. 固有値

因子負荷量

因子負荷量を確認すると、主成分 1 は、「防御率」、「被安打」、「被本塁打」、「失点」、「自責点」などの投手系の変数の相関が高いことがわかる。

また、「勝数」、「勝率」、「セーブ数」、「完封勝数」など勝ち系の変数にマイナスの符号がついており、「盗塁数」、「犠打数」などの攻撃系の変数もマイナスの符号になっている。

主成分 2 は、「順位」、「勝数」、「敗数」、「勝率」などの勝敗データとの相関が高いことがわかる。

主成分 3 は、「ホールド数」、「完投数」などの投手系の変数との相関が高いことがわかる。

No	変数名	主成分1	主成分2	主成分3	主成分4	主成分5	主成分6	主成分7	累積寄与率
1	固有値	10.596	6.938	2.149	1.539	1.406	1.208	1.059	
2	寄与率	0.365	0.239	0.074	0.053	0.048	0.042	0.037	
3	累積寄与率	0.365	0.605	0.679	0.732	0.780	0.822	0.858	
4	順位	0.278	-0.840	0.168	-0.232	0.057	0.149	-0.066	89.437
5	勝数	-0.168	0.968	0.005	0.097	0.003	-0.038	0.007	97.640
6	敗数	0.418	-0.885	-0.032	-0.065	-0.081	0.053	0.031	97.274
7	引当数	-0.753	-0.094	0.078	-0.079	0.225	-0.051	-0.108	85.282
8	勝率	-0.300	0.940	0.018	0.079	0.041	-0.046	0.002	98.380
9	得点	0.725	0.654	0.043	-0.042	0.049	-0.028	-0.030	96.144
10	失点	0.680	0.610	0.195	-0.090	-0.103	-0.063	-0.023	89.585
11	本塁打	0.615	0.544	-0.375	-0.229	0.237	-0.050	0.133	94.354
12	盗塁	-0.052	0.444	0.410	-0.409	-0.271	-0.080	0.121	63.020
13	犠打	-0.528	-0.126	0.129	0.381	-0.428	-0.402	-0.140	82.155
14	四球	0.556	0.315	0.018	0.532	0.243	-0.230	-0.210	84.728
15	死球	0.262	0.093	0.490	-0.226	0.239	-0.329	-0.497	78.078
16	打率	0.823	0.837	0.229	-0.035	-0.172	-0.075	-0.037	88.370
17	出塁率	0.694	0.615	0.213	0.206	-0.009	0.010	-0.163	97.391
18	長打率	0.704	0.657	-0.125	-0.146	0.072	-0.027	0.046	97.133
19	三振	0.315	0.060	-0.388	-0.220	0.527	0.236	-0.067	64.078
20	併殺打	0.582	-0.128	-0.097	0.596	-0.135	0.324	-0.040	84.485
21	防御率	-0.916	-0.311	-0.058	0.006	-0.085	-0.078	-0.031	95.354
22	セーブ	-0.803	0.424	-0.456	0.087	0.122	-0.131	-0.164	81.742
23	ホールド	-0.419	0.274	-0.654	0.064	-0.170	-0.030	-0.255	77.796
24	完投	-0.118	0.157	0.714	0.221	0.256	0.398	0.109	83.325
25	完封勝	-0.776	0.258	0.050	0.059	-0.148	0.044	0.030	69.965
26	奪三振	0.010	0.447	-0.100	-0.406	-0.417	0.509	0.010	80.788
27	被安打	0.875	-0.351	0.050	0.053	0.137	-0.173	-0.040	94.436
28	被本塁打	0.895	-0.115	-0.214	0.009	-0.096	-0.234	-0.033	97.067
29	与四球	0.895	0.050	-0.122	-0.000	-0.458	0.226	-0.070	78.528
30	与死球	0.338	0.152	-0.040	0.224	0.107	-0.286	0.739	82.637
31	失点	0.930	-0.298	-0.065	-0.030	-0.068	-0.077	-0.039	96.976
32	自責点	0.929	-0.276	-0.050	0.003	-0.081	-0.074	-0.036	95.446

図 35. 因子負荷量

因子負荷量散布図

主成分 1 と主成分 2 の因子負荷量の散布図を描き、軸の解釈を行う。

主成分 1 はプラス側に「長打率」、「得点」、「出塁率」、「被本塁打」、「失点」、「自責点」、「防御率」など、得点に関わる変数が並んでおり、マイナス側には「完封勝数」、「引分数」、「セーブ数」、「ホールド数」など、投手成績に関する変数が並んでいる。

これより、主成分 1 は、「投打」の軸と解釈することができる。

主成分 2 はプラス側に「勝数」、「勝率」、マイナス側に「敗数」が表れている。

これより、主成分 2 は「勝敗」の軸と解釈することができる。

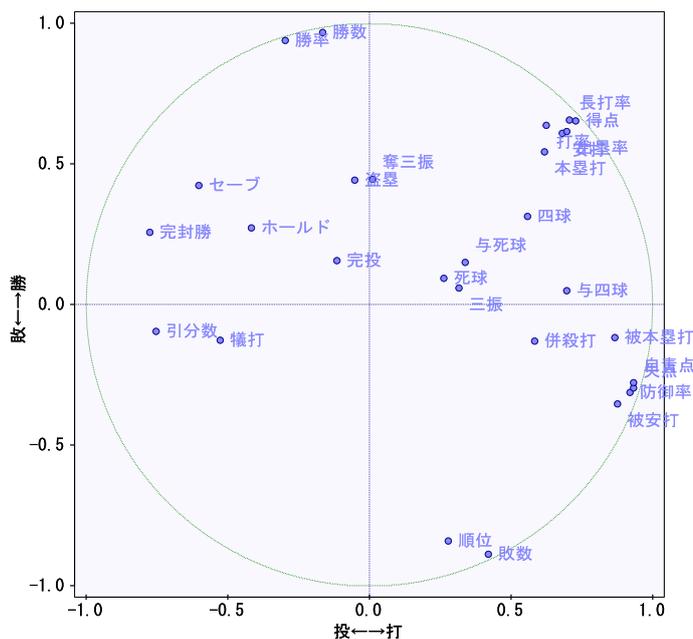


図 36. 因子負荷量散布図
(主成分 1: 投打軸と主成分 2: 勝敗軸)

次に、主成分 1 と主成分 3 の因子負荷量の散布図を描き、軸の解釈を行う。

主成分 3 はプラス側に「完投数」、マイナス側に「ホールド数」、「セーブ数」が表れている。

これより、主成分 3 は「完投型⇔継投型」の軸と解釈することができる。

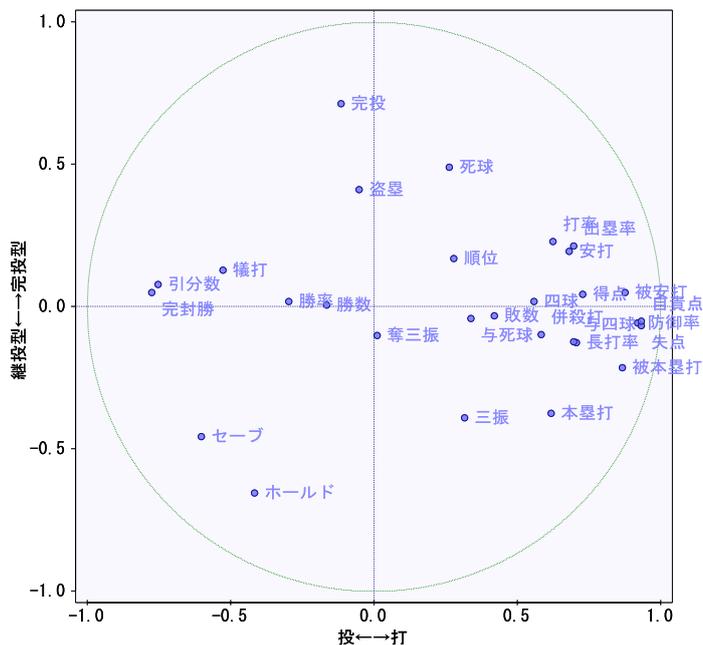


図 37. 因子負荷量散布図
(主成分 1: 投打軸と主成分 3: 完投・継投軸)

主成分得点散布図 (主成分 1(投打軸)・主成分 2(勝敗軸))

主成分 1(投打軸)と主成分 2(勝敗軸)の主成分得点の散布図を描き、各サンプルの傾向を検討する。

2011 年ソフトバンク, 2009 年巨人の強さが目立つ。

解釈しやすくするために、「年」、「チーム」、「リーグ」の各質的変数で層別した主成分得点の散布図を描く。

「年」で層別すると、2011 年は主成分 1(投打軸)において、明らかに投手系の変数に寄った傾向を示している。

さらに、2011 年に優勝したソフトバンクと中日はより「投」寄りの傾向を示しているので、2011 年は投手力のあるチームが上位になったと解釈できる。

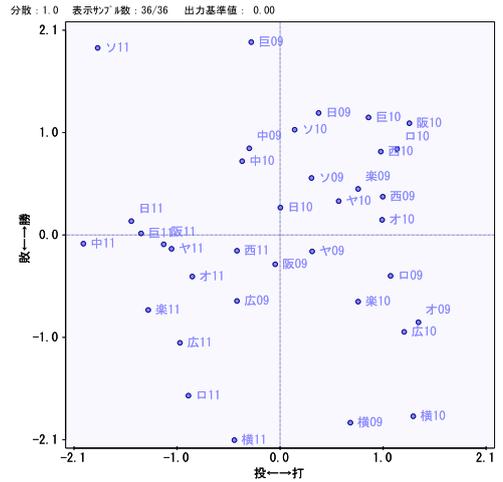


図 38. 主成分得点散布図
(主成分 1(投打軸)と主成分 2(勝敗軸))

層別した主成分得点散布図

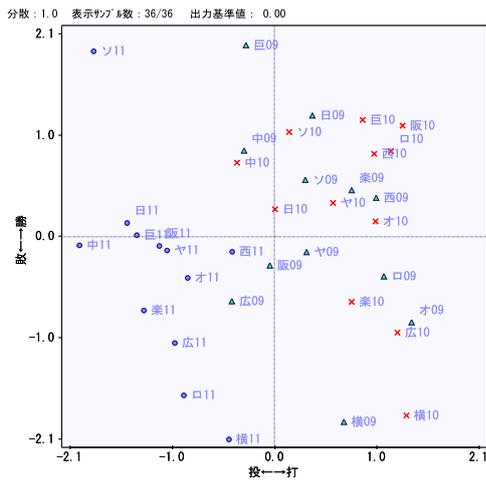


図 39. 主成分得点散布図(「年」層別)

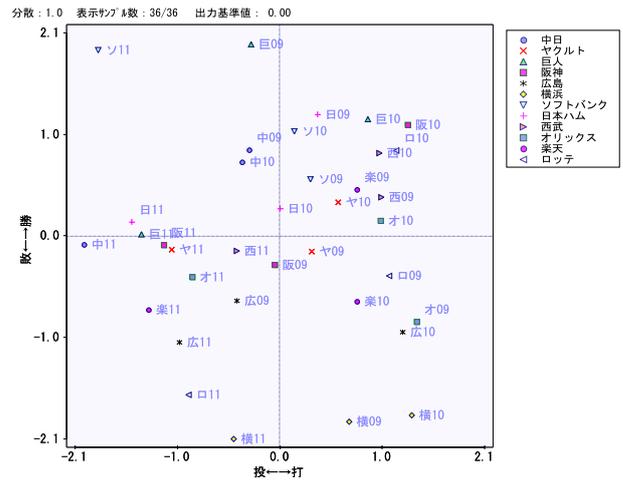


図 40. 主成分得点散布図(「チーム」層別)

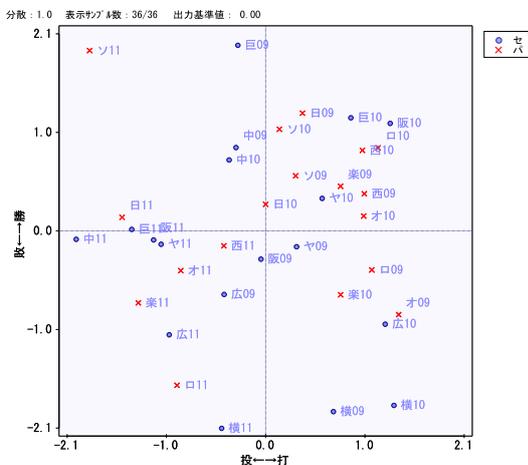


図 41. 主成分得点散布図(「リーグ」層別)

「チーム」層別すると、中日は「投」のチーム、西武、横浜は「打」のチームと言える。

「リーグ」で層別すると、リーグ別に特筆する特徴的な傾向は見られない。

主成分得点散布図 (主成分 1(投打軸)・主成分 3(完投・継投軸))

主成分 1(投打軸)と主成分 3(完投・継投軸)の主成分得点の散布図を描き、各サンプルの傾向を検討する。

完投型にはパリーグのチームが多く、継投型にはセリーグのチームが多いことがわかる。

解釈しやすくするために、「年」、「チーム」、「リーグ」の各質的変数で層別した主成分得点の散布図を描く。

「年」で層別すると、2009 年および 2010 年については打撃型のチームは完投型、投手型のチームは継投型の傾向が強い。それに対して、2011 年は、打撃型のチームは継投型、投手型のチームは完投型の傾向が強い。

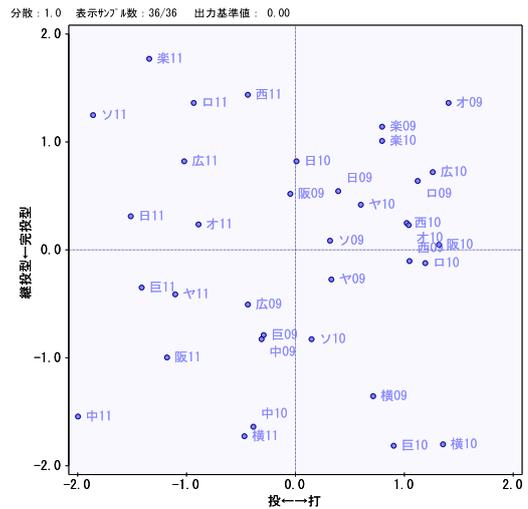


図 42. 主成分得点散布図
(主成分 1(投打軸)と主成分 3(完投・継投軸))

層別した主成分得点散布図

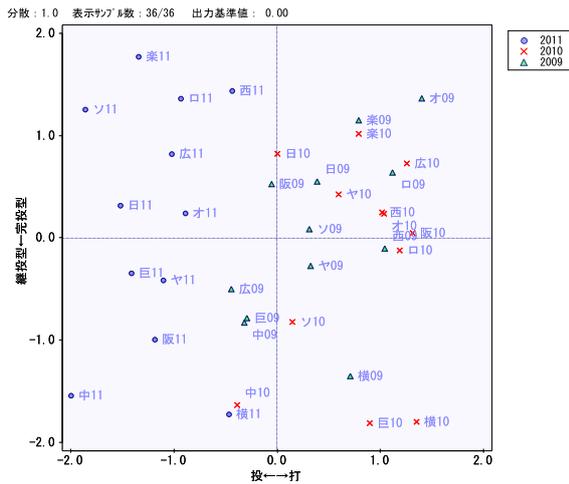


図 43. 主成分得点散布図(「年」層別)

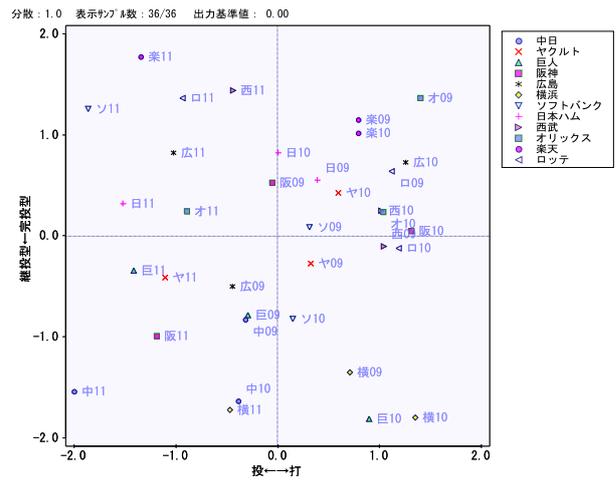


図 44. 主成分得点散布図(「チーム」層別)

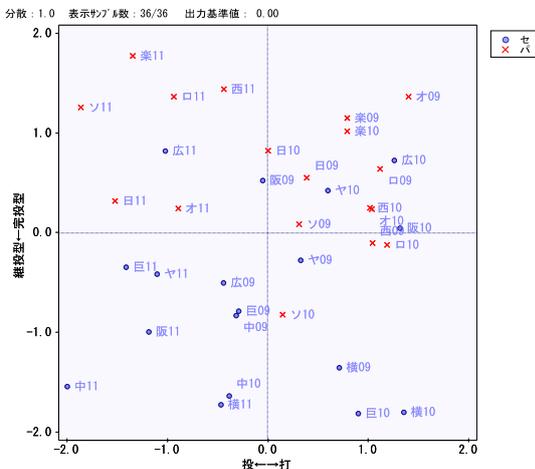


図 45. 主成分得点散布図(「リーグ」層別)

「チーム」で層別すると、中日、横浜は継投型のチームと言える。

「リーグ」で層別すると、パリーグは完投型、セリーグは継投型という傾向が顕著に確認できる。これは、パリーグにおける指名打者(DH)制が関係していると考えられる。

同時布置図

同時布置図の機能では、因子負荷量の情報と主成分得点の情報を同時に表示できる。

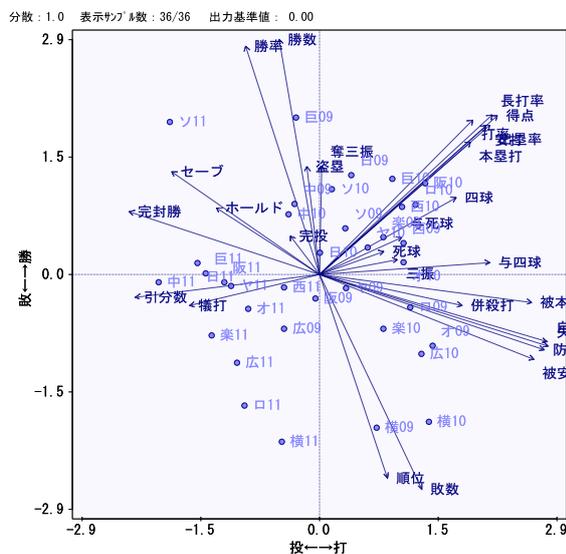


図 46. 同時布置図 (主成分 1 と主成分 2)

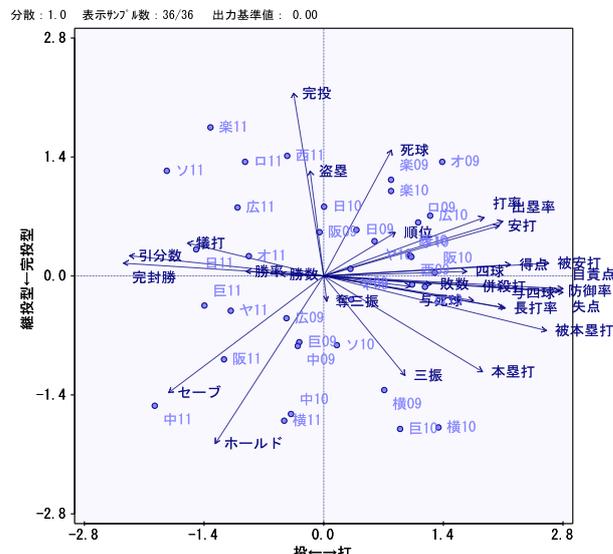


図 47. 同時布置図 (主成分 1 と主成分 3)

7. クラスタ分析

JUSE-StatWorks/V5 では、主成分分析で作られた主成分の情報を用いてクラスタ分析を行うことができる。

主成分得点, 主成分得点グラフ, 主成分得点散布図, 同時布置図の各画面において、「解析操作」メニューの「クラスタ分析」ボタンをクリックした後、クラスタ分析に用いる主成分を指定することによりクラスタ分析の解析画面(デンドログラム)が表示される。

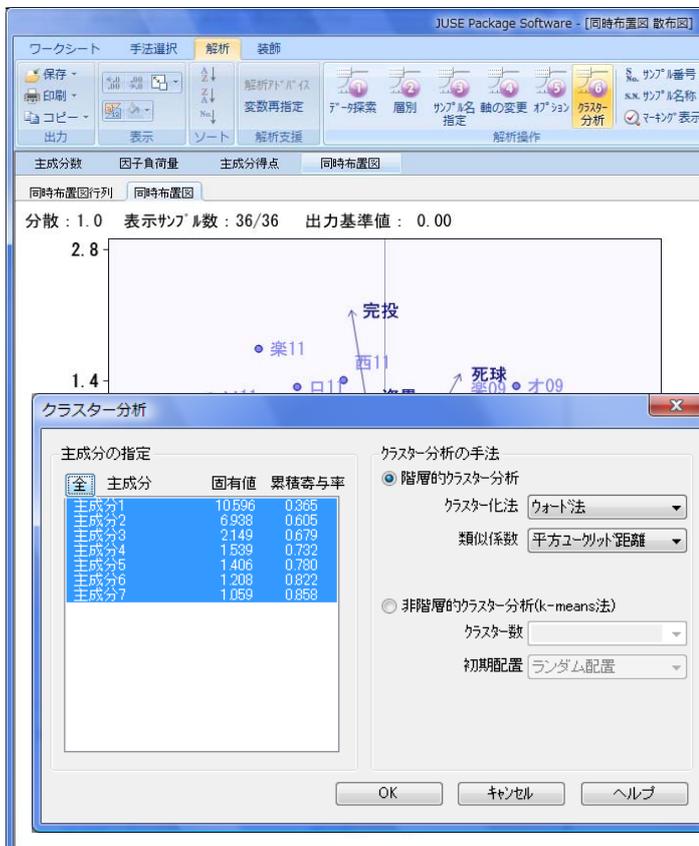


図 48. 主成分分析からクラスタ分析を実行

デンドログラム

デンドログラムは各クラスター間の結合レベルを樹形図で表したもので、凝集過程を視覚的に把握することができる。

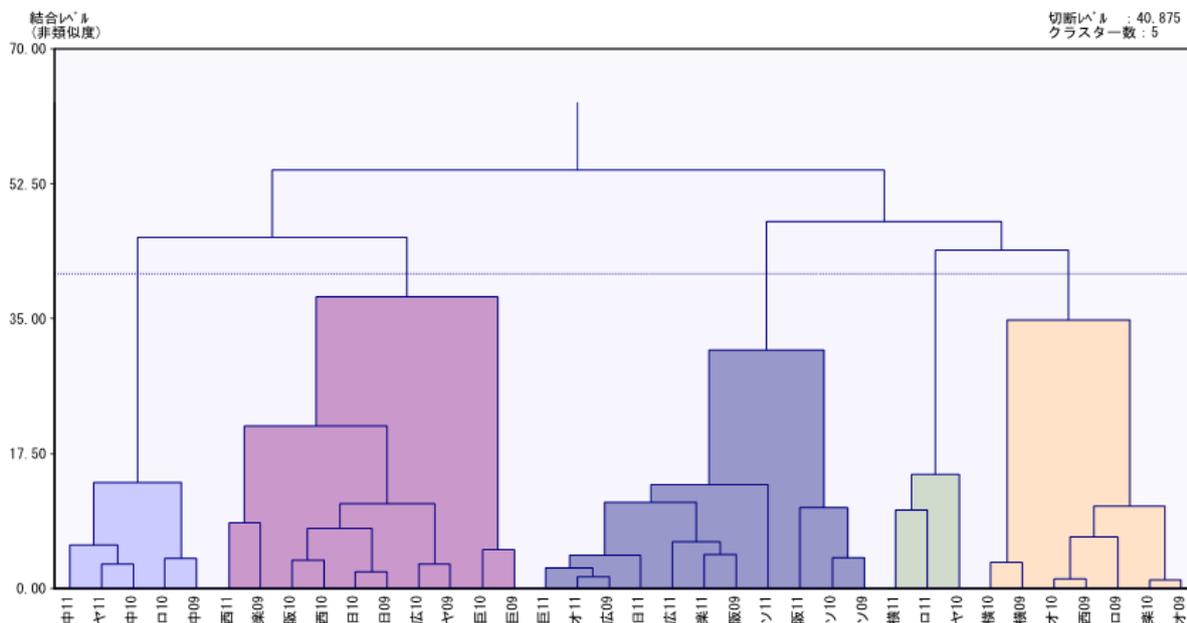


図 49. デンドログラム

デンドログラムの結果と固有技術的な知見より、以下のような傾向を読み取ることができる。

- 中日のデータは、3 年分のどの年のデータも同一のクラスターに属し、一貫したチームカラーを示していると言える。
- 2011 横浜, 2011 ロッテ, 2010 ヤクルトが属するクラスターは、成績不振のクラスターと解釈できる。

クラスター情報を主成分分析に反映する

デンドログラム画面において、「解析操作」メニューの「クラスター情報保存」ボタンをクリックすると、クラスター情報をワークシートへ登録できる。

クラスター情報を登録後、主成分分析の解析画面に戻り、クラスター情報を用いた解析ができる。主成分得点散布図に戻り、「層別」ボタンで「クラスター番号」で層別する。

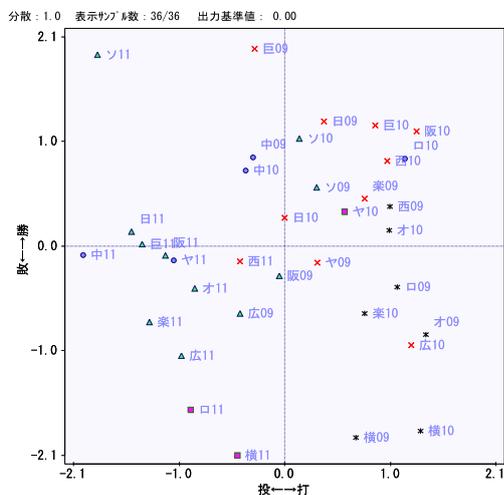


図 50. 主成分得点散布図 (主成分 1 と主成分 2)

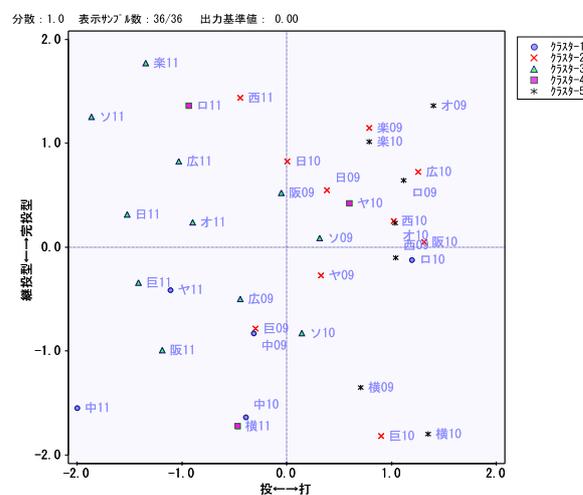


図 51. 主成分得点散布図 (主成分 1 と主成分 3)

クラスター番号で層別した主成分得点散布図より、各クラスターは以下のように解釈できる。

- クラスター1 は、投高打低の継投型
- クラスター2 は、バランス型
- クラスター3 は、投高打低の完投型
- クラスター4 は、成績不振(中途半端)型
- クラスター5 は、投低打高型

8. 主成分を説明変数にした重回帰分析

JUSE-StatWorks/V5 では、主成分分析で作られた主成分の情報を用いて重回帰分析を行うことができる。

因子負荷量、因子負荷量グラフ、主成分得点、主成分得点グラフの各画面において、「解析操作」メニューの「重回帰分析」ボタンをクリックした後、目的変数と説明変数の候補とする主成分を指定することにより、重回帰分析の解析(変数選択)画面が表示される。

ここでは、目的変数を「勝率」とし、主成分 1 から主成分 7 までを説明変数の候補として指定した後、[OK]ボタンをクリックして、重回帰分析の解析(変数選択)画面へ進む。



図 52. 主成分分析から重回帰分析を実行

変数選択

変数選択画面において、各主成分を説明変数として取り込むか否かを検討することにより重回帰モデルを構築する。

ここでは、「主成分 1(投打軸)」、「主成分 2(勝敗軸)」、「主成分 4」、「主成分 5」、「主成分 6」が説明変数として取り込まれ、自由度二重調整済寄与率 0.978、残差標準偏差 0.010、残差自由度 30 の回帰式が得られた。

変数選択	選択履歴	SE変化グラフ	偏回帰プロット一覧	偏回帰プロット	偏回帰残差一覧	
目的変数名	勝率	重相関係数	0.992	寄与率R ²	0.983	
		残差自由度	30	残差標準偏差	0.010	
				R* ²	0.981	
				R** ²	0.978	
vNo	説明変数名	分散比	P値 (上側)	偏回帰係数	標準偏回帰	トレランス
0	定数項	83768.9735	0.000	0.500		
2	投→打	163.2759	0.000	-0.022	-0.300	1.000
3	敗→勝	1605.6567	0.000	0.070	0.940	1.000
4	継投型→完	0.5531	0.483	+		
5	主成分4	11.3795	0.002	0.006	0.079	1.000
6	主成分5	3.0553	0.091	0.003	0.041	1.000
7	主成分6	3.9036	0.057	-0.003	-0.046	1.000
8	主成分7	0.0077	0.930	+		

図 53. 変数選択(主成分を用いた重回帰分析)

ここで、「主成分3(完投・継投軸)」は、説明変数として取り込まれなかった。このことにより、チームの投手事情が完投型であっても継投型であっても、勝率に対しては関係がないと考えられる。

また、ここまでに、主成分 1, 2, 3 について軸の解釈は行ってあったが、主成分 4 以降の軸の解釈は行っていなかった。説明変数として取り込まれた主成分 4, 5, 6 についても、軸の解釈を行う必要がある。

9. まとめ

以上の解析より、以下の通り確認できた。

① 「統一球」の導入による影響

2011 年は 2009 年、2010 年に比べて、投高打低の傾向が顕著に見られ、投手(守備側)に有利であったと言える。

また、チーム別には、2009 年、2010 年とも投手系の成績が良かった中日は、もともと投高打低のチームカラーであったと言え、2011 年統一球の影響を最も受けていないと考えられる。

これとは対照に、2009 年、2010 年打高のチームは、2011 年は苦しんだと言える。

② セ・パ両リーグの審判員の統合による影響

2009 年、2010 年に比べて、2011 年に四球の数が少ないのは、審判員統合によるストライクゾーンが広めになった影響と考えられる。これも 2011 年が投手有利になった原因の 1 つと言える。

③ 試合時間 3 時間 30 分ルールを導入による影響

2009 年、2010 年に比べて、2011 年は引分の数が多い。

しかし、引分数が多いチームが上位になっているとは認められず、引分数と順位の見いだせない。引分数が多いチームが有利とは言えない。

10. JUSE-StatWorks/V5 における主な機能強化項目

◆重回帰分析・数量化 I 類

No	解析手法	区分	強化項目・変更項目
1	重回帰分析・数量化 I 類	変更	解析可能な変数の数を拡張しました。 (説明変数: 1~255→1~999)
2	重回帰分析・数量化 I 類	変更	二重タブ化により出力画面を整理しました。
3	重回帰分析・数量化 I 類	新規	注目すべき値が表示されたセルを着色する強調表示機能を新規に追加しました。
4	重回帰分析・数量化 I 類	新規	「変数選択」画面に分散比の p 値, 標準偏回帰係数, トレランス等 を出力できるようにしました。また, 残差平方和はデフォルトで非 表示としました。
5	重回帰分析・数量化 I 類	新規	変数選択状態のデフォルトを選択できるようになりました(手動選 択(定数項のみ選択)/全変数選択)。
6	重回帰分析・数量化 I 類	新規	「選択履歴」画面に AIC を出力できるようにしました(デフォルトは 非表示)。
7	重回帰分析・数量化 I 類	新規	「確定モデル」画面を新規に追加しました。 「確定モデル」画面の出力項目は「変数選択」画面の出力項目と基 本的に同じですが, 「変数選択」画面で選択された変数のみを表 示する点が異なります。更に「確定モデル」画面には, 回帰式を式 の形で表わしたのも出力します。
8	重回帰分析・数量化 I 類	変更	「回帰係数」画面上にモデル評価用統計量(寄与率 R^2 など)は 出力しないようにしました。
9	重回帰分析・数量化 I 類	新規	「予測判定グラフ」画面に「寄与率 R^2 », 「 R^*2 », 「 R^{**2} 」を出力す るようにしました。
10	重回帰分析・数量化 I 類	新規	解析手法「重回帰分析」の機能統合に伴い, 「分散分析表」画面 を新規に追加しました。
11	重回帰分析・数量化 I 類	新規	解析手法「重回帰分析」の機能統合に伴い, 「残差一覧表」画面 で「残差」を「基準化残差」に切り替えることができました。
12	重回帰分析・数量化 I 類	変更	用語「t 値」を「残差 t 値」に変更しました。 また, 「残差一覧表」画面の上部に「残差 t 値: 外的スチューデント 化残差」の表示を新規に追加しました。
13	重回帰分析・数量化 I 類	変更	「残差一覧表」画面の「残差 t 値」に対する出力基準値のデフォ ルトを「2.0」に変更しました(V4.0 では「1.5」)。
14	重回帰分析・数量化 I 類	変更	「テコ比と残差 t 値」画面の補助線を, 「残差一覧表」画面の出力 基準値(全サンプルの場合はデフォルト値)と連動させるようにし ました。また, 補助線外のプロットはデフォルトでサンプル名称を 表示するようにしました。
15	重回帰分析・数量化 I 類	変更	「残差との連関図」画面の散布図, 層別ヒストグラムを拡大した場 合, 拡大後のグラフを, V4.0 では新規ウィンドウに表示しましたが, V5 では隣のタブ(「連関図(拡大)」タブ)に表示するようにし ました。
16	重回帰分析・数量化 I 類	変更	「予測」画面の予測条件・結果は, 回帰式が変わらない限りはタブ を切り替えても保持するようにしました。 これに伴い, リボンコントロール上に「データクリア」ボタンを新規に 追加しました。

◆主成分分析, 因子分析, 数量化Ⅲ類

No	解析手法	区分	強化項目・変更項目
1	主成分分析, 因子分析, 数量化Ⅲ類	変更	解析可能な変数の数を拡張しました。 (解析対象変数: 2~256→2~1000)
2	主成分分析, 因子分析, 数量化Ⅲ類	変更	二重タブ化により出力画面を整理しました。
3	主成分分析, 因子分析	新規	注目すべき値が表示されたセルを着色する強調表示機能を新規に追加しました。
4	主成分分析, 因子分析, 数量化Ⅲ類	変更	V4.0 の「基本表示」画面は廃止しました。 これは、V4.0 の「基本表示」画面の出力内容が他の画面と重複しており、しかも中途半端であることが理由です。
5	主成分分析, 因子分析, 数量化Ⅲ類	変更	初期表示画面を変更しました。 (主成分分析:「固有値」画面, 因子分析:「因子負荷量」画面, 数量化Ⅲ類:「変数スコア表」画面)
6	主成分分析	変更	デフォルトの主成分数を、①固有値が 1 以上の主成分の数以上、②累積寄与率が 80%以上、③主成分数は 5 以上、の 3 つの条件を満たす最小の主成分の数に変更しました。
7	因子分析	変更	デフォルトで規準化バリマックス回転後の結果を出力するようにしました。
8	主成分分析, 因子分析	新規	「出発行列」画面を新規に追加しました。
9	主成分分析	新規	「因子負荷量」画面に「固有値」、「寄与率」、「累積寄与率」、「累積寄与度」の出力を新規に追加しました。
10	主成分分析	新規	規準化バリマックス回転による軸の回転機能を新規に追加しました(回転する主成分軸は任意に選択できます)。
11	主成分分析, 因子分析	新規	主成分, 因子を説明変数とする回帰分析を直ぐに実行できるようにしました。
12	主成分分析, 因子分析, 数量化Ⅲ類	新規	因子負荷量, 変数スコアの棒グラフを出力する「因子負荷量グラフ」/「変数スコアグラフ」画面を新規に追加しました。
13	主成分分析, 因子分析, 数量化Ⅲ類	変更	因子負荷量散布図, 変数スコア散布図のプロット名称を、デフォルトで全プロットに対し表示するようにしました。
14	主成分分析, 因子分析, 数量化Ⅲ類	変更	「主成分得点」/「因子得点」/「サンプルスコア」グループでは、解析対象データのサンプル数によらず、デフォルトでは常に全サンプルを表示するようにしました。
15	主成分分析, 因子分析	新規	クラスター分析(階層的, 非階層的選択可)による主成分得点, 因子得点のクラスタリングを直ぐに実行できるようにしました。
16	主成分分析, 因子分析, 数量化Ⅲ類	新規	主成分得点, 因子得点, サンプルスコアの棒グラフを出力する「主成分得点グラフ」/「因子得点グラフ」/「サンプルスコアグラフ」画面を新規に追加しました。
17	主成分分析, 因子分析, 数量化Ⅲ類	変更	主成分得点散布図, 因子得点散布図, サンプルスコア散布図のプロット名称を、デフォルトで全プロットに対し表示するようにしました。
18	主成分分析, 因子分析, 数量化Ⅲ類	新規	「同時布置図」画面を新規に追加しました。 「同時布置図」画面には、主成分得点/因子得点散布図上に因子負荷量をベクトル表示したグラフを出力します。

◆判別分析・数量化Ⅱ類, 二次判別関数

No	解析手法	区分	強化項目・変更項目
1	判別分析・数量化Ⅱ類	変更	二重タブ化により出力画面を整理しました。
2	判別分析・数量化Ⅱ類	新規	各群の重心スコアの出力を新規に追加しました。
3	判別分析・数量化Ⅱ類	変更	2 群の判別での「変数選択」画面の用語を変更しました; $(D^2 \rightarrow \text{判別効率 } D^2, (D')^2 \rightarrow D^{*2}, (D'')^2 \rightarrow D^{**2}, D^2 \text{ の差} \rightarrow \text{変化量}, F \text{ 比} \rightarrow F \text{ 値}$
4	判別分析・数量化Ⅱ類	新規	2 群の判別での「変数選択」画面において, F 値に対する p 値の出力を新規に追加しました。
5	判別分析・数量化Ⅱ類	新規	2 群の判別の場合に変数増減法による逐次変数選択を行えるようにしました。
6	判別分析・数量化Ⅱ類	新規	2 群の判別の場合に「判別関数」画面を新規追加しました。 2 群の判別の場合の「判別関数」画面の出力項目は「変数選択」画面の出力項目と基本的に同じですが, 「変数選択」画面で選択された変数のみが表示する点が異なります。更に 2 群の判別の場合の「判別関数」画面には, 判別関数を式の形で表わしたのも出力します。
7	判別分析・数量化Ⅱ類	変更	一般的判定, ジャックナイフ判定での用語を変更しました; $D^2 \rightarrow \text{マハラビス距離}, \text{正答} \rightarrow \text{正答(実測値)}, \text{誤答} \rightarrow \text{誤答(実測値)}$
8	判別分析・数量化Ⅱ類	変更	一般的判定, ジャックナイフ判定の「サンプル表示」画面で, デフォルトで全サンプルを表示するようにしました。 また, 「予測」列, 「観測」列に群名称も表示するようにしました。
9	判別分析・数量化Ⅱ類	変更	スコアのヒストグラムの目盛を, 中心に「0.000」が表示されるように変更しました。 また, 「0.000」に補助線を表示するようにしました。
10	判別分析・数量化Ⅱ類	新規	「予測」画面に, 各群の重心からのマハラビスの距離の上側確率の出力を新規に追加しました(2 群の判別の場合はスコア・確率をデフォルトで表示, 3・4 群の判別の場合は確率をデフォルトで表示)
11	判別分析・数量化Ⅱ類	変更	「予測」画面の予測条件・結果は, 判別関数に変更されるまで, 他のタブに切り替えられても保持するように変更しました。 これに伴い, リボンコントロール上に「データクリア」ボタンを新規に追加しました。
12	二次判別関数	新規	解析手法「二次判別関数」を新規に追加しました。 解析手法「二次判別関数」の特徴は次の通りです; <ul style="list-style-type: none"> ● 解析可能なデータは, 目的変数が 2 群, かつ, 説明変数が量的変数のみとなります。 ● 等分散性の検定結果が表示されます。その結果を踏まえ, 線形判別関数に変更することも可能です。 ● 変数選択を行うことはできません(「変数の指定」ダイアログで指定された変数を全て使用します)。 ● 一般的判定のみが出力されます(ジャックナイフ判定は出力されません)

◆その他の多変量解析手法

No	解析手法	区分	強化項目・変更項目												
1	単回帰分析	新規	データに繰返しがある場合に対応しました(データに繰返しがある場合は分散分析表に要因「当てはまりの悪さ」を表示).												
2	単回帰分析	新規	傾き β_1 に対する一般仮説の検定の p 値, 及び, 切片 β_0 に対する検定の p 値の出力を新規に追加しました(デフォルトでは $H_0: \beta_1=0$ の検定の p 値のみを出力).												
3	単回帰分析	新規	原点を通る回帰直線のあてはめ機能を新規に追加しました.												
4	単回帰分析	新規	層別(2 カテゴリ)の分析機能を新規に追加しました(層毎の回帰線のあてはめ, 傾き β_1 ・切片 β_0 の差の検定の p 値の出力).												
5	直交多項式回帰	変更	二重タブ化により出力画面を整理しました. (グループ:変数選択, 確定モデル, 予測)												
6	ロジスティック回帰分析		(変更無し)												
7	重みつき回帰分析	新規	出力結果を解析手法「重回帰分析・数量化 I 類」と共通化しました.												
8	AID(多段層別分析)	変更	二重タブ化により出力画面を整理しました. (グループ:カテゴリ情報, 解析)												
9	AID(多段層別分析)	変更	目的変数が質的変数(2 群)の場合に, 各画面の文字列「2 群」を「第 2 群」に変更しました. また, 第 1 群, 第 2 群のカテゴリ名を「カテゴリ情報」画面に表示するようにしました.												
10	AID(多段層別分析)	変更	分割が行われていない状態で「2進木」タブが選択された場合は, 確認メッセージ「逐次選択による分割を実行しますか?」を表示するようにしました.												
11	非階層的クラスター分析 (k-means 法)	変更	初期クラスター配置のデフォルト設定を「ランダム配置」に変更しました.												
12	非階層的クラスター分析 (k-means 法)	変更	同じクラスター数に対し異なるクラスタリング結果が得られるように変更しました(乱数のシード値設定のタイミングの変更).												
13	階層的クラスター分析	新規	クラスタ化法, 類似係数のデフォルトを次のように設定しました; <table border="1" style="margin-left: 20px;"> <thead> <tr> <th>解析対象変数</th> <th>クラスタ化法</th> <th>類似係数</th> </tr> </thead> <tbody> <tr> <td>量的変数のみ</td> <td>ワード法</td> <td>平方ユークリッド距離</td> </tr> <tr> <td>量・質混在</td> <td></td> <td></td> </tr> <tr> <td>質的変数のみ</td> <td>群平均法</td> <td>ピアソンの一致係数</td> </tr> </tbody> </table>	解析対象変数	クラスタ化法	類似係数	量的変数のみ	ワード法	平方ユークリッド距離	量・質混在			質的変数のみ	群平均法	ピアソンの一致係数
解析対象変数	クラスタ化法	類似係数													
量的変数のみ	ワード法	平方ユークリッド距離													
量・質混在															
質的変数のみ	群平均法	ピアソンの一致係数													
14	SD 法(SD プロファイル)		(変更無し)												
15	集計表解析		(変更無し)												
16	SEM(構造方程式モデリング)		(変更無し)												
17	GM(グラフィカルモデリング)	新規	有向独立グラフから SEM のパス図を自動作成する機能を新規に追加しました.												

本著作物は原著作者の許可を得て、株式会社日本科学技術研修所（以下弊社）が掲載しています。本著作物の著作権については、制作した原著作者に帰属します。

原著作者および弊社の許可なく営利・非営利・イントラネットを問わず、本著作物の複製・転用・販売等を禁止します。

所属および役職等は、公開当時のものです。

■公開資料ページ

弊社ウェブページで各種資料をご覧ください <http://www.i-juse.co.jp/statistics/jirei/>

■お問い合わせ先

(株)日科技研 数理事業部 パッケージサポート係 <http://www.i-juse.co.jp/statistics/support/contact.html>