

应用统计



第13讲 多元回归和主成分分析



多元回归分析

事物间的联系常常是多方面的，一个应变量的变化可能受到其它多个自变量的影响。

多元回归分析研究多个变量之间关系的回归分析方法。

- **目的**：作出以多个自变量估计应变量的多元线性回归方程。
- **资料**：应变量为定量指标；自变量全部或大部分为定量指标，若有少量定性或等级指标需作转换。
- **用途**：解释和预报。

实例一

27名糖尿病人的血糖及有关变量的测量结果

序号 i	总胆固醇 (mmol/L) X_1	甘油三脂 (mmol/L) X_2	胰岛素 (μ U/ml) X_3	糖化血 红蛋白(%) X_4	血糖 (mmol/L) Y
1	5.68	1.90	4.53	8.2	11.2
2	3.79	1.64	7.32	6.9	8.8
3	6.02	3.56	6.95	10.8	12.3
4	4.85	1.07	5.88	8.3	11.6
5	4.60	2.32	4.05	7.5	13.4
6	6.05	0.64	1.42	13.6	18.3
7	4.90	8.50	12.60	8.5	11.1
.....					
21	5.78	3.36	2.96	8.0	13.6
22	5.43	1.13	4.31	11.3	14.9
23	6.50	6.21	3.47	12.3	16.0
24	7.98	7.92	3.37	9.8	13.2
25	11.54	10.89	1.20	10.5	20.0
26	5.84	0.92	8.61	6.4	13.3
27	3.84	1.20	6.45	9.6	10.4



多元回归模型

序号	X_1	X_2	\cdots	X_k	Y
1	X_{11}	X_{12}	\cdots	X_{1k}	Y_1
2	X_{21}	X_{22}	\cdots	X_{2k}	Y_2
\vdots	\vdots	\vdots	\cdots	\vdots	\vdots
n	X_{n1}	X_{n2}	\cdots	X_{nk}	Y_n

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2) \text{ 相互独立}$$



多元回归方程（最小二乘）

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$$

残差

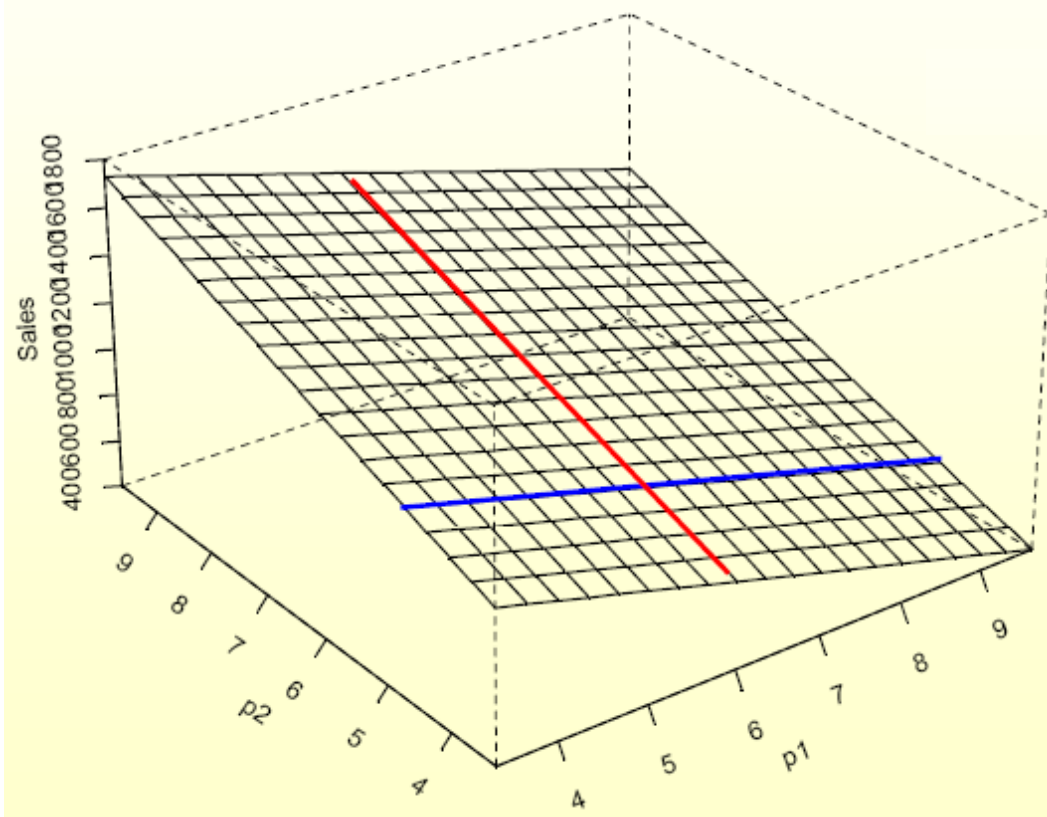
$$e_i = Y_i - \hat{Y}_i$$

选择 b_0, b_1, \dots, b_k , 使得 $\sum_{i=1}^n e_i^2$ 最小

Standard Error of the Regression :

$$s = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n - k - 1}}$$

二元回归平面





建立回归方程

$$X_R = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix}, \quad Y_R = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}, \quad \text{其中}$$
$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ji} \quad (i = 1, \dots, k)$$
$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j$$

$$L_{XX} = X_R^T X_R, \quad L_Y = X_R^T Y_R,$$

$$L_{XX} b = L_Y \Rightarrow b = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}, \quad b_0 = \bar{y} - b^T \bar{x} \Rightarrow \hat{Y} = b_0 + b_1 X_1 + \cdots + b_k X_k$$

$$\hat{Y} = 5.943 + 0.142X_1 + 0.351X_2 - 0.271X_3 + 0.638X_4 \quad (\text{实例一结果})$$

方差的分解

ANOVA (analysis of variance)

$$SST = L_{YY} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE \qquad \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

总偏差平方和：**SST**； 回归平方和：**SSR**； 剩余平方和：**SSE**

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-k-1)$$

$$\hat{\sigma}_e^2 = \frac{SSE}{n-k-1} \text{ 是 } \sigma^2 \text{ 的无偏估计}$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$\sigma^2 \text{ 的区间估计 } \left[\frac{SSE}{\chi_{1-\alpha/2}^2(n-k-1)}, \frac{SSE}{\chi_{\alpha/2}^2(n-k-1)} \right]$$

$$R_a^2 = 1 - \frac{SSE / (n-k-1)}{SST / (n-1)}$$



R^2 的解释

$0 \leq R^2 \leq 1$ ，说明自变量 X_1, X_2, \dots, X_k 能够解释 Y 变化的百分比，其值愈接近于 1，说明模型对数据的拟合程度愈好。

实例一
$$R^2 = \frac{133.711}{222.552} = 0.601$$

表明血糖含量变异的 60% 可由总胆固醇、甘油三脂、胰岛素和糖化血红蛋白的变化来解释。

回归方程的显著性检验, F检验

$$\frac{\text{SSE}}{\sigma^2} \sim \chi^2(n-k-1)$$

$$s = \sqrt{\frac{\sum_{i=1}^p e_i^2}{(n-k-1)}}, \quad s^2 \text{ 是 } \sigma^2 \text{ 的无偏估计}$$

$$H_0: \beta_1 = \cdots = \beta_k = 0 \quad \text{VS} \quad H_1: \beta_i (i=1, \dots, k) \text{ 中至少有一个不为 } 0$$

当 $H_0: \beta_1 = \cdots = \beta_k = 0$ 成立时,

$$\text{回归平方和} \quad \text{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \Rightarrow \frac{\text{SSR}}{\sigma^2} \sim \chi^2(k)$$

且SSE与SSR独立。

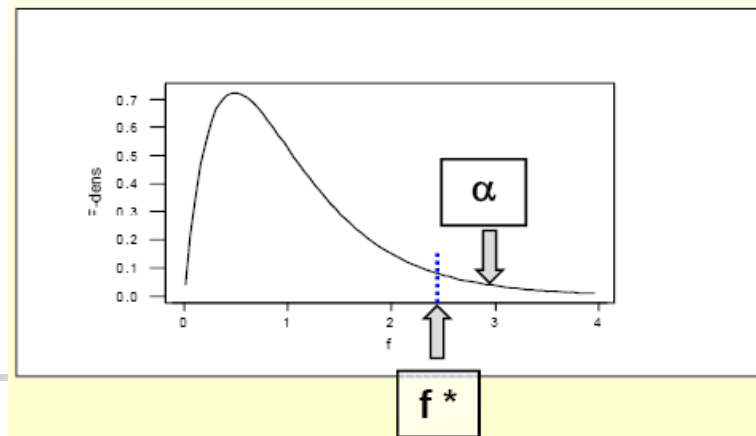
$$F = \frac{\text{SSR} / k}{\text{SSE} / (n-k-1)} \sim F(k, n-k-1)$$



方差分析表

平方和来源	平方和	自由度	
源于回归	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	k	SSR / k
源于残差	$SSE = \sum (Y_i - \hat{Y}_i)^2$	$n - k - 1$	$SSE / (n - k - 1)$
总平方和	$SST = \sum (Y_i - \bar{Y})^2$	$n - 1$	$SST / (n - 1)$

F检验 (单侧)



(1) $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$

$H_1 : \beta_1, \beta_2, \dots, \beta_k$ 不全为0

(2) 选择、（根据样本）计算统计量

$$F = \frac{SSR / k}{SSE / (n - k - 1)} \sim F(k, n - k - 1)$$

(3) 给出显著性水平 α ，查表，得 $F_\alpha(k, n - k - 1)$;

(4) 判断：若 $F \geq F_\alpha(k, n - k - 1)$, 拒绝原假设，
接受备择假设,。反之则反。



SPSS输出, R^2 拟合优度, 标准差

Variables Entered/Removed^b

Mode 	Variables Entered	Variables Removed	Method
1	糖化血红蛋白, 甘油三酯, 胰岛素, 总胆固醇 ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: 血糖

Model Summary

Mode 	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.775 ^a	.601	.528	2.00954

a. Predictors: (Constant), 糖化血红蛋白, 甘油三酯, 胰岛素, 总胆固醇



SPSS输出，方差分析表

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	133.711	4	33.428	8.278	.000 ^a
	Residual	88.841	22	4.038		
	Total	222.552	26			

a. Predictors: (Constant), 糖化血红蛋白, 甘油三酯, 胰岛素, 总胆固醇

b. Dependent Variable: 血糖



回归系数的分布

$$b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \bar{\mathbf{x}}^T L_{XX} \bar{\mathbf{x}}\right)\right), \quad b_i \sim N(\beta_i, \sigma^2 c_{ii}) \quad (i = 1, \dots, k)$$

其中 $L_{XX}^{-1} = [c_{ij}]_{p \times p}$

$$\frac{b_0 - \beta_0}{\sigma \sqrt{\frac{1}{n} + \bar{\mathbf{x}}^T L_{XX} \bar{\mathbf{x}}}} \sim N\left(0, \sigma^2\left(\frac{1}{n} + \bar{\mathbf{x}}^T L_{XX} \bar{\mathbf{x}}\right)\right), \quad \frac{b_i - \beta_i}{\sigma \sqrt{c_{ii}}} \sim N(0, 1) \quad (i = 1, \dots, k)$$



回归方程系数的 t 检验

$H_0 : \beta_1 = \cdots = \beta_k = 0$ VS $H_1 : \beta_i (i = 1, \cdots, k)$ 中至少有一个不为 0

当 $H_0 : \beta_i = 0$ 成立时, $\frac{b_i \sqrt{c_{ii}}}{\sigma} \sim N(0, 1)$, $\frac{b_i^2 c_{ii}}{\sigma^2} \sim \chi^2(1)$

$$F_i = \frac{\frac{b_i^2}{c_{ii}}}{\frac{\text{SSE}}{(n-k-1)}} \sim F(1, n-k-1)$$

$$T_i = \frac{\frac{b_i}{\sqrt{c_{ii}}}}{\sqrt{\frac{\text{SSE}}{(n-k-1)}}} \sim t(n-k-1)$$

F_i也叫偏回归平方和, 其值越大表明相应的自变量分量越重要

SPSS输出，回归系数的t检验

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	5.943	2.829		2.101	.047
	总胆固醇	.142	.366	.078	.390	.701
	甘油三酯	.351	.204	.309	1.721	.099
	胰岛素	-.271	.121	-.339	-2.229	.036
	糖化血红蛋白	.638	.243	.398	2.623	.016

a. Dependent Variable: 血糖

$$\hat{Y} = 5.943 + 0.142X_1 + 0.351X_2 - 0.271X_3 + 0.638X_4$$



多元回归方程（最小二乘）

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k$$

残差

$$e_i = Y_i - \hat{Y}_i$$

选择 b_0, b_1, \dots, b_k , 使得 $\sum_{i=1}^n e_i^2$ 最小

Standard Error of the Regression : $s = \sqrt{\frac{\sum_{i=1}^p e_i^2}{n - k - 1}}$



建立回归方程

$$X_R = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & \cdots & x_{2k} - \bar{x}_k \\ \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & \cdots & x_{nk} - \bar{x}_k \end{bmatrix}, \quad Y_R = \begin{bmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}, \quad \text{其中}$$
$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ji} \quad (i = 1, \dots, k)$$
$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j$$

$$L_{XX} = X_R^T X_R, \quad L_Y = X_R^T Y_R,$$

$$L_{XX} b = L_Y \Rightarrow b = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}, \quad b_0 = \bar{y} - b^T \bar{x} \Rightarrow \hat{Y} = b_0 + b_1 X_1 + \cdots + b_k X_k$$

$$\hat{Y} = 5.943 + 0.142X_1 + 0.351X_2 - 0.271X_3 + 0.638X_4 \quad (\text{实例一结果})$$



变量选择

- 回归方程包含的自变量越多，回归平方和越大，剩余的平方和越小，剩余均方也随之较小，预测值的误差也愈小，模拟的效果愈好。
- 但是方程中的变量过多，预报工作量就会越大，其中有些相关性不显著的预报因子会影响预测的效果。
- 因此在多元回归模型中，选择适宜的变量数目进行回归尤为重要。



全部比较法

1. 校正判定系数 R_a^2 选择法

2. C_p 选择法



1. 校正判定系数选择法

- 判定系数的定义：

$$SST = SSR + SSE \Rightarrow 1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{TSS}$$

- 意义：判定系数**越大**，自变量对因变量的解释程度**越高**，自变量引起的变动占总变动的**百分比高**。观察点在回归直线附近**越密集**。
- 取值范围：0-1



校正判定系数 R_a^2

■ 为什么要校正？

- 判定系数随解释变量个数的增加而增大。易造成错觉：要模型拟合得越好，就应增加解释变量。然而增加解释变量会降低自由度，减少可用的样本数。并且有时增加解释变量是不必要的。
- 导致解释变量个数不同模型之间对比困难。
- 判定系数只涉及平方和，没有考虑自由度。

■ 校正思路：

引进自由度校正所计算的平方和。



校正判定系数 R_a^2

$$R_a^2 = 1 - \frac{SSE / (n - k - 1)}{SST / (n - 1)}$$

校正判定系数和未校正的判定系数的关系：

$$(1) \quad R_a^2 = 1 - (1 - R^2) \frac{n - 1}{n - k - 1}$$

(2) $k > 1$ 时, $R_a^2 < R^2$, 且随着解释变量的增加两者的差距将越来越大。

也就是说校正的比未校正的判定系数增加得慢些！

(3) 判定系数 R^2 非负（取值在 $[0, 1]$ ）；

但是, R_a^2 取值可能为负, 这时规定 $R_a^2 = 0$



拟合优度 R^2 和F检验的关系

- (1) 都是对回归方程的显著性检验;
- (2) 都是把总平方和分解, 以构成统计量进行检验;
- (3) 两者同增同减, 具有一致性。

在数量上, 它们有如下关系

$$F = \frac{n-k-1}{k} \times \frac{R^2}{1-R^2}, \quad R_a^2 = 1 - \frac{n-1}{n-k-1+k \cdot F}$$



实例一的全部比较法结果

方程中的 自变量	R_a^2	C_p	方程中的 自变量	R_a^2	C_p
X_2, X_3, X_4	0.546	3.15	X_2, X_3	0.408	9.14
X_1, X_2, X_3, X_4	0.528	5.00	X_1, X_3	0.375	10.78
X_1, X_3, X_4	0.488	5.96	X_4	0.347	11.63
X_1, X_2, X_4	0.447	7.97	X_1	0.284	14.92
X_1, X_4	0.441	7.42	X_1, X_2	0.275	15.89
X_2, X_4	0.440	7.51	X_3	0.231	17.77
X_3, X_4	0.435	7.72	X_2	0.179	20.53
X_1, X_2, X_3	0.408	9.88			



选择变量的逐步回归法

1. **前进法**，回归方程中的自变量从无到有、从少到多逐个引入回归方程。**此法已基本淘汰。**

2. **后退法**，先将全部自变量选入方程，然后逐步剔除无统计学意义的自变量。

剔除自变量的方法是在方程中选一个偏回归平方和最小的变量，作 F 检验决定它是否剔除，若无统计学意义则将其剔除，然后对剩余的自变量建立新的回归方程。重复这一过程，直至方程中所有的自变量都不能剔除为止。理论上最好，建议使用采用此法。

3. **逐步回归法**，逐步回归法是在前述两种方法的基础上，进行**双向筛选**的一种方法。该方法本质上是前进法。



逐步回归法SPSS输出结果

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	糖化血红蛋白	.	Stepwise (Criteria: Probability-of- F-to-enter \leq . 050, Probability-of- F-to-remove \geq .100).
2	总胆固醇	.	Stepwise (Criteria: Probability-of- F-to-enter \leq . 050, Probability-of- F-to-remove \geq .100).

a. Dependent Variable: 血糖

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.610 ^a	.372	.347	2.36506
2	.696 ^b	.484	.441	2.18672

a. Predictors: (Constant), 糖化血红蛋白

b. Predictors: (Constant), 糖化血红蛋白, 总胆固醇

ANOVA^c

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	82.714	1	82.714	14.788	.001 ^a
	Residual	139.837	25	5.593		
	Total	222.552	26			
2	Regression	107.790	2	53.895	11.271	.000 ^b
	Residual	114.762	24	4.782		
	Total	222.552	26			

a. Predictors: (Constant), 糖化血红蛋白

b. Predictors: (Constant), 糖化血红蛋白, 总胆固醇

c. Dependent Variable: 血糖



回归系数

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.006	2.364		1.272	.215
	糖化血红蛋白	.978	.254	.610	3.845	.001
2	(Constant)	1.310	2.308		.568	.576
	糖化血红蛋白	.732	.259	.456	2.833	.009
	总胆固醇	.678	.296	.369	2.290	.031

a. Dependent Variable: 血糖

选择淘汰的过程

Excluded Variables^c

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	总胆固醇	.369 ^a	2.290	.031	.423	.828
	甘油三酯	.341 ^a	2.269	.033	.420	.952
	胰岛素	-.347 ^a	-2.222	.036	-.413	.891
2	甘油三酯	.210 ^b	1.112	.278	.226	.599
	胰岛素	-.274 ^b	-1.785	.088	-.349	.834

a. Predictors in the Model: (Constant), 糖化血红蛋白

b. Predictors in the Model: (Constant), 糖化血红蛋白, 总胆固醇

c. Dependent Variable: 血糖

调整选择淘汰的显著性水平

Variables Entered/Removed^a

Mo...	Variables Entered	Variables Removed	Method
1	糖化血红蛋白	.	Stepwise (Criteria: Probability-of-F-to-enter \leq .100, Probability-of-F-to-remove \geq .150).
2	总胆固醇	.	Stepwise (Criteria: Probability-of-F-to-enter \leq .100, Probability-of-F-to-remove \geq .150).
3	胰岛素	.	Stepwise (Criteria: Probability-of-F-to-enter \leq .100, Probability-of-F-to-remove \geq .150).
4	甘油三酯	.	Stepwise (Criteria: Probability-of-F-to-enter \leq .100, Probability-of-F-to-remove \geq .150).
5	.	总胆固醇	Stepwise (Criteria: Probability-of-F-to-enter \leq .100, Probability-of-F-to-remove \geq .150).

a. Dependent Variable: 血糖

选择淘汰过程

Excluded Variables^e

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	总胆固醇	.369 ^a	2.290	.031	.423	.828
	甘油三酯	.341 ^a	2.269	.033	.420	.952
	胰岛素	-.347 ^a	-2.222	.036	-.413	.891
2	甘油三酯	.210 ^b	1.112	.278	.226	.599
	胰岛素	-.274 ^b	-1.785	.088	-.349	.834
3	甘油三酯	.309 ^c	1.721	.099	.344	.562
5	总胆固醇	.078 ^d	.390	.701	.083	.458

a. Predictors in the Model: (Constant), 糖化血红蛋白

b. Predictors in the Model: (Constant), 糖化血红蛋白, 总胆固醇

c. Predictors in the Model: (Constant), 糖化血红蛋白, 总胆固醇, 胰岛素

d. Predictors in the Model: (Constant), 糖化血红蛋白, 胰岛素, 甘油三酯

e. Dependent Variable: 血糖

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.006	2.364		1.272	.215
	糖化血红蛋白	.978	.254	.610	3.845	.001
2	(Constant)	1.310	2.308		.568	.576
	糖化血红蛋白	.732	.259	.456	2.833	.009
	总胆固醇	.678	.296	.369	2.290	.031
3	(Constant)	4.309	2.776		1.552	.134
	糖化血红蛋白	.635	.253	.396	2.507	.020
	总胆固醇	.545	.293	.297	1.861	.076
	胰岛素	-.219	.122	-.274	-1.785	.088
4	(Constant)	5.943	2.829		2.101	.047
	糖化血红蛋白	.638	.243	.398	2.623	.016
	总胆固醇	.142	.366	.078	.390	.701
	胰岛素	-.271	.121	-.339	-2.229	.036
	甘油三酯	.351	.204	.309	1.721	.099
5	(Constant)	6.500	2.396		2.713	.012
	糖化血红蛋白	.663	.230	.413	2.880	.008
	胰岛素	-.287	.112	-.360	-2.570	.017
	甘油三酯	.402	.154	.354	2.612	.016

a. Dependent Variable: 血糖



R^2 与 R_a^2 比较

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.610 ^a	.372	.347	2.36506
2	.696 ^b	.484	.441	2.18672
3	.740 ^c	.547	.488	2.09351
4	.775 ^d	.601	.528	2.00954
5	.773 ^e	.598	.546	1.97213

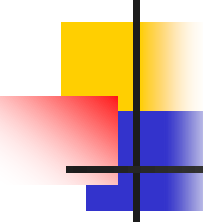
a. Predictors: (Constant), 糖化血红蛋白

b. Predictors: (Constant), 糖化血红蛋白, 总胆固醇

c. Predictors: (Constant), 糖化血红蛋白, 总胆固醇, 胰岛素

d. Predictors: (Constant), 糖化血红蛋白, 总胆固醇, 胰岛素, 甘油三酯

e. Predictors: (Constant), 糖化血红蛋白, 胰岛素, 甘油三酯

ANOVA^f


Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	82.714	1	82.714	14.788	.001 ^a
	Residual	139.837	25	5.593		
	Total	222.552	26			
2	Regression	107.790	2	53.895	11.271	.000 ^b
	Residual	114.762	24	4.782		
	Total	222.552	26			
3	Regression	121.748	3	40.583	9.260	.000 ^c
	Residual	100.804	23	4.383		
	Total	222.552	26			
4	Regression	133.711	4	33.428	8.278	.000 ^d
	Residual	88.841	22	4.038		
	Total	222.552	26			
5	Regression	133.098	3	44.366	11.407	.000 ^e
	Residual	89.454	23	3.889		
	Total	222.552	26			

a. Predictors: (Constant), 糖化血红蛋白

b. Predictors: (Constant), 糖化血红蛋白, 总胆固醇

c. Predictors: (Constant), 糖化血红蛋白, 总胆固醇, 胰岛素

d. Predictors: (Constant), 糖化血红蛋白, 总胆固醇, 胰岛素, 甘油三酯

e. Predictors: (Constant), 糖化血红蛋白, 胰岛素, 甘油三酯

f. Dependent Variable: 血糖

影响我国区域科技资源配置 效率要素的定量分析^{*}

李石柱¹, 李冬梅², 唐五湘²

(1 北京科学技术委员会发展计划处, 北京; 2 北京机械工业学院工商管理分院, 北京 100085)

摘要: 本文运用经济计量学中的回归分析法, 对影响我国区域科技资源配置效率水平的要素进行定量分析, 研究确定科技资源配置效率的主要影响因素, 为制定合理的科技资源配置政策提供依据。

关键词: 科技资源配置效率; 要素; 回归分析

中图分类号: F204 文献标识码: A 文章编号: 1004- 115X(2003) 02- 0060- 04

数据

表 1 2000 年各地区科技状况统计有关数据一览表

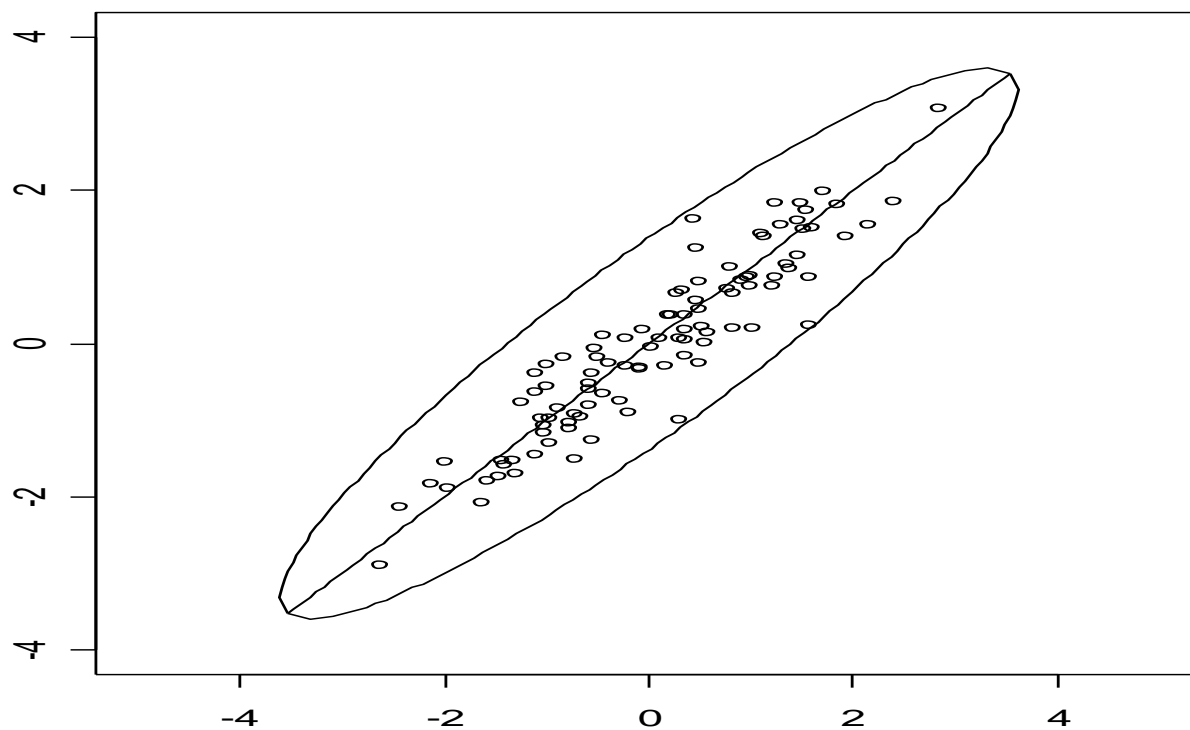
省 份	总效率值	阶段比例			使用结构比例			经费投入产业的比例					社会环境	
		基础研究	应用研究	实验发展	科研机构	高等院校	企 业	加 工 制造业	邮 电 通信业	计算机应 用服务业	科 学 研究业	综合技术 服务业	市场指数	人均 GDP
		X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁	X ₁₂	X ₁₃
北 京	0.7164	0.1567	0.3117	0.5316	0.5667	0.1148	0.3185	0.2707	0.0210	0.0438	0.6268	0.0376	6.3	1.7936
天 津	1.0268	0.0673	0.3452	0.5875	0.2157	0.1127	0.6717	0.7067	0.0015	0.0278	0.2576	0.0064	6.58	1.6377
河 北	0.5953	0.1613	0.3507	0.4879	0.1691	0.1258	0.7052	0.7671	0.0026	0.0012	0.2282	0.0009	6.7	0.7546
山 西	0.5594	0.0890	0.4568	0.4542	0.2500	0.0641	0.6859	0.7063	0.0177	0.0013	0.2731	0.0016	4.57	0.4986
内蒙古	0.8101	0.1512	0.4583	0.3905	0.2282	0.0906	0.6812	0.6842	0.0227	0.0355	0.2528	0.0047	3.45	0.5897
辽 宁	0.6717	0.0978	0.2859	0.6163	0.2671	0.0595	0.6734	0.6657	0.0031	0.0363	0.2877	0.0071	5.6	1.1017
吉 林	0.7986	0.1643	0.4485	0.3871	0.3261	0.0824	0.5915	0.5656	0.0073	0.0346	0.3751	0.0174	4.51	0.6676
黑龙江	0.6601	0.0887	0.5862	0.3251	0.1600	0.1328	0.7072	0.7639	0.0038	0.0156	0.2037	0.0130	3.97	0.8818
上 海	0.9448	0.1433	0.4688	0.3878	0.3131	0.1028	0.5841	0.5559	0.0358	0.0381	0.3622	0.0079	6.59	2.7188



主成分与因子分析

- 人们常常会遇到有**很多变量**的数据。
- 比如全国或各个地区的带有许多经济和社会变量的数据；各个学校的研究、教学等各种变量的数据等等。
- 这些数据的共同特点是变量很多，在如此多的变量之中，有很多是相关的。人们希望能够找出它们的**少数“代表”**来对它们进行描述。
- 两种常用的把变量维数降低以便于描述、理解和分析的方法：**主成分分析**（principal component analysis）和**因子分析**（factor analysis）。
- 实际上**主成分分析**可以说是因子分析的一个特例。

二维数据（椭圆的长短轴）



Size and shape variation in the painted turtle. A principal component analysis. JOLICOEUR, P; MOSIMANN, J E, **Growth**, Vol 24, 339-354, 1960.
Cited times: 284

SIZE AND SHAPE VARIATION						
Principal axes	24 Males			24 Females		
	1st (major)	2nd (inter-mediate)	3rd (minor)	1st (major)	2nd (inter-mediate)	3rd (minor)
Magnitude of variance	195.28	3.69	1.10	680.40	6.50	2.86
% of total	97.61	1.84	0.55	98.64	0.94	0.41
Nature of variation	<i>Size joint variation of all dimensions</i>	<i>Shape contrast of length vs. width mostly</i>	<i>Shape contrast of length vs. height mostly</i>	<i>Size joint variation of all dimensions</i>	<i>Shape contrast of length vs. width mostly</i>	<i>Shape contrast of length and width vs. height</i>



主成分分析理论推导

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}, \quad D(X) = \Sigma,$$

找一个 X 的线性组合 $Y_1 = a_{11}X_1 + a_{21}X_2 + \cdots + a_{p1}X_p$, 使得

Y_1 的方差达到最大。于是提出如下优化问题:
$$\begin{cases} \max Var(Y_1) \\ s.t. a_1^T a_1 = 1 \end{cases}$$



主成分分析理论推导

$$\begin{cases} \max Var(Y_1) \\ s.t. a_1^T a_1 = 1 \end{cases}$$

用 Lagrange 乘子法求解，令

$$L(a_1, \lambda) = Var(a_1^T X) - \lambda(a_1^T a_1 - 1) = a_1^T \Sigma a_1 - \lambda(a_1^T a_1 - 1)$$

$$\begin{cases} \frac{\partial L}{\partial a_1} = 2(\Sigma - \lambda I)a_1 = 0 \\ \frac{\partial L}{\partial \lambda} = a_1^T a_1 - 1 = 0 \end{cases} \Rightarrow |\Sigma - \lambda I| = 0$$



定理：设 $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$ 是 p 维随机变量， $D(X) = \Sigma$ ， Σ 的特征值为

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ ， $A = (a_1 \ a_2 \ \cdots \ a_p)$ 为相应的单位正交

特征向量，则 X 的第 i 个主成分为 $Y_i = a_i^T X$ ($i = 1, 2, \cdots, p$)。

对任意非零向量 a ， $\lambda_p \leq \frac{a^T \Sigma a}{a^T a} \leq \lambda_1$ ，当 $a = a_1$ 时达到最大值 λ_1 。

对 $r = 2, 3, \cdots, p$ ，记 $\Gamma_r = \text{span}(a_r, \cdots, a_p)$ ，可证明

$\max_{a \neq 0, a \in \Gamma_r} \frac{a^T \Sigma a}{a^T a} = \lambda_r$ ，且最大值在 $a = a_r$ 时达到。



主成分分析

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} = A^T X$$

$$\text{Var}(Y_i) = a_i^T \Sigma a_i$$

$$\text{Cov}(Y_i, Y_j) = a_i^T \Sigma a_j = 0$$

$$X = \begin{bmatrix} \frac{x_{11} - \bar{x}_1}{s_1} & \dots & \frac{x_{1p} - \bar{x}_p}{s_p} \\ \frac{x_{21} - \bar{x}_1}{s_1} & \dots & \frac{x_{2p} - \bar{x}_p}{s_p} \\ \vdots & \ddots & \vdots \\ \frac{x_{n1} - \bar{x}_1}{s_1} & \dots & \frac{x_{np} - \bar{x}_p}{s_p} \end{bmatrix}, \quad \text{其中}$$
$$\bar{x}_i = \frac{1}{n} \sum_{j=1}^n x_{ji} \quad (i = 1, \dots, p)$$
$$s_i^2 = \frac{1}{n-1} \sum_{j=1}^n (x_{ji} - \bar{x}_i)^2$$

$$R = \frac{X^T X}{n-1}, \quad R = V D V^T = \begin{bmatrix} a_1 & \dots & a_p \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix} \cdot \begin{bmatrix} a_1^T \\ \vdots \\ a_p^T \end{bmatrix}$$



主成分得分

$$R = \begin{bmatrix} a_1 & \cdots & a_p \end{bmatrix} \cdot \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix} \cdot \begin{bmatrix} a_1^T \\ \vdots \\ a_p^T \end{bmatrix} = \sum_{i=1}^p \lambda_i a_i a_i^T$$

$$y_1 = a_1^T x = a_{11}x_1 + \cdots + a_{p1}x_p$$

$$y_2 = a_2^T x = a_{12}x_1 + \cdots + a_{p2}x_p$$

...

$$y_p = a_p^T x = a_{1p}x_1 + \cdots + a_{pp}x_p$$



主成分分析数据处理步骤

- 先将数据归一化。即每个变量的样本值变换为：样本均值为0，样本方差为1
- 计算协方差矩阵，即相关系数矩阵
- 计算相关系数矩阵的特征值和特征向量
- 计算贡献率
- 得到主成分
- 计算主成分得分

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

主成分分析在中国上市公司综合评价中的作用

田波平^{1,2}, 王 勇¹, 郭文明¹, 葛喜鹏¹

(1. 哈尔滨工业大学数学系, 黑龙江 哈尔滨 150001)

(2. 哈尔滨工业大学管理学院, 黑龙江 哈尔滨 150001)

摘要: 主要介绍了主成分分析在上市公司综合评价中的作用. 主成分分析作为一种客观赋权的方法, 权数是能随着宏观经济环境变化而变化的动态权数体系, 因为主成分分析所应用的数据来源于上市公司年度报告所提供的财务指标. 它主要对所选择的 40 只绩优股, 进行横向的比较, 并根据第一主成分得分进行排序, 给广大的投资者提供参考.

《商业周刊》是美国一家以财经报道为主的、在这方面具有广泛影响力的综合性的国际出版物。该周刊十多年来一直致力于对上市公司的评价和排名,它既有单指标排名,又有多指标综合评价排名。单指标排名根据美元计价的上市公司市价总值大小排出全球市价总值最大的 1000 家上市公司。而综合评价排名是对纳入标准普尔 500 指数中的 500 家上市公司,根据 8 个指标的综合结果加以排名,评出前 50 名为“《商业周刊》最佳 50 家上市公司”。其综合评价方法选取的 8 个指标,分别是 ① 当年的总收益; ② 三年来的总收益; ③ 一年来的销售总额增长率(%); ④ 三年来的销售总额增长率(%); ⑤ 一年来的利润增长率(%); ⑥ 三年来的利润增长率(%); ⑦ 一年来的净利润; ⑧ 一年来的净资产收益率(%). 综合的方法是对这 8 项指标分别进行排名,取得分值最大的 20% 为 A 级,其次的 20% 为 B 级,依次类推,最后的 20% 为 E 级。各上市公司的总得分是其 8 个指标的评分之和,然后按总得分排名。



3 主成分分析方法在证券个股评价中的实证研究^[3, 4, 7]

3.1 样本股票的选择

对于样本股票的选择,我们主要是通过金融街([www. jrj com. cn](http://www.jrj.com.cn))和中信证券网上股票交易系统网站上市公司统计年报检索系统,从沪深两市中挑选了具有各自代表性的 40 只股票。这里所谓的代表性是指:① 行业代表性;② 板块代表性;③ 业绩代表性。这 40 只股票分别代表了“信息产业”、“旅游行业”、“医药行业”、“金融、保险行业”、“采掘业”、“交通运输行业”和“制造行业”。其中 ① x_1 表示主营业务收入;② x_2 表示主营业务利润;③ x_3 表示利润总额;④ x_4 表示净利润;⑤ x_5 表示总资产;⑥ x_6 表示固定资产;⑦ x_7 表示净资产收益率;⑧ x_8 表示每股净资产;⑨ x_9 表示每股资本公积金;⑩ x_{10} 表示每股收益。这 10 个指标中既有反映企业投入产出规模的总资产、主营业务收入、净利润指标,也有反映企业投资者投入效率的每股收益、净资产收益率指标,所选样本股票的原始数据见金融街([www. jrj com. cn](http://www.jrj.com.cn))和中信证券网上股票交易系统网站。

一般若前 k 个主成分的累计贡献率达到 85% 以上, 就表明取前 k 个主成分基本包含了全部测量指标所具有的信息, 因为前三个主成分的累计贡献率 $p = 93\%$, 已经大于 85%, 则取前 3 个主成分变量代替原来 10 个标量, 并根据特征向量来求主成分得分, 提出使用第一主成分来评价个股业绩, 利用第一主成分来对样本股票进行排序, 得到所选样本股票 2002 年的经营状况, 给广大投资者一个理性的参考作用

前 3 个主成分对应的特征向量列于表 2

表 2 特征向量

主成分	1	2	3
X_1	0.401048	0.094703	0.089112
X_2	0.403139	0.090295	0.080681
X_3	0.414576	0.085301	0.039037
X_4	0.406954	0.095834	0.067097
X_5	0.345556	0.021592	0.005027
X_6	0.408464	0.045950	0.060557
X_7	0.024043	0.393296	-0.646526
X_8	-0.160771	0.468642	0.473934
X_9	-0.151813	0.490794	0.447068
X_{10}	-0.058561	0.590469	-0.365173

表3 按第一主成分之值降序排列

第一主成分名次	个股名称	第一主成分	第二主成分	第三主成分	第一主成分名次	个股名称	第一主成分	第二主成分	第三主成分
1	中国石化	13.4017	1.2511	1.1297	21	四川长虹	-0.5419	-0.2892	0.8742
2	中国联通	3.1959	-0.9146	0.9783	22	新农开发	-0.5923	-0.9564	-0.9792
3	招商银行	2.6160	-0.0311	-0.2109	23	首旅股份	-0.5938	-0.3389	-1.1511
4	宝钢股份	2.2731	0.0271	0.3667	24	亚星客车	-0.6366	-0.7988	-0.2368
5	深发展A	0.6350	-0.8643	0.2277	25	南京新百	-0.6685	-0.2314	-0.1528
6	东方航空	0.2594	-1.9742	-0.3807	26	澳柯玛	-0.6685	-0.6592	-0.3892
7	TCL 通讯	-0.2631	3.7395	0.0729	27	上菱电器	-0.6700	0.8461	0.0472
8	白云山A	-0.2639	-1.3660	1.5709	28	万科A	-0.6952	1.0604	0.5275
9	哈空调	-0.3192	-2.0780	0.1971	29	东软股份	-0.7185	-0.0452	-0.3955
10	长江投资	-0.3237	-1.6834	-0.9318	30	古井贡A	-0.7917	-0.1552	0.0537
11	五粮液	-0.3390	-0.2908	1.0363	31	中兴通讯	-0.8109	1.0072	-5.6019
12	工大首创	-0.3497	-1.5703	-0.7419	32	燕京啤酒	-0.8390	0.5939	-0.2033
13	一汽轿车	-0.3628	-0.9179	2.6450	33	金地集团	-0.8670	0.9162	0.2771
14	上海机场	-0.3871	0.0854	0.2315	34	春兰股份	-0.8869	0.3322	-0.2729
15	国旅联合	-0.4326	-1.5952	1.2882	35	新华医疗	-0.8870	0.4139	-0.6256
16	鲁能泰山	-0.4416	-1.0236	2.0592	36	海王生物	-0.9252	0.1794	-0.4305
17	东阿阿胶	-0.4911	-1.2382	0.6094	37	新大陆	-1.0070	0.7514	-0.7509
18	万杰高科	-0.4985	-0.8176	-0.2349	38	鄂尔多斯	-1.0222	0.5215	-0.0613
19	丽珠集团	-0.5140	-0.7549	0.8344	39	波导股份	-1.0543	3.5375	0.2003
20	哈药股份	-0.5334	0.5633	-0.3279	40	用友软件	-1.9846	4.7682	-1.1483

从表 3 可以认为第一主成分代表总的业绩水平, 亦即综合财务业绩。从第一主成分的计算公式可以看到第一主成分较大的系数是在 x_1 、 x_2 、 x_3 、 x_4 、 x_5 及 x_6 上, 也就是说这几个指标最能代表综合财务业绩。由此, 我们按第一主成分之值排序, 得到了表 3, 它是按第一主成分的数值由大到小, 即综合财务业绩由强到弱而列出的。此结果显示: 中国石化 ($F_1 = 13.4017$) 明显居于首位, 中国联通 ($F_1 = 3.1959$)、招商银行 ($F_1 = 2.6160$)、宝钢股份 ($F_1 = 2.2731$) 次之。可以看出, 综合财务业绩的排序与 x_1 、 x_2 、 x_3 、 x_4 、 x_5 及 x_6 的总的顺序大体保持一致, 但与其中每一个并不总保持完全一致。这说明了综合财务业绩并不能用任何一个指标完全代替, 即使是采用最有代表性的主营业务收入 x_1 也不行, 而是通过考虑了多个指标的综合情况来得到的。从上面第二主成分的计算公式可以看出, 它在变量 x_7 、 x_8 、 x_9 和 x_{10} 上有较大的正值系数, 其含义是有较多的净资产收益率、每股净资产、每股资本公积和每股收益将获得较大的数值。用友软件 ($F_2 = 4.7682$) 名列第一, TCL 通讯 ($F_2 = 3.7395$)、波导股份 ($F_2 = 3.5375$)、中国石化 ($F_2 = 1.2511$) 次之。而哈空调 ($F_2 = -2.0780$) 排在最后。第二主成分是衡量各股稳健发展的象征, 从第二主成分的得分可以

看出,名列前茅的多是科技含量较高、资本雄厚、经营比较稳健的股票。它们一般属于信息技术产业和钢铁石油产业。对于此类股票中长线投资价值较大。总之,面对上市公司财务报表中的众多指标,使用多元统计分析中的主成分分析法,可以用很少的综合指标代替原来众多的原始指标。以第一主成分的得分作为个股综合业绩的度量,能够真实的反映原始变量的主要信息,结果可靠。再加上第二主成分,第三主成分,广大的投资者就可以比较清楚的了解上市公司的财务状况了。

4 结 论

本文首先通过对沪深两市所选 40 样本股票的主成分得分的实证研究。应用一种客观赋权的动态评价方法,此方法包含了多指标的综合性和权数的动态可变性。利用上市公司年度报告所提供的财务指标,应用多元统计中的主成分分析方法对多指标表现进行综合。此外,本方法所运用的权数体系来源于特定年份的实际数据,所以它是能与宏观环境相匹配的动态权数体系,而在已介绍的国内外评价方法的主要特点就是单指标或固定权数综合。然后对所选样本股票来自 2002 年年报的财务指标,利用 SA S 统计软件进行处理,算出各只股票的第一主成分得分,并利用得分的大小进行降序排名,从而给广大的投资者一个显性的参考。