МГУ "Статистика-2" Курсовая работа Рудик 2000г.

1. Формулировка исследовательской задачи и подготовка данных

1.1. Выбор переменных

Задача состоит в объяснении способа построения результирующих признаков по имеющимся данным. В качестве результирующего признака V_S выберем переменную, обозначенную как «индекс человеческого развития» (V 8.35 Human Development Index).

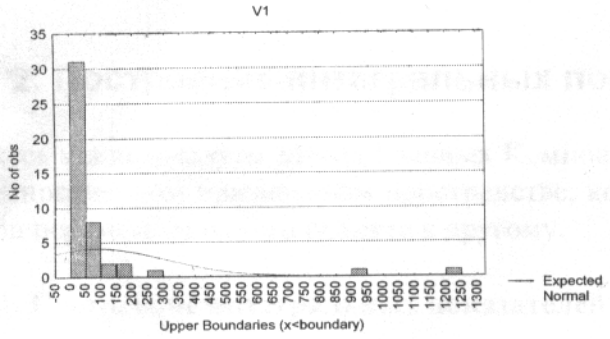
В качестве же детерминантов выберем следующие переменные:

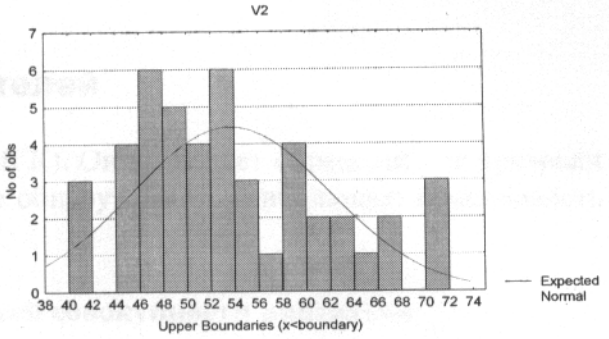
V_1	V 8.01	Население
V_2	V 8.02	Относительный показатель числа иждивенцев
V_3	V 8.05	Средняя продолжительность жизни человека
V_4	V 8.16	Занятость (% населения)
V_5	V 8.17	Среднее количество часов работы в год
V_6	V 8.18	Безработица (% рабочей силы)
V_7	V 8.21	Кол-во школьников (%)
V_8	V 8.22	Кол-во студентов (%)
V_9	V 8.26	Неграмотность (% населения от 15 лет)
V_{10}	V 8.34	Доля собственников (%)

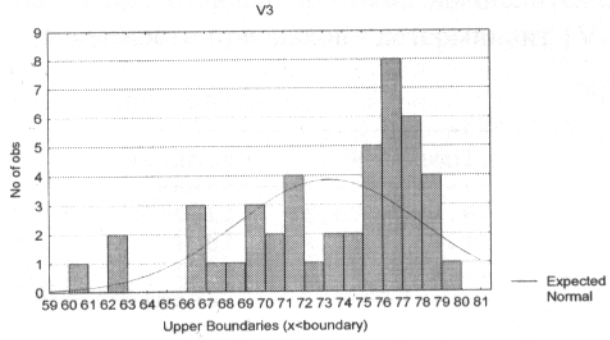
1.2. Параметры стандартизации признаков - детерминант:

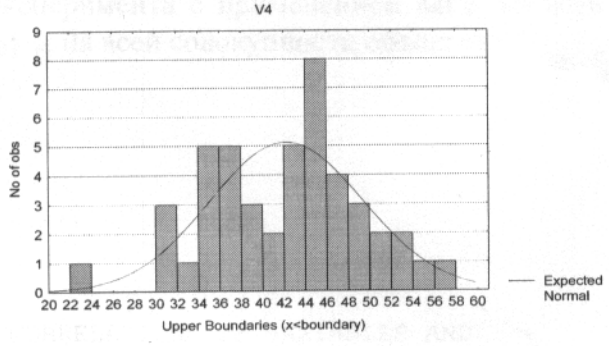
ПРИЗНАК	СРЕДНЕЕ	СТАНДАРТНОЕ ОТКЛОНЕНИЕ	МИНИМУМ	МАКСИМУМ
V_1	89.01	223.83	0.27	1231.00
V_2	53.67	8.23	40.40	71.80
V_3	73.26	4.77	60.40	79.50
V_4	42.32	7.31	23.06	56.29
V_5	1.92	0.22	1.33	2.40
V_6	9.16	8.45	1.79	50.00
V_7	80.11	22.54	33.00	121.00
V_8	31.85	17.72	2.00	99.00
V_9	6.58	9.05	0.00	48.00
V_{10}	61.75	13.89	16.60	80.50

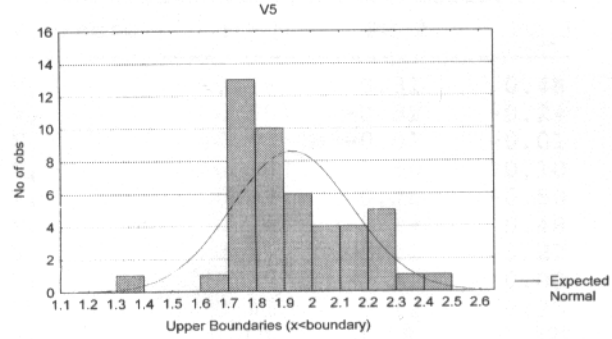
Дальше представляются диаграммы плотностей признаков:

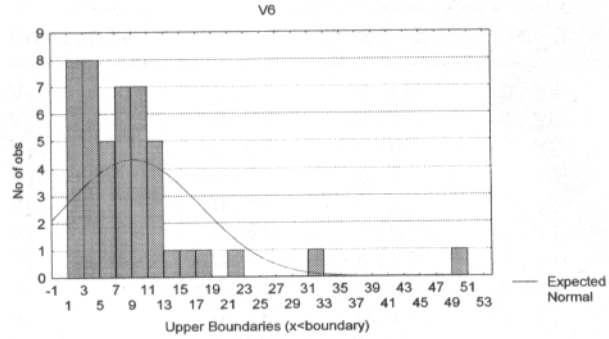

V1
No of obs
Upper Boundaries (x<boundary)
Expected
Normal

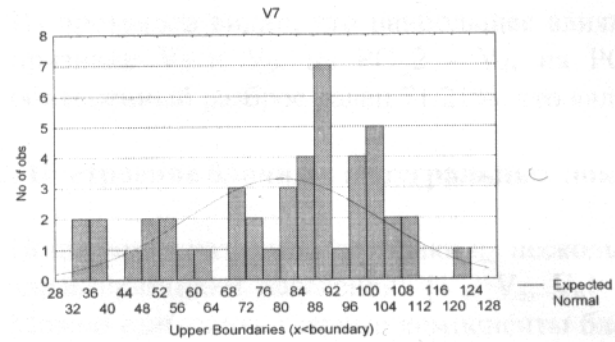

V2
No of obs
Upper Boundaries (x<boundary)
Expected
Normal

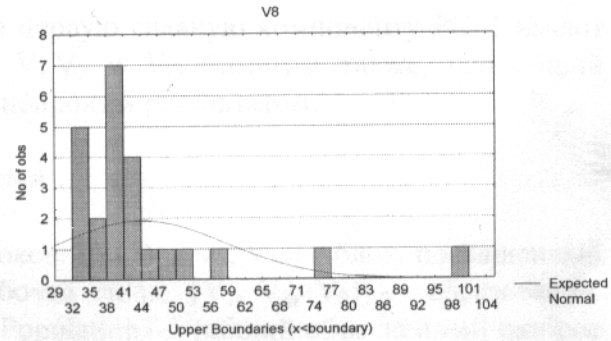

V3
No of obs
Upper Boundaries (x<boundary)
Expected
Normal

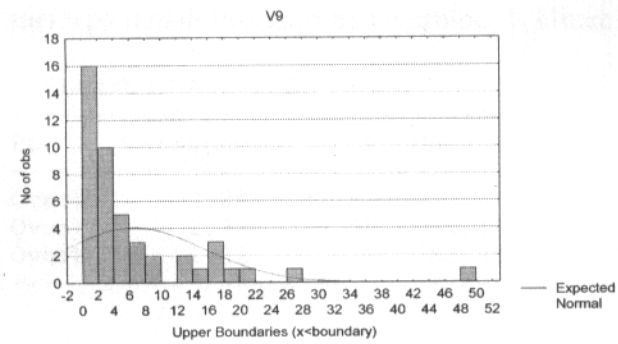

V4
No of obs
Upper Boundaries (x<boundary)
Expected
Normal

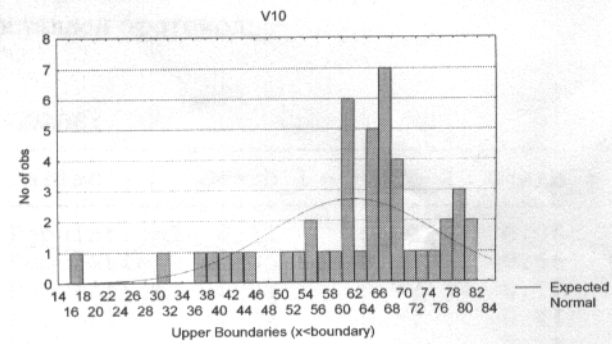

V5
No of obs
Upper Boundaries (x<boundary)
Expected
Normal


V6
No of obs
Upper Boundaries (x<boundary)
Expected
Normal


V7
No of obs
Upper Boundaries (x<boundary)
Expected
Normal


V8
No of obs
Upper Boundaries (x<boundary)
Expected
Normal


V9
No of obs
Upper Boundaries (x<boundary)
Expected
Normal


V10
No of obs
Upper Boundaries (x<boundary)
Expected
Normal

2. Построение интегральных показателей

Здесь мы используем Метод Главных Компонент (МГК). Он позволяет определить те признаки в многомерном признаковом пространстве, которые обнаруживают наибольшую изменчивость при переходе от одного объекта к другому.

2.1. **Построение интегральных показателей для всей совокупности признаков**

Далее представлен протокол вычислительного эксперимента с применением МГК ко всей совокупности признаков - детерминант $\{V_1, \ldots V_{10}\}$ и на всей совокупности объектов:

EXPLAINED VARIANCE:

PC	Variance	%	Cumul.
PC_1	4.17	41.67	41.67
PC_2	1.67	16.74	58.41
PC_3	1.28	12.79	71.21

LOADS:

Variable	PC_1	PC_2	PC_3
V_1	0.28	0.31	0.48
V_2	0.35	-0.32	-0.24
V_3	-0.43	-0.01	-0.01
V_4	-0.19	0.60	0.10
V_5	0.22	0.25	-0.50
V_6	0.24	-0.44	0.48
V_7	-0.39	-0.22	0.22
V_8	-0.35	-0.22	0.08
V_9	0.43	0.12	0.25
V_{10}	0.10	-0.28	-0.32

CORRELATIONS OF VARIABLES AND PRINCIPAL COMPONENTS:

Variable	PC_1	PC_2	PC_3
V_1	0.57	0.40	0.54
V_2	0.72	-0.41	-0.26
V_3	-0.87	-0.02	-0.02
V_4	-0.39	0.78	0.10
V_5	0.46	0.32	-0.58
V_6	0.49	-0.57	0.54
V_7	-0.80	-0.29	0.24
V_8	-0.71	-0.29	0.08
V_9	0.88	0.16	0.27
V_{10}	0.21	-0.36	-0.39

Из протокола видно, что наибольшее влияние на первую главную компоненту PC_1 имеют признаки V_3 и V_9, на PC_2 - V_4, на PC_3 - V_1,V_5 и V_6. Заметим также, что общий объясненный разброс равен 71.21%, что является неплохим результатом.

2.2. **Построение блочных интегральных показателей**

Попытаемся разделить признаки на несколько блоков: $\{V_1, V_2, V_3, V_{10}\}$ - блок, посвященный характеристикам населения; $\{V_4, V_5, V_6\}$ --- рабочей силы; $\{V_7, V_8, V_9\}$ --- образования. Можно вычислить главные компоненты блоков: Population1-3 (общий объясненный разброс составил 92.6%), Labour1-2 (84.65%), Education1-2 (88.57%). Затем построим надблочный интегральный показатель Overbloc_1. Ниже предоставлен протокол:

EXPLAINED VARIANCE:

PC	Variance	%	Cumul.
Overb_1	2.28	32.51	32.51
Overb_2	1.33	18.98	51.48
Overb_3	1.22	17.37	68.85

LOADS:

Variable	Overb_1	Overb_2	Overb_3
Population1	0.61	0.01	0.06
Population2	0.05	0.39	-0.56
Population3	-0.02	0.72	0.17
Labour1	-0.32	0.32	-0.44
Labour2	-0.41	0.04	0.11
Education1	-0.57	-0.07	0.34
Education2	-0.16	-0.47	-0.58

3. Применение метода линейной регрессии для исследования зависимости между результирующим признаком и признаками-детерминантами.

Используем метод линейной регресии на всем признаковом пространстве. Вот протокол:

```
REGRESSION QUALITY:

R*2  (Determination coeff.)  = 0.93
R    (Correlation)           = 0.97
SEE* (St. Error of Estimate) = 0.027
MAE  (Mean Absolute Error)   = 0.016
MAPE (Mean Abs. Perc. Error) = 1.99 %


LINEAR REGRESSION FORMULA:

V_S = 0.306 + -1.35E-4*V_1 + -0.001*V_2 + 0.011*V_3 + -0.001*V_4 + -0.04*V_5 + -0.001*V_6
+ -3.80E-4*V_7 + 3.918E-4*V_8 + -0.002*V_9 + 0.001*V_10
    (2.11)    (-4.81)      (-1.78)   (7.06)     (-1.69)     -1.67)    (-1.46)
(-1.17)         (1.19)       (-2.42)    (1.74)


COEFFICIENTS OF LINEAR REGRESSION:
-----------------------------------------------------------
Variable    Coeff.    StErr    t-Stat  Var.Mean Var.StDev
-----------------------------------------------------------
V_S                                     0.86      0.10
INTERCEPT    0.306    0.145     2.11 *
V_1          0.000    0.000    -4.81 *   89.01    223.83
V_2         -0.001    0.001    -1.78     53.67      8.23
V_3          0.011    0.002     7.06 *   73.26      4.77
V_4         -0.001    0.001    -1.69     42.21      7.18
V_5         -0.040    0.024    -1.67      1.93      0.21
V_6         -0.001    0.001    -1.46      9.16      8.45
V_7          0.000    0.000    -1.17     80.43     22.34
V_8          0.000    0.000     1.19     31.80     17.70
V_9         -0.002    0.001    -2.42 *    6.58      9.05
V_10         0.001    0.000     1.74 ?   61.87     13.61
-----------------------------------------------------------
Degrees of freedom for t-Statistic = 35
```

Это лучший результат (если брать в качестве критерия функционал качества R^2) который можно получить, использую всего лишь 10 признаков. Если посмотреть на t-статистику, то можно заметить, что результирующий признак V_S наименее чувствителен к V_7 и V_8, наиболее чувствителен к V_1 и V_3. Кстати, в следующем эксперименте мы оставили только эти 2 признака и результат практически не изменился:

```
REGRESSION QUALITY:

R*2  (Determination coeff.)  = 0.90
R    (Correlation)           = 0.95
SEE* (St. Error of Estimate) = 0.032
MAE  (Mean Absolute Error)   = 0.024
MAPE (Mean Abs. Perc. Error) = 2.99 %


LINEAR REGRESSION FORMULA:

V_S =   -0.162 + -2.04E-4*V_1 + 0.014*V_3
      (-1.95)    (-8.53)     (12.68)
```

```
COEFFICIENTS OF LINEAR REGRESSION:
-----------------------------------------------------
Variable     Coeff.     StErr    t-Stat  Var.Mean Var.StDev
-----------------------------------------------------
V_S                                        0.86      0.10
INTERCEPT    -0.162     0.083    -1.95
V_1           0.000     0.000    -8.53     89.01    223.83
V_3           0.014     0.001    12.68     73.26      4.77
-----------------------------------------------------
Degrees of freedom for t-Statistic = 43
```

Мы также можем исследовать зависимость между результирующим признаком V_S и вычисленными ранее интегральными показателями. Возьмем для начала первые три главные компоненты всего пространства признаков:

```
REGRESSION QUALITY:

R*2  (Determination coeff.)  = 0.89
R    (Correlation)           = 0.95
SEE* (St. Error of Estimate) = 0.034
MAE  (Mean Absolute Error)   = 0.022
MAPE (Mean Abs. Perc. Error) = 2.68 %


LINEAR REGRESSION FORMULA:

V_S =  0.864 + -0.045*PC_1 + -0.015*PC_2 + -0.023*PC_3
     (174.19) (-18.22)       (-3.80)        (-5.26)


COEFFICIENTS OF LINEAR REGRESSION:
-----------------------------------------------------
Variable     Coeff.     StErr    t-Stat  Var.Mean Var.StDev
-----------------------------------------------------
V_S                                        0.86      0.10
INTERCEPT     0.864     0.005   174.19
PC_1         -0.045     0.002   -18.22      0.00      2.04
PC_2         -0.015     0.004    -3.80      0.00      1.29
PC_3         -0.023     0.004    -5.26      0.00      1.13
-----------------------------------------------------
Degrees of freedom for t-Statistic = 42
```

Видно, что основная нагрузка лежит на первой главной компоненте PC_1. Если мы уберем остальные компоненты, то результат изменится несильно ($R^2 = 0.79$).

Если рассматривать множество блочных интегральных показателей, то неплохой результат получится на подмножестве {Population1-2,Education1}:

```
REGRESSION QUALITY:

R*2  (Determination coeff.)  = 0.90
R    (Correlation)           = 0.95
SEE* (St. Error of Estimate) = 0.033
MAE  (Mean Absolute Error)   = 0.024
MAPE (Mean Abs. Perc. Error) = 3.04 %


LINEAR REGRESSION FORMULA:

V_S = 0.864 + -0.053*Population1 + -0.034*Population2 + 0.013*Education1
   (177.95) (-8.97)               (-6.56)              (2.35)
```

```
COEFFICIENTS OF LINEAR REGRESSION:
-----------------------------------------------------------
Variable     Coeff.     StErr    t-Stat   Var.Mean Var.StDev
-----------------------------------------------------------
Vs                                          0.86      0.10
INTERCEPT     0.864     0.005    177.95
Population1  -0.053     0.006     -8.97      0.00      1.37
Population2  -0.034     0.005     -6.56      0.00      1.05
Education1    0.013     0.006      2.35      0.00      1.49
-----------------------------------------------------------
Degrees of freedom for t-Statistic = 42
```

И, наконец, протокол исследования зависимости между V_S и надблочными интегральными показателями с помощью метода линейной регресии:

```
REGRESSION QUALITY:

R*2  (Determination coeff.)  = 0.82
R    (Correlation)           = 0.91
SEE* (St. Error of Estimate) = 0.044
MAE  (Mean Absolute Error)   = 0.031
MAPE (Mean Abs. Perc. Error) = 3.96 %


LINEAR REGRESSION FORMULA:

Vs = 0.864 + -0.058*Overb_1 + -0.023*Overb_2 + 0.019*Overb_3
   (134.54) (-13.41)          (-4.08)           (3.27)


COEFFICIENTS OF LINEAR REGRESSION:
-----------------------------------------------------------
Variable     Coeff.     StErr    t-Stat   Var.Mean Var.StDev
-----------------------------------------------------------
Vs                                          0.86      0.10
INTERCEPT     0.864     0.006    134.54
Overb_1      -0.058     0.004    -13.41      0.00      1.51
Overb_2      -0.023     0.006     -4.08      0.00      1.15
Overb_3       0.019     0.006      3.27      0.00      1.10
-----------------------------------------------------------
Degrees of freedom for t-Statistic = 42
```

Результат только для первой главной компоненты Overb_1: $R^2 = 0.72$. Нетрудно заметить, что надблочный интегральный показатель коррелируется с результирующим признаком существенно хуже, чем первая главная компонента всего пространства признаков. Видимо, это связано с наличием «посредников» - блочных интегральных показателей.

4. Классификация объектов в пространстве признаков - детерминантов

У нас будут 4 класса, так как уменьшение количества классов приведет к уменьшению доли объясненного разброса почти в 2 раза. Ниже приведен протокол классификации объектов:

```
Number of classes = 4
Explained scatter = 50.81 %  (5.08 / 10.00)
```

CLASSES:

Class	Example	Objects	%	Explned	Scatter	%
1	INDIA	1	2.17	75.29	75.29	100.00
2	CHINA	1	2.17	42.37	42.37	100.00
3	COLOMBIA	12	26.09	72.03	148.36	48.55
Swamp	AUSTRALIA	32	69.57	44.07	193.98	22.72
Total		46	100.00	233.75	460.00	50.81

CENTERS OF CLASSES (STANDARDIZED):

Variable	1	2	3	Swamp
V_1	3.83	5.16	-0.10	-0.24
V_2	1.49	-0.38	1.20	-0.49
V_3	-2.73	-1.01	-1.01	0.50
V_4	-0.09	1.21	-0.67	0.22
V_5	0.30	1.58	0.81	-0.36
V_6	4.88	-0.75	0.04	-0.14
V_7	-1.65	-1.33	-1.24	0.56
V_8	-1.47	-1.70	-0.72	0.37
V_9	4.63	2.25	0.56	-0.42
V_{10}	0.83	-0.61	0.23	-0.09

CONTRIBUTIONS:

Variable	1	2	3	Swamp	Total
V_1	14.71	26.61	0.11	1.92	43.35
V_2	2.22	0.14	17.33	7.54	27.23
V_3	7.43	1.02	12.27	7.87	28.59
V_4	0.01	1.45	5.38	1.50	8.33
V_5	0.09	2.48	7.86	4.19	14.62
V_6	23.86	0.56	0.02	0.67	25.11
V_7	2.72	1.77	18.39	9.94	32.82
V_8	2.17	2.90	6.26	4.38	15.71
V_9	21.39	5.05	3.76	5.77	35.98
V_{10}	0.68	0.38	0.66	0.28	2.00
Total	75.29	42.37	72.03	44.07	233.75

CENTERS OF CLASSES (REAL):

Variable	1	2	3	Swamp
V_1	938.00	1231.00	67.71	34.78
V_2	65.80	50.60	63.45	49.72
V_3	60.40	68.50	68.49	75.60
V_4	41.57	50.77	37.46	43.75
V_5	1.99	2.26	2.10	1.85
V_6	50.00	2.90	9.49	7.95
V_7	44.00	51.00	53.08	92.75
V_8	6.00	2.00	19.17	38.28
V_9	48.00	26.70	11.59	2.77
V_{10}	73.00	53.60	65.02	60.60

CLASSIFICATION (LATENT NOMINAL VARIABLE):

Grade	Objects	Distance	Contrib.
1	INDIA	0.00	75.29
2	CHINA	0.00	42.37
3	COLOMBIA	0.71	5.35
	MEXICO	1.26	6.29
	VENEZUELA	1.61	5.22
	TURKEY	1.63	8.14
	MALAYSIA	1.91	7.86
	PHILIPPINES	2.08	6.45
	ARGENTINA	2.26	4.24
	BRAZIL	2.27	5.80
	CHILE	2.33	3.81
	THAILAND	3.30	3.76
	SOUTH AFRICA	3.95	9.25
	INDONESIA	4.35	5.85
Swamp	AUSTRALIA	0.85	1.42
	UN,KINGDOM	0.97	0.77
	AUSTRIA	1.09	2.30
	FRANCE	1.32	1.71
	NEW ZEALAND	1.36	1.28
	LUXEMB	1.37	2.13
	BELGIUM	1.46	1.58
	NETHERLANDS	1.50	2.22
	DENMARK	1.63	2.48
	JAPAN	1.75	2.34
	ITALY	1.79	1.21
	CZECH REP	1.80	0.28
	NORWAY	1.88	2.10
	SWEDEN	1.94	1.52
	GERMANY	1.99	2.17
	PORTUGAL	2.04	-0.36
	GREECE	2.11	1.51
	ISRAEL	2.15	-0.08
	ICELAND	2.19	2.04
	FINLAND	2.20	2.57
	POLAND	2.20	-0.61
	IRELAND	2.25	0.82
	TAIWAN	2.28	0.59
	HUNGURY	2.33	-0.16
	USA	2.56	1.78
	KOREA	2.72	0.86
	SWITZERLAND	2.78	2.32
	SPAIN	2.82	1.36
	HONG KONG	2.83	1.16
	RUSSIA	3.16	0.79
	SINGAPORE	3.20	0.75
	CANADA	3.55	3.19