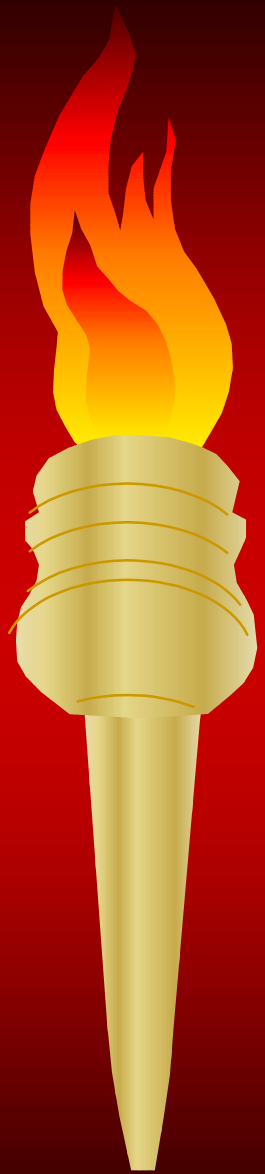
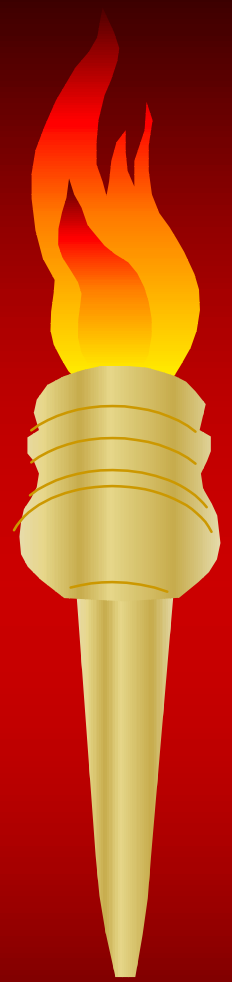
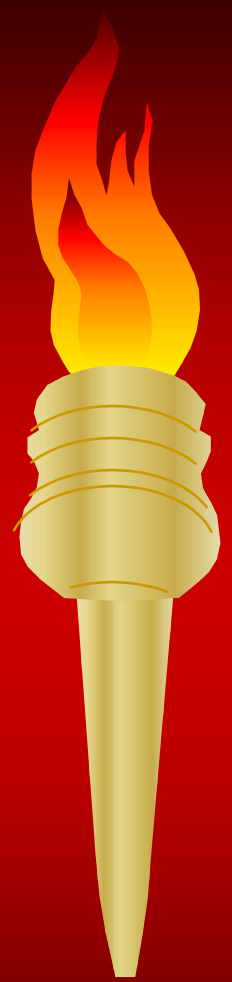


Многомерный статистический анализ



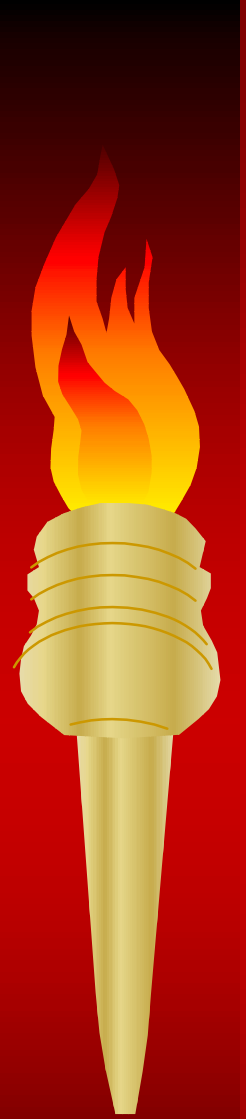
Многмерный статистический анализ

- Зачем он нужен?
  - При тестировании моделей
    - Необходимо исключить влияние некоторых показателей
    - Установить наличие совместного действия факторов
    - Установить наличие скрытых связей
  - При исследовательском анализ данных
    - Снизить размерность используемых данных
    - Сегментировать данные



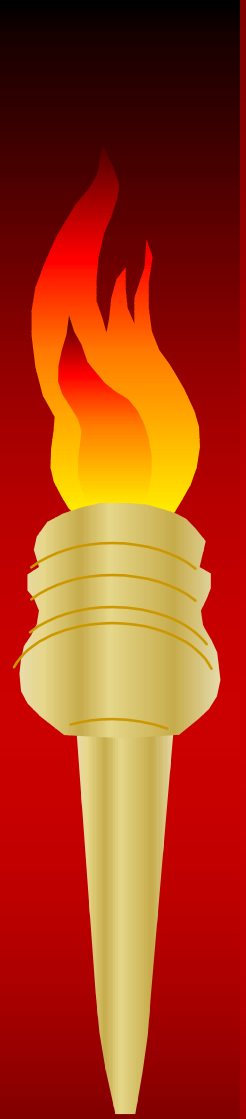
Разработка данных (Data Mining)

- Относительно новый подход, появление которого связано с накоплением больших объемов данных в компьютеризированной форме (data warehouses)
- Нахождение внутренних закономерностей в данных, а не тестирование статистических гипотез



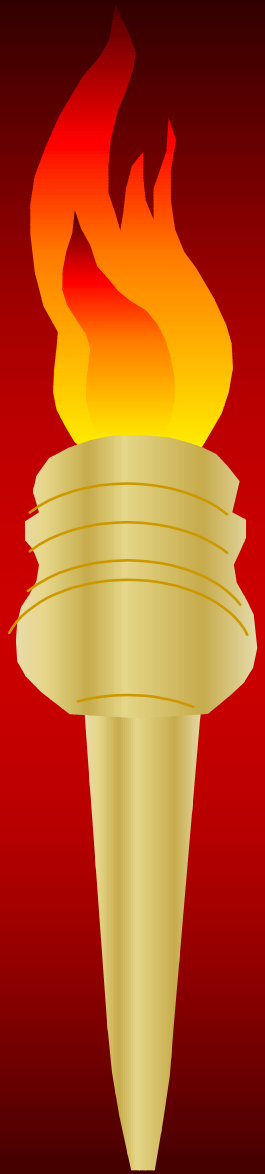
Разработка данных (Data Mining)

- Наиболее популярные методы
  - OLAP (On-Line Analytical Data Processing) - построение многомерных таблиц
  - Нейронные сети (neuronal networks)
  - Сегментирование данных (древовидное моделирование - classification trees)
- Более распространенные методы
  - Факторный анализ
  - Корреспондентский анализ

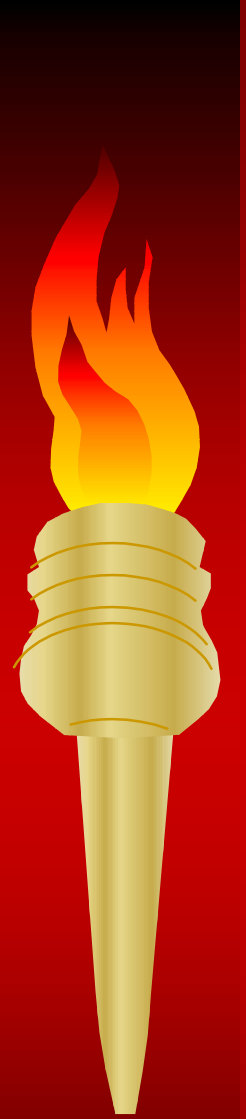


Разработка данных (Data Mining)

- Требуют очень больших массивов данных
  - не менее десяти наблюдений на переменную в факторном анализе (но более ста наблюдений)
  - не менее десяти наблюдений на связь при нейросетевом моделировании
- В общем сотни, лучше тысячи наблюдений

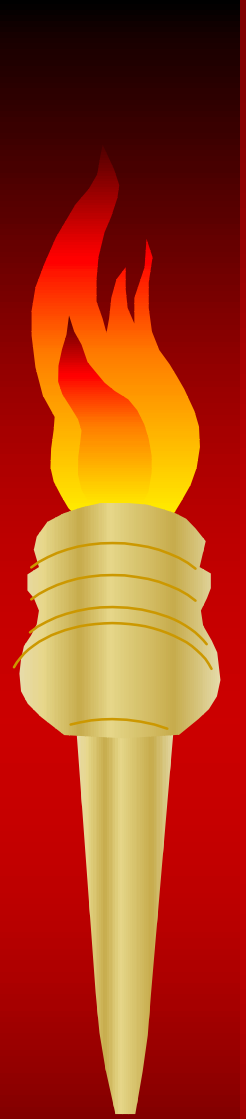


Тестирование моделей



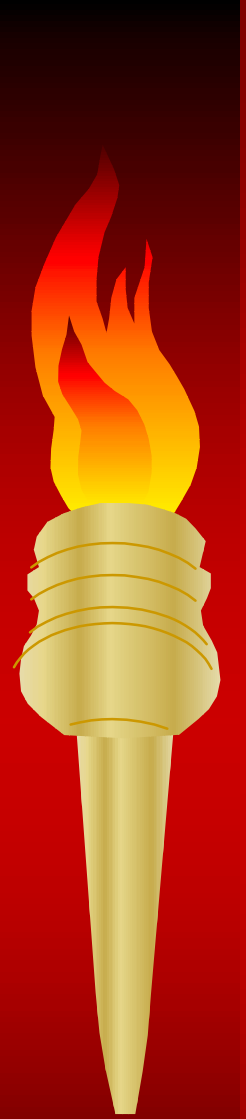
Необходимо определить тип зависимой переменной

- Качественная
- Количественная



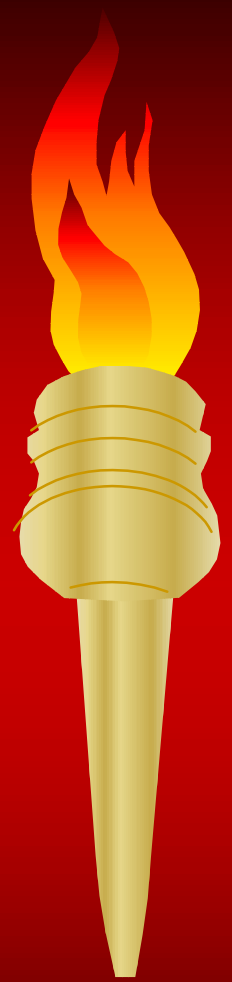
Количественная зависимая переменная

- Общая линейная модель (GLM)
  - Какие независимые переменные участвуют в анализе?
    - Только качественные - MANOVA
    - Количественные и качественные - MANCOVA
    - Только количественные - MLR



MANOVA

- Многофакторный дисперсионный анализ (не путать с факторным анализом!)
- Изучение влияния нескольких качественных факторов и их сочетаний на значения количественных переменных
- На самом деле, сравнение средних значений в таблицах

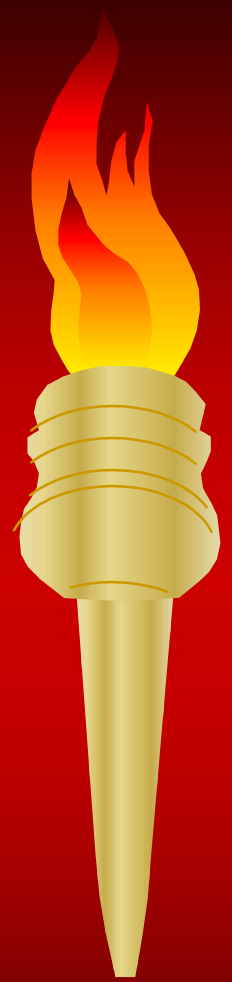


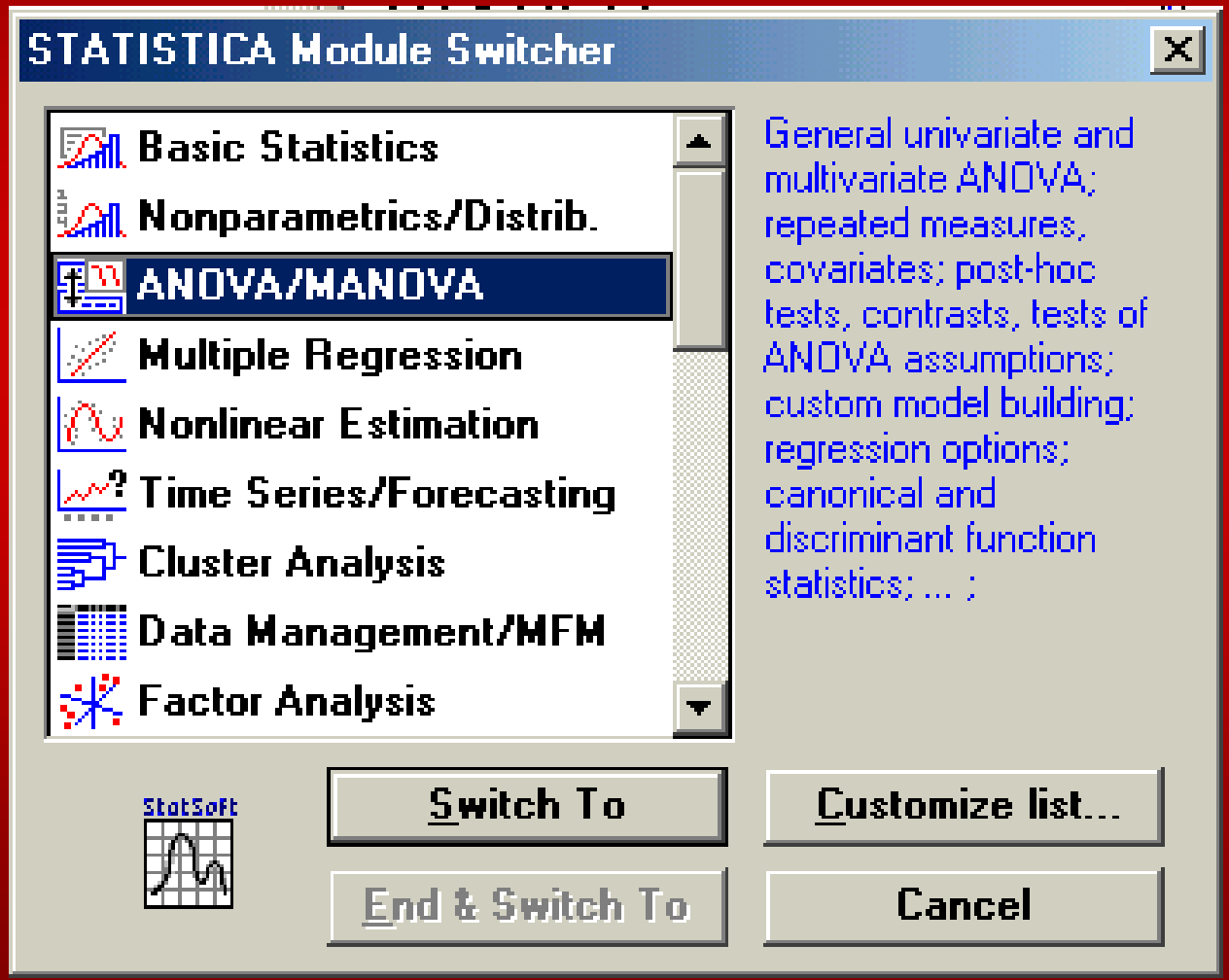
MANOVA

A1						A2					
B1		B2		B3		B1		B2		B3	
C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
x1	x5	x9	x13	x17	x21	x25	x29	x33	x37	x41	x45
x2	x6	x10	x14	x18	x22	x26	x30	x34	x38	x42	x46
x3	x7	x11	x15	x19	x23	x27	x31	x35	x39	x43	x47
x4	x8	x12	x16	x20	x24	x28	x32	x36	x40	x44	x48

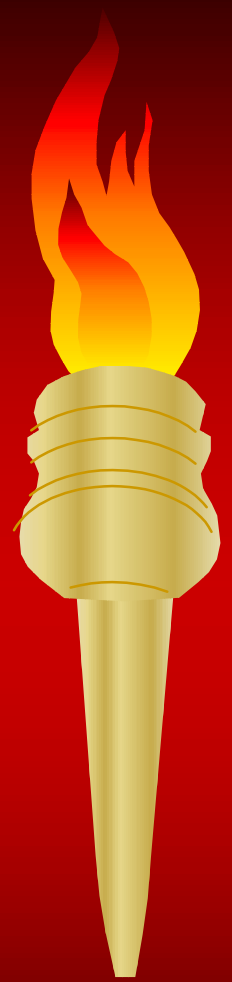
- Ответит на вопросы:
  - Одинаковы ли средние всех 12 групп
  - Ести ли различия между средними групп, образованными фактором A, B или C по отдельности
  - Имеется ли суммарное влияние факторов A, B и C

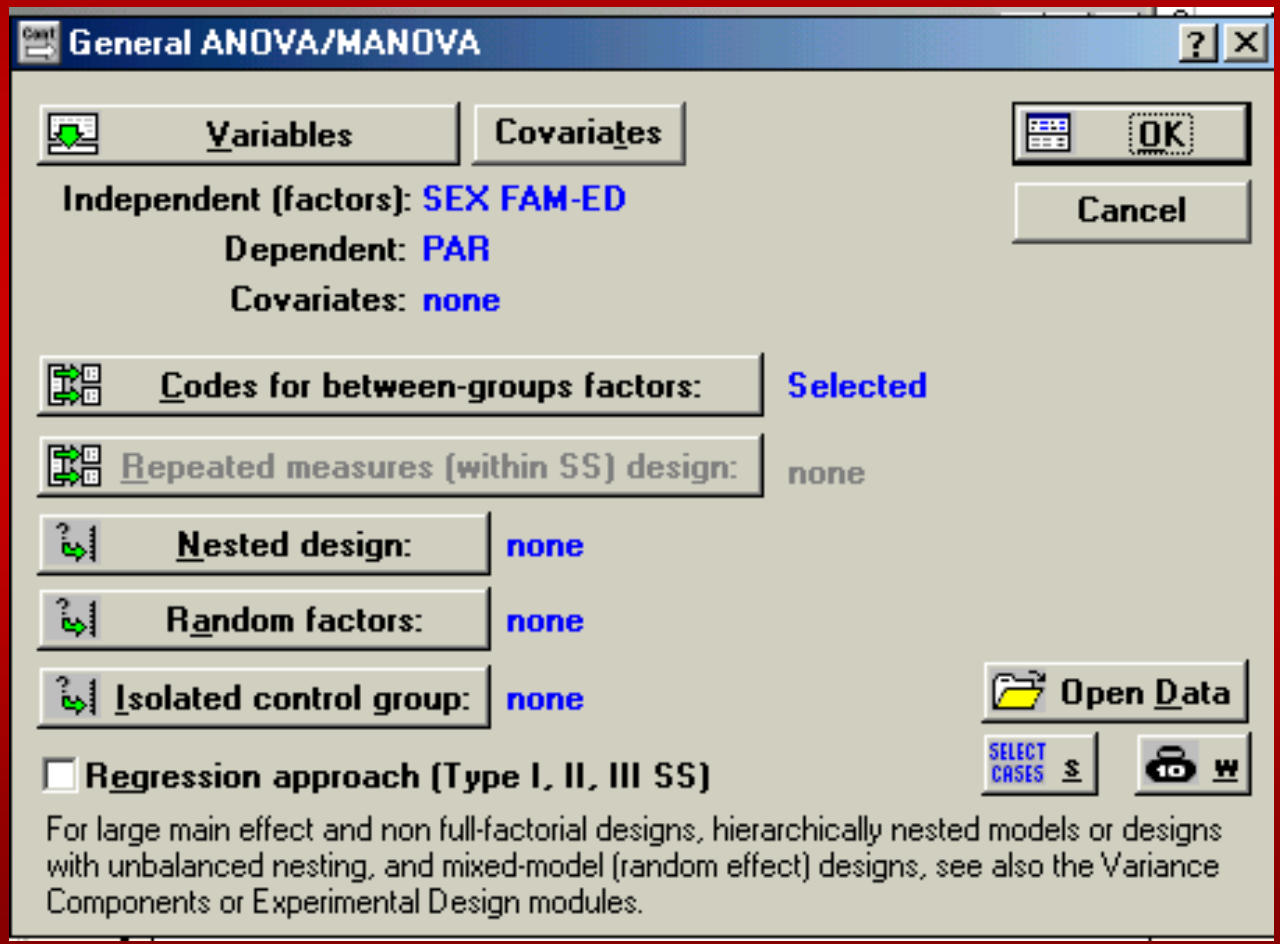
MANOVA



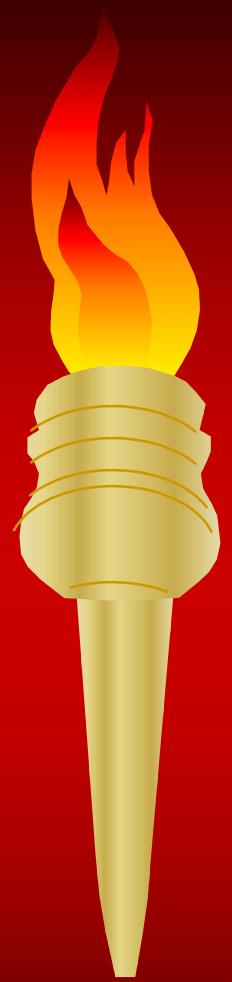

STATISTICA Module Switcher
Basic Statistics
Nonparametrics/Distrib.
ANOVA/MANOVA
Multiple Regression
Nonlinear Estimation
Time Series/Forecasting
Cluster Analysis
Data Management/MFM
Factor Analysis
General univariate and multivariate ANOVA; repeated measures, covariates; post-hoc tests, contrasts, tests of ANOVA assumptions; custom model building; regression options; canonical and discriminant function statistics; ... ;
StatSoft
Switch To
Customize list...
End & Switch To
Cancel

MANOVA




General ANOVA/MANOVA
Variables
Covariates
OK
Cancel
Independent (factors): SEX FAM-ED
Dependent: PAR
Covariates: none
Codes for between-groups factors:
Selected
Repeated measures (within SS) design:
none
Nested design:
none
Random factors:
none
Isolated control group:
none
Open Data
SELECT CASES
Regression approach (Type I, II, III SS)
For large main effect and non full-factorial designs, hierarchically nested models or designs with unbalanced nesting, and mixed-model (random effect) designs, see also the Variance Components or Experimental Design modules.

MANOVA



ANOVA Results

DESIGN: 3 - way ANOVA, fixed effects

DEPENDENT: 1 variable: PAR

BETWEEN: 1-SEX (2): 1 2
2-FAM (2): 1 2
3-ED (4): 1 2 3 4

WITHIN: none

All effects | Means/Graphs | Post hoc comparisons | Cancel

Specific effect/M

Planned comparis

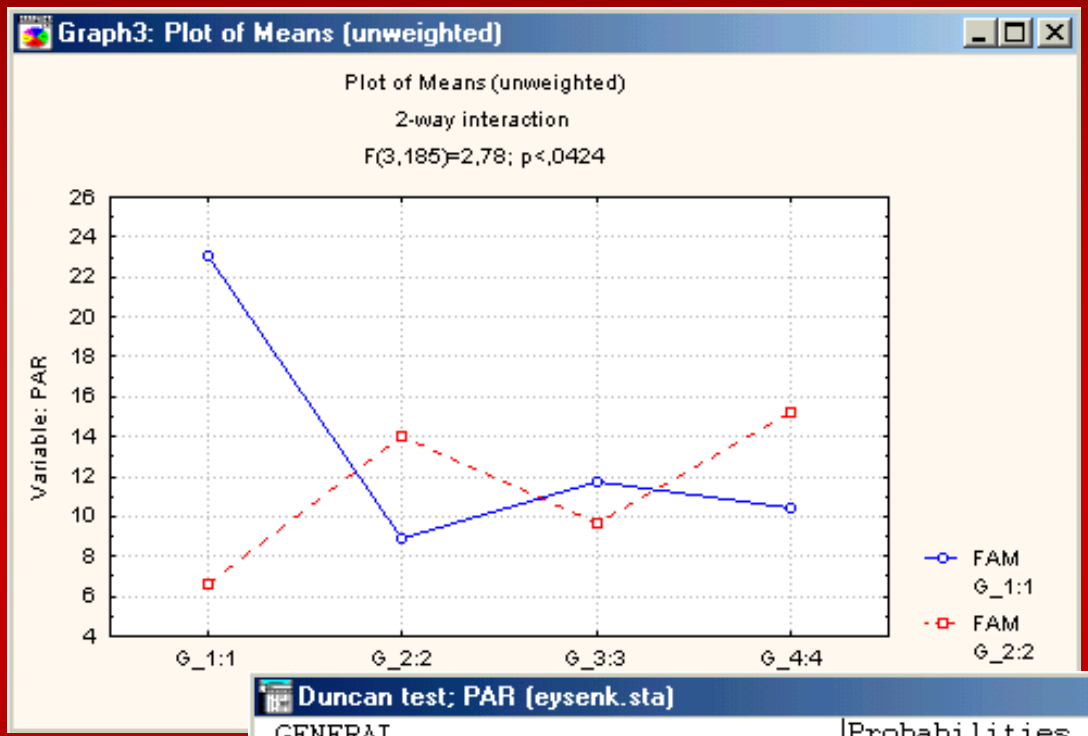
Pooled effect/

Summary of all Effects; design: (eysenk.sta)

Continue... 1-SEX, 2-FAM, 3-ED

Effect	df Effect	MS Effect	df Error	MS Error	F	p-level
1	1	294,1616	185	156,7267	1,876908	,172346
2	1	87,7365	185	156,7267	,559805	,455289
3	3	80,3911	185	156,7267	,512938	,673854
12	1	39,1748	185	156,7267	,249956	,617700
13	3	195,1141	185	156,7267	1,244932	,294824
23	3	435,8036	185	156,7267	2,780660	,042410
123	3	177,6508	185	156,7267	1,133507	,336836

MANOVA

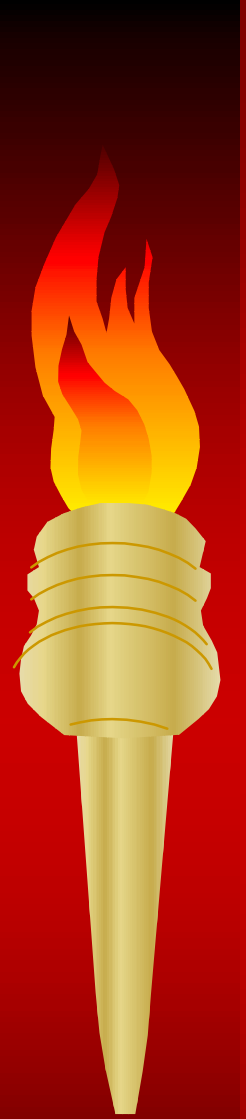


Duncan test; PAR (eysenk.sta)

GENERAL
MANOVA

Probabilities for Post Hoc Tests
INTERACTION: 2 x 3

SEX	FAM	ED		{1} 23,100	{2} 8,8750	{3} 11,700	{4} 10,421	{5} 6,5714	{6} 13,971	{7} 9,6490
. . .	1	1	{1}		,01	,03	,02	,00	,08	,01
. . .	1	2	{2}	,01		,61	,77	,64	,36	,87
. . .	1	3	{3}	,03	,61		,79	,36	,64	,70
. . .	1	4	{4}	,02	,77	,79		,48	,50	,87
. . .	2	1	{5}	,00	,64	,36	,48		,19	,56
. . .	2	2	{6}	,08	,36	,64	,50	,19		,43
. . .	2	3	{7}	,01	,87	,70	,87	,56	,43	
. . .	2	4	{8}	,11	,27	,50	,38	,13	,80	,32



MLR

- Множественный линейный регресионный анализ
- Оценка зависимости количественной переменной от одной или более количественных переменных

MLR (основы)

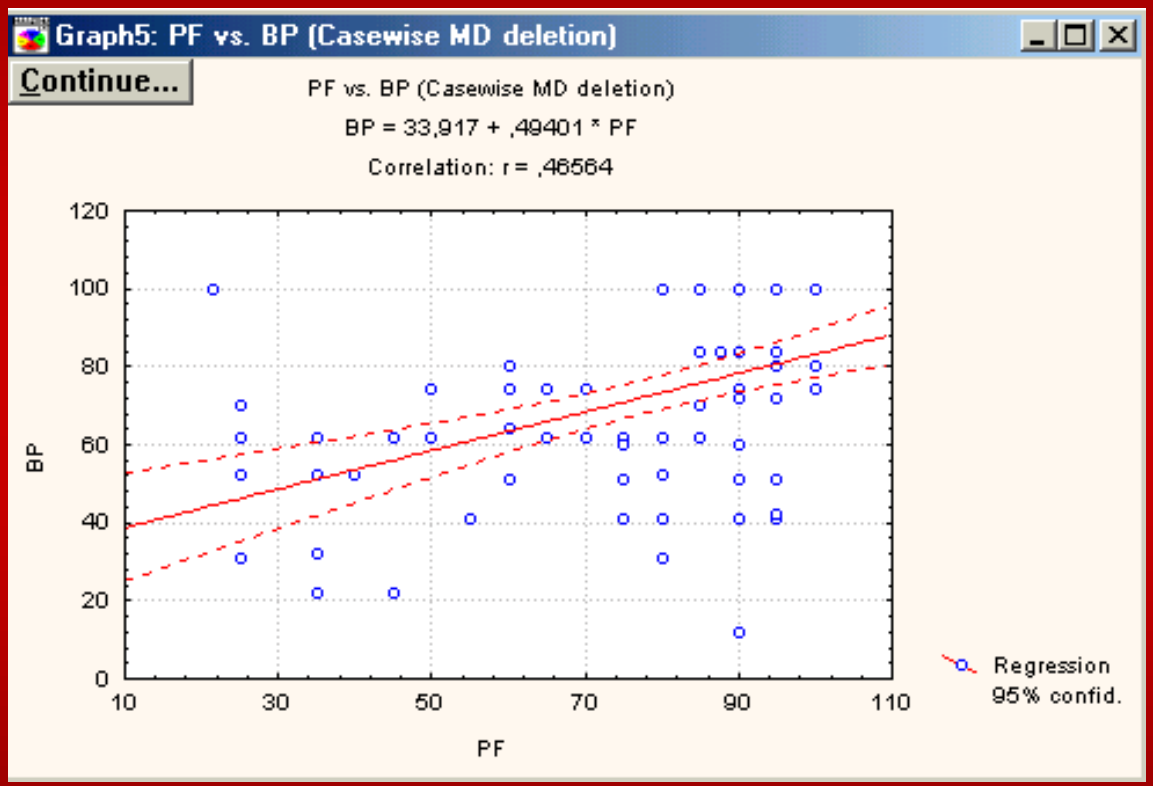
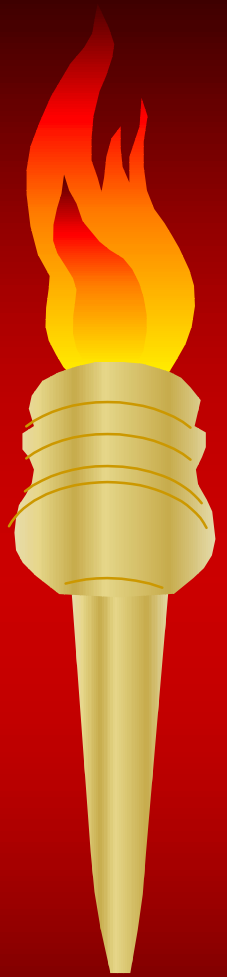


Диаграмма рассеяния + регрессионная линия
пытаемся провести ее так, чтобы расстояния от нее до наблюдаемых точек на графике были минимальными

MLR



Multiple Regression

Variables:

Independent: BP
Dependent: PF

OK
Cancel

Input file: Raw Data
MD deletion: Casewise
Mode: Standard

☑ Perform default (non-stepwise)
☐ Review descr. stats, corr. matri
☐ Extended precision computatio
☐ Batch processing/printing
☐ Print residual analysis

Specify all variables for the analysis; add
later. For stepwise regression etc. desele

Multiple Regression Results

Multiple Regression Results

Dep. Var. : PF Multiple R : ,46564420 F = 25,19363
R²: ,21682452 dF = 1,91
No. of cases: 93 adjusted R²: ,20821819 p = ,000003
Standard error of estimate: 19,603567779
Intercept: 45,863772341 Std.Error: 6,635755 t(91) = 6,9116 p < ,0000

BP beta=,466

(significant beta's are highlighted)

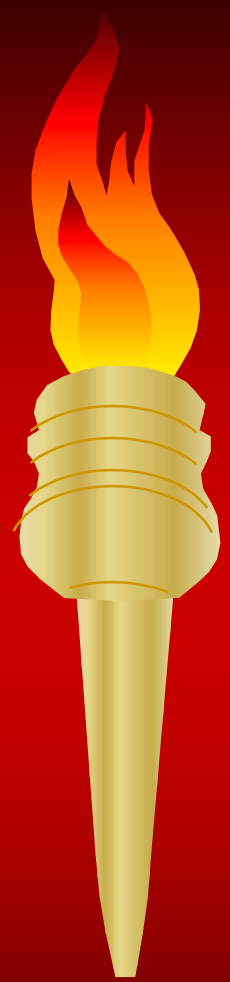
Regression summary
Predict dependent var.
Compute confidence limits
OK
Cancel
Residual analysis
Correlations & desc. stats
Alpha (display): .05
Apply

Regression Summary for Dependent Variable: PF (sf36l.sta)

Continue...

R= ,46564420 RI= ,21682452 Adjusted RI= ,20821819
F(1,91)=25,194 p<,00000 Std.Error of estimate: 19,606

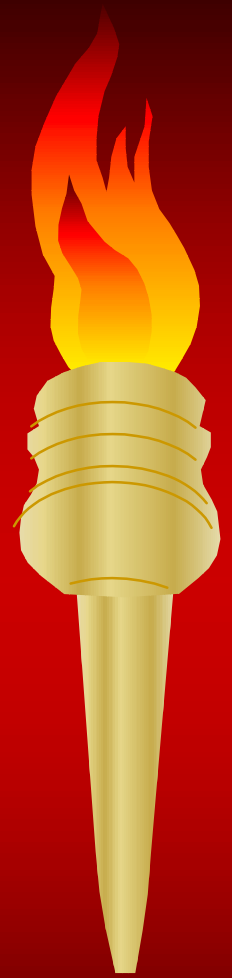
N=93	BETA	St. Err. of BETA	B	St. Err. of B	t(91)	p-level
Intercpt			45,86377	6,635755	6,911613	,000000
BP	,465644	,092770	,43891	,087444	5,019325	,000003



MLR (анализ остатков)

- Остатки - это значения, которые не могут быть объяснены независимой переменной
- Если у нас несколько переменных, мы берем первую, рассчитываем остатки и повторяем анализу уже на остатках (исключив влияние первой переменной) и т.д.
- Поэтому коэффициенты, получаемые в результате регрессионного анализа показывают ***независимое*** влияние переменных
- Если модель адекватна, остатки будут распределены по нормальному закону и разброс, соответственно, будет не очень сильным

MLR (анализ остатков)



Residual Analysis

Dep. Var. : PF — Multiple R : ,46564420 — F = 25,19363
R²: ,21682452 — df = 1,91
No. of cases: 93 — adjusted R²: ,20821819 — p = ,000003
Standard error of estimate: 19,605567779
Intercept: 45,863772341 Std.Error: 6,635755 t(91) = 6,9116 p < ,0000

Statistics
- Correlations & descr. (1)
- Regression summary (2)
- Display residuals & pred. (3)
- Durbin-Watson stat (4)
- Save residuals & pred. (5)

Scatter Plots
- Pred. & residuals (D)
- Pred. & squared resids (E)
- Pred. & observed (F)

Probability Plots
- Normal plot of resids (M)

Cancel

Casewise Plots
- Plots of residuals (A)
- Plots of outliers (B)
- Plots of predicted (C)

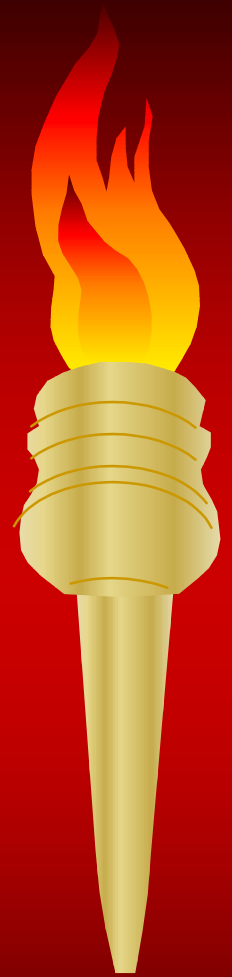
Standard Residual (sf36l.sta)

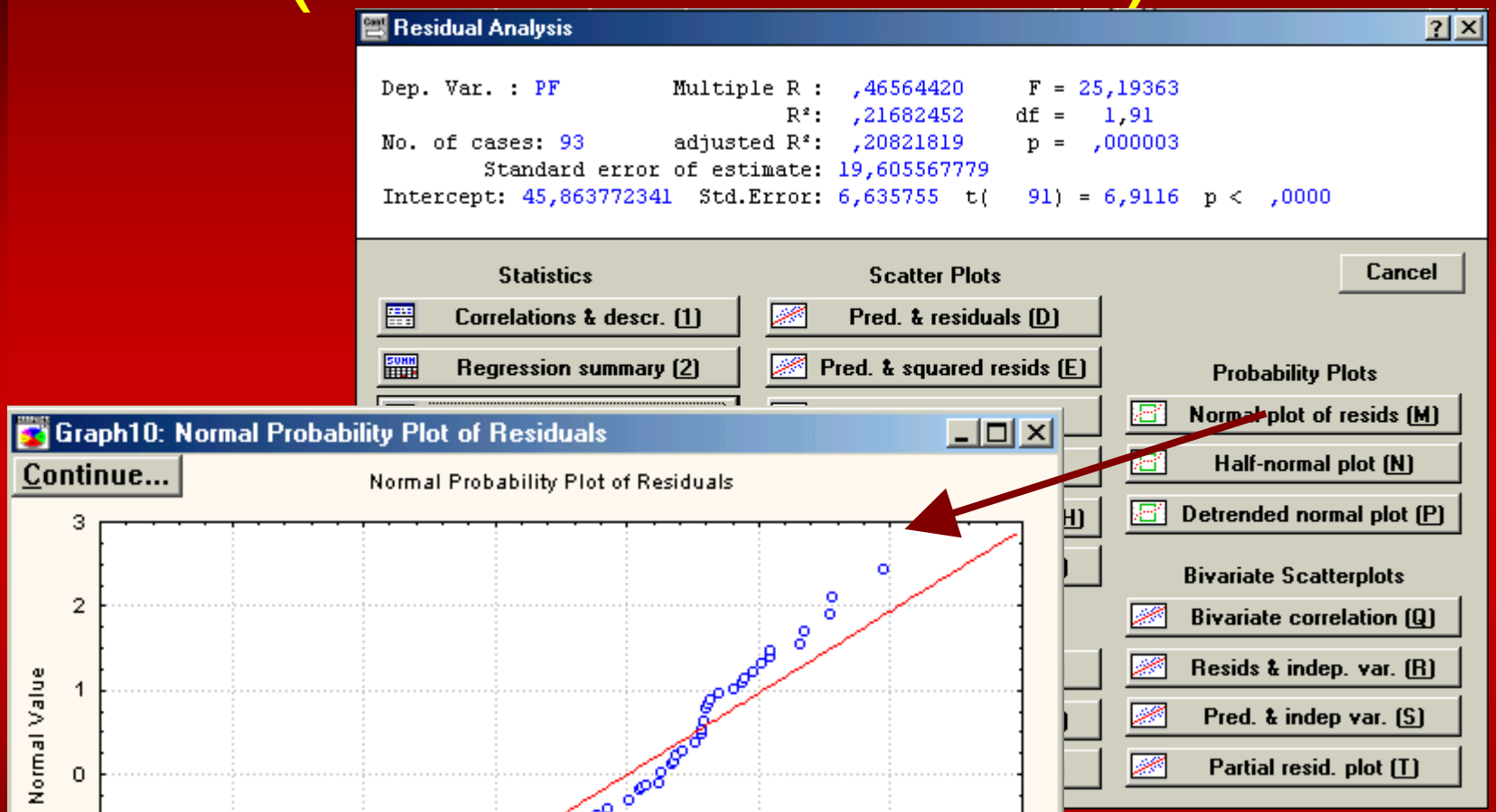
Continue...

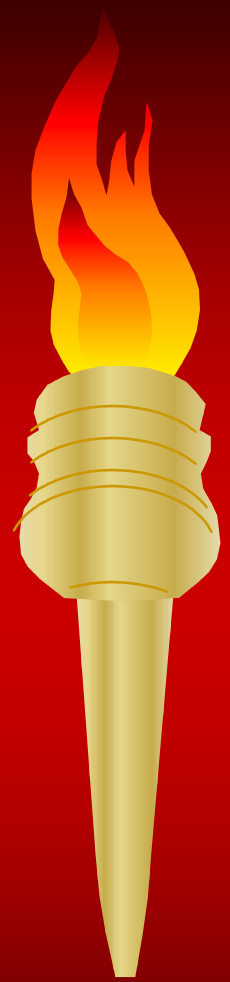
Dependent variable: PF

Case	Standard Residuals −3,5 … 1,98	Observed Value	Predictd Value	Residual
17		90,0000	89,75478	,2452
18		21,4286	89,75478	−68,3262
19		90,0000	72,19838	17,8016
20		90,0000	89,75478	,2452
21		75,0000	63,85909	11,1409
22		95,0000	89,75478	5,2452
23		25,0000	68,68710	−43,6871
24		65,0000	73,07619	−8,0762
25		70,0000	73,07619	−3,0762
26		60,0000	80,97658	−20,9766
27		100,0000	89,75478	10,2452
28		50,0000	78,34312	−28,3431
29		95,0000	89,75478	5,2452
30		95,0000	68,24818	26,7518
31		95,0000	89,75478	5,2452

MLR (анализ остатков)

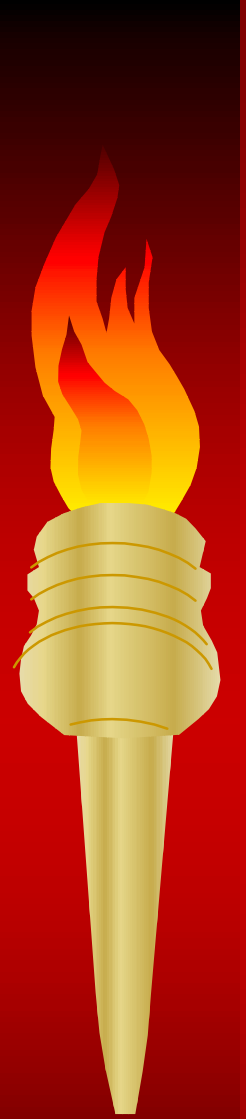






MLR (модели)

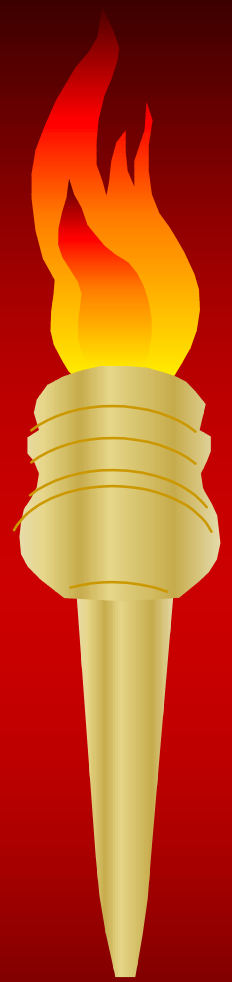
- Полная (включаются все переменные)
- Пошаговый отбор переменных (stepwise)
  - Forward
  - Backward



MANCOVA

- Многофакторный дисперсионный анализ с ковариантами
  - Комбинация из множественного линейного регрессионного анализа и дисперсионного анализа. Этапы
    - Построение модели с учетом только количественных переменных при помощи MLR
    - Дисперсионный анализ на остатках

MANCOVA



General ANOVA/MANOVA

Variables | Covariates | OK | Cancel

Independent (factors): SEX FAM-ED

Dependent: PAR

Covariates: AGE VNP

Codes for between-groups factors: Selected

Repeated measures (within SS) design: none

Nested design:

Random factors:

Isolated control grou

Regression approach (

For large main effect and non f
with unbalanced nesting, and
Components or Experimental D

ANOVA Results

```
DESIGN: 3 - way ANCOVA, fixed effects

DEPENDENT: 1 variable: PAR
COVARIATE: 2 variables: AGE VNP
  BETWEEN: 1-SEX      (  2): 1 2
           2-FAM      (  2): 1 2
           3-ED       (  4): 1 2 3 4

   WITHIN: none
```

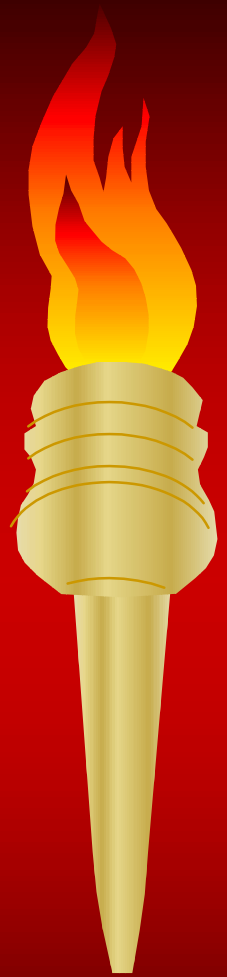
All effects | Means/Graphs | Post hoc comparisons | Cancel

Specific effect/Means/Graphs | Descriptive stats & graphs | Print design

Planned comparisons | Options | Within-cell regression

Pooled effect/error term | Output options

MANCOVA



Summary of all Effects; design: (eysenk.sta)

Continue... 1-SEX, 2-FAM, 3-ED

Effect	df Effect	MS Effect	df Error	MS Error	F	p-level
1	1	85,1137	183	122,7064	,693637	,406016
2	1	55,9069	183	122,7064	,455615	,500532
3	3	15,3134	183	122,7064	,124797	,945351
12	1	137,5578	183	122,7064	1,121033	,291092
13	3	239,7315	183	122,7064	1,953701	,122558
23	3	317,7355	183	122,7064	2,589397	,054347
123	3	214,2954	183	122,7064	1,746408	,159103

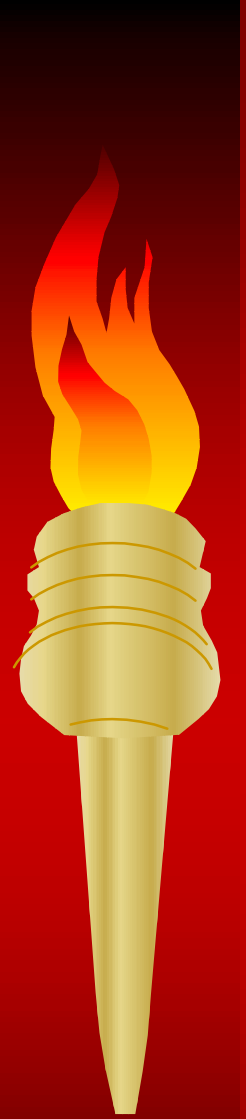
После коррекции по значениям переменных VNP и AGE эффект FAM*ED исчез (на самом деле, это эффект VNP)

Summary of all Effects; design: (eysenk.sta)

Continue... 1-SEX, 2-FAM, 3-ED

Effect	df Effect	MS Effect	df Error	MS Error	F	p-level
1	1	572,1018	184	143,0188	4,000185	,046966
2	1	91,7552	184	143,0188	,641560	,424179
3	3	201,7854	184	143,0188	1,410901	,241055
12	1	43,1038	184	143,0188	,301386	,583681
13	3	260,6784	184	143,0188	1,822686	,144560
23	3	481,9720	184	143,0188	3,369990	,019717
123	3	232,0732	184	143,0188	1,622676	,185640

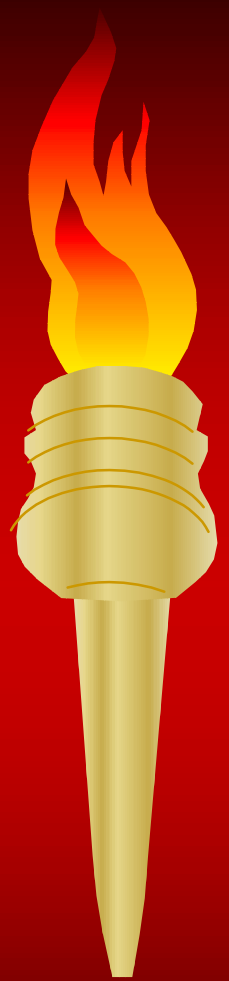
Без коррекции по VNP



MANCOVA

- Наиболее общая форма анализа данных, однако в системе Statistica в соответствующем модуле невозможно выполнять MLR
- SAS позволяет выполнять все виды анализа в одном модуле

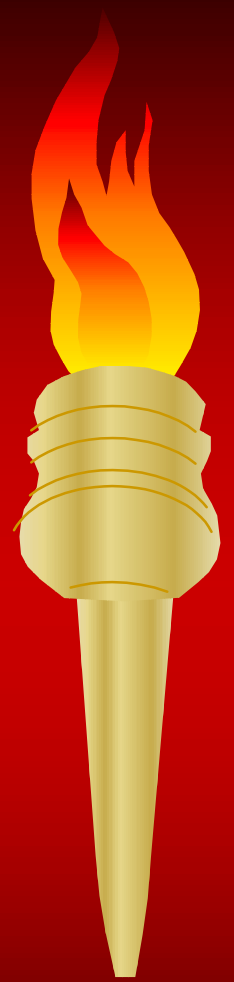
MANCOVA (SAS)



```
PROGRAM EDITOR
Command ===>

00001 PROC GLM DATA=mydat.mu99;
00002 CLASS SMPR SMPA EDUC MORT99;
00003 MODEL CH= SMPR SMPA|EDUC|MORT99 HDL;
00004 LSMEANS MORT99 EDUC /STDERR
00005 MEANS MORT99|EDUC / DUNCAN;
00006 RUN; QUIT;
```

```
                    SAS        18:41 Monday, June 6,

      General Linear Models Procedure
          Class Level Information

        Class     Levels    Values

        SMPR           2    0 1

        SMPA           2    0 1

        EDUC           3    1 2 3

        MORT99         2    0 1

   Number of observations in data set = 3907
```

MANCOVA (SAS)



General Linear Models Procedure

Dependent Variable: CH V1 TOTAL CHOLESTEROL

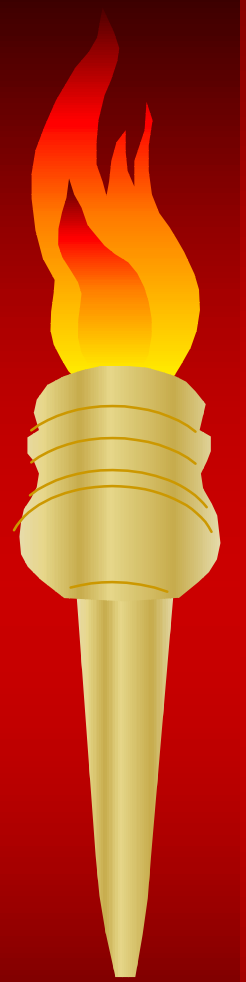
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	97326.42633	7486.64818	4.80	0.0001
Error	3110	4847719.72060	1558.75232		
Corrected Total	3123	4945046.14693			

R-Square	C.V.	Root MSE	CH Mean
0.019682	18.05328	39.48104	218.691741

Dependent Variable: CH V1 TOTAL CHOLESTEROL

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SMPR	1	516.03680	516.03680	0.33	0.5651
SMPA	1	1265.23144	1265.23144	0.81	0.3677
EDUC	2	29229.53245	14614.76623	9.38	0.0001
SMPA*EDUC	2	8520.60084	4260.30042	2.73	0.0652
MORT99	1	32009.07352	32009.07352	20.54	0.0001
SMPA*MORT99	1	385.45248	385.45248	0.25	0.6190
EDUC*MORT99	2	6965.28601	3482.64301	2.23	0.1072
SMPA*EDUC*MORT99	2	5992.86857	2996.43428	1.92	0.1464
HDL	1	12442.34421	12442.34421	7.98	0.0048

MANCOVA (SAS)



General Linear Models Procedure

Duncan Grouping	Mean	N	MORT99
A	221.279	1520	1
B	216.240	1604	0

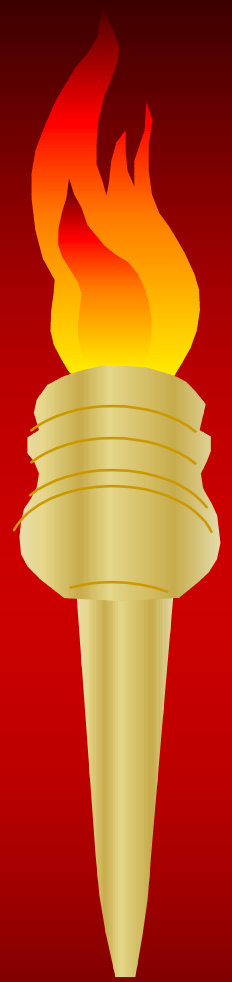
General Linear Models Procedure

Duncan Grouping	Mean	N	EDUC
A	222.773	984	3
B	218.652	830	2
B			
B	215.651	1310	1

Level of EDUC	Level of MORT99	CH N	CH Mean	CH SD
1	0	522	214.013410	38.1368610
1	1	788	216.736041	39.7042289
2	0	440	213.586364	36.8341329
2	1	390	224.366667	43.2912230
3	0	642	219.869159	38.5875225
3	1	342	228.225146	41.8455358

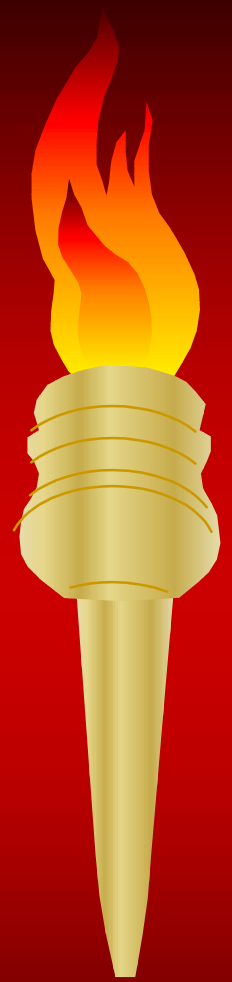
MANCOVA (SAS)

MORT99	CH LSMEAN	Std Err LSMEAN	Pr > |T| H0:LSMEAN=0
0	215.933556	1.207323	0.0
1	222.908094	1.612393	0.0



Качественная зависимая переменная

- Независимые переменные
  - Качественные
    - Стратифицированный анализ Mantel-Haenszel
    - Логлинейный анализ
  - Качественные и количественные
    - Логистическая регрессия
  - Количественные
    - Дискриминантный анализ



Связь нескольких качественных переменных с качественной

- Метод Mantel-Haenszel
  - Используется в рамках т.н. стратифицированного анализа (например, влияние курение на частоту развития ИМ с учетом пола, возраста - 5-летние группы, социального статуса
  - Часто используется в мультицентровых исследованиях
  - Может использоваться для анализа выживаемости (год дожития - отдельная переменная)

Метод Mantel-Haenszel

- Выбираем пару базовых переменных (например, ИМ и курение)
- Строим отдельные четырехпольные таблицы для каждого уровня остальных факторов:

ПОЛ=мужской

	Курит	Не курит
ИМ+	A1	B1
ИМ-	C1	D1

ПОЛ=женский

	Курит	Не курит
ИМ+	A2	B2
ИМ-	C2	D2

$$OR_{MH} = \frac{\sum A_i \bullet D_i / N_i}{\sum B_i \bullet C_i / N_i}$$

$$x = \sum A_i$$

$$E(x) = \sum \frac{R_{1i} \bullet C_{1i}}{N_i}$$

$$Var(x) = \sum \frac{R_{1i} \bullet C_{1i} \bullet R_{2i} \bullet C_{2i}}{N_i^2 \bullet (N_i - 1)}$$

$$\chi^2 = \frac{(|x - E(x)| - 0.5)^2}{Var(x)}$$

Метод Mantel-Haenszel

В системе Statistica можно использовать только для данных по выживаемости (требуется времена дожития)
В SAS для любых таблиц

```
Command ===>

    1    PROC FREQ DATA=MYDAT.MU99;
    2    TABLES EDUC*MORT99*SMPA/ CMH;
    3    RUN;
NOTE: The PROCEDURE FREQ used 1.00 seconds.
PROGRAM EDITOR
Command ===>

00001 PROC FREQ DATA=MYDAT.MU99;
00002 TABLES EDUC*MORT99*SMPA/ CMH;
00003 RUN;
```

Метод Mantel-Haenszel (SAS)

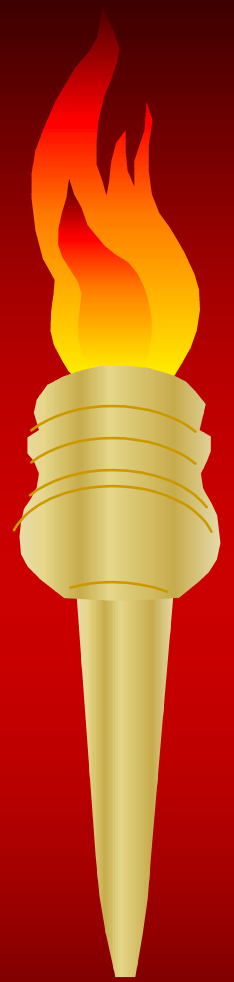


TABLE 3 OF MORT99 BY SMPA
CONTROLLING FOR EDUC=3

MORT99	SMPA(SMOKED IN THE PAST)		
Frequency Percent Row Pct Col Pct	0	1	Total
0	247 24.46 37.42 78.16	413 40.89 62.58 59.51	660 65.35
Total	316 31.29	694 68.71	1010 100.00

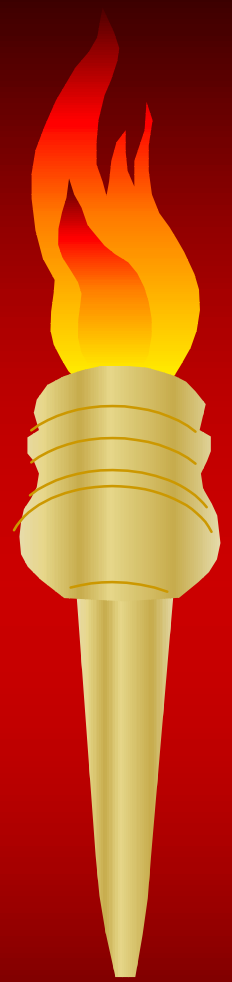
(Continued)

MORT99	SMPA(SMOKED IN THE PAST)		
Frequency Percent Row Pct Col Pct	0	1	Total
1	69 6.83 19.71 21.84	281 27.82 80.29 40.49	350 34.65
Total	316 31.29	694 68.71	1010 100.00

SUMMARY STATISTICS FOR MORT99 BY SMPA
CONTROLLING FOR EDUC

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	83.951	0.000
2	Row Mean Scores Differ	1	83.951	0.000
3	General Association	1	83.951	0.000

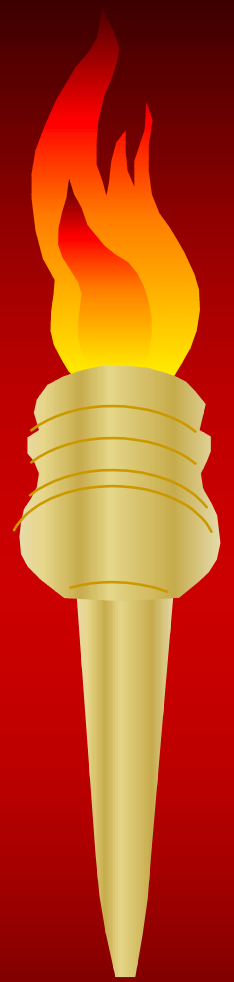


Метод Mantel-Haenszel (SAS)

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Bounds	
Case-Control (Odds Ratio)	Mantel-Haenszel	2.371	1.971	2.851
	Logit	2.372	1.966	2.862
Cohort (Col1 Risk)	Mantel-Haenszel	1.978	1.709	2.289
	Logit	1.980	1.700	2.306
Cohort (Col2 Risk)	Mantel-Haenszel	0.850	0.820	0.880
	Logit	0.857	0.828	0.887

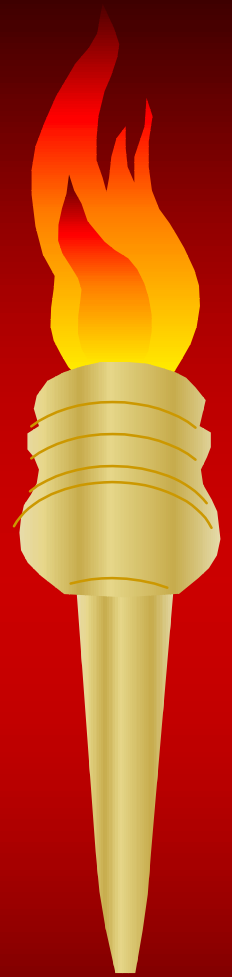
The confidence bounds for the M-H estimates are test-based.



Логлинейный анализ

- Аналог дисперсионного анализа для многомерных таблиц
- Анализируется совместное влияние нескольких переменных на частоты в многомерной таблице
- Если модель соответствует данным (fit), то переменные не влияют на частоты в таблице - модель можно подогнать зная только краевые частоты (R1, R2, C1, C2)

Логлинейный анализ



Log-Linear: 2x2x2 - Microsoft Internet Explorer

Файл Правка Вид Избранное Сервис Справка

Назад Поиск Избранное Журнал

Адрес http://faculty.vassar.edu/lowry/log.html

Data Entry

	A_1		A_2	
	C_1	C_2	C_1	C_2
B_1	1	1	5	5
B_2	2	2	10	10

df [ABC] = 4

G^2 [ABC] = 0

p = ns

Calculate

Reset

Note that G^2 is distributed approximately as chi-square. p values are derived by referring calculated values of G^2 to the appropriate sampling distributions of chi-square.

Data Entry

	A_1		A_2	
	C_1	C_2	C_1	C_2
B_1	10	1	25	5
B_2	2	2	10	10

df [ABC] = 4

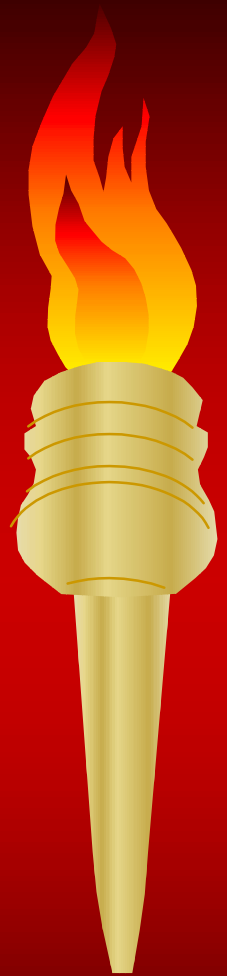
G^2 [ABC] = 10.61

Calculate

Reset

Note that G^2 is distributed approximately as chi-square. p values are derived by referring calculated values of G^2 to the appropriate sampling distributions of

Логлинейный анализ



AB Table

	B_1	B_2	
A_1	11	4	15
A_2	30	20	50
	41	24	65

AC Table

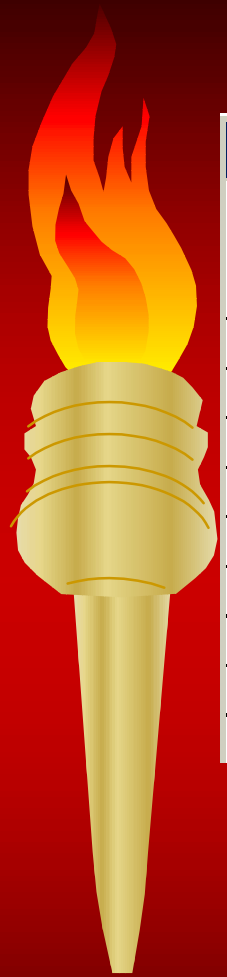
	C_1	C_2	
A_1	12	3	15
A_2	35	15	50
	47	18	65

BC Table

	C_1	C_2	
B_1	35	6	41
B_2	12	12	24
	47	18	65

df [AB] = 1	df [AC] = 1	df [BC] = 1
G^2 [AB] = 0.91	G^2 [AC] = 0.6	G^2 [BC] = 9.29
p = ns	p = ns	p = p<.005
df [ABC-AB] = 3	df [ABC-AC] = 3	df [ABC-BC] = 3
G^2 [ABC-AB] = 9.7	G^2 [ABC-AC] = 10.01	G^2 [ABC-BC] = 1.32
p = p<.025	p = p<.025	p = ns
df [ABC-AC-BC] = 2	df [ABC-AB-BC] = 2	df [ABC-AB-AC] = 2
G^2 [ABC-AC-BC] = 0.72	G^2 [ABC-AB-BC] = 0.41	G^2 [ABC-AB-AC] = 9.1
p = ns	p = ns	p = p<.025

Логлинейный анализ



Data: NEW.STA 10v * 10c

NUM VAL	1 A	2 B	3 C	4 WEIG
1	1,000	1,000	1,000	10,000
2	1,000	1,000	2,000	1,000
3	1,000	2,000	1,000	2,000
4	1,000	2,000	2,000	2,000
5	2,000	1,000	1,000	25,000
6	2,000	1,000	2,000	5,000
7	2,000	2,000	1,000	10,000
8	2,000	2,000	2,000	10,000
9				

Analysis

e analyzed:

A B C

2 x 2 x 2

Input file: Frequencies with coding variables

OK

Cancel

Variables: A-C

Variable containing frequencies: WEIG

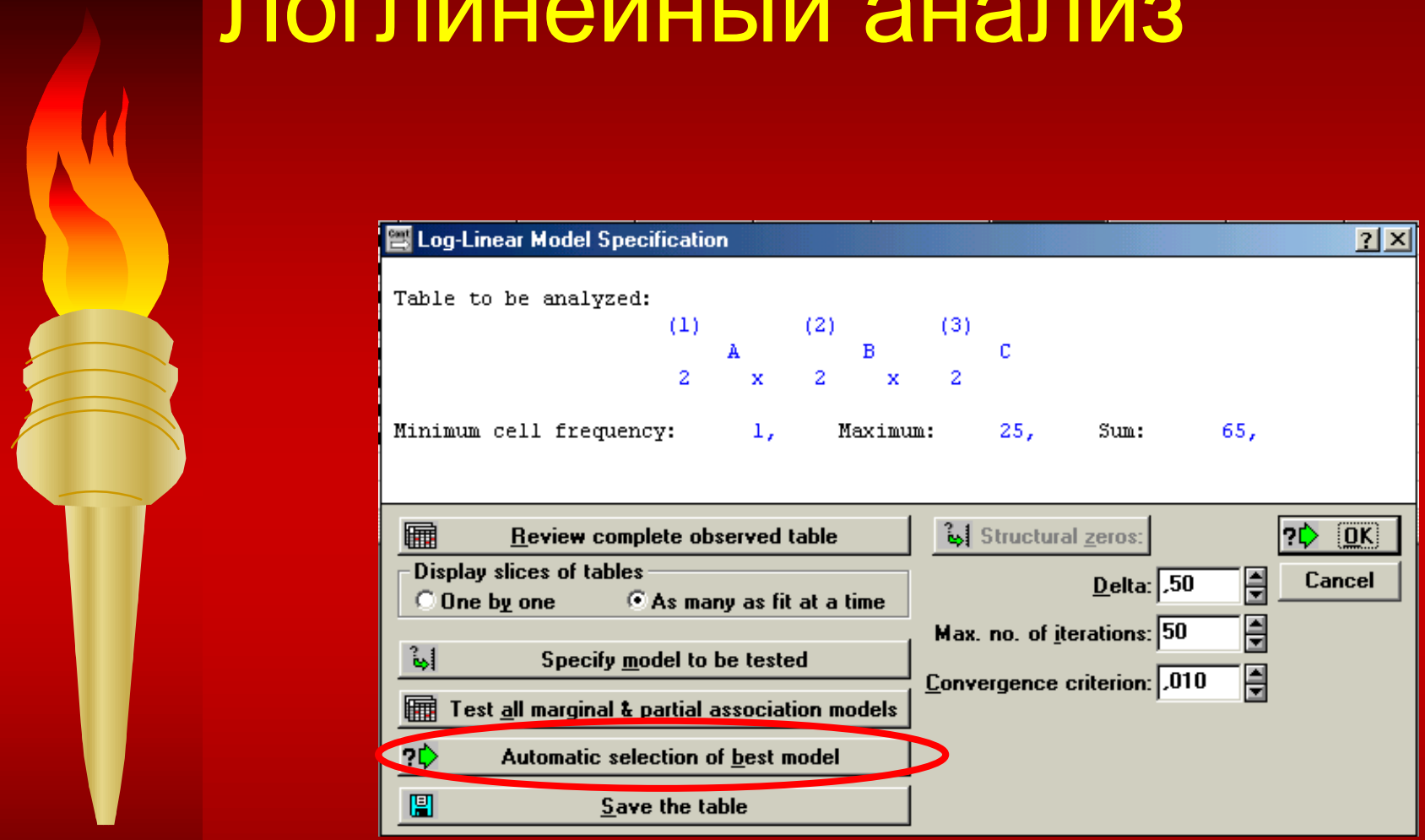
Open Data

Select codes Selected

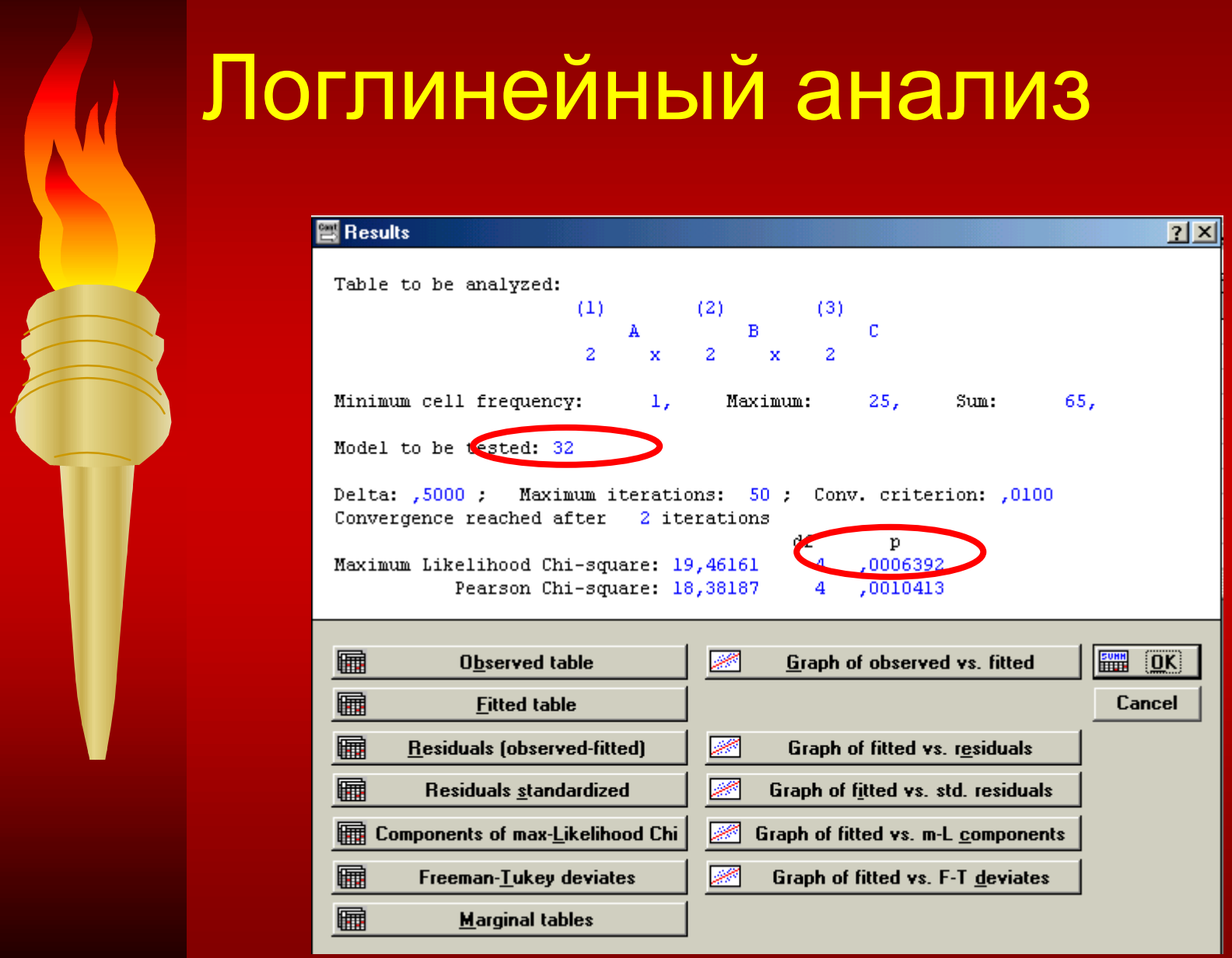
SELECT CASES s

w

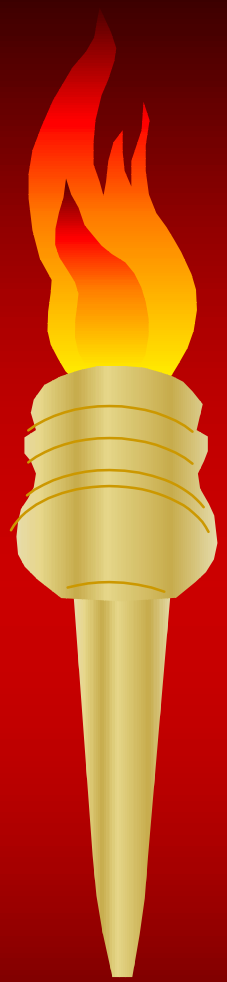
Логлинейный анализ



Логлинейный анализ



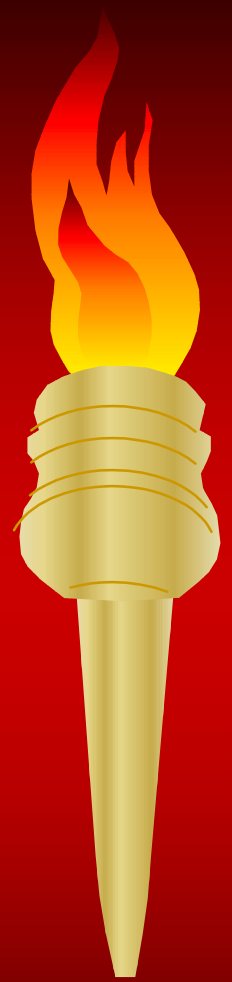
Логлинейный анализ (SAS)



```
PROGRAM EDITOR
Command ===>

00001 DATA loglin; INPUT A B C W; CARDS;
00002 1 1 1 10
00003 1 1 2 1
00004 1 2 1 2
00005 1 2 2 2
00006 2 1 1 25
00007 2 1 2 5
00008 2 2 1 10
00009 2 2 2 10
00010 ;
00011 RUN;
00012 PROC CATMOD DATA=loglin;
00013 MODEL  C*A*B =_RESPONSE_ /ML;
00014 LOGLIN A|B|C;
00015 WEIGHT W;
00016 RUN;
```

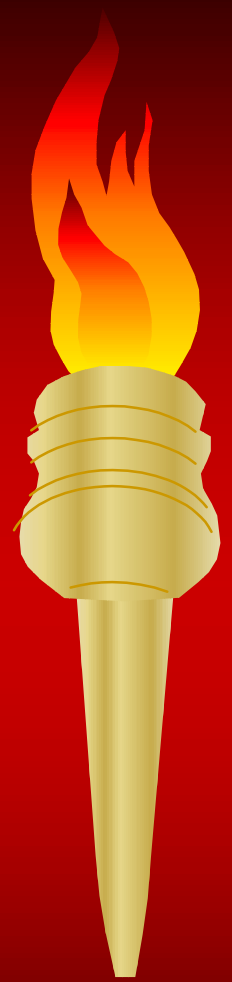
MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

Source	DF	Chi-Square	Prob
A	1	12.99	0.0003
B	1	0.51	0.4746
A*B	1	0.19	0.6636
C	1	6.03	0.0141
C*A	1	0.19	0.6636
C*B	1	6.03	0.0141
C*A*B	1	0.19	0.6636
LIKELIHOOD RATIO	0	.	.



Логлинейный анализ (SAS)

- На самом деле необходимо проанализировать несколько моделей:
  1. “Независимую” (LOGLIN A B C)
  2. «С исключением верхнего уровня» (LOGLIN A B C B*C)
  3. «С двойными взаимодействиями (LOGLIN A B C A*B A*C B*C)
  4. «Полную» (LOGLIN A|B|C)
- Записать остатки (LIKLEHOOD RATIO)
- Посчитать выигрыш при усложнении модели (хи^2 и df)



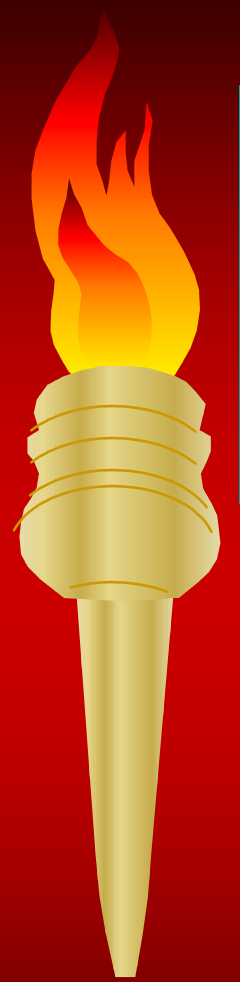
Логлинейный анализ (SAS)

- Составить таблицу

модель	df	LR	p	dChi	DeltaDF
1	4	10.61	0.0313		
2	3	1.31	0.7258	9.30	1
3	1	0.19	0.6568	1.12	1
4	0	0	0	0.19	1

- Видно, что наибольший выигрыш при переходе от модели 1 к модели 2, все остальные хи^2 не достоверны
- Значит выбираем модель 2

Логлинейный анализ (SAS)



MAXIMUM LIKELIHOOD ANALYSIS OF VARIANCE TABLE

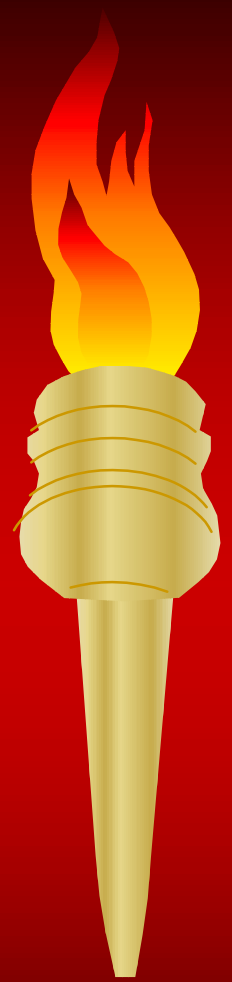
Source	DF	Chi-Square	Prob
A	1	16.73	0.0000
B	1	0.39	0.5305
C	1	8.59	0.0034
C*B	1	8.59	0.0034
LIKELIHOOD RATIO	3	1.31	0.7258

ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
RESPONSE	1	-0.6020	0.1472	16.73	0.0000
	2	0.0943	0.1504	0.39	0.5305
	3	0.4409	0.1504	8.59	0.0034
	4	0.4409	0.1504	8.59	0.0034

NOTE: _RESPONSE_ = A B C C*B

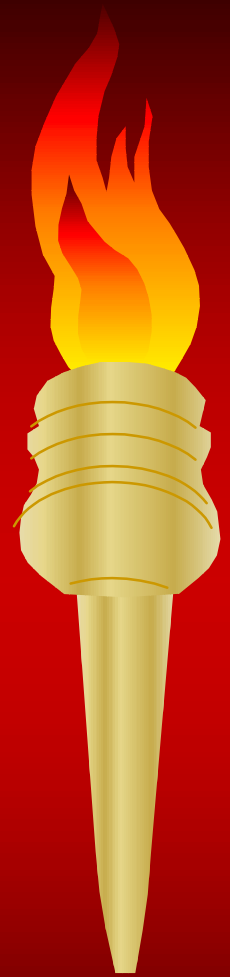
Мы можем оценить эффекты. Например, коэффициент для A -0,6. Беря антилогарифм ($e^{-0.6}$=0,55) имеем, что значений первого уровня A почти в 2 раза меньше, чем второго. Глядя на 4 строку, можно сказать, что C1 значительно больше при B1, и т.д.

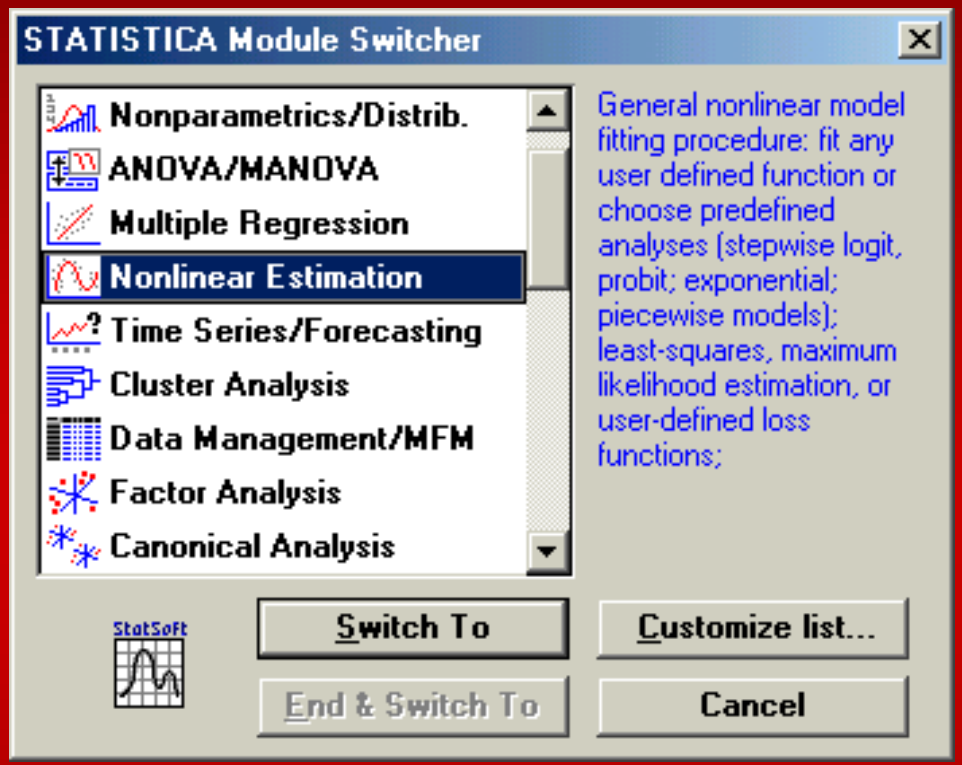


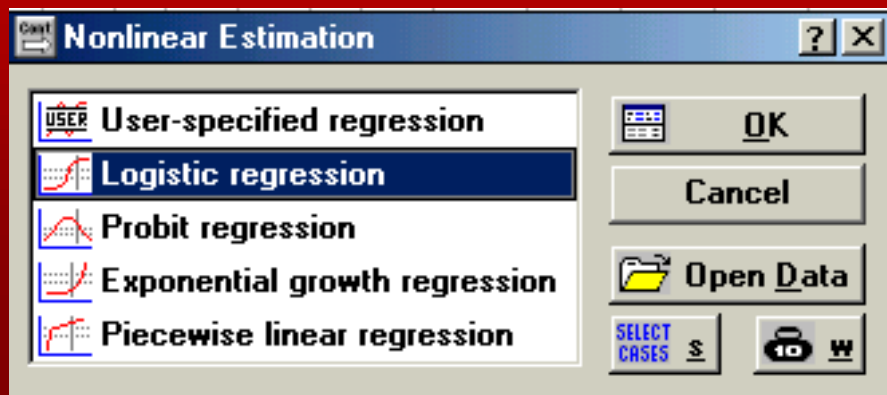
Логистическая регрессия

- На самом деле, наиболее общий случай (более общий, чем логлинейный анализ), правда оптимизирована для работы с бинарными зависимыми переменными
- Может принимать для анализа как качественные, так и количественные переменные
- В принципе, если логлинейный анализ аналог ANOVA, то логистическая регрессия - аналог ANCOVA (в SAS можно даже использовать одну и ту же процедуру)

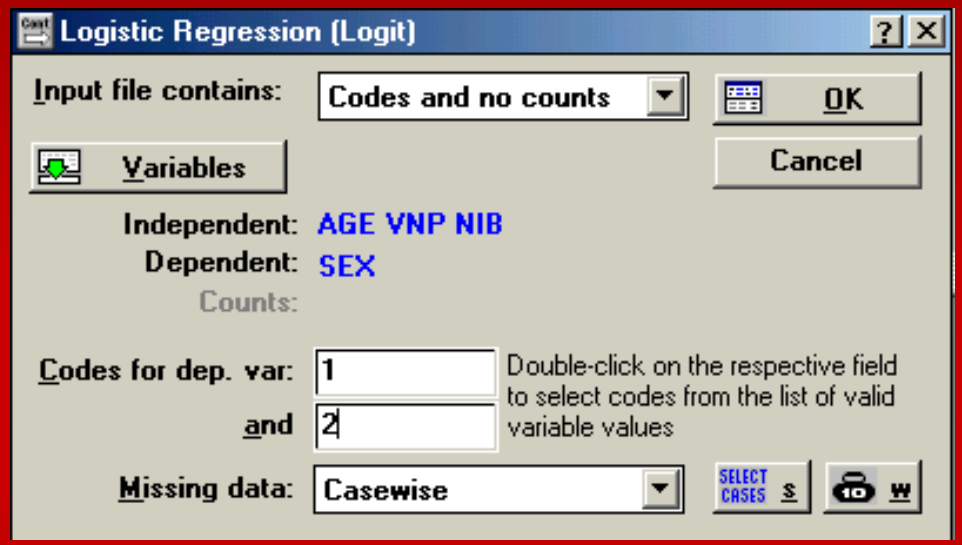
Логистическая регрессия





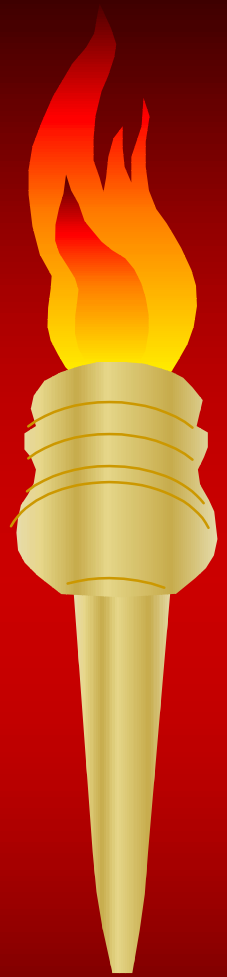


Логистическая регрессия



- При проведении логистического регрессионного анализа качественные переменные с более чем двумя уровнями должны быть превращены в т.н. dummy переменные (пустышки) с двумя уровнями. Например, образование с 3 будет превращено в 2 переменных, например высшее (есть/нет) и ниже среднего (да/нет).

Логистическая регрессия



Results

Model is: logistic regression (logit) No. of 0's: 69,00000 (34,32836%)
No. of 1's: 132,0000 (65,67164%)
Dependent variable: SEX Independent variables: 3
Loss function is: maximum likelihood Final value: 119,61091998
-2*log(Likelihood): for this model = [illegible] intercept only: 258,5622
Chi-square = 19,34033, df = 3, p = ,0002331

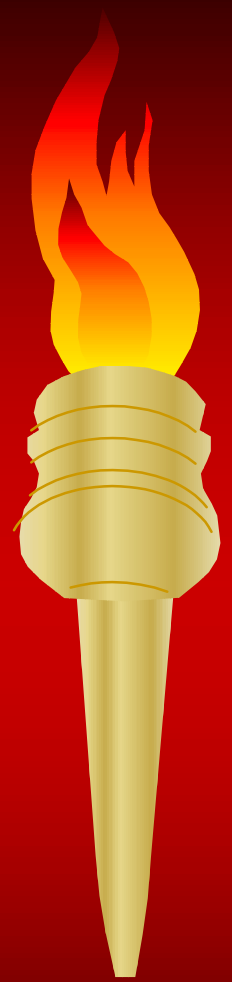
Parameters & standard errors
Cov./corrs. of parameters
Scale MSE to 1 Cnf. interval: 95,0 %
Residuals
Predicted values
Observed values
Classification of case
Means & standard
Difference (previ
Save predicted and r
Fitted 2D function & observed vals
Fitted 3D function & observed vals
Distribution of residuals
Normal probability plot of residuals
Half-normal probability plot
OK
Cancel

Model: Logistic regression (logit) N of 0's:69 1's:132

Continue...

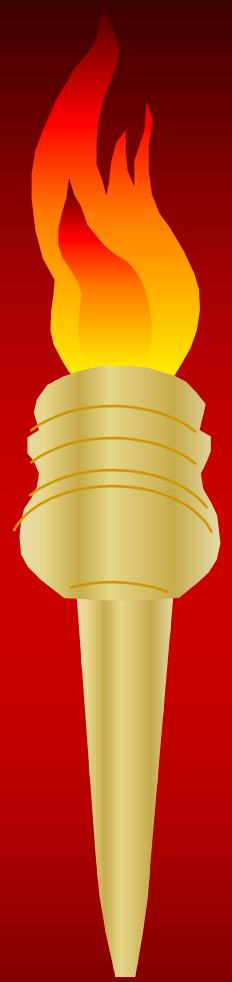
Dep. var: SEX Loss: Max likelihood (MS-err. scaled to 1)
Final loss: 119,61091998 ChiI(3)=19,340 p=,00023

N=201	Const.B0	AGE	VNP	NIB
Estimate	-5,24838	,00834	,3242	-,17996
Standard Error	1,54970	,03663	,0843	,45949
t(197)	-3,38671	,22763	3,8445	-,39165
p-level	,00085	,82017	,0002	,69574
-95%CL	-8,30450	-,06390	,1579	-1,08611
+95%CL	-2,19225	,08058	,4905	,72619
Wald's Chi-square	11,46980	,05181	14,7799	,15339
p-level	,00071	,81994	,0001	,69532
Odds ratio (unit ch)	,00526	1,00837	1,3829	,83530
-95%CL	,00025	,93810	1,1711	,33753
+95%CL	,11167	1,08391	1,6332	2,06719
Odds ratio (range)		1,27355	35,3878	,83530
-95%CL		,15675	5,6800	,33753
+95%CL		10,34733	220,4760	2,06719



Логистическая регрессия

- Для качественных бинарных показателей мы можем взять OR прямо из таблицы (это будет OR, откорректированный на влияние других факторов риска)
- Для количественных показателей необходимо взять регресионный коэффициент (estimate), умножить его на стандартное отклонение для этого показателя и взять от произведения антилогарифм.
- Например, для возраста в описанном ранее примере estimate=0,008, SD=5,4. Соответственно $OR=e^{0,008*5,4}=1,04$



Логистическая регрессия (SAS)

- Можно использовать несколько процедур - CATMOD и LOGISTIC. Первая может также использоваться для мультиноминальных (не бинарных) типов анализа
- CATMOD очень требовательна к памяти

```
00001 PROC LOGISTIC DATA=mydat.mu99;
00002 MODEL MORT99=SBP EDHIGH EDLOW SMPA SMPR;
00003 RUN;
00004 PROC CATMOD DATA=mydat.mu99;
00005 DIRECT SBP;
00006 MODEL MORT99=SBP EDHIGH EDLOW SMPA SMPR/ NOGLS ML NOPROFILE;
00007 RUN;
00008 QUIT;
```

Логистическая регрессия (SAS)

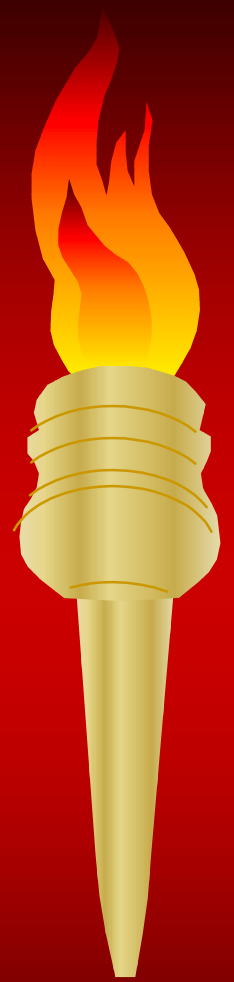
The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	5.2032	0.2993	302.2534	0.0001	.
SBP	-0.0313	0.00197	251.6423	0.0001	-0.385224
EDHIGH	0.3162	0.1025	9.5254	0.0020	0.080750
EDLOW	-0.3936	0.0938	17.6021	0.0001	-0.107249
SMPA	-0.4710	0.1208	15.1963	0.0001	-0.104752
SMPR	-0.6347	0.0951	44.5831	0.0001	-0.173424

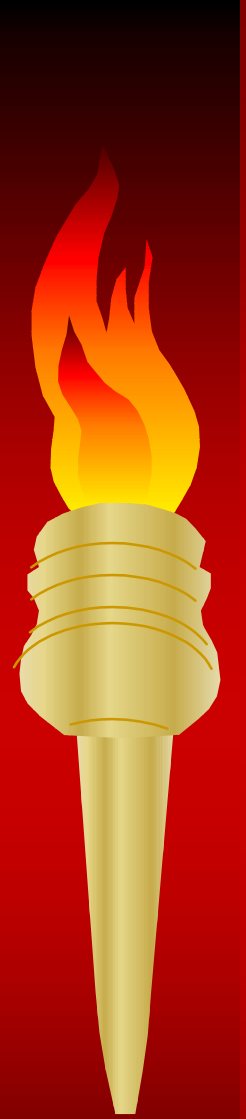
ANALYSIS OF MAXIMUM LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	4.6116	0.2787	273.88	0.0000
SBP	2	-0.0313	0.00197	251.64	0.0000
EDHIGH	3	-0.1581	0.0512	9.53	0.0020
EDLOW	4	0.1968	0.0469	17.60	0.0000
SMPA	5	0.2355	0.0604	15.20	0.0001
SMPR	6	0.3174	0.0475	44.58	0.0000



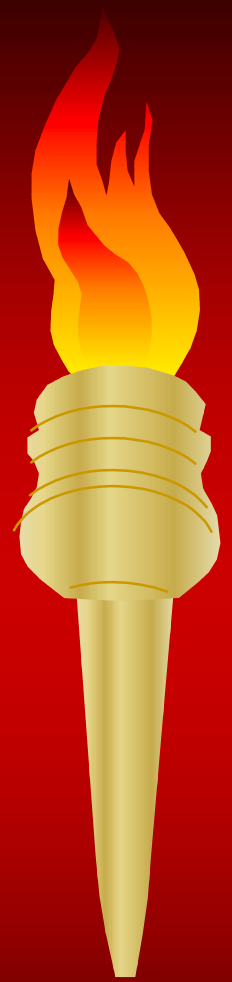
Дискриминантный анализ

- Методика создания классифицирующих правил
- Пытается создать серию уравнений, которые позволили бы правильно классифицировать группы
- В принципе, является аналогом ANOVA, но только «перевернутым» (рост позволяет отличить мужчин от женщин)



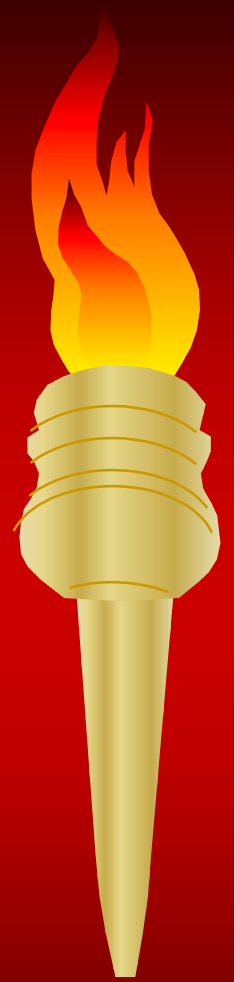
Особый случай - анализ выживаемости

- Нечто среднее между анализом качественных и количественных показателей (качественный показатель - цензурирование, количественный - время под наблюдением)
- Основа унивариантного анализа - построение таблиц дожития (life tables) и кривых дожития (Kaplan-Meier)
- Многомерные методики
  - Только качественные независимые переменные
    - Стратифицированный анализ Mantel-Haenszel
    - Коррекция кривых дожития
  - Качественные и количественные независимые переменные
    - Моделирование выживаемости (Weibull и др.)
    - Модель пропорционального риска (Cox)



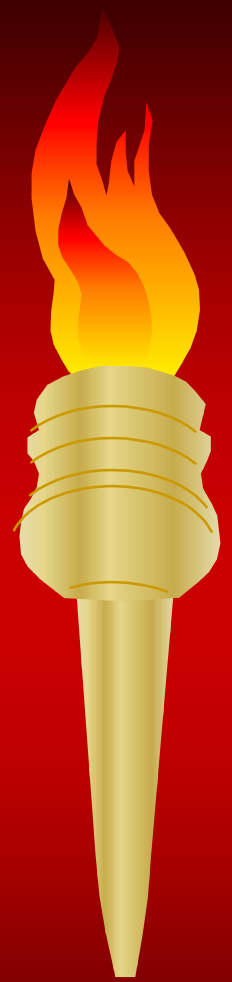
Снижение размерности

- Количественные переменные
  - Анализ главных компонент
  - Факторный анализ
  - Многомерное шкалирование
  - Иерархический кластерный анализ
- Качественные переменные
  - Корреспондентский анализ
- Объединение наблюдений
  - Кластерный анализ



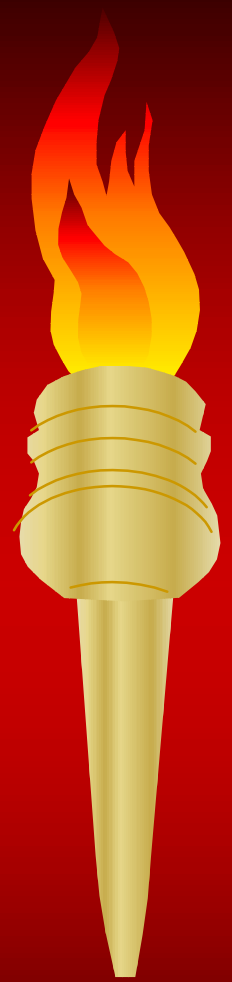
Факторный анализ

- Вместе с анализом главных компонент - основной метод снижения расзмерности
- Пытается найти новые переменные, которые бы комбинировали в себе несколько старых (рост, измеренный в метрах, футах и локтях)
- Отличие от метода главных компонент - отсутствие “уникальных” факторов - факторов, представленных одной переменной
- Единицей анализа является корреляционная матрица



Факторный анализ

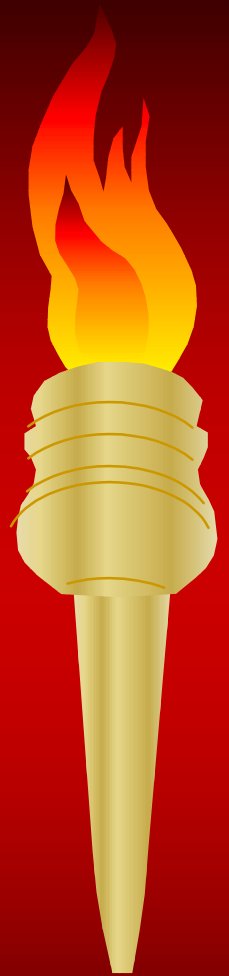
- Выделяют неколько вариантов
  - Стандартный (R) - факторный анализ переменных
  - Обратный (Q) - факторный анализ наблюдений
  - Временной (T) - факторный анализ временных тенденций (у разных людей по одной переменной)
  - Обратный временной (S) - факторный анализ наблюдений в зависимости от времени
  - Одного наблюдения (O) - факторный анализ временных тенденций по разным показателям у одного субъекта

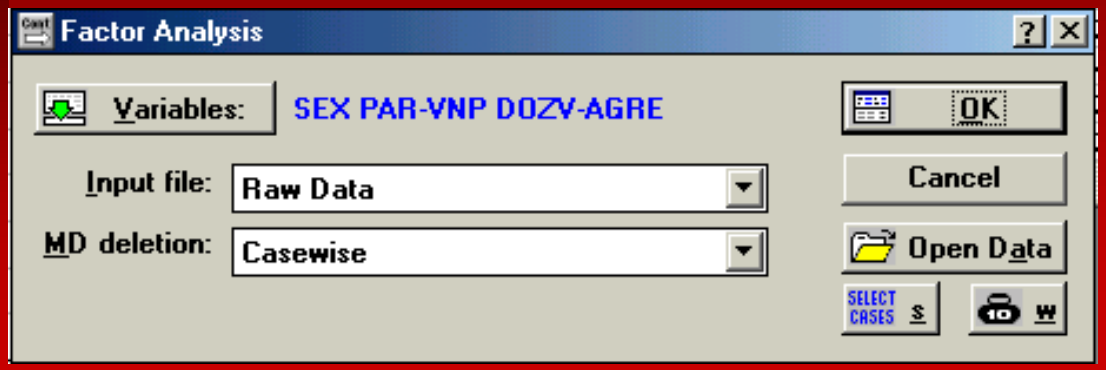


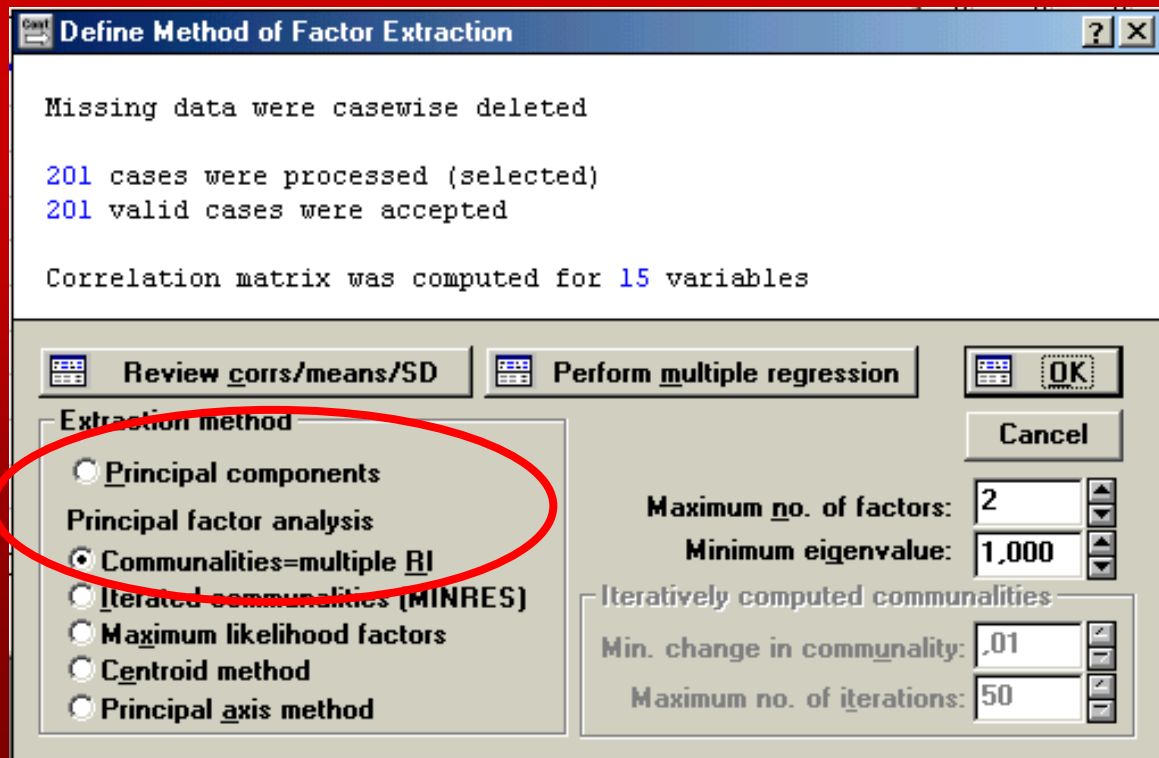
Факторный анализ

- Требует не менее 10 (5) наблюдений на каждую включенную в анализ переменную, но общее число наблюдений должно быть больше 100
- Факторный анализ отличается от компонентного тем, что находится в диагональных элементах корреляционной матрицы (1 в компонентном, иное число - максимальное значение r в строке, множ. Коэффициент корреляции и т.п. - в факторном)

Факторный анализ

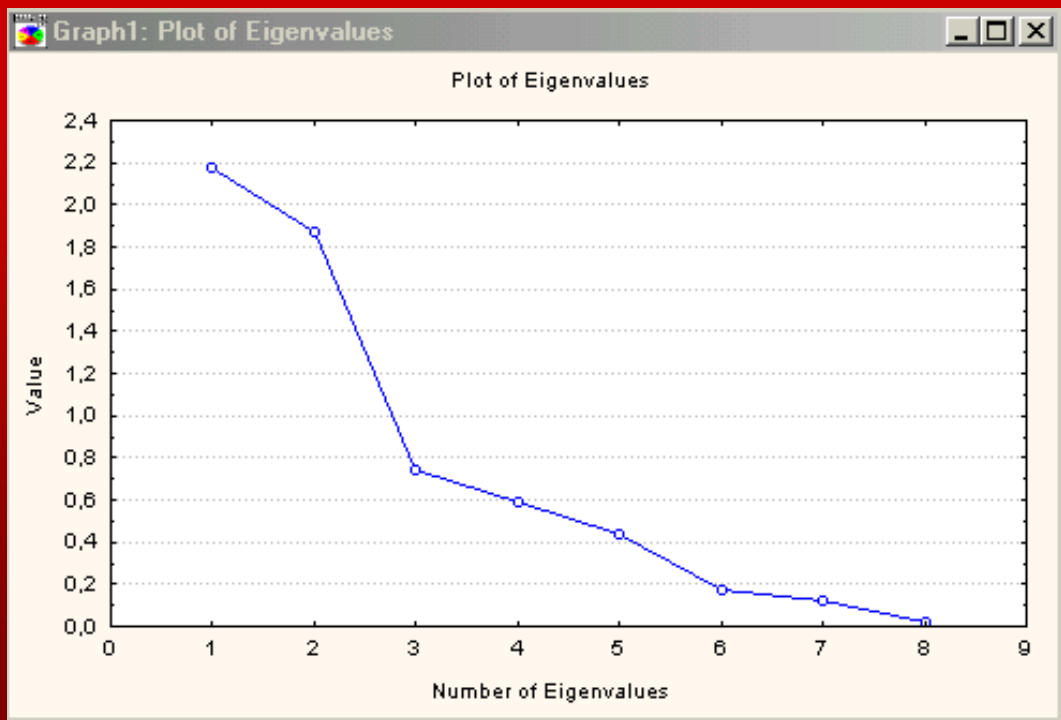




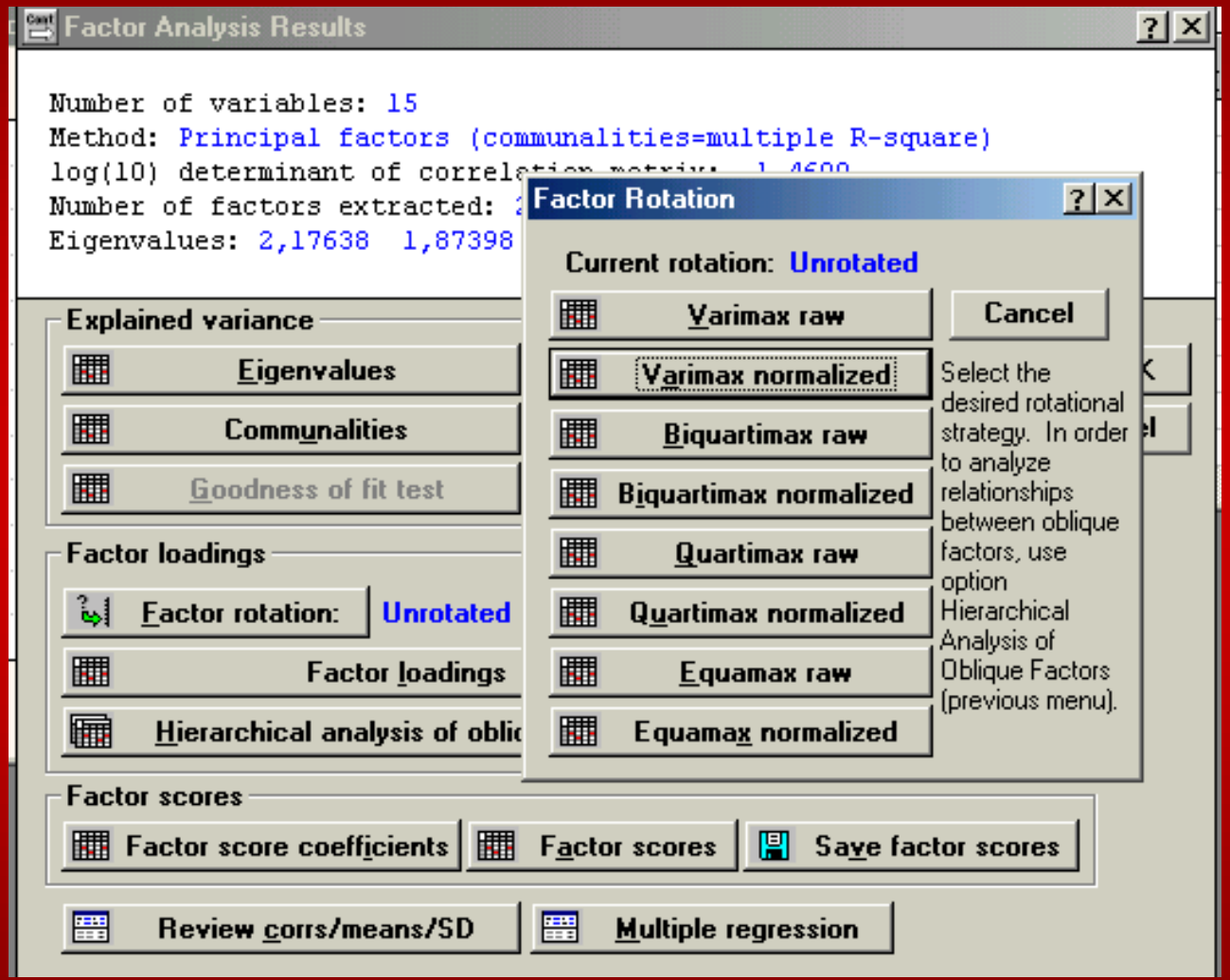


Факторный анализ

- Методы определения числа факторов
  - Критерий Кайзера (собственное значение фактора - Eigenvalue больше 1)
  - Критерий Жоффе (Eigenvalue>0.7)
  - Процент объясненной дисперсии (90% и более)
  - Scree plot

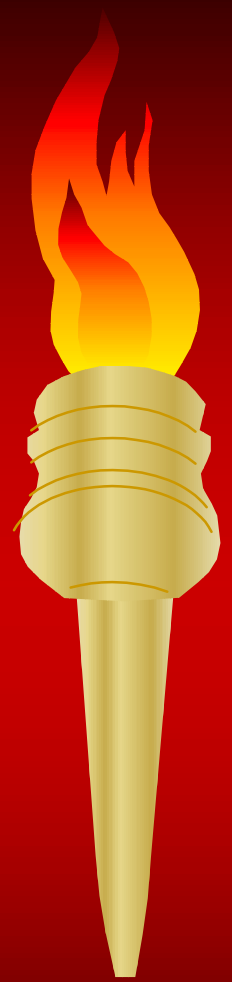


Факторный анализ



Вращение факторов - Varimax или Oblimin - необходимо для интерпретируемости результатов - простая структура

Факторный анализ



Factor Loadings (Varimax normalized) (eysenk.sta)

Continue...

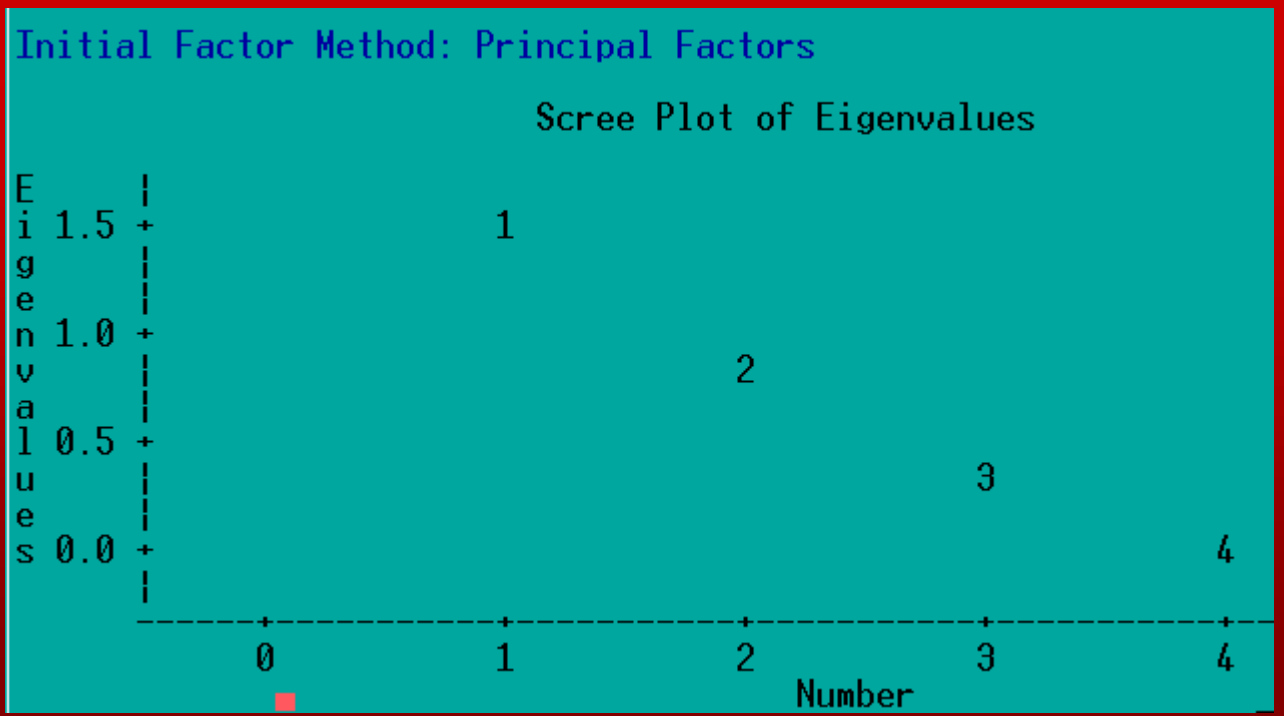
Extraction: Principal factors (comm.=multiple R-square)
(Marked loadings are > ,400000)

Variable	Factor 1	Factor 2
SEX	-,434748	,091442
PAR	,414157	-,033518
VNP	-,442744	,101523
DOZV	,437745	-,147297
UDOV	,266659	-,660534
KONF	,080965	,742981
BEZL	,557509	,391287
PORN	,308654	,161431
ZAST	-,249146	,612586
STID	-,184641	,081979
DOMI	,035320	-,543752
OTVR	-,492575	-,038430
VOZB	,668622	,055499
BIOL	,195436	-,062284
AGRE	,020808	,312492
Expl.Var	2,057647	1,992715
Prp.Totl	,137176	,132848

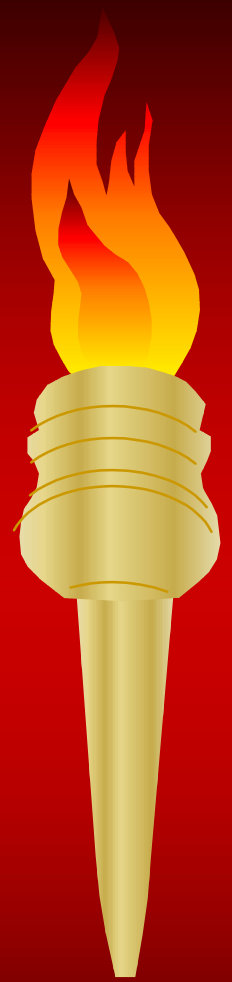
Факторный анализ (SAS)

```
PROGRAM EDITOR
Command ===>

00001 PROC FACTOR DATA=mydat.mu99 PRIORS=MAX SCREE ROTATE=V;
00002 VAR CH HDL TG SBP DBP;
00003 RUN;
```

Initial Factor Method: Principal Factors



Факторный анализ (SAS)



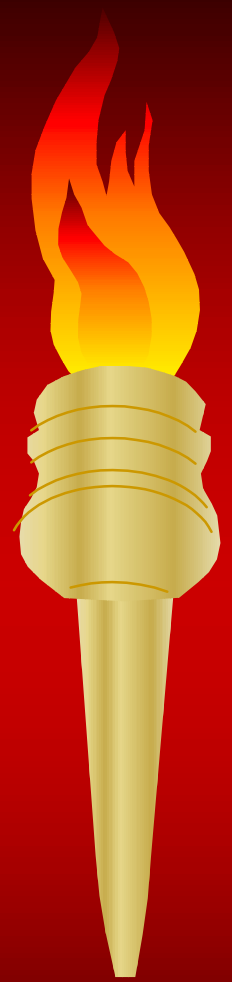
Rotation Method: Varimax

Rotated Factor Pattern

	FACTOR1	FACTOR2	FACTOR3	
CH	0.08055	0.63287	0.04089	V1 TOTAL CHOLESTEROL
HDL	0.06757	-0.04088	0.52926	V1 HDL CHOLESTEROL
TG	0.08077	0.53919	-0.36173	TRIG1 MINUS TRIGBLK1
SBP	0.85181	0.08656	0.11457	
DBP	0.85557	0.10789	0.00277	

Variance explained by each factor

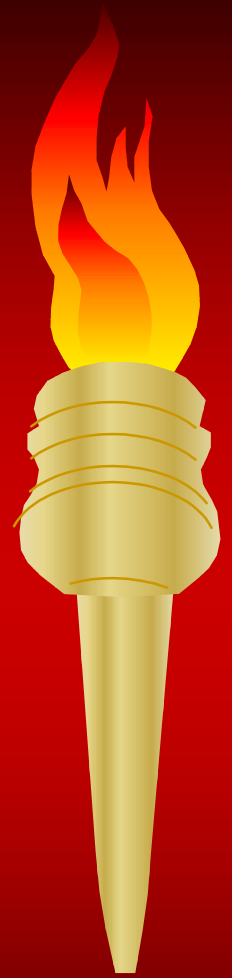
FACTOR1	FACTOR2	FACTOR3
1.475154	0.712047	0.425780

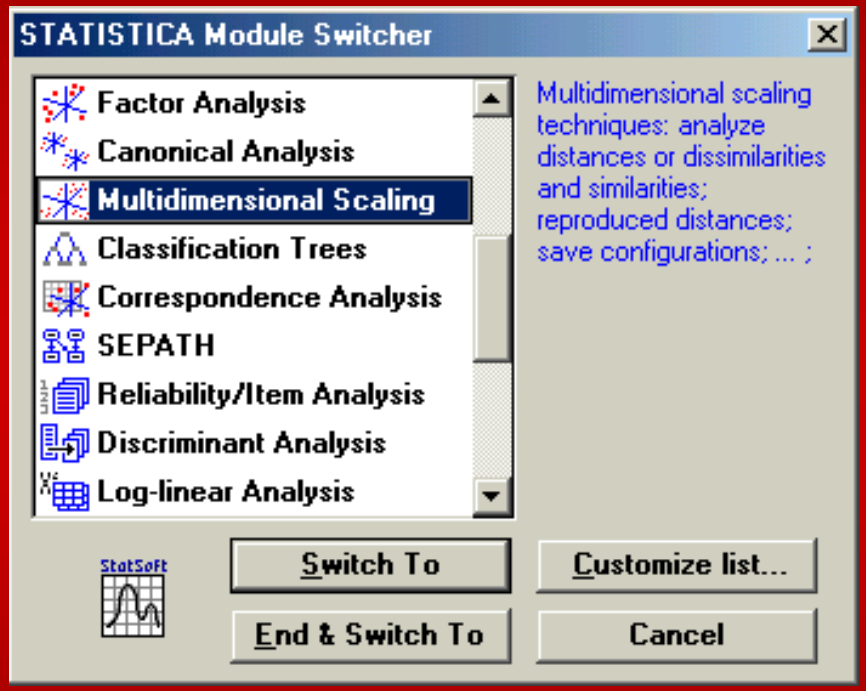


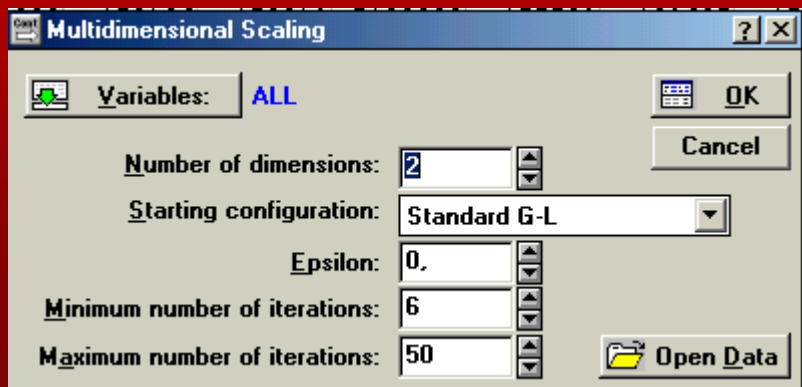
Многомерное шкалирование

- Метод анализа матрицы дистанций
  - Истинные дистанции
  - Коэффициенты корреляции (1-r)
  - Другие индексы
- Аналог географической карты

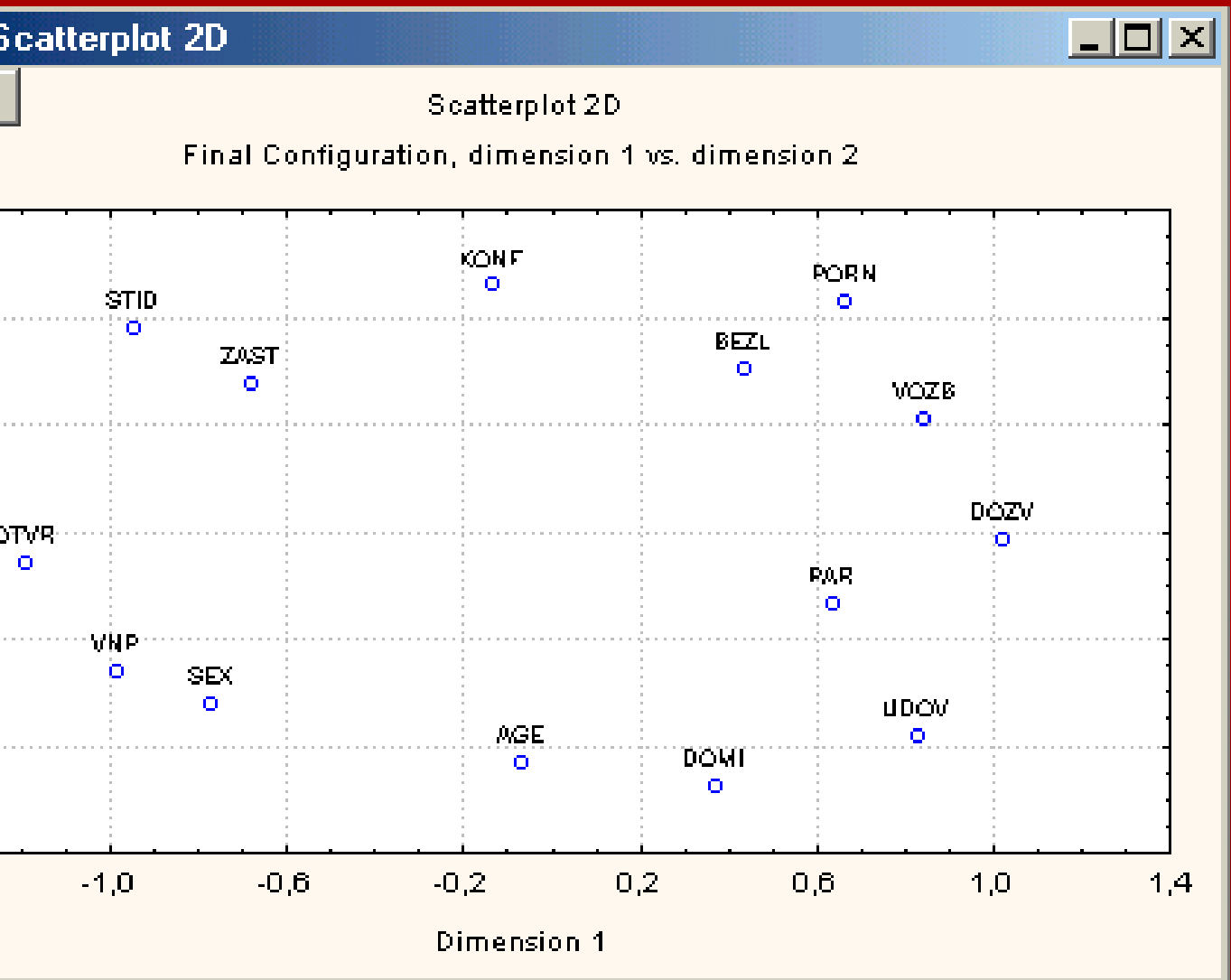
Многомерное шкалирование

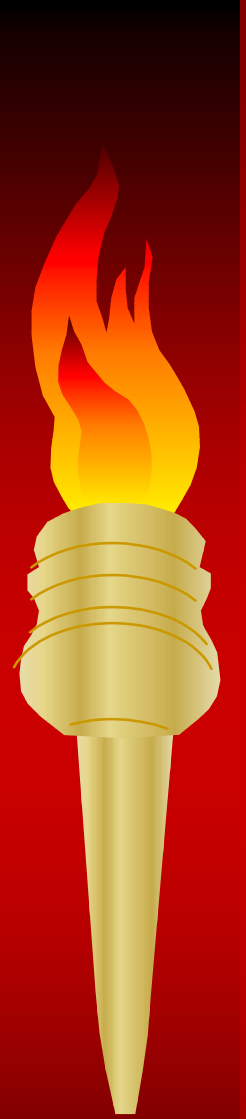






Многомерное шкалирование

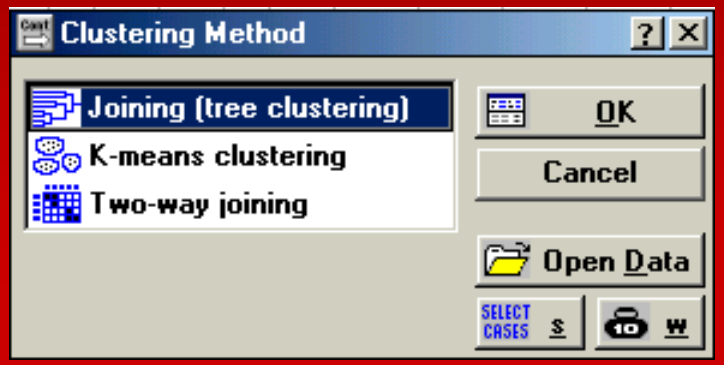


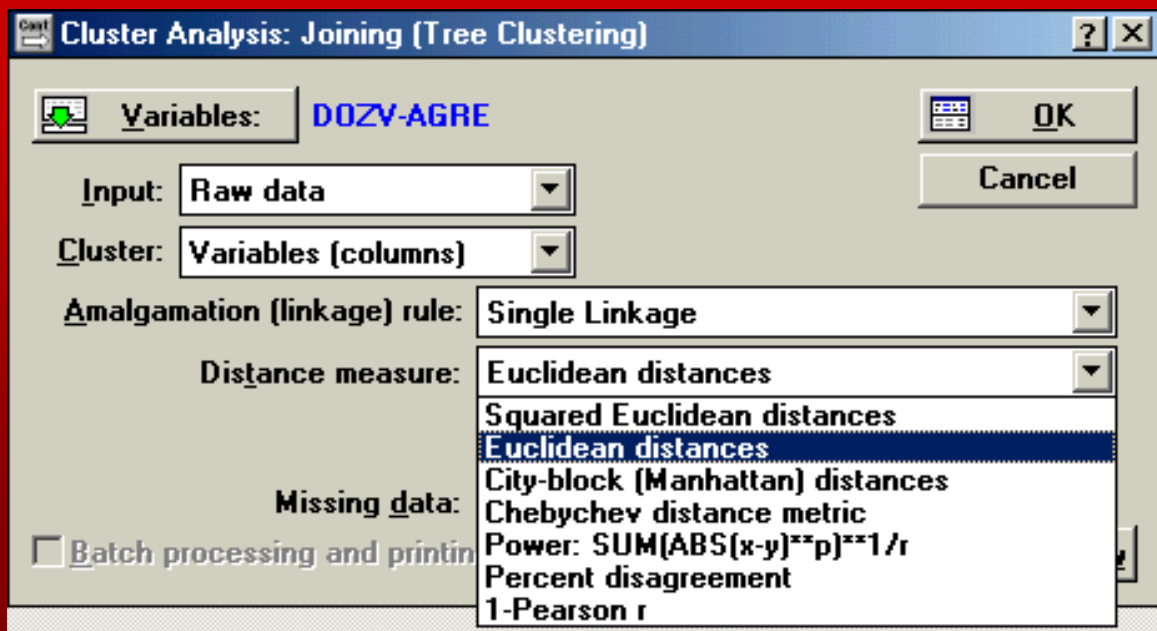


Кластерный анализ

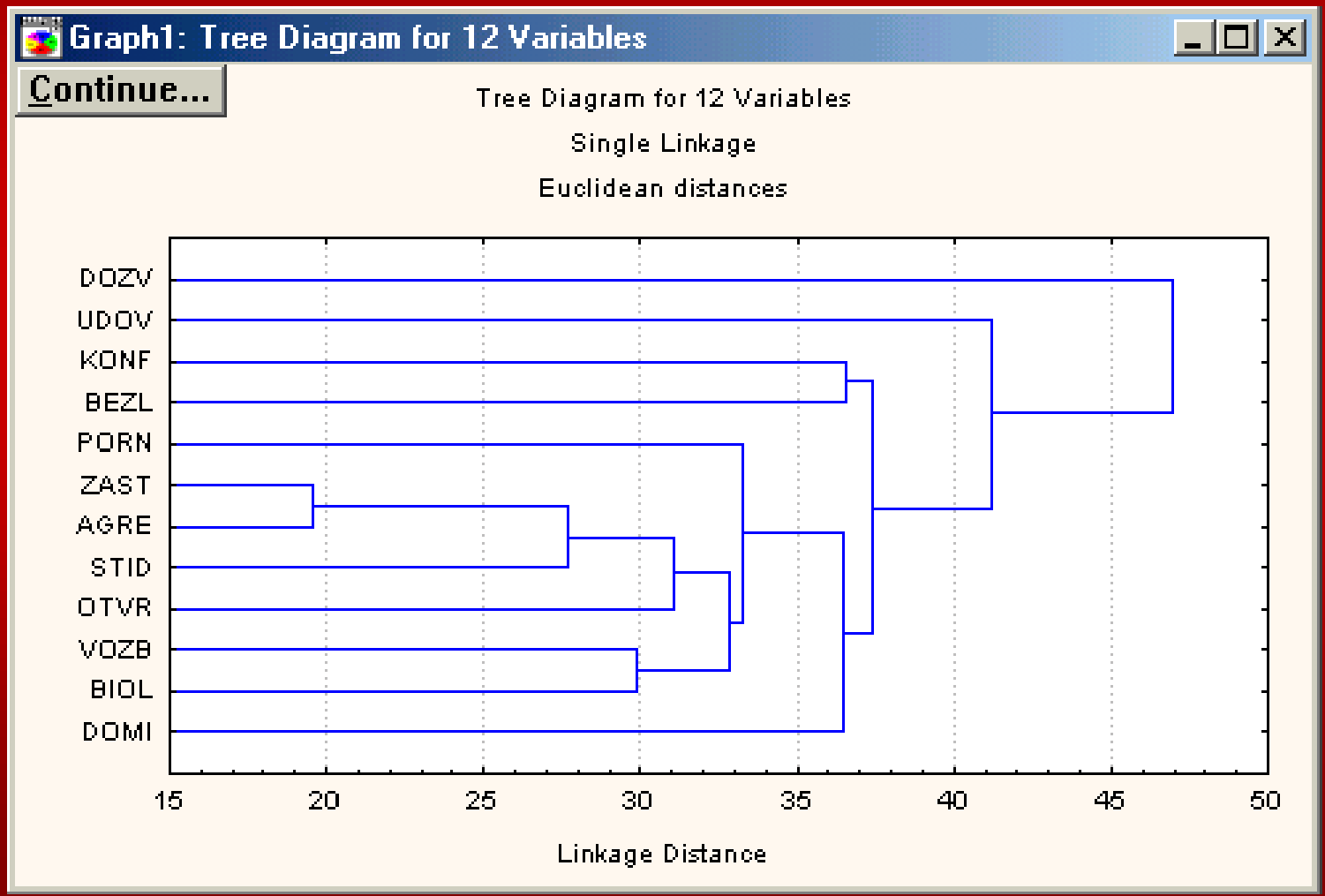
- Может использоваться для снижения размерности/классификации переменных (иерархический анализ)
- Может использоваться для объединения наблюдений на основании ряда показателей (групп - кластеров-пациентов) k-means

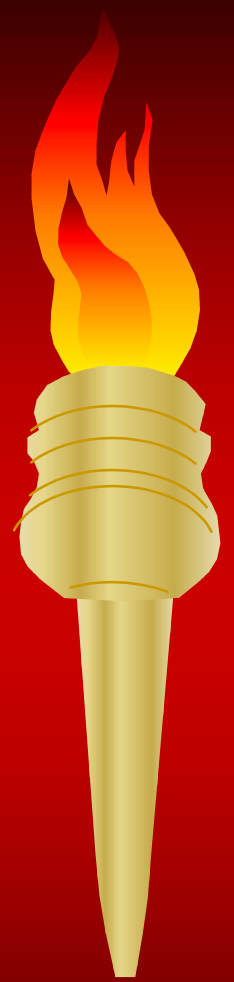
Кластерный анализ





Кластерный анализ

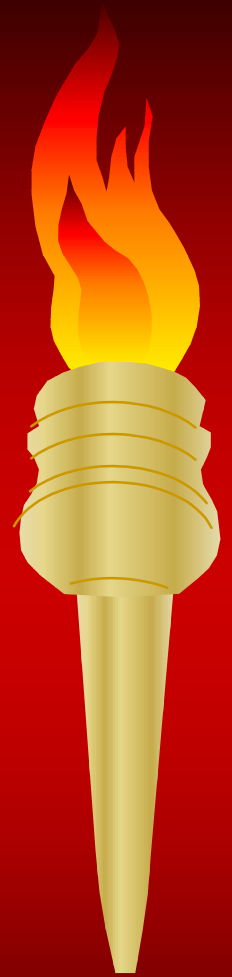


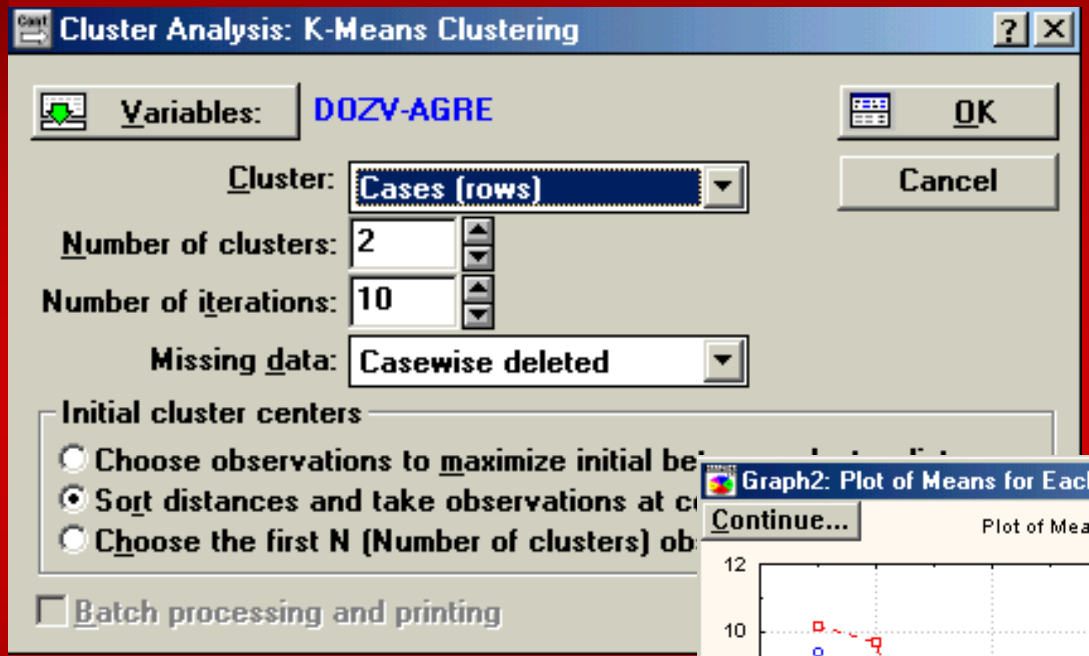


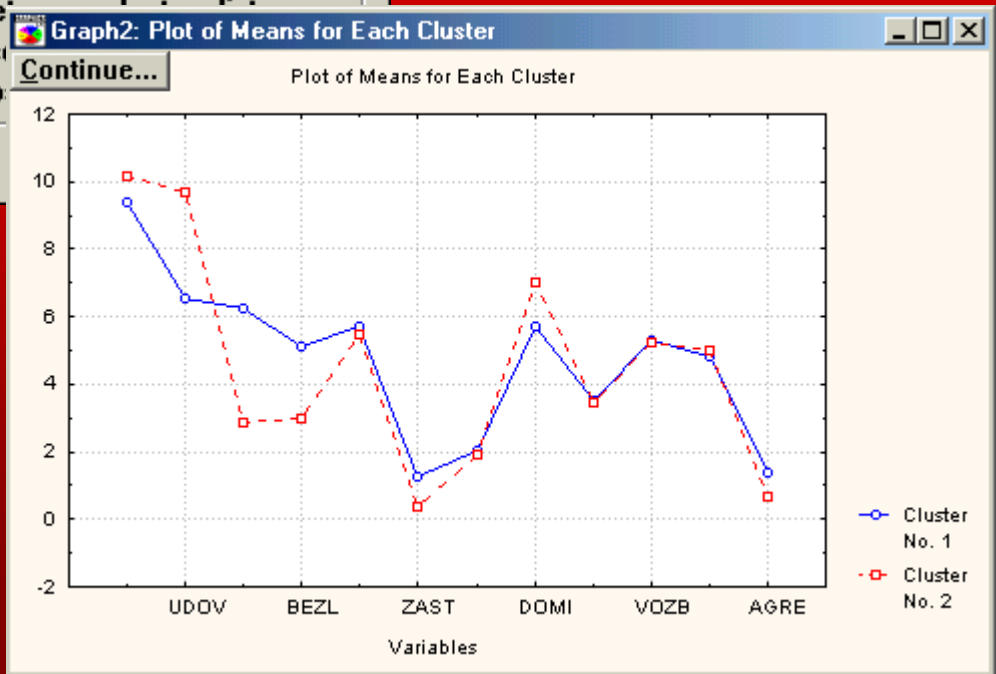
Кластерный анализ (k-means)

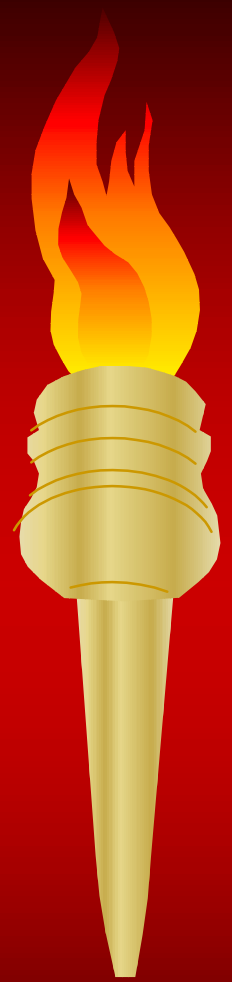
- Необходимо заранее решить, какое количество кластеров следует выделить
- Работает как ANOVA наоборот:
  - наблюдения объединяются таким образом, чтобы расстояния между группами (между средними групп) были максимальными
  - поэтому группы почти всегда будут статистически достоверно отличаться по включенным параметрам

Кластерный анализ (k-means)





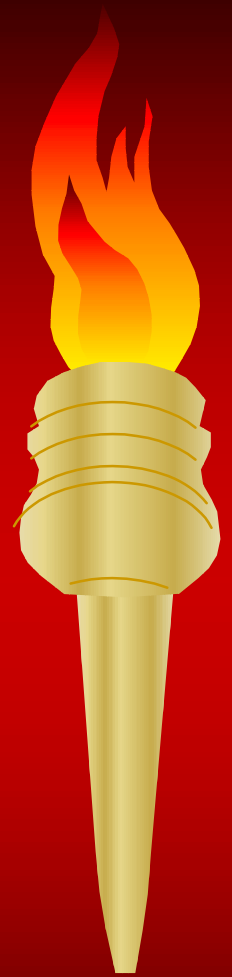


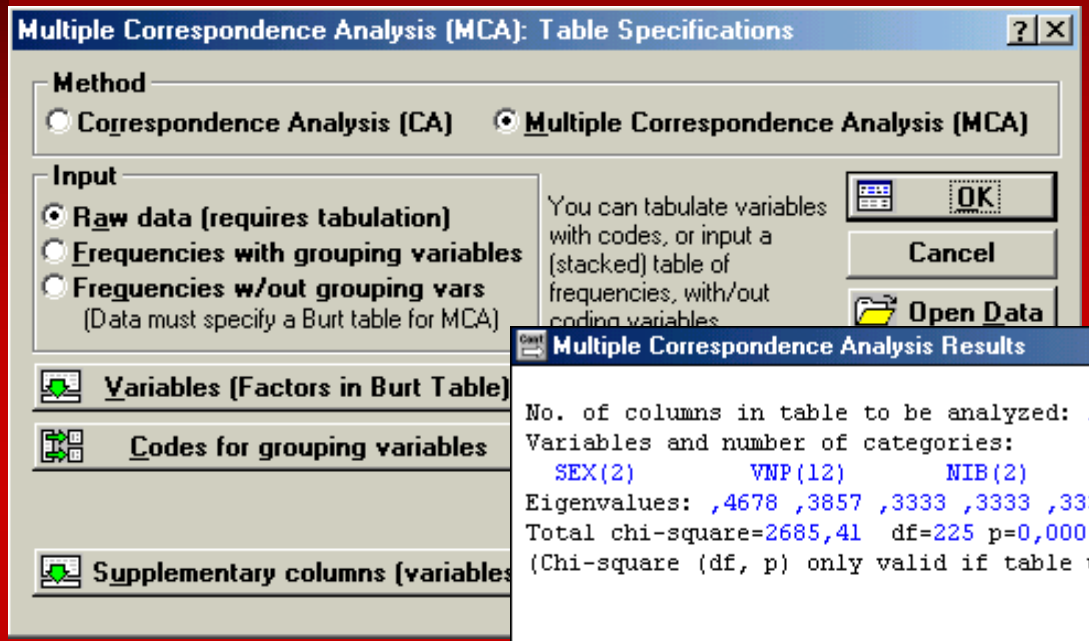


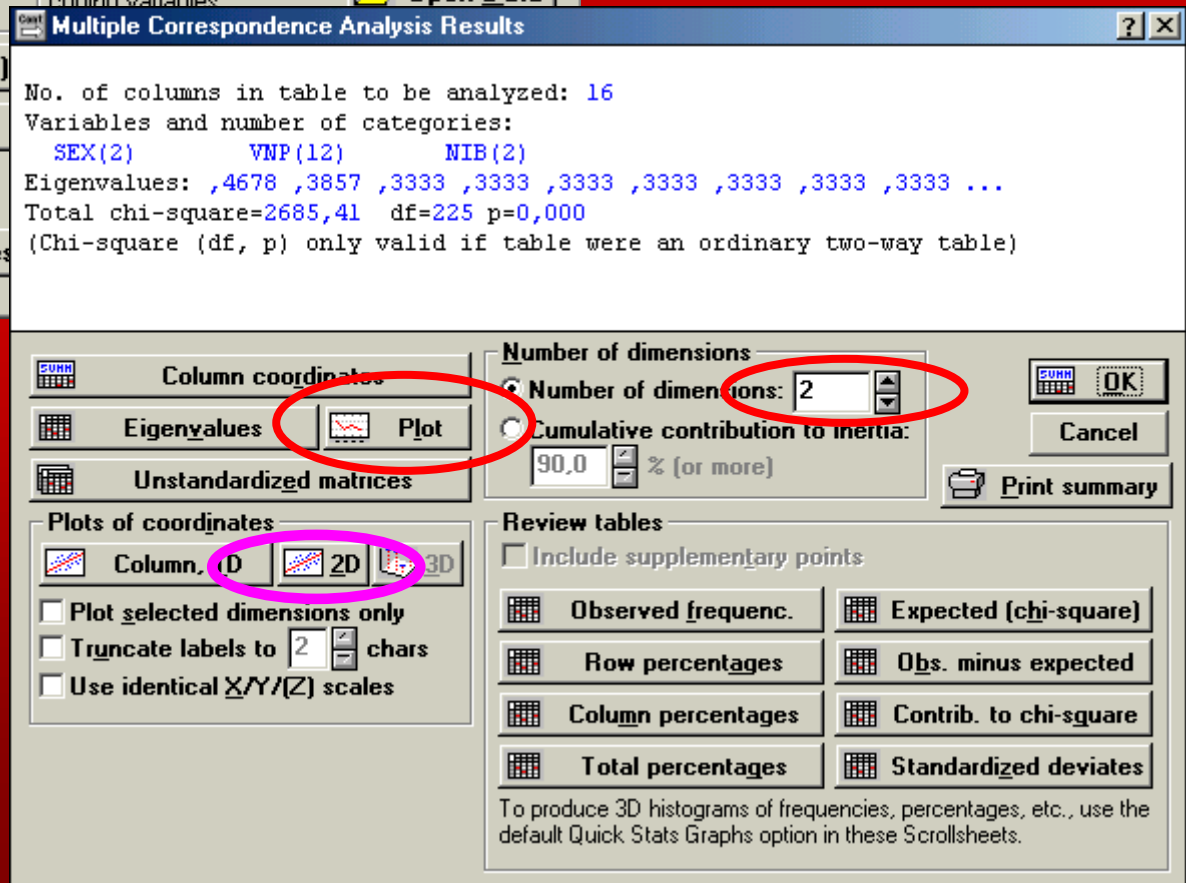
Корреспондентский анализ

- Аналог факторного анализа для качественных переменных
- Результаты также сходны с многомерным шкалированием
- Был разработан для лингвистического анализа

Корреспондентский анализ







Корреспондентский анализ