Multivariate Data Analysis Using Statgraphics Centurion: Part 3

Dr. Neil W. Polhemus
Statpoint Technologies, Inc.

Multivariate Statistical Methods

The simultaneous observation and analysis of more than one response variable.

*Primary Uses

1. Data reduction or structural simplification
2. Sorting and grouping
3. Investigation of the dependence among variables
4. Prediction
5. Hypothesis construction and testing

*Johnson and Wichern, Applied Multivariate Statistical Analysis

Methods

- Correspondence Analysis
  - Example: research funding proposals
- Multiple Correspondence Analysis
  - Example: survey questions
- Multivariate Analysis of Variance (MANOVA)
  - Example: designed experiment with 3 responses
- Partial Least Squares (PLS)
  - Example: stock portfolio

1. Correspondence Analysis

- Similar to principal component analysis, except that it applies to categorical data.
- Traditionally applied to contingency tables.
- Creates a map in a low dimensional space that provides insights into the relationships between row and column categories.

Sample Data (funding.sgd)

C:\Webinar\funding.sgd

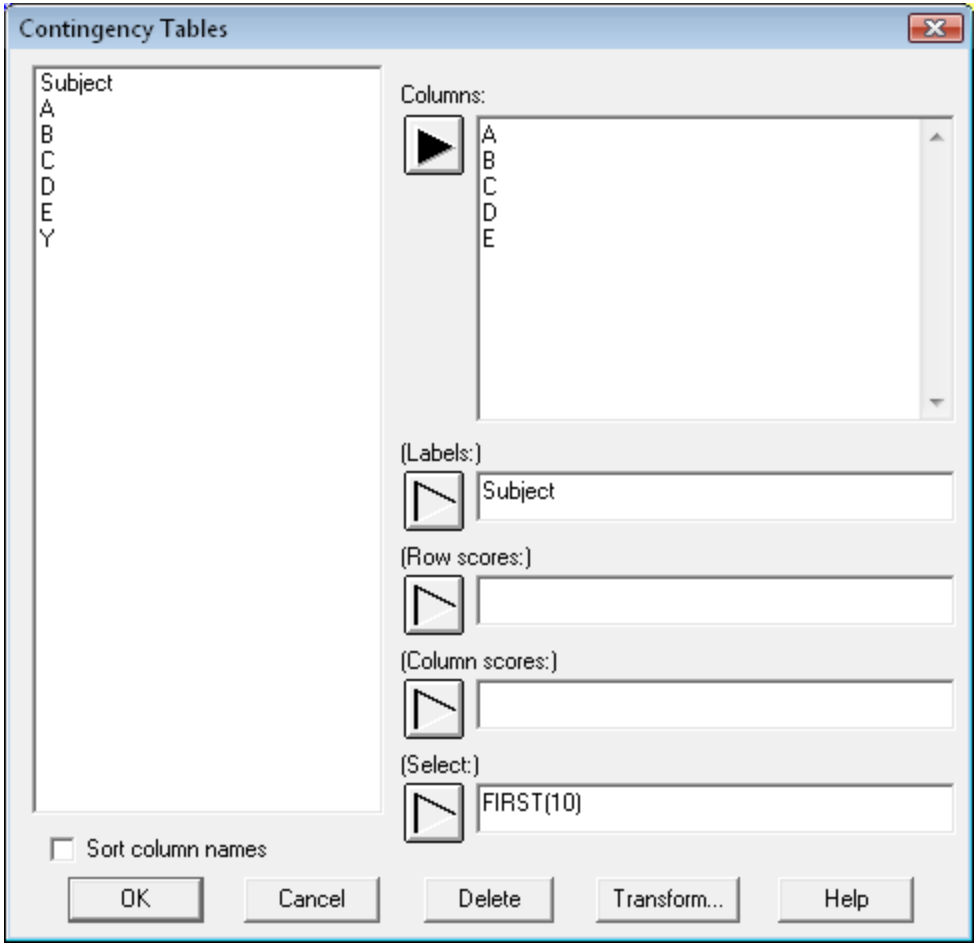
	Subject	A	B	C	D	E	Y
1	Geology	3	19	39	14	10	0
2	Biochemistry	1	2	13	1	12	1
3	Chemistry	6	25	49	21	29	0
4	Zoology	3	15	41	35	26	0
5	Physics	10	22	47	9	26	1
6	Engineering	3	11	25	15	34	1
7	Microbiology	1	6	14	5	11	1
8	Botany	0	12	34	17	23	1
9	Statistics	2	5	11	4	7	0
10	Mathematics	2	11	37	8	20	1
11	Museums	4	12	11	19	7	
12	Math Sciences	4	16	48	12	27	

funding | B | C

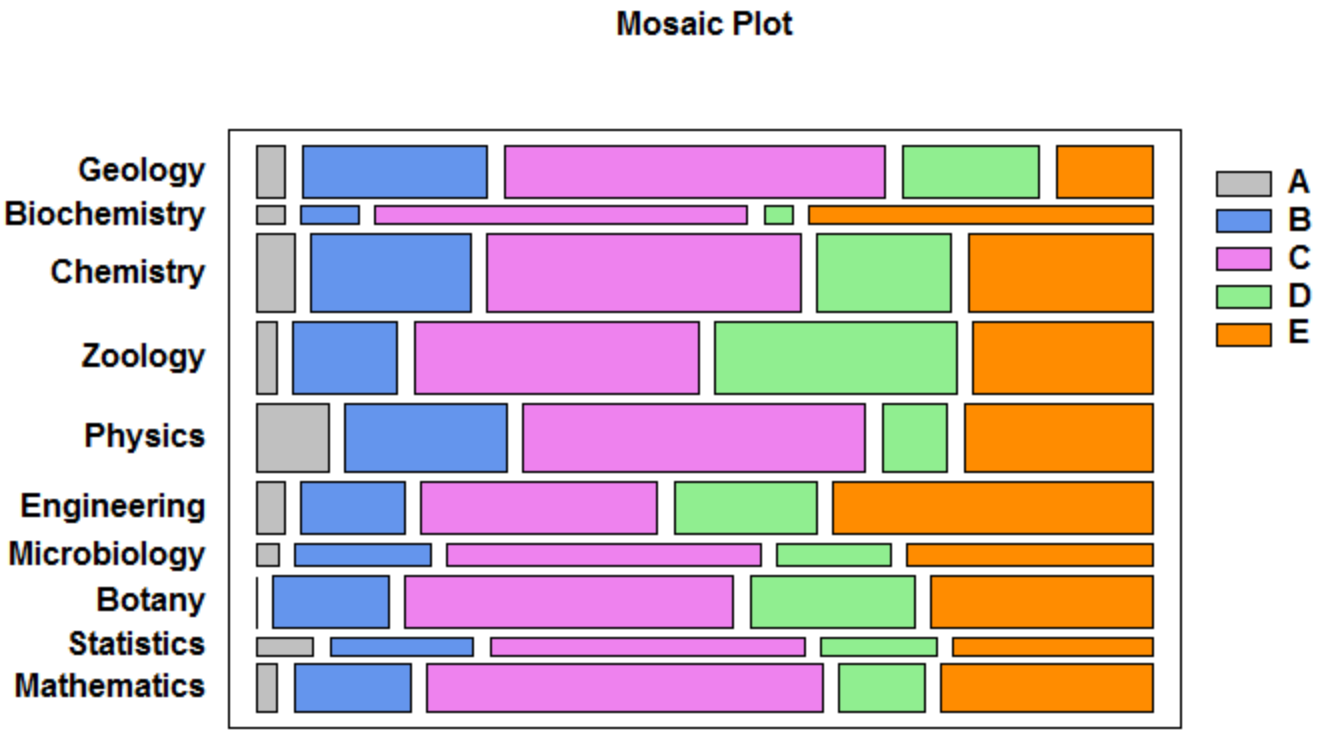
796 research proposals submitted to a research agency.
A-E = funding level (A=most funded; D=least funded; E=not funded).

Source: Correspondence Analysis in Practice by Michael Greenacre.

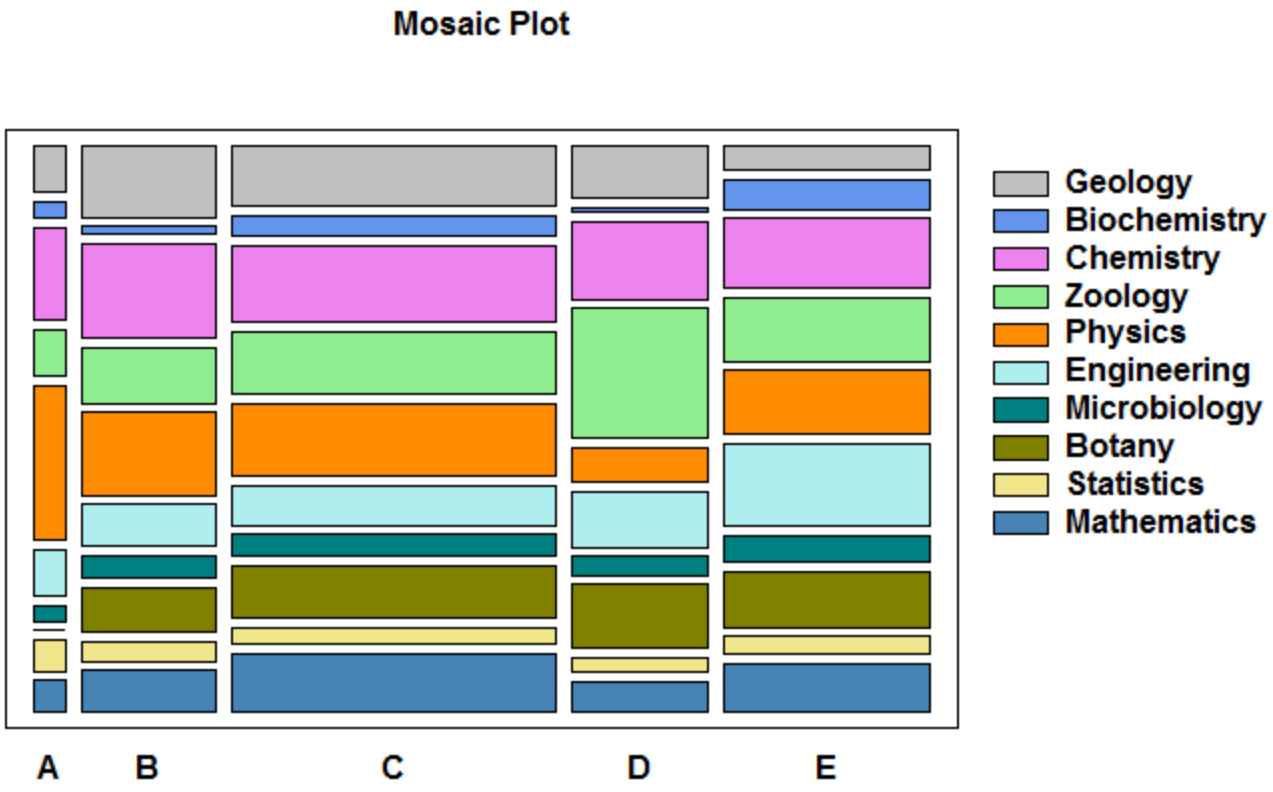
Contingency Table



Mosaic Plot – Row Profiles



Mosaic Plot – Column Profiles



Chi-Square Test of Independence

Tests of Independence

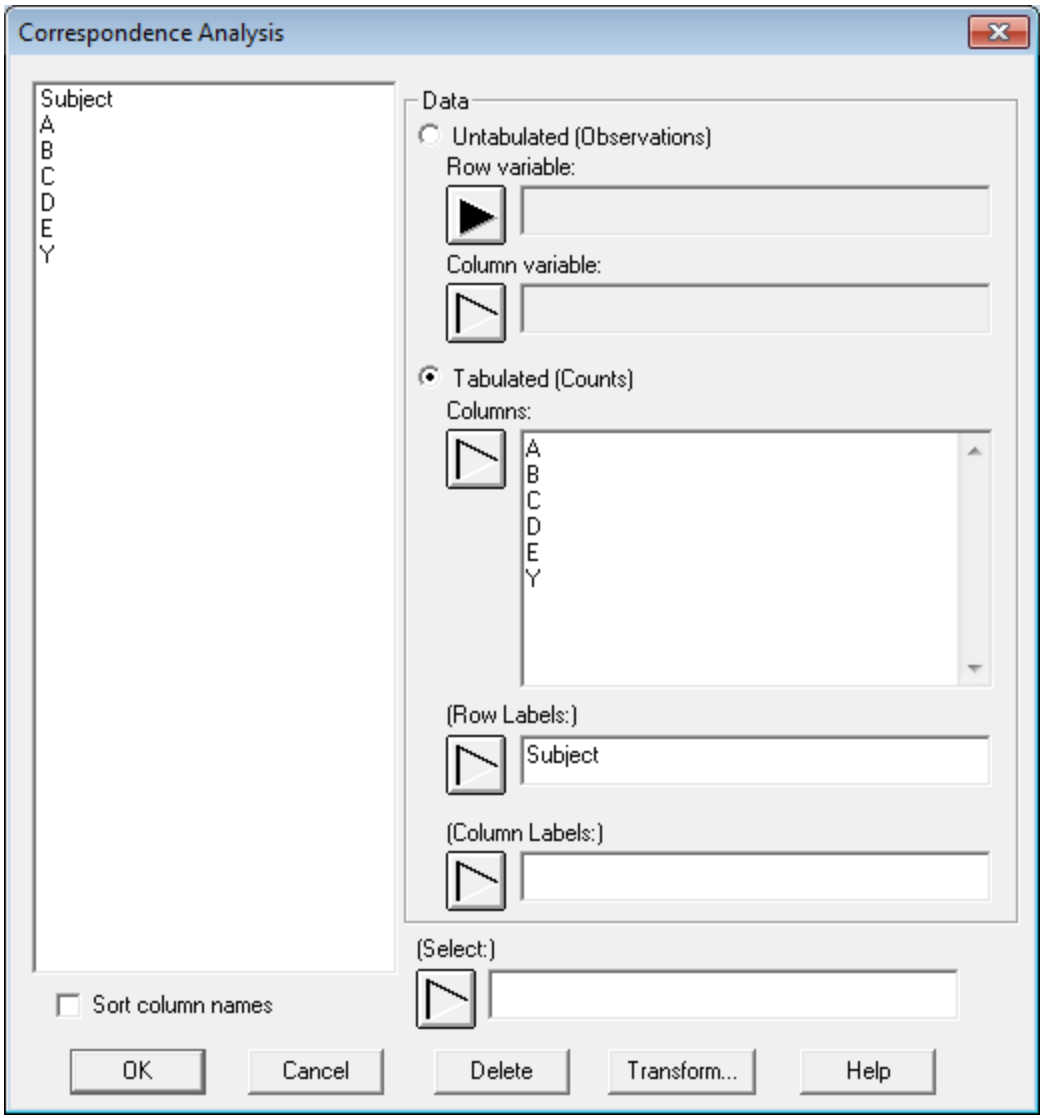
Test	*Statistic*	*Df*	*P-Value*
Chi-Square	65.972	36	0.0017

Warning: some expected cell counts < 5.

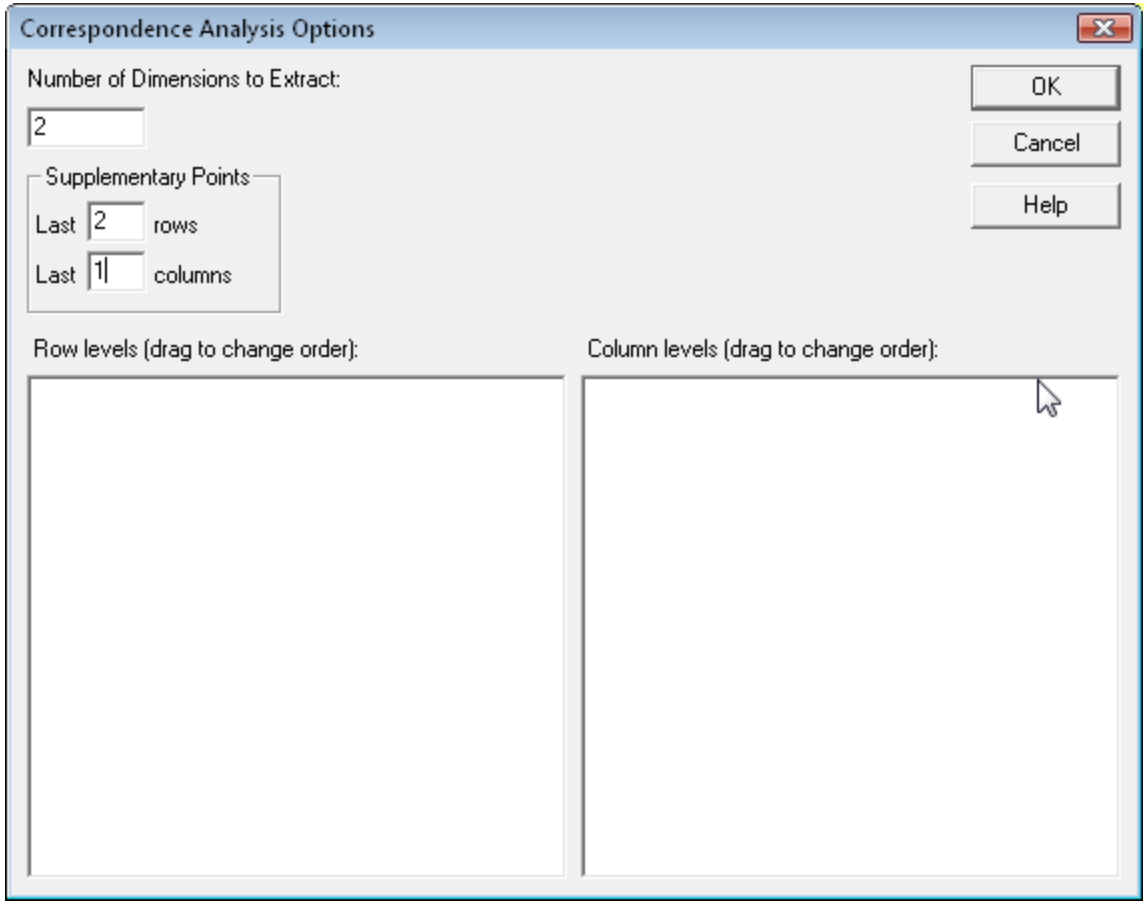
$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

A small P-value (large X^2) indicates a dependency between the row and column classifications.

Correspondence Analysis – Data Input



Analysis Options



Contingency Table

Correspondence Analysis

Contingency Table

	A	B	C	D	E	TOTAL
Geology	3	19	39	14	10	85
Biochemistry	1	2	13	1	12	29
Chemistry	6	25	49	21	29	130
Zoology	3	15	41	35	26	120
Physics	10	22	47	9	26	114
Engineering	3	11	25	15	34	88
Microbiology	1	6	14	5	11	37
Botany	0	12	34	17	23	86
Statistics	2	5	11	4	7	29
Mathematics	2	11	37	8	20	78
TOTAL	31	128	310	129	198	796

Row and Column Profiles

Row and Column Profiles

Row Profiles

	A	B	C	D	E	MASS
Geology	0.035	0.224	0.459	0.165	0.118	0.107
Biochemistry	0.034	0.069	0.448	0.034	0.414	0.036
Chemistry	0.046	0.192	0.377	0.162	0.223	0.163
Zoology	0.025	0.125	0.342	0.292	0.217	0.151
Physics	0.088	0.193	0.412	0.079	0.228	0.143
Engineering	0.034	0.125	0.284	0.170	0.386	0.111
Microbiology	0.027	0.162	0.378	0.135	0.297	0.046
Botany	0.000	0.140	0.395	0.198	0.267	0.108
Statistics	0.069	0.172	0.379	0.138	0.241	0.036
Mathematics	0.026	0.141	0.474	0.103	0.256	0.098
MASS	0.039	0.161	0.389	0.162	0.249	

Column Profiles

	A	B	C	D	E	MASS
Geology	0.097	0.148	0.126	0.109	0.051	0.107
Biochemistry	0.032	0.016	0.042	0.008	0.061	0.036
Chemistry	0.194	0.195	0.158	0.163	0.146	0.163
Zoology	0.097	0.117	0.132	0.271	0.131	0.151
Physics	0.323	0.172	0.152	0.070	0.131	0.143
Engineering	0.097	0.086	0.081	0.116	0.172	0.111
Microbiology	0.032	0.047	0.045	0.039	0.056	0.046
Botany	0.000	0.094	0.110	0.132	0.116	0.108
Statistics	0.065	0.039	0.035	0.031	0.035	0.036
Mathematics	0.065	0.086	0.119	0.062	0.101	0.098
MASS	0.039	0.161	0.389	0.162	0.249	

Inertia and Chi-Squared Decomposition

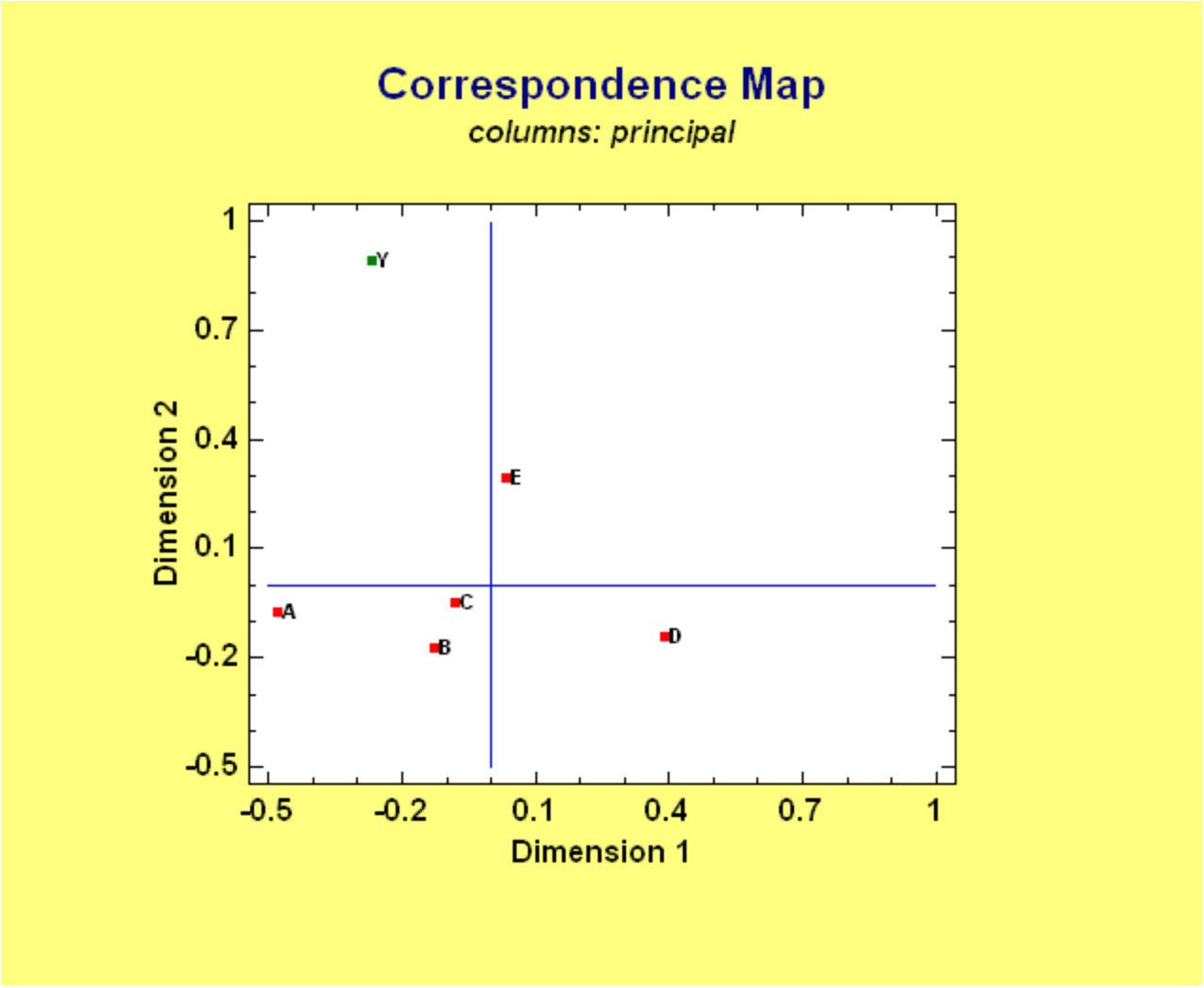
Inertia and Chi-Square Decomposition

	Singular		*Chi-*		*Cumulative*	
Dimension	*Value*	*Inertia*	*Square*	*Percentage*	*Percentage*	*Histogram*
1	0.1978	0.0391	31.1368	47.1973	47.1973	**********
2	0.1743	0.0304	24.1831	36.6569	83.8542	**********
3	0.1043	0.0109	8.6519	13.1146	96.9688	****
4	0.0501	0.0025	1.9997	3.0312	100.0000	*
TOTAL		0.0829	65.971			

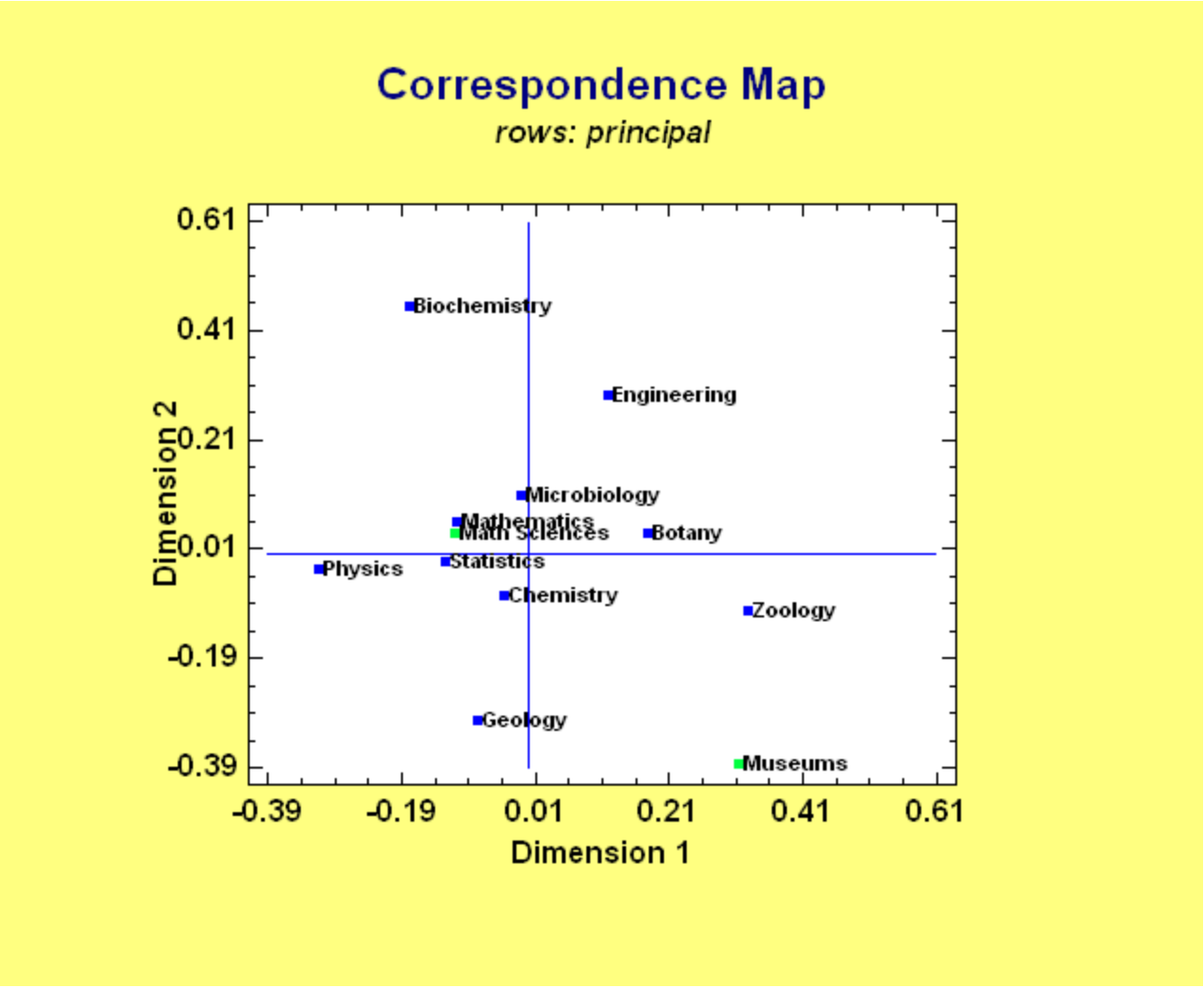
Chi-Square column shows the contribution of each dimension to the chi-square statistic.

Inertia measures the amount of variability along a principal dimension (equals chi-square divided by sample size).

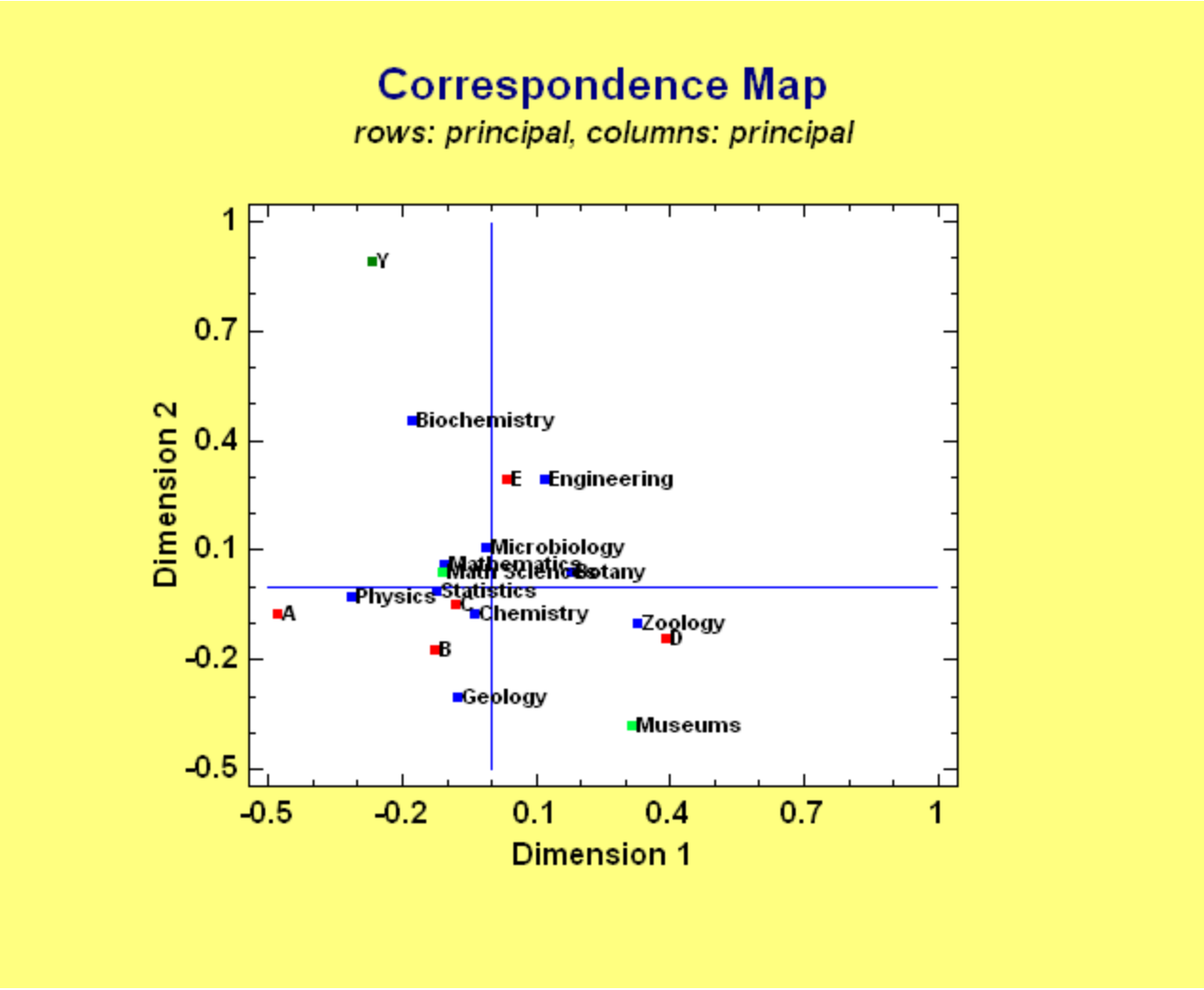
Correspondence Map - Columns



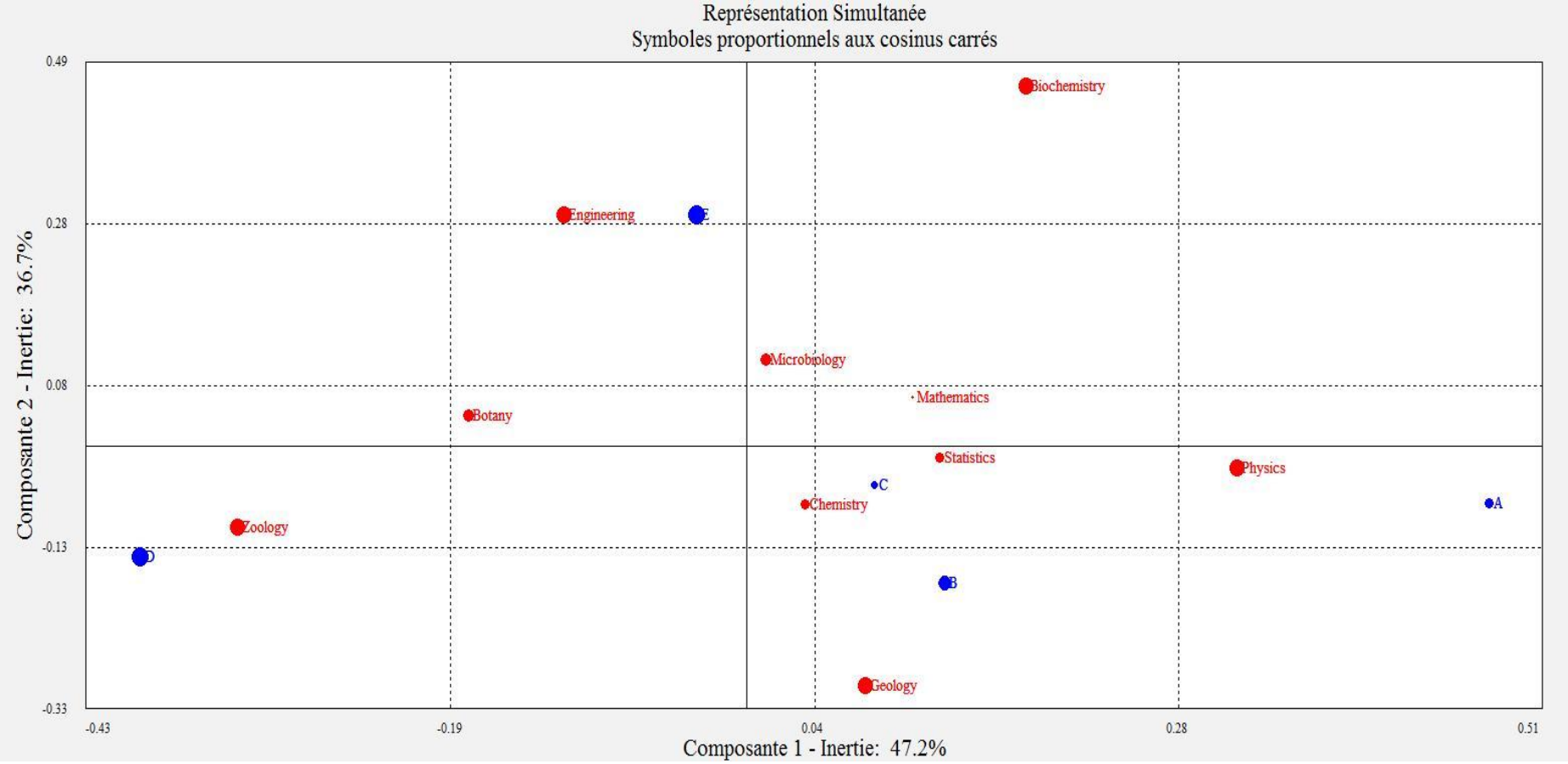
Correspondence Map - Rows



Correspondence Map - Both



Uniwin Plus – Additional Output



Size of point symbols scaled by the squared cosines, which are related to the quality of the projection.

2. Multiple Correspondence Analysis

- Deals with the associations within one set of variables
- The goal is to understand how strongly and in what way the variables are related

Sample Data (survey.sgd)

C:\DocData16Hold\survey.sgd

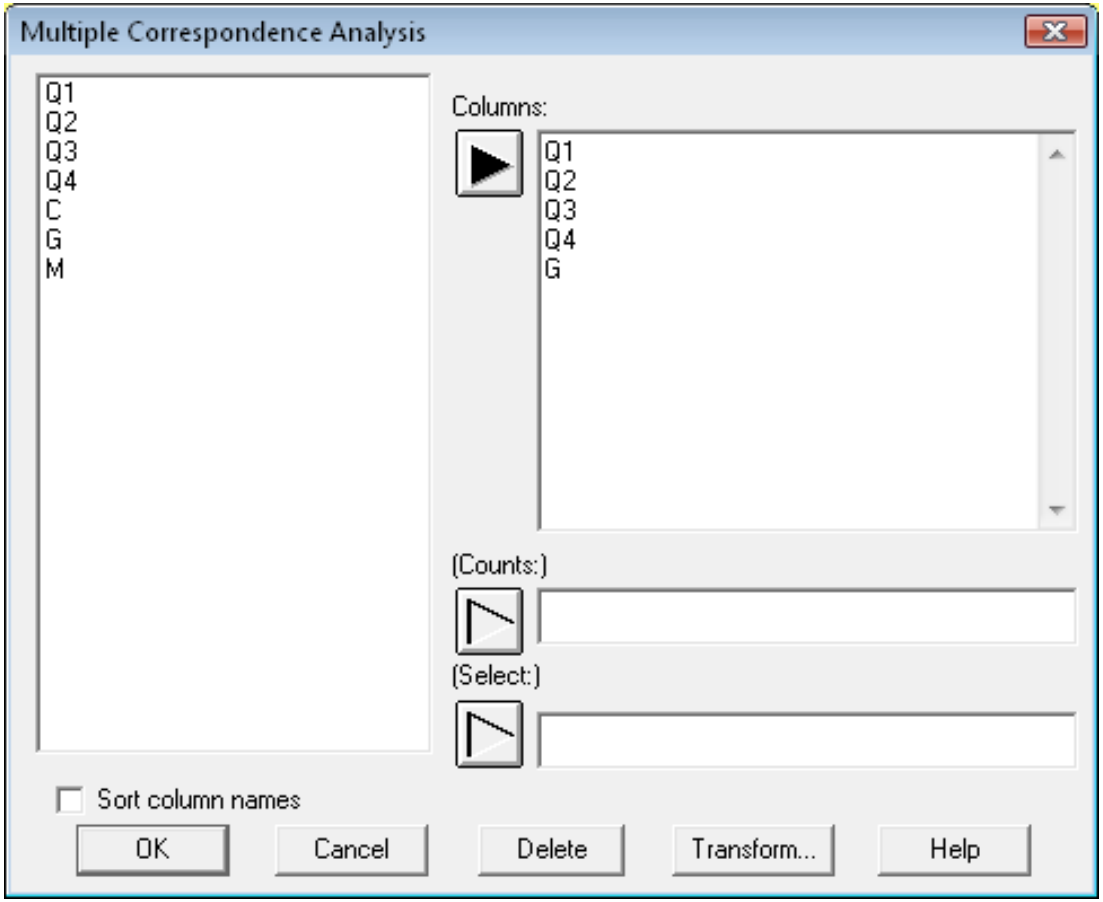
	Q1	Q2	Q3	Q4	C	G	M
					country	gender	marital status
1	W	H	w	w	DW	M	ma
2	w	H	H	w	DW	M	ma
3	?	H	H	w	DW	M	ma
4	?	?	?	?	DW	F	si
5	?	?	?	?	DW	F	si
6	W	H	w	W	DW	M	ma
7	?	H	H	?	DW	M	ma
8	?	?	w	?	DW	F	si
9	W	H	H	w	DW	F	ma
10	?	H	H	?	DW	M	ma

survey B C

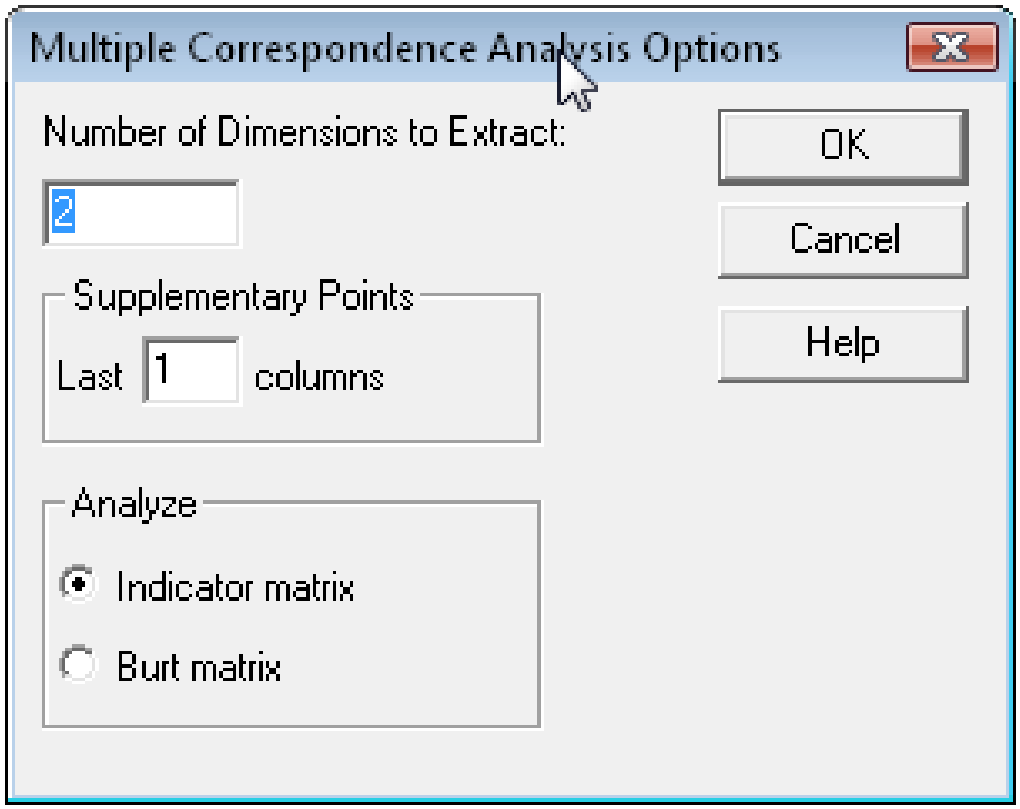
3,418 respondents answered 4 questions regarding women in the workplace.
4 possible responses to each question, coded as: W, w, H, or ?

Source: Correspondence Analysis in Practice by Michael Greenacre.

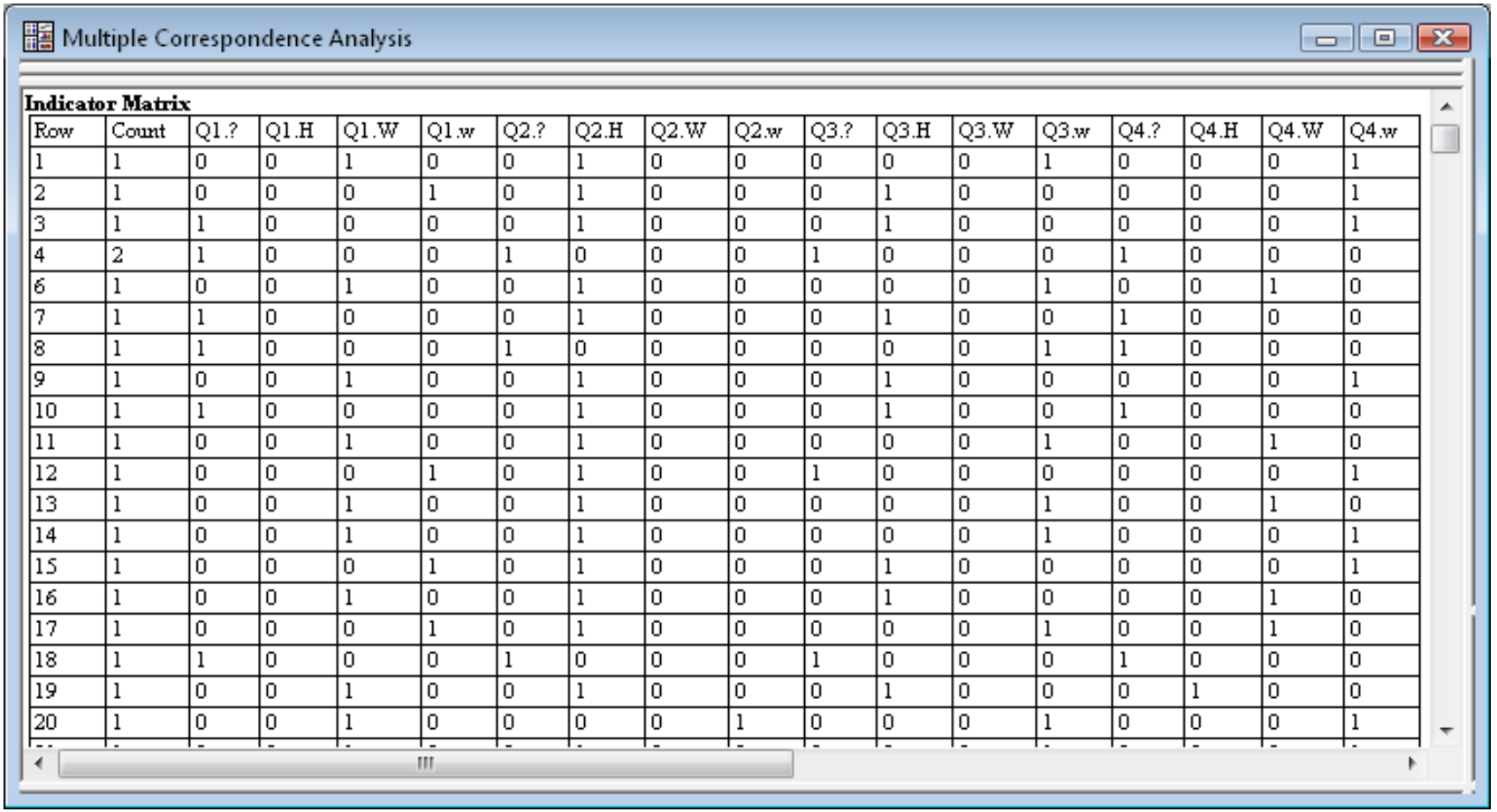
Data Input


Multiple Correspondence Analysis
Q1
Q2
Q3
Q4
C
G
M
Columns:
Q1
Q2
Q3
Q4
G
(Counts:)
(Select:)
Sort column names
OK
Cancel
Delete
Transform...
Help

Analysis Options



Indicator Matrix



Indicator Matrix

Row	Count	Q1.?	Q1.H	Q1.W	Q1.w	Q2.?	Q2.H	Q2.W	Q2.w	Q3.?	Q3.H	Q3.W	Q3.w	Q4.?	Q4.H	Q4.W	Q4.w
1	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1
2	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1
3	1	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1
4	2	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
6	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0
7	1	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0
8	1	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
9	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	0	1
10	1	1	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0
11	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0
12	1	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	1
13	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0
14	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1
15	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	0	1
16	1	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0
17	1	0	0	0	1	0	1	0	0	0	0	0	1	0	0	1	0
18	1	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0
19	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0
20	1	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	1

Burt Table

Burt Table

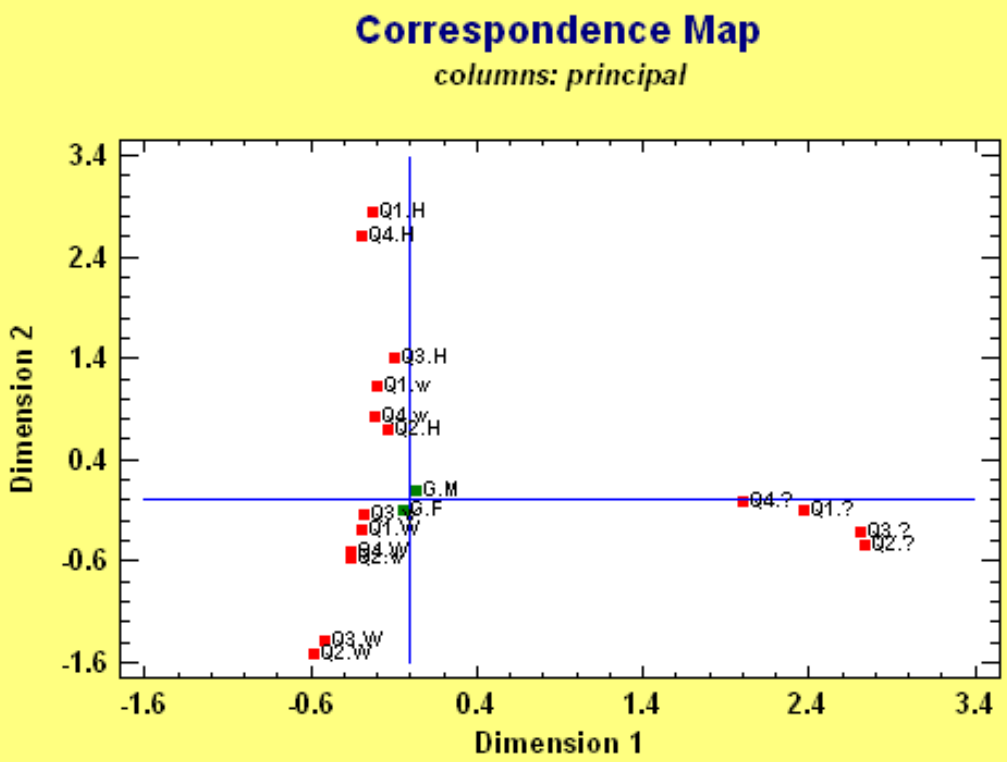
	Q1.?	Q1.H	Q1.W	Q1.w	Q2.?	Q2.H	Q2.W	Q2.w	Q3.?	Q3.H	Q3.W	Q3.w	Q4.?	Q4.H	Q4.W	Q4.w
Q1.?	362	0	0	0	196	108	1	57	204	55	7	96	264	2	51	45
Q1.H	0	79	0	0	0	72	1	6	0	61	1	17	6	38	14	21
Q1.W	0	0	2501	0	91	1131	172	1107	91	345	355	1710	157	40	1766	538
Q1.w	0	0	0	476	5	335	7	129	18	181	16	261	38	17	128	293
Q2.?	196	0	91	5	292	0	0	0	229	4	9	50	203	0	62	27
Q2.H	108	72	1131	335	0	1646	0	0	60	573	24	989	186	84	760	616
Q2.W	1	1	172	7	0	0	181	0	2	4	127	48	1	0	165	15
Q2.w	57	6	1107	129	0	0	0	1299	22	61	219	997	75	13	972	239
Q3.?	204	0	91	18	229	60	2	22	313	0	0	0	234	0	49	30
Q3.H	55	61	345	181	4	573	4	61	0	642	0	0	81	73	202	286
Q3.W	7	1	355	16	9	24	127	219	0	0	379	0	4	1	360	14
Q3.w	96	17	1710	261	50	989	48	997	0	0	0	2084	146	23	1348	567
Q4.?	264	6	157	38	203	186	1	75	234	81	4	146	465	0	0	0
Q4.H	2	38	40	17	0	84	0	13	0	73	1	23	0	97	0	0
Q4.W	51	14	1766	128	62	760	165	972	49	202	360	1348	0	0	1959	0
Q4.w	45	21	538	293	27	616	15	239	30	286	14	567	0	0	0	897

Inertia and Chi-Square Decomposition

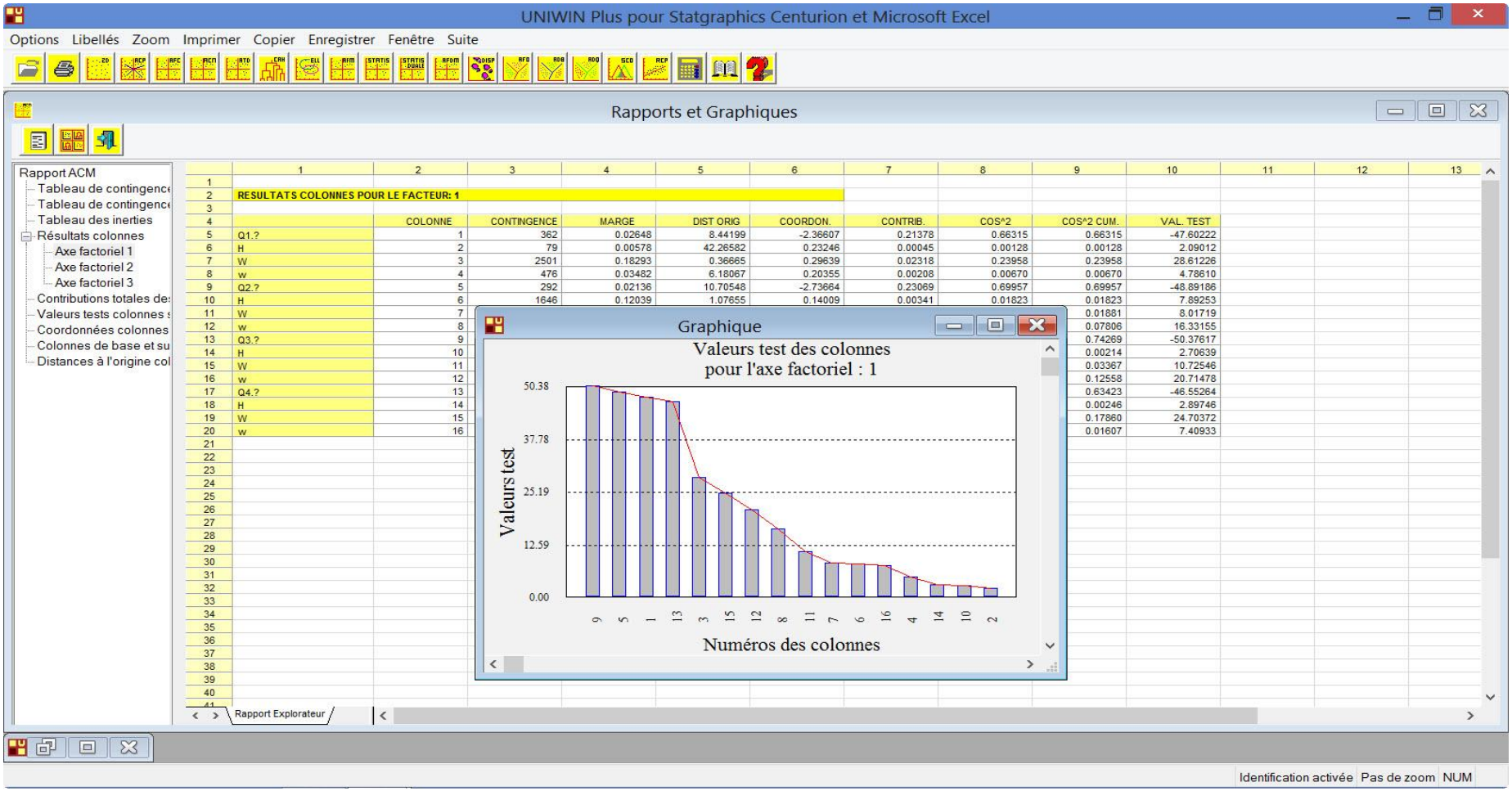
Inertia and Chi-Square Decomposition

	Singular		*Chi-*		*Cumulative*	
Dimension	*Value*	*Inertia*	*Square*	*Percentage*	*Percentage*	*Histogram*
1	0.8327	0.6934	9479.6304	23.1120	23.1120	********
2	0.7164	0.5132	7016.5099	17.1068	40.2188	******
3	0.6039	0.3647	4986.1386	12.1566	52.3754	****
4	0.5544	0.3074	4202.8549	10.2469	62.6222	****
5	0.4665	0.2176	2975.1574	7.2537	69.8759	***
6	0.4261	0.1815	2481.7591	6.0507	75.9266	**
7	0.4059	0.1648	2252.7962	5.4925	81.4191	**
8	0.3782	0.1430	1955.0782	4.7666	86.1857	**
9	0.3692	0.1363	1863.7876	4.5441	90.7297	**
10	0.3371	0.1137	1553.9048	3.7885	94.5183	**
11	0.3170	0.1005	1373.7982	3.3494	97.8677	**
12	0.2529	0.0640	874.5848	2.1323	100.0000	*
TOTAL		3.0000	41016.000			

Correspondence Map



Uniwin Plus – Additional Output



Test values greater than 3 indicate columns that are important to the analysis.

3. Multivariate Analysis of Variance (MANOVA)

- Extends univariate ANOVA to the case of multiple dependent variables.

- Tests hypotheses concerning a vector of means:

$$\mu = \begin{pmatrix} \mu_{Y_1} \\ \mu_{Y_2} \\ \dots \\ \mu_{Yp} \end{pmatrix}$$

Example: Designed Experiment with Multiple Response Variables

C:\Program Files\Statgraphics\STATGRAPHICS Centurion XVI.II\Data\film.sgd

	Rate of extrusion	Amount of additive	Tear resistance	Gloss	Opacity
1	-10	1	6.5	9.5	4.4
2	-10	1	6.2	9.9	6.4
3	-10	1	5.8	9.6	3
4	-10	1	6.5	9.6	4.1
5	-10	1	6.5	9.2	0.8
6	-10	1.5	6.9	9.1	5.7
7	-10	1.5	7.2	10	2
8	-10	1.5	6.9	9.9	3.9
9	-10	1.5	6.1	9.5	1.9
10	-10	1.5	6.3	9.4	5.7
11	10	1	6.7	9.1	2.8
12	10	1	6.6	9.3	4.1
13	10	1	7.2	8.3	3.8
14	10	1	7.1	8.4	1.6
15	10	1	6.8	8.5	3.4
16	10	1.5	7.1	9.2	8.4
17	10	1.5	7	8.8	5.2
18	10	1.5	7.2	9.7	6.9
19	10	1.5	7.5	10.1	2.7
20	10	1.5	7.6	9.2	1.9

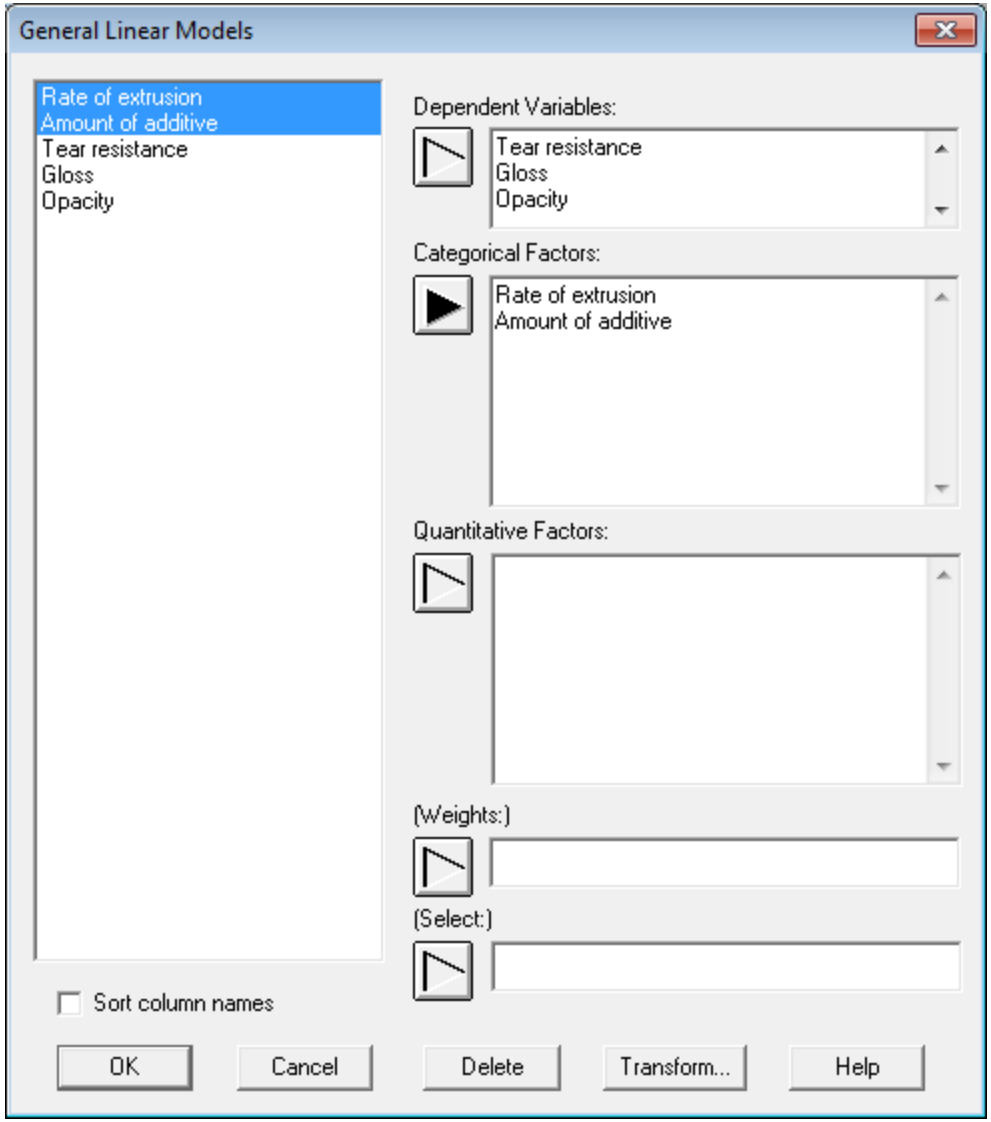
film B C

Source: Johnson and Wichern, Applied Multivariate Statistical Analysis

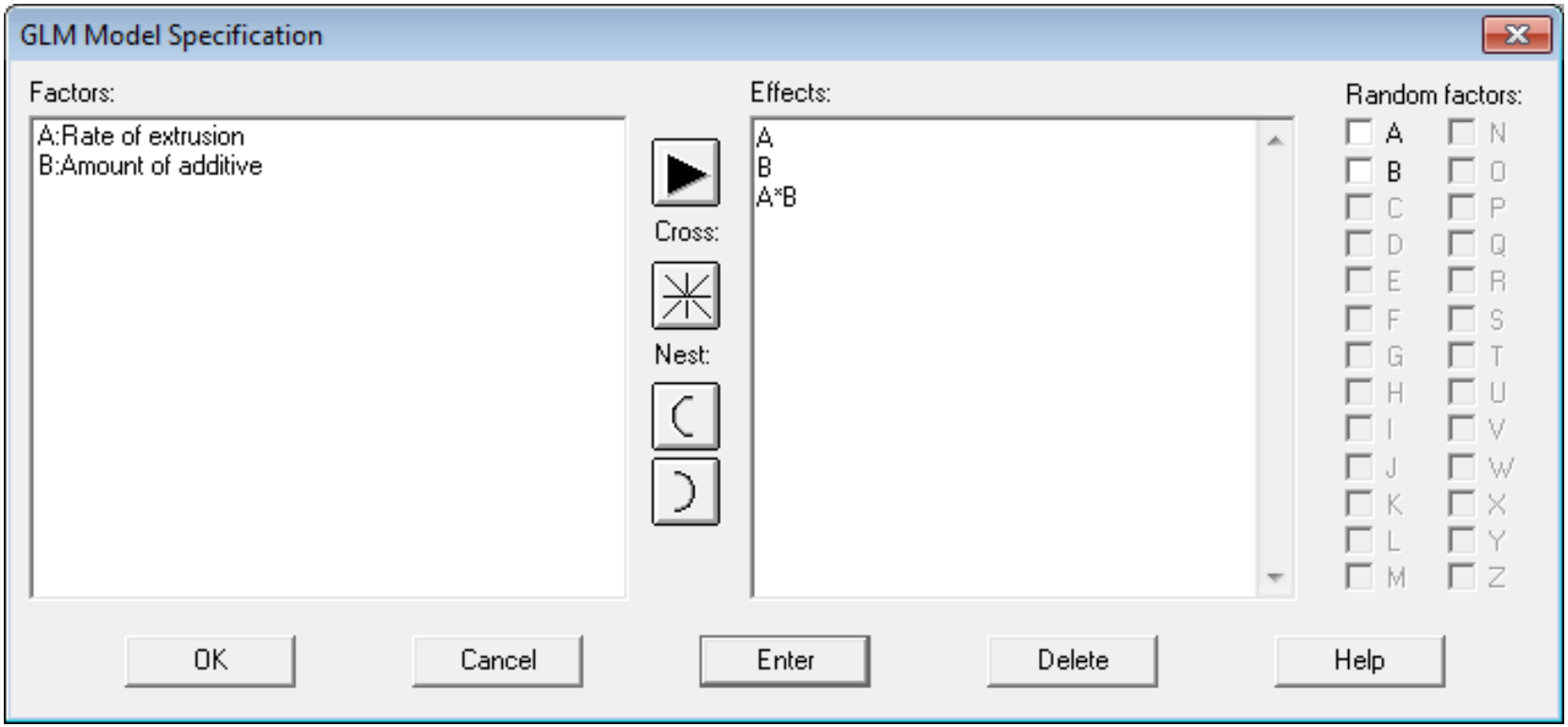
Why use MANOVA?

- Avoids inflation of the Type I error which occurs when testing each dependent variable separately.
- Occasionally detects dependencies that would not be detected otherwise.
- Most effective when response variables are moderately correlated (0.4-0.7).
- Good approach to analyzing repeated measures designs.

GLM: Data Input



GLM: Model Specification



GLM – Univariate Analyses

General Linear Models

Number of dependent variables: 3
Number of categorical factors: 2
A=Rate of extrusion
B=Amount of additive
Number of quantitative factors: 0

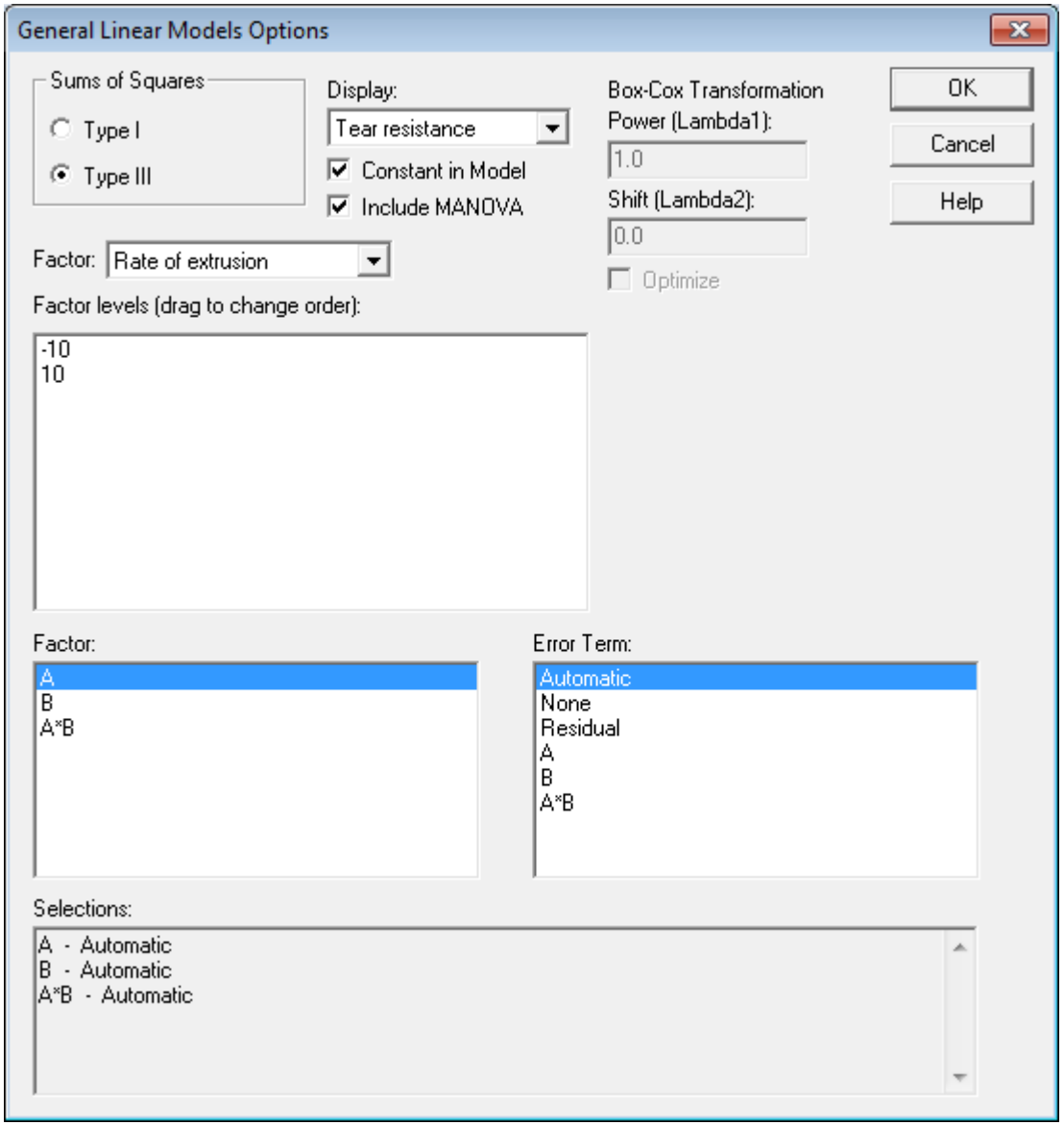
Analysis of Variance for Tear resistance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Model	2.5015	3	0.833833	7.56	0.0023
Residual	1.764	16	0.11025		
Total (Corr.)	4.2655	19			

Type III Sums of Squares

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Rate of extrusion	1.7405	1	1.7405	15.79	0.0011
Amount of additive	0.7605	1	0.7605	6.90	0.0183
Rate of extrusion*Amount of additive	0.0005	1	0.0005	0.00	0.9471
Residual	1.764	16	0.11025		
Total (corrected)	4.2655	19			

GLM: Analysis Options



GLM: Analysis Summary

Adds MANOVA tests for each effect. Based on sums of squares and cross-product matrices for the hypothesis H (no effect) and the error E.

MANOVA for A

Wilks' lambda = 0.381858 F = 7.55427 P-value = 0.00303404
Pillai trace = 0.618142 F = 7.55427 P-value = 0.00303404
Hotelling-Lawley trace = 1.61877 F = 7.55427 P-value = 0.00303404
Roy's greatest root = 1.61877 s = 1 m = 0.5 n = 6.0

Hypothesis Matrix H

	Tear resistance	*Gloss*	*Opacity*
Tear resistance	1.7405	-1.5045	0.8555
Gloss	-1.5045	1.3005	-0.7395
Opacity	0.8555	-0.7395	0.4205

Error Matrix E

	Tear resistance	*Gloss*	*Opacity*
Tear resistance	1.764	0.02	-3.07
Gloss	0.02	2.628	-0.552
Opacity	-3.07	-0.552	64.924

GLM: Test of Interaction

MANOVA for A*B

Wilks' lambda = 0.777106 F = 1.33852 P-value = 0.301782

Pillai trace = 0.222894 F = 1.33852 P-value = 0.301782

Hotelling-Lawley trace = 0.286826 F = 1.33852 P-value = 0.301782

Roy's greatest root = 0.286826 s = 1 m = 0.5 n = 6.0

Hypothesis Matrix H

	Tear resistance	*Gloss*	*Opacity*
Tear resistance	0.0005	0.0165	0.0445
Gloss	0.0165	0.5445	1.4685
Opacity	0.0445	1.4685	3.9605

Error Matrix E

	Tear resistance	*Gloss*	*Opacity*
Tear resistance	1.764	0.02	-3.07
Gloss	0.02	2.628	-0.552
Opacity	-3.07	-0.552	64.924

Wilk's lambda

Compares the between groups covariance matrix to the within groups covariance matrix:

$$\Lambda^* = \frac{|E|}{|E+H|}$$

Extends the univariate F tests to handle multivariate hypotheses (that all level means are equal for each of the response variables).

4. PLS (Partial Least Squares)

PLS is a procedure for finding the relationship between two matrices: a set of predictor variables X and a set of response variables Y.

$$Y = X\beta + E$$

Unlike Multiple Regression, the number of observations n may be less than the number of predictor variables p.

Basic Process

Looks for latent variables (combinations of the X's) that explain a large proportion of the variance in the Y's.

Requires selecting c components where $c < n$.

Example (plsstocks.sgd)

n = 17 common stocks

stockmetrics.sgd

	Ticker	Company	Sector	Industry	P/E	Fwd P/E	PEG	P/S	
1	AAPL	Apple Inc.	Technology	Personal Comput	9.730	8.610	0.510	2.450	3.
2	CAT	Caterpillar Inc	Industrial Gooc	Farm & Construc	10.090	9.130	0.720	0.850	3.
3	CRR	CARBO Ceramics	Basic Materials	Oil & Gas Equip	19.170	15.740	1.920	3.150	2.
4	CSX	CSX Corp.	Services	Railroads	13.520	11.800	1.160	2.100	2.
5	CVS	CVS Caremark Cc	Services	Drug Stores	17.940	12.370	1.340	0.550	1.
6	FCX	Freeport-McMoRa	Basic Materials	Copper	10.150	6.770	3.380	1.710	1.
7	GE	General Electri	Industrial Gooc	Diversified Mac	16.600	12.480	1.470	1.620	1.
8	GOOG	Google Inc.	Technology	Internet Inform	24.670	14.980	1.670	5.260	3.
9	HAL	Halliburton Cor	Basic Materials	Oil & Gas Equip	14.360	10.130	0.880	1.310	2.
10	HON	Honeywell Inter	Industrial Gooc	Diversified Mac	20.090	13.510	1.930	1.540	4.
11	MDT	Medtronic, Inc.	Healthcare	Medical Appliar	14.270	12.120	2.120	2.890	2.
12	ORCL	Oracle Corporat	Technology	Application Sof	15.070	11.060	1.410	4.110	3.
13	PG	Procter & Gambl	Consumer Goods	Personal Produc	19.920	17.820	2.420	2.550	3.
14	SLB	Schlumberger Li	Basic Materials	Oil & Gas Equip	18.230	12.810	1.070	2.330	2.
15	SYK	Stryker Corp.	Healthcare	Medical Appliar	19.270	13.870	2.130	2.850	2.
16	T	AT&T, Inc.	Technology	Telecom Service	30.790	13.750	5.310	1.610	2.
17	WFC	Wells Fargo & C	Financial	Money Center Ba	10.990	9.490	1.200	4.020	1.

stockmetrics B C

Source: finviz.com

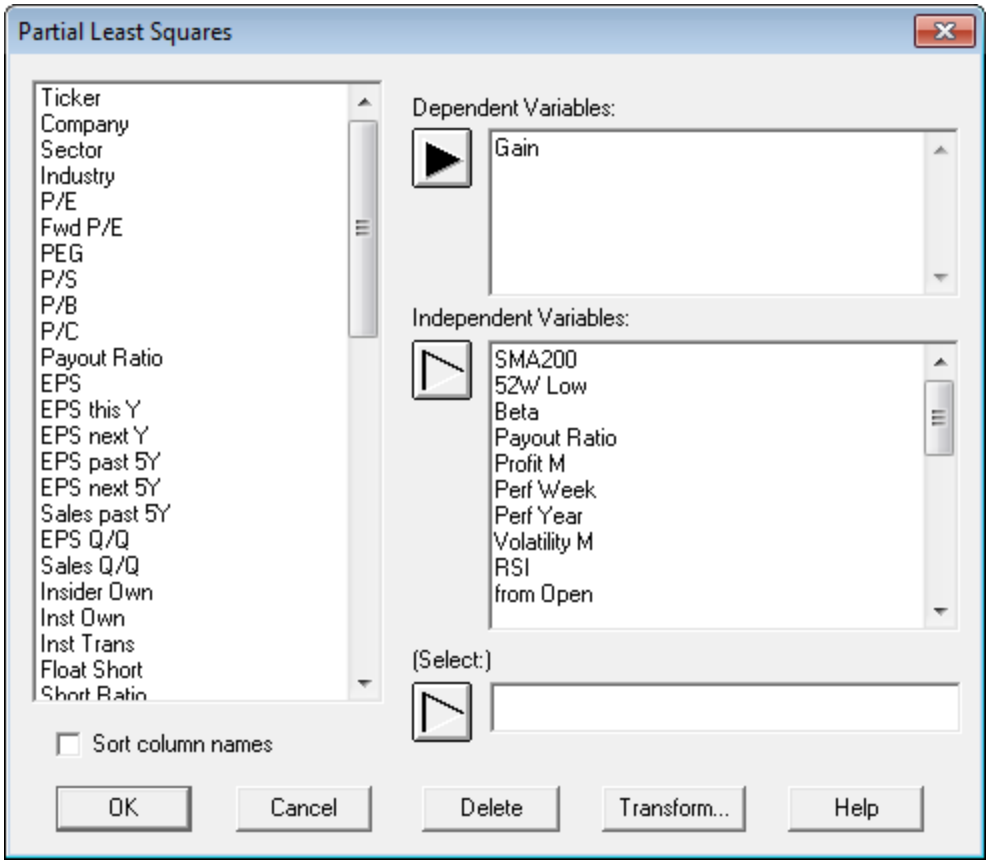
Variables

- Y: percent change in stock price between April 4, 2013 and January 8, 2014.

- X: collection of metrics obtained on April 4, 2013.

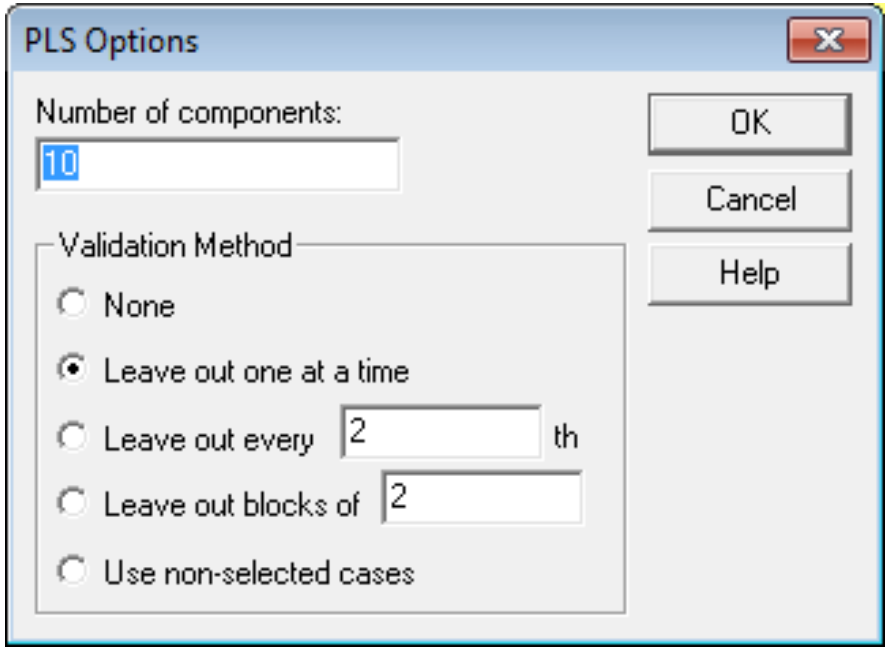
Goal: To develop a model that will predict the change in the stock price over the subsequent 8 months based upon the metrics available on April 4.

Data Input Dialog Box



n = 17 samples, p = 26 predictor variables

Analysis Options



Specify:

- maximum number of components to extract
- validation method used to help select model

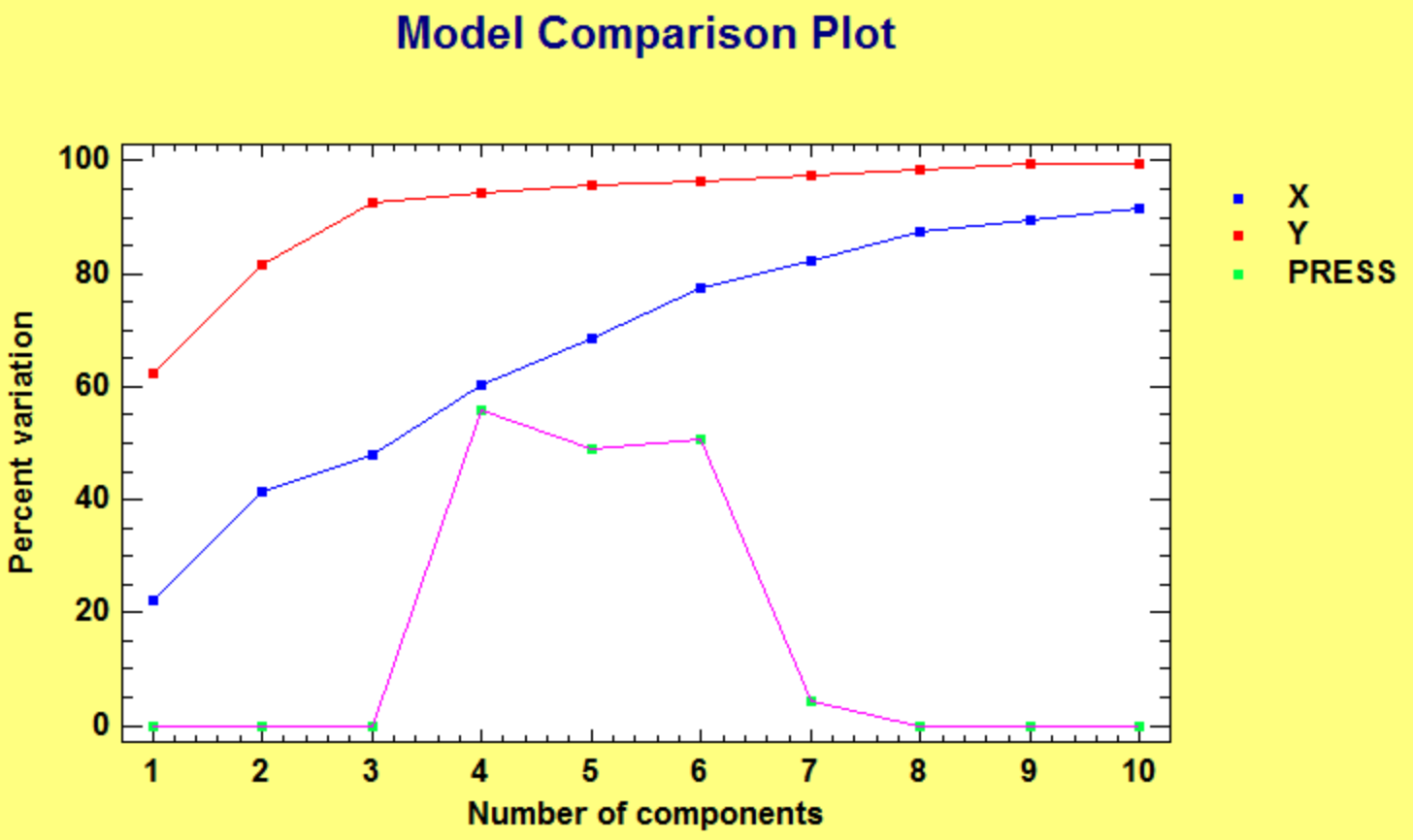
Selecting Number of Components

Independent and Dependent Variables

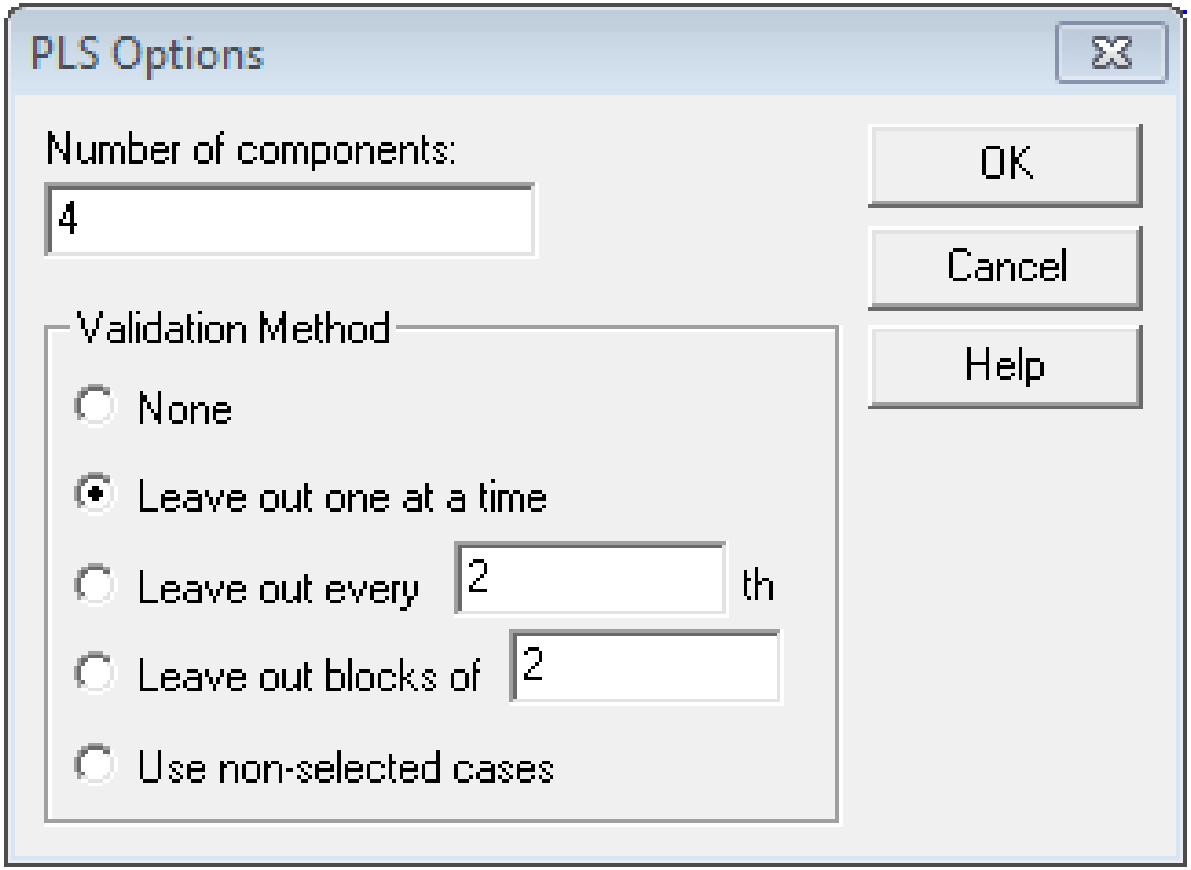
	% Variation	*Cumulative %*	*% Variation*	*Cumulative %*	*Average Prediction*
Component	*in X*	*of X*	*in Y*	*of Y*	*R-Squared*
1	22.0154	22.0154	62.5255	62.5255	0.0
2	19.31	41.3254	18.9418	81.4673	0.0
3	6.57748	47.9029	11.0395	92.5069	0.0
4	12.2584	60.1612	1.7923	94.2992	55.9479
5	8.42438	68.5856	1.25669	95.5559	48.9511
6	8.98554	77.5711	0.776974	96.3328	50.6795
7	4.56788	82.139	1.09186	97.4247	4.4203
8	5.34263	87.4816	0.867377	98.2921	0.0
9	1.88191	89.3636	1.03079	99.3229	0.0
10	2.16042	91.524	0.280125	99.603	0.0

Average Prediction R-Squared measures how well the model predicts the observations that were withheld from the fit.

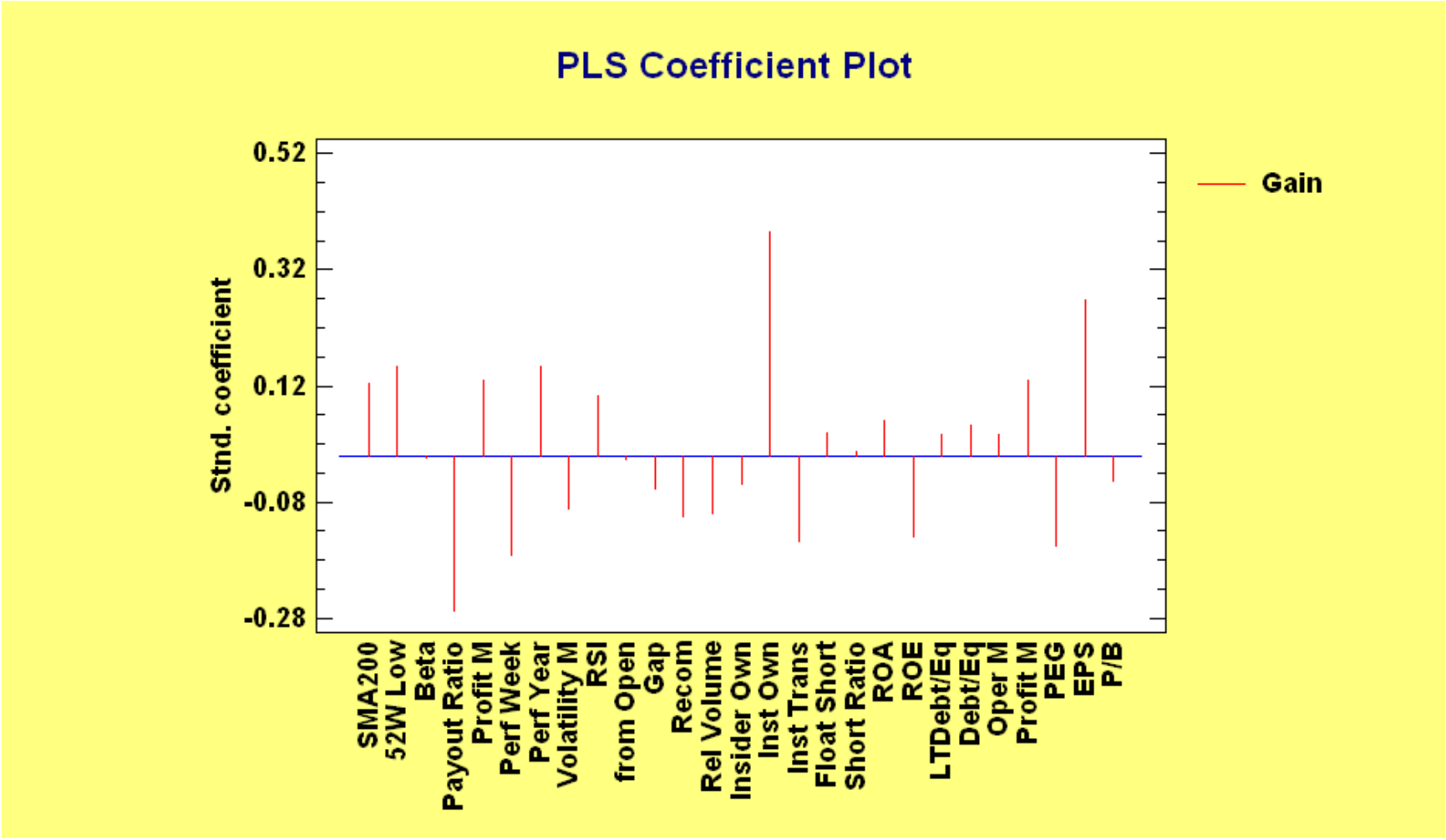
Model Comparison Plot



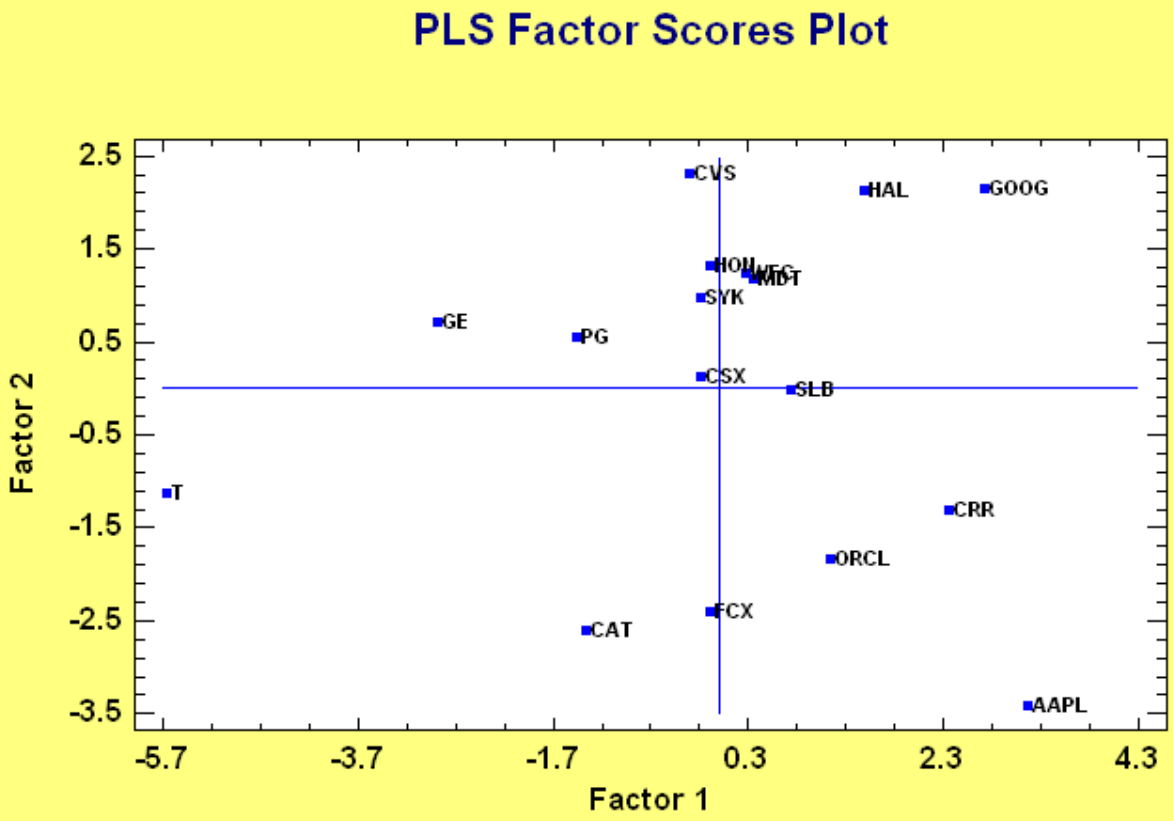
Final Model


PLS Options
Number of components:
4
OK
Cancel
Help
Validation Method
None
Leave out one at a time
Leave out every
2
th
Leave out blocks of
2
Use non-selected cases

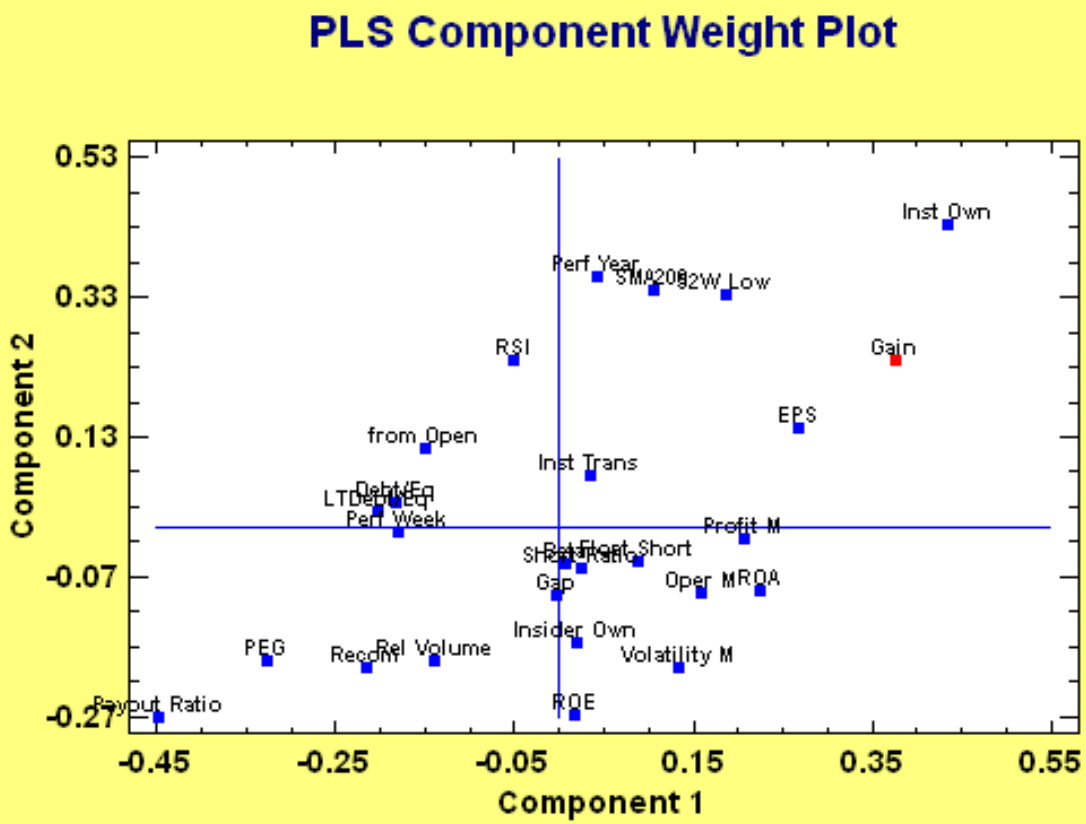
Coefficient Plot



Factor Plot


PLS Factor Scores Plot
Factor 2
Factor 1
2.5
1.5
0.5
-0.5
-1.5
-2.5
-3.5
-5.7
-3.7
-1.7
0.3
2.3
4.3
CVS
HAL
GOOG
SYK
GE
PG
CSX
SLB
T
CRR
ORCL
FCX
CAT
AAPL

Component Weights Plot



Residual Analysis

Predictions and Residuals

				Standardized
Row	*Gain*	*Predicted*	*Residual*	*Residual*
1	26.71	26.967	-0.256987	-0.107928
2	4.09	3.36088	0.729116	0.267997
3	26.61	28.8279	-2.21793	-0.825462
4	16.65	18.2366	-1.58661	-0.48625
5	27.95	28.5082	-0.558191	-0.188135
6	11.8	11.3349	0.465082	0.151451
7	17.89	15.3021	2.58791	1.01891
8	42.44	41.3354	1.10455	0.412914
9	23.97	24.6278	-0.657788	-0.251198
10	21.67	20.252	1.41796	0.446329
11	30.0	24.6337	5.36631	1.66675
12	16.38	14.864	1.51604	0.533656
13	3.27	10.8702	-7.60019	-2.41174
14	17.51	16.291	1.21902	0.399293
15	17.89	15.7965	2.09345	0.677537
16	-8.08	-8.41592	0.335916	0.171076
17	24.34	28.2977	-3.95767	-1.29128

More Information

Statgraphics Centurion: www.statgraphics.com

Uniwin Plus: www.statgraphics.fr or www.sigmaplus.fr

Or send e-mail to info@statgraphics.com

Join the Statgraphics Community on:

Follow us on