RUG1

university of groningen

Loglinear Models for Contingency Tables

Seminar in Methodology and Statistics

Karin Beijering

K.Beijering@rug.nl

www.rug.nl/staff/k.beijering

Slide 1

RUG1

To set the date:
* >Insert >Date and Time
* At Fixed: fill the date in format mm-dd-yy
* >Apply to All

RUG; 30-8-2007

Outline

- Introduction
- Data
- Running Loglinear Analysis
- Output / Results
- Concluding remarks

Introduction

- Study the relationship between categorical variables
  - Chi-Square
  - L**oglinear Models**

- Loglinear Analysis is an extension of Chi-Square

- Modeling of cell counts in contingency tables

- Robust analysis of complicated contingency tables involving several variables

- Describe associations and interaction patterns among a set of categorical variables

Introduction

- Loglinear models are "ANOVA-like" models for the log-expected cell counts of contingency tables

- Loglinear models are logarithmic versions of the general linear

$$\text{Outcome}_i = (\text{Model}_i) + \text{error}_i$$

- The logarithm of the cell frequencies is a linear function of the logarithms of the components:

$$\ln(O_i) = \ln(\text{Model}_i) + \ln(\varepsilon_i)$$

Introduction

› Assumptions (Chi-Square and Loglinear Analysis)
- categorical data
- each categorical variable is called a factor
- every case should fall into only one cross-classification category
- all expected frequencies should be greater than 1, and not more than 20% should be less than 5.
  1. collapse the data across one of the variables
  2. collapse levels of one of the variables
  3. collect more data
  4. accept loss of power
  5. add a constant (0,5) to all cells of the table

Data

- Random samples of Danish, Norwegian and Swedish declarative main clauses containing the word ‘maybe’ (resp. *måske*, *kanskje*, *kanske*)

- Three possible structures:

- V2
-! XP MAYBE …
- MAYBE (that) S …

Data – clause types

› **V2**

- Olle har **kanske** inte sovit inatt

  Olle has **maybe** not slept last.night
- **Kanske** har Olle inte sovit inatt

  Maybe has Olle not slept last.night

› **XP maybe … (non-V2)**

- Olle **kanske** inte har sovit inatt*

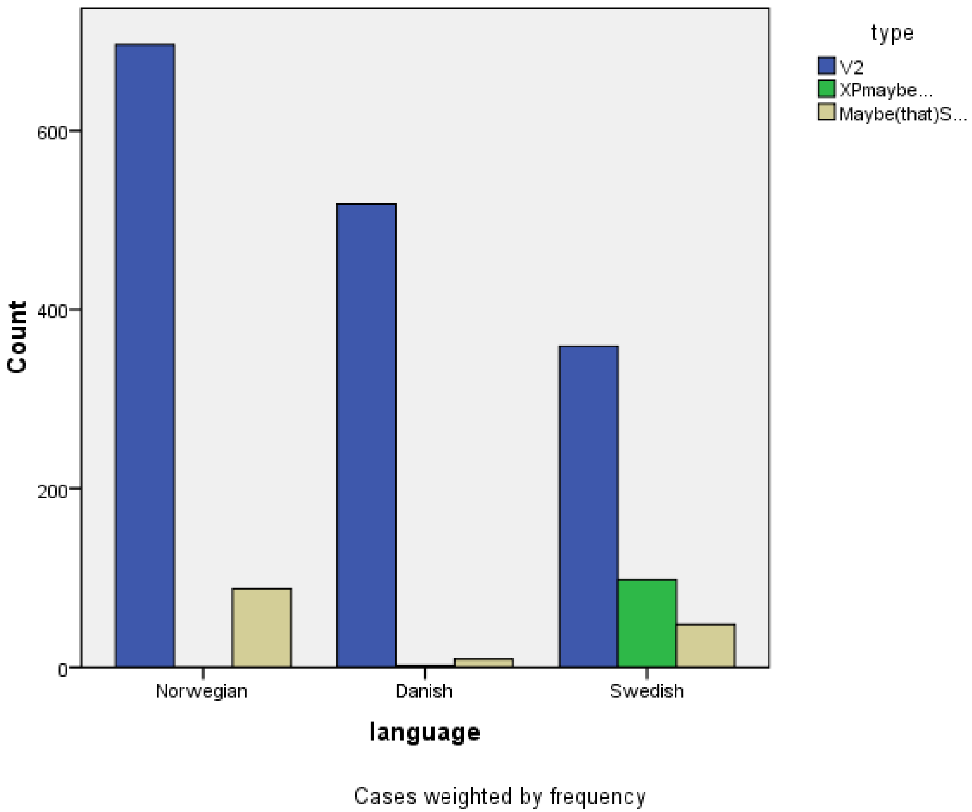
  Olle **maybe** not has slept last.night

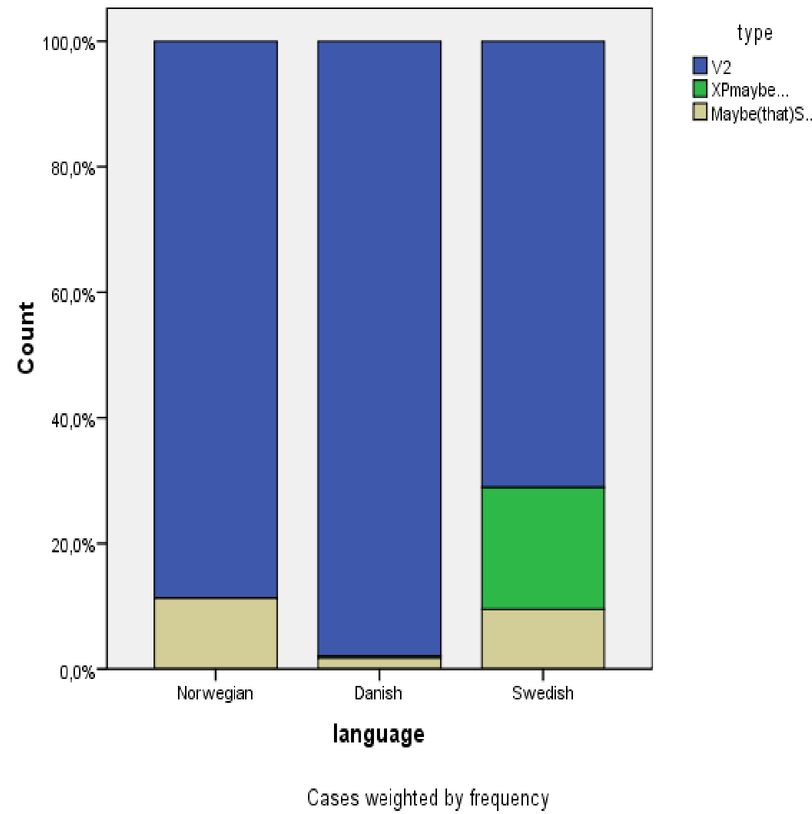
› **Maybe (that) S … (non-V2)**

- **Kanske** (att) Olle inte har sovit inatt

  Maybe (that) Olle not has slept last.night

Data – bar charts





Data – two-way (3 x 3) contingency table

language * type Crosstabulation

			type			
			V2	XPmaybe...	Maybe(that) S...	Total
language	Norwegian	Count	696	0	88	784
		Expected Count	678,3	43,1	62,5	784,0
		% within language	88,8%	,0%	11,2%	100,0%
		% within type	44,2%	,0%	60,7%	43,1%
		% of Total	38,3%	,0%	4,8%	43,1%
	Danish	Count	518	2	9	529
		Expected Count	457,7	29,1	42,2	529,0
		% within language	97,9%	,4%	1,7%	100,0%
		% within type	32,9%	2,0%	6,2%	29,1%
		% of Total	28,5%	,1%	,5%	29,1%
	Swedish	Count	359	98	48	505
		Expected Count	436,9	27,8	40,3	505,0
		% within language	71,1%	19,4%	9,5%	100,0%
		% within type	22,8%	98,0%	33,1%	27,8%
		% of Total	19,7%	5,4%	2,6%	27,8%
	Total	Count	1573	100	145	1818
		Expected Count	1573,0	100,0	145,0	1818,0
		% within language	86,5%	5,5%	8,0%	100,0%
		% within type	100,0%	100,0%	100,0%	100,0%
		% of Total	86,5%	5,5%	8,0%	100,0%

Data – two-way (3 x 3) contingency table

- The crosstabulation does not tell whether the distributional differences are real or due to chance variation. Chi-square measures the difference between the observed cell counts and expected cell counts (the frequencies you would expect if the rows and columns were unrelated).
- H_0: no association between variables (observed counts = expected counts)
- H_a: association between variables (oberved counts ≠ expected counts)

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	3,062E2	4	,000
Likelihood Ratio	308,442	4	,000
Linear-by-Linear Association	14,819	1	,000
N of Valid Cases	1818		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 27,78.

Data – two-way (3 x 3) contingency table

› Chi-Square is useful for determining relationships between categorical variables, however, it does not provide information about the strength and direction of the relationship.

- **Symmetric measures** quantify the strength of an association
- **Directional measures** quantify the reduction in the error of predicting the row variable value when the column variable value is known, or vice versa.
- The values of the measures of association are between 0 and 1.

  0= no relationship

  1= perfect relationship
- NB **Odds Ratios** are more suitable to measure effect size (2 x 2 tables).

Data – two-way (3 x 3) contingency table

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,410	,000
	Cramer's V	,290	,000
	Contingency Coefficient	,380	,000
N of Valid Cases		1818	

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	,077	,007	10,178	,000
		language Dependent	,095	,009	10,178	,000
		type Dependent	,000	,000	.[c]	.[c]
	Goodman and Kruskal tau	language Dependent	,079	,004		,000[d]
		type Dependent	,081	,010		,000[d]
	Uncertainty Coefficient	Symmetric	,108	,009	10,579	,000[e]
		language Dependent	,079	,007	10,579	,000[e]
		type Dependent	,174	,013	10,579	,000[e]

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Cannot be computed because the asymptotic standard error equals zero.

d. Based on chi-square approximation

e. Likelihood ratio chi-square probability.

Loglinear analysis

› Three procedures are available for using loglinear models to study relationships between categorical variables:

- **Model Selection Loglinear Analysis**
- General Loglinear Analysis
- Logit Loglinear Analysis

Model Selection Loglinear Analysis

- Identify models for describing the relationship between categorical variables.
- Find out which categorical variables are associated
- Find the "Best" Model
- Fits hierarchical loglinear models to multi-dimensional crosstabulations using an iterative proportional-fitting algorithm.

Models and parameters

› Independence model

$$\log \mu_{ij} = \lambda + \lambda_i^{1} + \lambda_j^{2}$$

› Saturated model

$$\log \mu_{ij} = \lambda + \lambda_i^{1} + \lambda_j^{2} + \lambda_{ij}^{12}$$

› Hierarchical model

$\log \mu_{ij}$ = log of the expected cell frequency of the cases for cell ij

λ = constant

^{123} = variables

ijk = categories within the variables

λ_i^{1} = main effect for variable 1

λ_j^{2} = main effect for variable 2

λ_{ijk}^{123} = interaction effect for variables 1, 2 and 3

$$\log \mu_{ij} = \lambda + \lambda_i^{1} + \lambda_j^{2} + \lambda_k^{3} + \lambda_{ij}^{12} + \lambda_{ik}^{13} + \lambda_{jk}^{23} + \lambda_{ijk}^{123}$$

Similarities to regression and ANOVA

general linear model:

$$\text{Outcome}_i = (\text{Model}_i) + \text{error}_i$$

multiple regression:

$$Y_i = (b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n) + \varepsilon_i$$

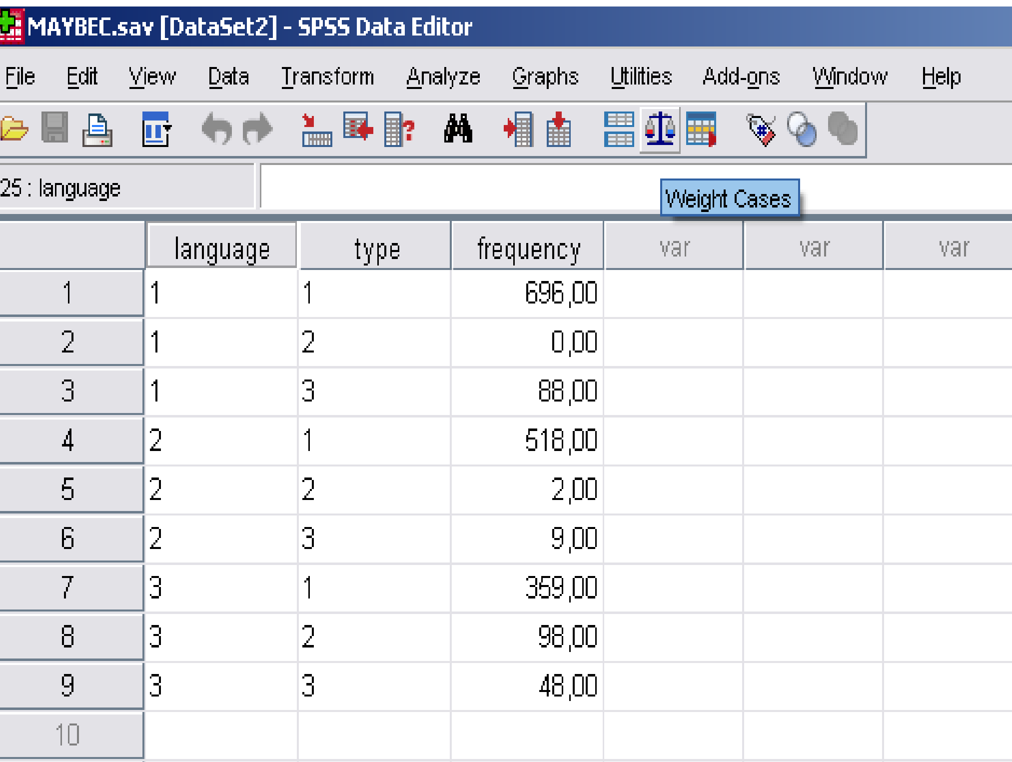
ANOVA:

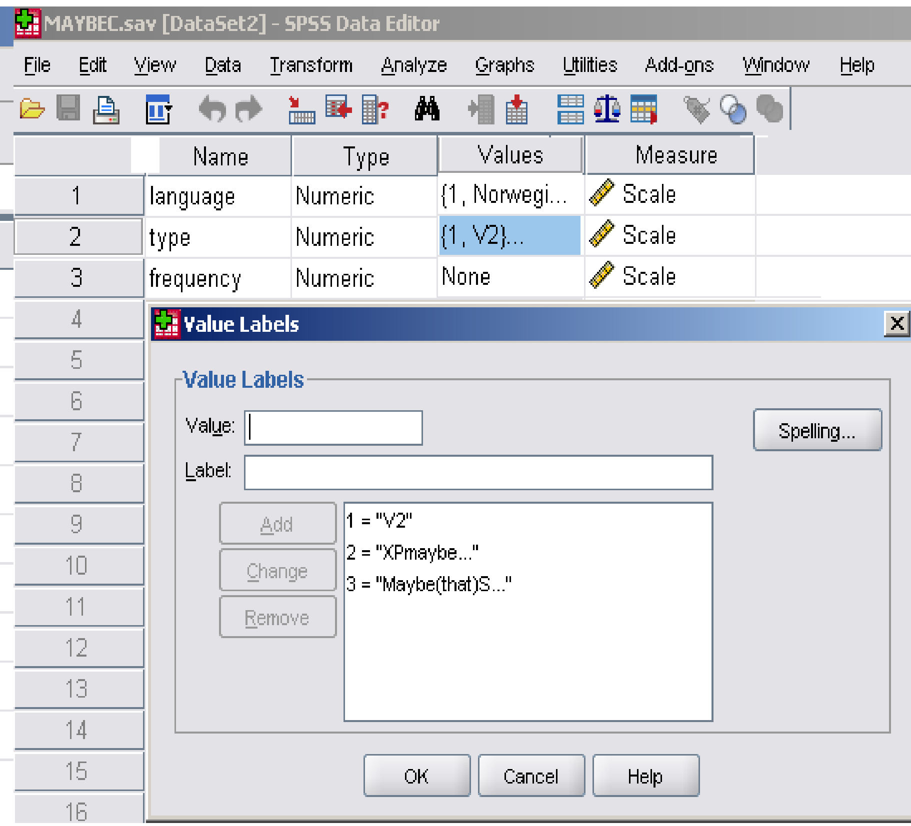
$$\text{Outcome}_i = (b_0 + b_1A_i + b_2B_i + b_3AB_i) + \varepsilon_i$$

Loglinear model:

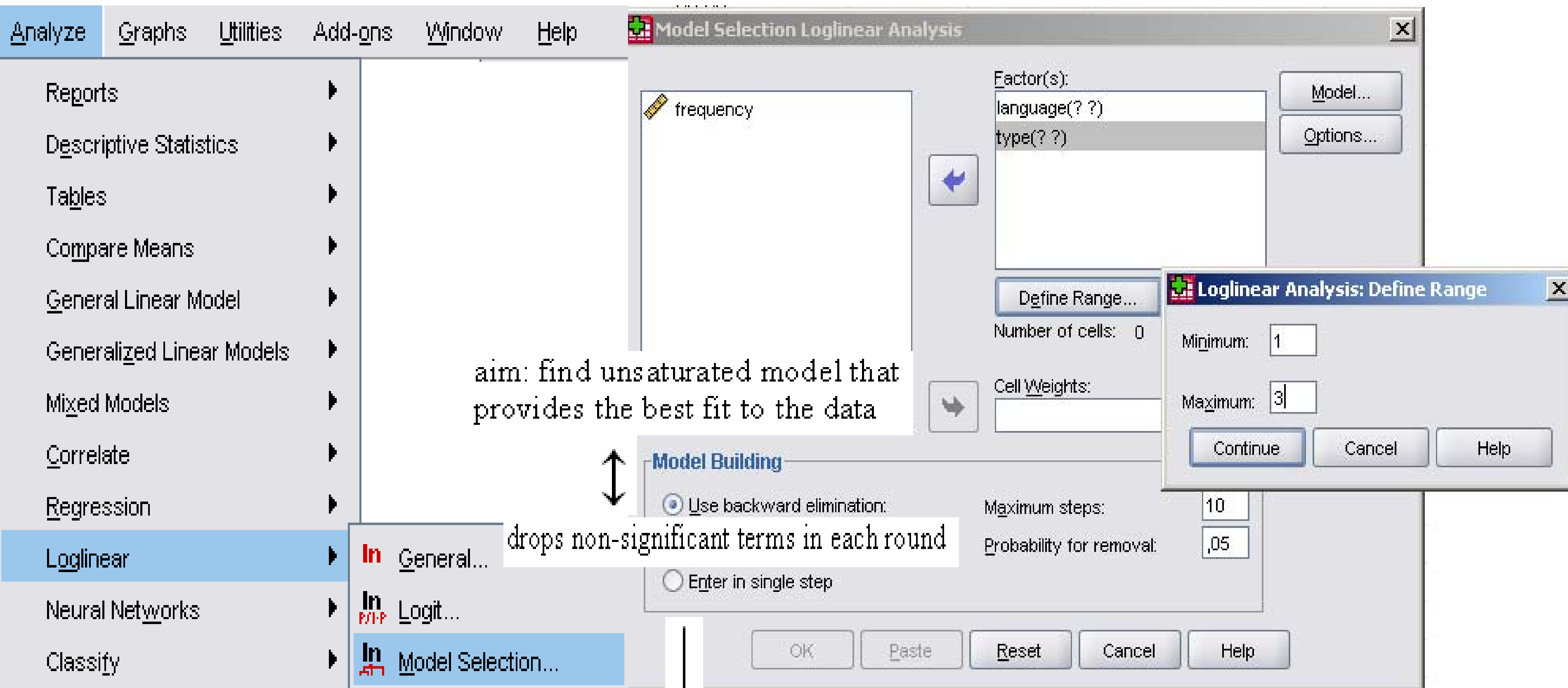
$$\ln(\mathcal{O}_i) = \ln(\text{Model}_i) + \ln(\varepsilon_i)$$
$$\ln(\mathcal{O}_{ij}) = (b_0 + b_1A_i + b_2B_j + b_3AB_{ij}) + \ln(\varepsilon_{ij})$$

Running Model Selection Loglinear Analysis





Running Model Selection Loglinear Analysis

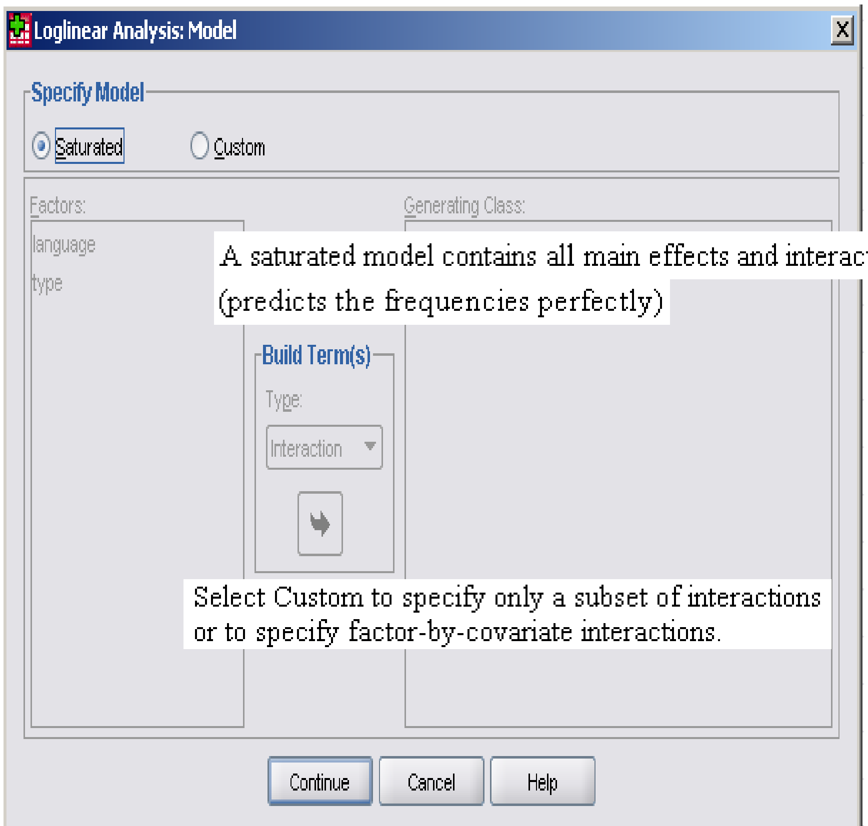


aim: find unsaturated model that provides the best fit to the data

drops non-significant terms in each round

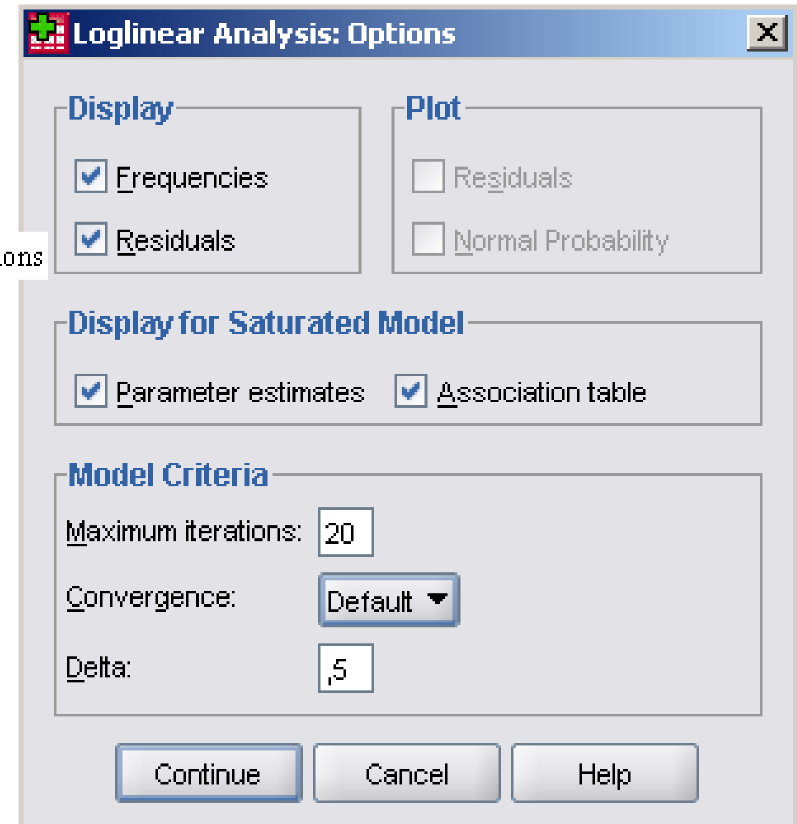
non-hierarchical model (not recommended)

Running Model Selection Loglinear Analysis



A saturated model contains all main effects and interactions (predicts the frequencies perfectly)

Select Custom to specify only a subset of interactions or to specify factor-by-covariate interactions.



Output Model Selection Loglinear Analysis

- Cell Counts and Residuals (saturated model)
- Convergence Information
- K-Way and Higher-Order Effects
- Parameter Estimates
- Partial Associations
- Backward Elimination Statistics
- Goodness-of-Fit-Tests

Convergence Information

Convergence Information

Generating Class	language*type
Number of Iterations	1
Max. Difference between Observed and Fitted Marginals	,000
Convergence Criterion	484,416

Convergence Information[a]

Generating Class	language*type
Number of Iterations	0
Max. Difference between Observed and Fitted Marginals	,000
Convergence Criterion	484,416

a. Statistics for the final model after Backward Elimination.

K-Way and Higher-Order Effects

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects[a]	1	8	2610,088	,000	2644,168	,000	0
	2	4	308,442	,000	306,153	,000	2
K-way Effects[b]	1	4	2301,646	,000	2338,015	,000	0
	2	4	308,442	,000	306,153	,000	0

a. Tests that k-way and higher order effects are zero.

b. Tests that k-way effects are zero.

Parameter Estimates

-

For Design 1, at least one cell count is zero. The parameter estimates for this saturated model are therefore not computed.

- Add 0,5 to each cell in case of structural zero's (empty cells in the crosstabulation)

Parameter Estimates

Effect	Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
language*type	1	,536	,237	2,260	,024	,071	1,000
	2	-1,681	,465	-3,611	,000	-2,593	-,768
	3	,702	,187	3,759	,000	,336	1,068
	4	-,121	,348	-,348	,728	-,804	,561
language	1	-,217	,236	-,920	,358	-,680	,246
	2	-,678	,185	-3,655	,000	-1,042	-,314
type	1	2,333	,136	17,143	,000	2,066	2,599
	2	-1,998	,261	-7,659	,000	-2,509	-1,487

Partial Associations

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
language	2	75,887	,000	2
type	2	2225,759	,000	1

Backward Elimination Statistics

Step Summary

Step[a]		Effects	Chi-Square[c]	df	Sig.	Number of Iterations
0	Generating Class[b]	language*type	,000	0	.	
	Deleted Effect 1	language*type	308,442	4	,000	2
1	Generating Class[b]	language*type	,000	0	.	

a. At each step, the effect with the largest significance level for the Likelihood Ratio Change is deleted, provided the significance level is larger than ,050

b. Statistics are displayed for the best model at each step after step 0.

c. For 'Deleted Effect', this is the change in the Chi-Square after the effect is deleted from the model.

- Step 0. The model generated by the two-way interaction of factors; that is, the saturated model, is considered. This model also contains the main effects. The two-way interaction is tested for significance by deleting it from the model. The change in chi-square from the saturated model to the model without the two-way interaction is tested and found to be significant (significance value < 0.05). Thus, this interaction term cannot be dropped from the model.

- Step 1. Since the two-way interaction could not be removed from the model, there are no more terms to test. Thus, the final model includes the two-way interaction and the main effects.

Goodness-of-Fit-Tests

- The goodness-of-fit table presents two tests of the null hypothesis that the final model adequately fits the data. If the significance value is small (<0.05), then the model does not adequately fit the data. The goodness-of-fit statistics are based on the cell counts and residuals.Here, the model perfectly predicts the data.

Goodness-of-Fit Tests

	Chi-Square	df	Sig.
Likelihood Ratio	,000	0	.
Pearson	,000	0	.

Multi-way tables

- Cross tables can be extended/refined, i.e. more factors can be added to the table.

- In addition to language and type, information about other epistemic elements in the clause (auxiliaries, adverbs, particles etc.), the finite verb (modal or not), the type of subject (pronoun or not), etc. can be added.

- 2 x 2 x 2 table

  language (Danish / Norwegian) * type (V2 / NV2) * Vf (modal / other)

Three-way (2 x 2 x 2) contingency table

Vf * type * language Crosstabulation

language				type V2	NV2	Total
norwegian	Vf	modal	Count	128	10	138
			Expected Count	122,5	15,5	138,0
		other	Count	568	78	646
			Expected Count	573,5	72,5	646,0
		Total	Count	696	88	784
			Expected Count	696,0	88,0	784,0
danish	Vf	modal	Count	118	3	121
			Expected Count	117,8	3,2	121,0
		other	Count	400	11	411
			Expected Count	400,2	10,8	411,0
		Total	Count	518	14	532
			Expected Count	518,0	14,0	532,0

Convergence Information

Convergence Information

Generating Class	language*type*Vf
Number of Iterations	1
Max. Difference between Observed and Fitted Marginals	,000
Convergence Criterion	,568

Convergence Information[a]

Generating Class	language*type, language*Vf
Number of Iterations	0
Max. Difference between Observed and Fitted Marginals	,000
Convergence Criterion	,568

a. Statistics for the final model after Backward Elimination.

K-Way and Higher-Order Effects

test whether removing terms significantly affects the fit of the model (p=0.05)

K-Way and Higher-Order Effects

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher Order Effects[a]	1	7	1720,006	,000	1840,511	,000	0
	2	4	45,642	,000	42,343	,000	2
	3	1	,401	,527	,428	,513	3
K-way Effects[b]	1	3	1674,364	,000	1798,167	,000	0
	2	3	45,241	,000	41,915	,000	0
	3	1	,401	,527	,428	,513	0

a. Tests that k-way and higher order effects are zero.

b. Tests that k-way effects are zero.

Parameter Estimates

Parameter Estimates

Effect	Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
language*type*Vf	1	,069	,088	,781	,435	-,104	,243
language*type	1	-,324	,088	-3,656	,000	-,497	-,150
language*Vf	1	-,136	,088	-1,542	,123	-,310	,037
type*Vf	1	,062	,088	,701	,483	-,111	,235
language	1	,431	,088	4,875	,000	,258	,605
type	1	1,445	,088	16,326	,000	1,271	1,618
Vf	1	-,738	,088	-8,343	,000	-,912	-,565

Partial Associations

Partial Associations

Effect	df	Partial Chi-Square	Sig.	Number of Iterations
language*type	1	36,288	,000	2
language*Vf	1	4,085	,043	2
type*Vf	1	2,542	,111	2
language	1	48,555	,000	2
type	1	1106,776	,000	2
Vf	1	519,033	,000	2

Backward Elimination Statistics

Backward Elimination Statistics

Step Summary

Step[a]			Effects	Chi-Square[c]	df	Sig.	Number of Iterations
0	Generating Class[b]		language*type*Vf	,000	0	.	
	Deleted Effect	1	language*type*Vf	,401	1	,527	3
1	Generating Class[b]		language*type, language*Vf, type*Vf	,401	1	,527	
	Deleted Effect	1	language*type	36,288	1	,000	2
		2	language*Vf	4,085	1	,043	2
		3	type*Vf	2,542	1	,111	2
2	Generating Class[b]		language*type, language*Vf	2,943	2	,230	
	Deleted Effect	1	language*type	37,451	1	,000	2
		2	language*Vf	5,248	1	,022	2
3	Generating Class[b]		language*type, language*Vf	2,943	2	,230	

a. At each step, the effect with the largest significance level for the Likelihood Ratio Change is deleted, provided the significance level is larger than ,050

b. Statistics are displayed for the best model at each step after step 0.

c. For 'Deleted Effect', this is the change in the Chi-Square after the effect is deleted from the model.

Backward Elimination Statistics

- Step 0. This model includes all interactions and main effects. The three-way interaction is tested for significance by deleting it from the model. The change in chi-square from the saturated model to the model without the three-way interaction is tested and found to be not significant (significance value > 0.05). Thus, the three-way interaction term can be dropped from the model.

- Step 1. The model generated by all two-way interactions is considered. This model also includes the main effects. Each two-way interaction is tested for significance by deleting it from the model. Since the significance value for the change in chi-square for the effects language*type and language*Vf is less than 0.05, these terms should be kept in the model. The effect type*Vf can be dropped.

- Step 2. The retained two-way interactions language*type and language*Vf are considered. None of them can be removed from the model (significance value < 0.05), there are no more terms to test.

- Step 3. The final model includes the main effects and the two-way interaction terms language*type and language*Vf.

Goodness-of-Fit-Tests

language	type	Vf	Residuals	Std. Residuals
norwegian	V2	modal	5,490	,496
		other	-5,490	-,229
	NV2	modal	-5,490	-1,395
		other	5,490	,645
danish	V2	modal	,184	,017
		other	-,184	-,009
	NV2	modal	-,184	-,103
		other	,184	,056

Residuals should consist of positive and negative values of approximately equal magnitudes and should be smaller than 2 (standardized residuals).

small values of chi-square statistics indicate a good model

Goodness-of-Fit Tests

	Chi-Square	df	Sig.
Likelihood Ratio	2,943	2	,230
Pearson	2,674	2	,263

The goodness-of-fit table presents two tests of the null hypothesis that the final model adequately fits the data. If the significance value is small (<0.05), then the model does not adequately fit the data. The goodness-of-fit statistics are based on the cell counts and residuals. Here, the final (unsaturated) model fits the data well, i.e., the difference between observed counts [data] and expected counts [model] is not significant (p >0.05).

Related procedures

Model Selection Loglinear Analysis is useful for identifying an initial model for further analysis in General Loglinear Analysis or Logit Loglinear Analysis.

› **General Loglinear Analysis** uses loglinear models without specifying response or predictor variables. It has more input and output options, and is useful for examining the final model produced by Model Selection Loglinear Analysis. Either a Poisson or a multinomial distribution can be analyzed.

› **Logit Loglinear Analysis** models the values of one or more categorical variables given one or more categorical predictors using logit-expected cell counts of crosstabulation tables. It treats one or more categorical variables as responses (independent), and tries to predict their values given the other (explanatory/dependent) categorical variables.

Related procedures

- If there is one dependent variable, you can alternately use **Multinomial Logistic Regression**.
- If there is one dependent variable and it has just two categories, you can alternately use **Logistic Regression**.
- If there is one dependent variable and its categories are ordered, you can alternately use **Ordinal Regression**.

Concluding remarks

- \+ suitable to analyse complicated multiway-tables
- \+ robust “ANOVA-like” analysis of complicated contingency tables
- \+ interactions and main effects of factors
- \+ parameter estimates / partial associations
- \- individual effect of values of factors cannot be determined
- \- structural zero’s
- \- no distinction between dependent / independent variables
- \- specification of many variables with many levels can lead to a situation where many cells have small numbers of observations.

References

- Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. Wiley: New York.
- Everitt, B.S. 1992. *The Analysis of Contingency Tables*. Chapman & Hall: London.
- Field, A. 2005. *Discovering Statistics Using SPSS*. Sage Publications: London.
- SPSS 16.
  - Online Help: loglinear analysis
  - Tutorial: Loglinear Modeling