university of groningen

Log-linear Modeling

Seminar in Methodology and Statistics
Edgar Weiffenbach (with Nick Ruiz)
S1422022

Overview

› Our project
› Log-linear modeling
› Log-linear modeling in the field
› Summary
› References

Our project

› The difference between size reading and gradable readig:

- That sure is a big ship. (size reading)
- He sure is a big idiot. (gradable reading)

Our project

- Lassy corpus
- Adjective+noun pairs
- Three adjectives:
  - Reusachtig
  - Gigantisch
  - Kolossaal
- Three other variables
  - Position in sentence (e.g.: subject, object)
  - Determiner (definite/indefinite)
  - Gradable/size reading

Our project

- Do these variables play a role in the choice between on of the the three adjectives?

Log-linear modeling

- A way of modeling the cell count of contingecy tables with categorical data (like Chi-square).

- No distinction between dependent and independent variables.

- Assumes Poisson-distributed data (like data obtained from a corpus).

Log-linear modeling

- Remember Chi-Square?
  - F^e = (row total x column total) / total

	Y Yes	No	total
X Yes	20 (37,5)	40 (22,5)	60
No	130 (112,5)	50 (67,5)	180
total	150	90	240

Log-linear modeling

- F^e = (row total x column total) / total

- $F_{ij}^e = (F_{i.}^o \times F_{.j}^o) / N$

- Log-linear modeling uses the natural logarithm (ln) to transform the data. When using ln, the following rules apply:
  - $\ln (a \times b) = \ln a + \ln b$
  - $\ln (a / b) = \ln a - \ln b$

Log-linear modeling

- $F_{ij}^{e} = (F_{i.}^{o} \times F_{.j}^{o}) / N$

- $\ln F_{ij}^{e} = \ln F_{i.}^{o} + \ln F_{.j}^{o} - \ln N$

- "the terms which were originally multiplied are replaced by a linear combination of logarithmic terms: a log-linear model" (Rietveld & van Hout: 1993)

Log-linear modeling

- $F_{ij}^{e} = (F_{i.}^{o} \times F_{.j}^{o}) / N$
  $= (150 \times 180) / 240$
  $= 112{,}5$

	Y Yes	No	total
X Yes	20 (37,5)	40 (22,5)	60
No	**130** **(112,5)**	50 (67,5)	**180**
total	**150**	90	**240**

- $\ln F_{ij}^{e} = \ln F_{i.}^{o} + \ln F_{.j}^{o} - \ln N$
  $= \ln 150 + \ln 180 - \ln 240$
  $= 5.193 + 5.011 - 5.481 = 4.723$

  $F_{ij}^{e} = e^{4.723}$ (ANTILOG)
  $= 112.5$

Log-linear modeling

- Having transformed the data, you can now think of the contingency table as reflecting various main effects and interacting effects that are added together in <u>a linear fashion</u> to create the observed table of frequencies.

- $\ln F_{ij}^{e} = \mu + \lambda_i^{A} + \lambda_j^{B} + \lambda_{jj}^{AB}$
  - μ = overall mean of the natural log of the expected frequencies
  - λ = represents an "effect" that the variable(s) has(/have) on the cell frequencies
  - A & B = the variables
  - i&j = categories within the variables (rows & columns)

Log-linear modeling

- $\text{Ln } F_{ij}^{e} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{jj}^{AB}$
  - μ = overall mean of the natural log of the expected frequencies
  - λ = represents an “effect” that the variable(s) has(/have) on the cell frequencies
  - A & B = the variables
  - i&j = categories within the variables (rows & columns)

  - λ_i^A = main effect for variable A
  - λ_j^B = main effect for variable B
  - λ_{jj}^{AB} = interaction effect for variables A & B

Log-linear modeling

- Remember:
  - Log-linear modeling is a way of modeling the cell count of contingecy tables with categorical data.
- $\ln F_{ij}^{e} = \mu + \lambda_i^{A} + \lambda_j^{B} + \lambda_{jj}^{AB}$
  - Is called the "saturated model".
    - It has as many effects as the contingency table has cells.
    - Therefore it has no degrees of freedom
    - So it fits the data perfectly ($F^e = F^o$)
    - But the data is a sample (=/= population), so the model overfits the data.

Log-linear modeling

- Fortunately the effects are combined additively, so it is easy to remove an effect and test if the model still fits the data.
  - This is called the Model Selecting Log-linear Analysis.
  - The goal is to find the most parsimonious (≈ simple) model that does not differ significantly from the saturated model (and thus from the observed frequencies).

Log-linear modeling

- Model Selecting Log-linear Analysis.
  - Is mostly done hierarchicaly:
    - λ_{jj}^{AB} is made up out of λ_i^A and λ_j^B, therefore λ_i^A and λ_j^B must be in the model when λ_{jj}^{AB} is.

	Backward deletion	$\chi^2=$
1.	$\text{Ln } F_{ij}^e = \mu + \lambda_i^A + \lambda_j^B + \lambda_{jj}^{AB}$	0
2.	$\text{Ln } F_{ij}^e = \mu + \lambda_i^A + \lambda_j^B$	?
3.	$\text{Ln } F_{ij}^e = \mu + \lambda_i^A$	?
4.	$\text{Ln } F_{ij}^e = \mu$	?

Log-linear modeling

› This may not be the best approach for a 2x2 contingency table, but it is a very easy statistic for analyzing tables with more dimensions.

- For instance a 3x3 contingency table
  - $\text{Ln } F_{ij}^{e} = \mu + \lambda_{i}^{A} + \lambda_{j}^{B} + \lambda_{k}^{c} + \lambda_{jj}^{AB\,+}\,\lambda_{jk}^{AC\,+}\,\lambda_{jk}^{BC} + \lambda_{jjk}^{ABC}$
- Extra dimensions (variables) leed to a large increase in main and higherorder (=interactional) effects and with log-linear modeling you can easily find out which effects help create the observed frequencies and which can be left out of the model.

Log-linear modeling in the field

- De Haan & van Hout - Statistics and Corpus Analysis: A Loglinear Analysis of Syntactic Constraints on Postmodifying Clauses (1986).
- Bell, Dirks, Levitt & Dubno - Log-Linear Modeling of Consonant Confusion Data (1986).
- Girard & Larmouth - Log-Linear Statistical Models: Explaining the Dynamics of Dialect Diffusion (1988).

Summary

› Log-linear modeling
  - Is a way of modeling the cell count of contingecy tables with categorical data.
  - Replaces originally multiplied terms by a linear combination of logarithmic terms.
  - Tries to find the most parsimonious model that does not differ significantly from the saturated model.

university of groningen

References

- Toni Rietveld and Roeland van Hout (1993) *Statistical Techniques for the Study of Language and Language Behavior.* Mouton De Gruyter: Berlin.

- Alan Agresti (1996) *An Introduction to Categorical Data Analysis.* Wiley: New York.

- Ronald Christensen (1997) *Log-Linear Models and Logistic Regression.* Springer-Verlag: New York.