Optimizing (Socio-)linguistic Analysis: Language Variation Suite Toolkit

Dr. Olga Scrivner

Research Scientist, CNS, SICE, IU
Corporate Faculty, Data Analytics Graduate Program, HU
CEWIT Faculty Fellow

April 12, 2018
Provide researchers with a variety of quantitative methods to advance language variation studies.
Objectives

1. Introduce a novel (socio)linguistic toolkit
2. Develop practical skills
3. Understand and interpret advanced statistical models
Language Variation Suite

It is a Shiny web application designed for data analysis in sociolinguistic research.

It can be used for:

- Processing spreadsheet data
- Reporting in tables and graphs
- Analyzing means, regression, conditional trees ...
  (and much more)
**Background**

LVS is built in R using Shiny package:

1. **R** - a free programming language for statistical computing and graphics

2. **Shiny App** - a web application framework for R

Computational power of R + Web interactivity
Background

http://littleactuary.github.io/blog/Web-application-framework-with-Shiny/
Data Preparation

Important things to consider before data entry:

- **File format:**
  - Comma separated value (CSV) - faster processing
  - Excel format will slow processing

- **Column names should not contain spaces**
  - Permitted: non-accented characters, numbers, underscore, hyphen, and period

- **One column must contain your dependent variable**

- **The rest of the columns contain independent variables**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>Number</td>
<td>R.Use</td>
<td>Lexical.Item</td>
<td>Style</td>
<td>Store</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>1</td>
<td>8</td>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
</tbody>
</table>
Workspace

Browser

- Chrome, Firefox, Safari - recommendable
- Explorer may cause instability issues

Accessibility

- PC, Mac, Linux
  - Data files will be uploaded from any location on your computer
- Smart Phone
  - Data files must be on a cloud platform connected to your phone account (e.g. dropbox)
Since LVS is hosted on a server, Shiny idle time-out settings may stop application when it is left inactive (it will grey out).

Solution: Click **reload** and re-upload your csv file
a. **Categorical** - non-numerical data with **two** values
   - yes - no; deletion - retention; perfective - imperfective

b. **Continuous** - numerical data
   - duration, age, chronological period

c. **Multinomial** - non-numerical data with **three or more** values
   - deletion - aspiration - retention

d. **Ordinal** - scale: currently not supported
Terminology Review

a. **Categorical** - non-numerical data with **two** values
   - yes - no; deletion - retention; perfective - imperfective
b. **Continuous** - numerical data
   - duration, age, chronological period
c. **Multinomial** - non-numerical data with **three or more** values
   - deletion - aspiration - retention
d. **Ordinal** - scale: currently not supported
Workshop Files

https://languagevariationsuite.wordpress.com/

1. **categoricaldata.csv**: categorical dependent - Labov New York 1966 study

2. **continuousdata.csv**: continuous dependent - Intervocalic /d/ in Caracas corpus (Díaz-Campos et al.)

3. **LVS web site**: www.languagevariationsuite.com
Language Variation Suite - Structure

1. **Data**
   - Upload file, data summary, adjust data, cross tabulation

2. **Visual Analysis**
   - Plotting, cluster classification

3. **Inferential Statistics**
   - Modeling, regression, conditional trees, random forest
Language Variation Suite - Structure

Language Variation Suite (LVS)

1. Data
   - Upload file, data summary, adjust data, cross tabulation

2. Visual Analysis
   - Plotting, cluster classification

3. Inferential Statistics
   - Modeling, regression, conditional trees, random forest
Upload file `movie_metadata.csv`
Excel Format

**1** Slow processing

or Step1: Upload Excel File

Choose EXCEL File (Will take long to upload)

Browse... No file selected

Step3: Select excel sheet

Type the name of your excel sheet (ex. sheet1)

Type here

**2** Requires the name of your excel sheet
To optimize speed - **Save as CSV** prior upload
Upload categoricaldata.csv

Step 1: Upload CSV File

Choose CSV File

Browse... categoricaldata.csv

Upload complete

- Header

Separator
- Comma
- Semicolon
- Tab

Quote
- None
- Double Quote
- Single Quote
The data content is imported as a table and allows for sorting columns.

<table>
<thead>
<tr>
<th>R.Use</th>
<th>Lexical.Item</th>
<th>Style</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
<tr>
<td>retention</td>
<td>Fourth</td>
<td>normal</td>
<td>Saks</td>
</tr>
</tbody>
</table>
Summary provides a quantitative summary for each variable, e.g. frequency count, mean, median.

<table>
<thead>
<tr>
<th>R.Use</th>
<th>Lexical.Item</th>
<th>Style</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>deletion: 499</td>
<td>Floor: 347</td>
<td>emphatic: 271</td>
<td>Kleins: 216</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Saks: 178</td>
</tr>
</tbody>
</table>
Data Structure

1. Total number of observations (rows)
2. Number of variables (columns)
3. Variable types
   - **Factor** - categorical values
   - **Num** - numeric values (0.95, 1.05)
   - **Int** - integer values (1, 2, 3)
Cross-tabulation examines the relationship between variables.
Cross-Tabulation: One Dependent and One Independent Variables

Which column contains your dependent variable?

- NULL
- R.Use
- Lexical.Item
- Style
- Store

Variable for Column

- NULL
- R.Use
- Lexical.Item
- Style
- Store
Cross-Tabulation Output

Raw frequency / Proportion by column / Proportion across row

<table>
<thead>
<tr>
<th></th>
<th>Floor/Col%</th>
<th>Row%</th>
<th>Fourth/Col%</th>
<th>Row%</th>
<th>RowSum</th>
</tr>
</thead>
<tbody>
<tr>
<td>deletion</td>
<td>204/59/41</td>
<td></td>
<td>295/77/59</td>
<td></td>
<td>499</td>
</tr>
<tr>
<td>retention</td>
<td>143/41/62</td>
<td></td>
<td>88/23/38</td>
<td></td>
<td>231</td>
</tr>
<tr>
<td>ColumnSum</td>
<td>347</td>
<td></td>
<td>383</td>
<td></td>
<td>730</td>
</tr>
</tbody>
</table>

Mosaic plot: R.Use and Lexical item

59 %

41 %

38 %

62 %
Language Variation Suite - Structure

1. Data
   - Upload file, data summary, adjust data, cross tabulation

2. Visual Analysis
   - Plotting, cluster classification

3. Inferential statistics
   - Modeling, regression, varbrul analysis, conditional trees, random forest
Shiny pages are fluid and reactive.

Continuous Dependent Variable (mean) at

Dependent

Action | Animation | Biography | Comedy | Drama

Independent

To adjust plot display, place cursor at the right edge of browser and stretch it to the right.
One Variable Plot
Two Variables Plot

Barplot: R.Use and Style

- Normal
- Emphatic

Count

Deletion

Retention

Style

Emphatic

Normal

Data Preparation

Language Variation Suite

Working with Data

Visual Analytics

Inferential Analysis

Mixed Effects

Appendix

References
Saving Plot

1. Right click on plot
2. Save Image As

- Open Image in New Tab
- Save Image As...
- Copy Image
- Copy Image Address
- Search Google for Image
Three Variables Plot

- Floor
  - Macy's
  - Kleins
  - Saks

- Fourth
  - Kleins
  - Saks
  - Macy's

Axes:
- X-axis: Deletion, Retention
- Y-axis: Prog
Cluster Plot

- Classification of data into **sub-groups** is based on **pairwise similarities**

- Groups are clustered in the form of a **tree-like dendrogram**

Options:

- One Variable Plot
- Two Variables Plot
- Three Variables Plot
- Cluster Plot
- Frequency Plot
Variable must contain at least three values to be clustered.

Your dependent variable
NULL

One independent variable for cluster
NULL

Your dependent variable
R.Use
NULL
Lexical.Item
Style
Store

One independent variable for cluster
Store
NULL
R.Use
Lexical.Item
Style
Store
Cluster Analysis for R.Use and Store

Saks (upper middle-class store), Macy’s (middle-class store), Kleins (working-class)
Inferential Statistics

TIME FOR THE
SERIOUS STUFF
Language Variation Suite - Structure

1. Data
   - Upload file, data summary, adjust data, cross tabulation

2. Visual Analysis
   - Plotting, cluster classification

3. Inferential statistics
   - Modeling, regression, varbrul analysis, conditional trees, random forest
How to Create a Regression Model

Step 1  **Modeling** - create a model with dependent and independent variables

Step 2  **Regression** - specify the type of regression (fixed, mixed) and type of dependent variable (binary, continuous, multinomial)

Step 3  **Stepwise Regression** - compare models (Log-likelihood, AIC, BIC)

Step 4  **Conditional Trees** - apply non-parametric tests to the model
Modeling

Select one dependent variable

Choose one column:
- R.Use
- NULL

Choose columns:
- Lexical.Item
- Style
- Store

Reference Level
- NULL
- deletion
- retention

base level
We are interested in RETENTION = Application
Regression Types

- Model
  - a.) Fixed effect
  - b.) Mixed effect - individual speaker/token variation (within group)

- Type of Dependent Variable
  - a.) Binary/categorical (only two values)
  - b.) Continuous (numeric)
  - c.) Multinomial - categorical with more than two values
Model Output

Call:
glm(formula = as.formula(paste(y, paste(listfactors, collapse = "i"), sep = "~")), family = binomial, data = plotData(), na.action = na.omit)

Deviance Residuals:
  Min      10     Median      3Q     Max
-1.4534  -0.8549  -0.5164   1.0493  2.4455

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6276   0.2596  -6.269  3.64e-10 ***
Lexical.ItemFourth  -0.9912   0.1749  -5.666   1.46e-08 ***
Stylenormal      -0.3197   0.1787  -1.789    0.0736 .
StoreMacys        1.8004   0.2615   6.884   5.81e-12 ***
StoreSaks         2.2564   0.2817   8.011  1.13e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 911.27  on 729  degrees of freedom
Residual deviance: 791.82  on 725  degrees of freedom
AIC: 801.82

Number of Fisher Scoring iterations: 5
Model Output

Call:
glm(formula = as.formula(paste(y, paste(listfactors, collapse = "|"), sep = "~")), family = binomial, data = plotData(), na.action = na.omit)

Deviance Residuals:
         Min       1Q   Median       3Q      Max
-1.4534  -0.8549  -0.5164   1.0493   2.4455

Coefficients:              Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.6276    0.2596  -6.269 3.64e-10 ***
Lexical.ItemFourth -0.9912    0.1749  -5.666 1.46e-08 ***
Storenormal       -0.3197    0.1787  -1.789  0.0736 .
StoreMacys         1.8004    0.2615   6.884 5.81e-12 ***
StoreSaks          2.2564    0.2817   8.011 1.13e-15 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 911.27  on 729  degrees of freedom
Residual deviance: 791.82  on 725  degrees of freedom
AIC: 801.82

Number of Fisher Scoring iterations: 5
### Interpretation

| Coefficient       | Estimate | Std. Error | z value | Pr(>|z|)   |
|-------------------|----------|------------|---------|------------|
| (Intercept)       | -1.6276  | 0.2596     | -6.269  | 3.64e-10   |
| Lexical.ItemFourth| -0.9912  | 0.1749     | -5.666  | 1.46e-08   |
| Stylenormal       | -0.3197  | 0.1787     | -1.789  | 0.0736     |
| StoreMacys        | 1.8004   | 0.2615     | 6.884   | 5.81e-12   |
| StoreSaks         | 2.2564   | 0.2817     | 8.011   | 1.13e-15   |

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

- **Deletion** is the reference value
- Positive coefficient - positive effect
- Negative coefficient - negative effect
Lexical item **Fourth** has a negative effect on retention compared to Floor and is significant.

**Normal** style has a slightly negative effect on retention but its coefficient is not significant.

**Macy's** and **Saks** have a positive and significant effect on retention. Saks (upper middle class store) is more significant than Macy’s (middle class store).

http://www.free-online-calculator-use.com/scientific-notation-converter.html
Lexical item **Fourth** has a negative effect on **retention** compared to Floor and is significant.

**Normal** style has a slightly negative effect on **retention** but its coefficient is not significant.

**Macy’s** and **Saks** have a positive and significant effect on **retention**. Saks (upper middle class store) is more significant than Macy’s (middle class store).
**Conditional tree**: a simple non-parametric regression analysis, commonly used in social and psychological studies

- Linear regression: all information is combined linearly
- Conditional tree regression: visual splitting to capture interaction between variables

Recursive splitting (tree branches)
1. The distribution of *was/were* is split in two groups by individuals.
2. The variant *were* occurs significantly more frequently with the first group.
1. Polarity is relevant to the second group of individuals.
2. The variant *were* occurs significantly more often with negative polarity.
1. **Affirmative Polarity** is conditioned by **Age**.

2. The variant **was** is produced significantly more often by Individuals of 46 and younger.
Conditional Tree

This method builds a tree by splitting on the values of your independent variables.

First, you need to select one dependent variable and independent variables in "Modeling" and "Regression" type.

Select Apply

- none
- apply

[1] "Dependent Variable: R.Use Independent Variables: Lexical.Item"
[2] "Dependent Variable: R.Use Independent Variables: Style"
1. **Store** is the most significant factor for R-use
   - **Kleins** (working class store) - more R-deletion

2. R-use in Macy’s and Saks is conditioned by **lexical item**:
   - **Floor** shows more R-retention than **Fourth**

3. **Style** is not significant
Random Forest

1. Variable importance for predictors
2. Robust technique with \textit{small n large p} data
3. All predictors considered jointly (allows for inclusion of correlated factors)
Random Forests determine which variables are important in the variable classification. See references for more details.

Select Apply
- none
- apply

Predictors to right of dashed vertical line are significant. If the number of variables is very large, forests can be run once with all the variables, then run again using only predictors from the first run.
Random Forest

1. Store is the most important predictor
2. Lexical Item is the second predictor
3. Style is irrelevant: close to zero and red dotted line (cut-off value).
Fixed and Mixed Models

Fixed Effects Model: All predictors are treated independent. Underlying assumption - no group-internal variation between speakers or tokens.

Mixed Effects Model: Allows for evaluation of individual- and group-level variation.
Fixed and Mixed Models

Fixed Regression Model - ignoring individual variations (speakers or words) may lead to Type I Error: “a chance effect is mistaken for a real difference between the populations”

Mixed Regression Model - prone to Type II Error: “if speaker variation is at a high level, we cannot discern small population effects without a large number of speakers” (Johnson 2009, 2015)
Mixed Effect Regression

**Mixed Model** = fixed effects + random effects

**Fixed-effect factor** - “repeatable and a small number of levels”

**Random-effect factor** - “a non-repeatable random sample from a larger population”  (Wieling 2012)

- walk, sleep, study, finish, eat, etc
- event verb, stative verb
- speaker1, speaker3, speaker3, etc
- male, female
Mixed Effect Regression

Mixed Model = fixed effects + random effects

Fixed-effect factor - “repeatable and a small number of levels”

Random-effect factor - “a non-repeatable random sample from a larger population” (Wieling 2012)

- walk, sleep, study, finish, eat, etc
- event verb, stative verb
- speaker1, speaker3, speaker3, etc
- male, female
Preparing for Mixed Model

1. Download *continuousdata.csv*

2. Upload this file on LVS
### Uploaded Dataset

<table>
<thead>
<tr>
<th>Dependent</th>
<th>Subjects</th>
<th>Sex</th>
<th>Age</th>
<th>Class</th>
<th>token</th>
<th>TokenFrequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.97</td>
<td>CA1HA</td>
<td>m</td>
<td>20-34</td>
<td>1</td>
<td>mudamos</td>
<td>45</td>
</tr>
<tr>
<td>0.98</td>
<td>CA1HA</td>
<td>m</td>
<td>20-34</td>
<td>1</td>
<td>edad</td>
<td>149</td>
</tr>
<tr>
<td>0.96</td>
<td>CA1HA</td>
<td>m</td>
<td>20-34</td>
<td>1</td>
<td>florida</td>
<td>20</td>
</tr>
<tr>
<td>0.95</td>
<td>CA1HA</td>
<td>m</td>
<td>20-34</td>
<td>1</td>
<td>edad</td>
<td>149</td>
</tr>
<tr>
<td>0.98</td>
<td>CA1HA</td>
<td>m</td>
<td>20-34</td>
<td>1</td>
<td>distanciados</td>
<td>2</td>
</tr>
<tr>
<td>0.98</td>
<td>CA1HA</td>
<td>m</td>
<td>20-34</td>
<td>1</td>
<td>cada</td>
<td>331</td>
</tr>
</tbody>
</table>
NULL when the dependent variable is continuous
Mixed Effect Modeling

Mixed Effects - group-internal variation

Select Random Variable for Mixed Model (ex. Subjects or Tokens)

Subjects, token

NULL
Dependent
Sex
Age
Class
PrecedingContext
FollowingContext
Regression Results

Scaled residuals:

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-4.7906</td>
<td>-0.4281</td>
<td>0.1440</td>
<td>0.6619</td>
<td>1.8390</td>
</tr>
</tbody>
</table>

Random effects:

- Groups Name: Variance: 7.436e-06, Std.Dev: 0.002727
- token (Intercept): Variance: 1.455e-04, Std.Dev: 0.012064
- Subjects (Intercept): Variance: 9.616e-04, Std.Dev: 0.031010

Number of obs: 517, groups: token, 301; Subjects, 12

Fixed effects:

- (Intercept): Estimate: 9.591e-01, Std.Err: 7.495e-03, df: 127.964, t value: 1.31e-14, Pr(>|t|): 0.00000000000000001
- Sexm: Estimate: 4.018e-03, Std.Err: 7.490e-03, df: 8.030e+00, t value: 0.537, Pr(>|t|): 0.6061
- Age35-54: Estimate: 6.121e-04, Std.Err: 9.167e-03, df: 8.007e+00, t value: 0.067, Pr(>|t|): 0.9484
- Age55+: Estimate: -1.643e-02, Std.Err: 9.172e-03, df: 8.024e+00, t value: -1.791, Pr(>|t|): 0.1110
- TokenFrequency: Estimate: 1.082e-05, Std.Err: 3.853e-06, df: 6.046e+00, t value: 2.807, Pr(>|t|): 0.00306

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
## Regression Results

<table>
<thead>
<tr>
<th>Scaled residuals:</th>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-4.7906</td>
<td>-0.4281</td>
<td>0.1440</td>
<td>0.6619</td>
<td>1.8390</td>
</tr>
</tbody>
</table>

Random effects:

- Groups Name | Variance | Std. Dev.
- token (Intercept) | 7.436e-06 | 0.002727
- Subjects (Intercept) | 1.455e-04 | 0.012064
- Residual | 9.616e-04 | 0.031010

Number of obs: 517, groups: token, 301; Subjects, 12

Fixed effects:

|                  | Estimate | Std. Error | df | t value | Pr(>|t|) |
|------------------|----------|------------|----|---------|---------|
| (Intercept)      | 9.591e-01 | 7.495e-03 | 127.964 | 1.31e-14 ** |
| Sexm             | 4.018e-03 | 7.490e-03 | 8.030e+00 | 0.537 | 0.6061 |
| Age35–54         | 6.121e-04 | 9.167e-03 | 8.007e+00 | 0.067 | 0.9484 |
| Age55+           | -1.643e-02 | 9.172e-03 | 8.024e+00 | -1.791 | 0.1110 |
| TokenFrequency   | 1.082e-05 | 3.853e-06 | 6.046e+00 | 2.807 | 0.0306 * |

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Interpretation - Random Effects

1. **Standard Deviation**: a measure of the variability for each random effect (speakers and tokens)

2. **Residual**: random variation that is not due to speakers or tokens (residual error)
Interpretation - Fixed Effects

1. **Estimate/coefficient**: reported in log-odds (negative or positive)

2. **P-value**: tells you if the level is significant
Frequency Plot

Visual Data Exploration

Select a column with tokens list

Select a column with token frequency data

one variable plot
two variables plot
three variables plot
cluster plot

Frequency Plot

Select a column with tokens list

Select a column with token frequency data

Sex
Age
Class
token
PrecedingContext
FollowingContext
totalDuration
TokenFrequency

Sex
Age
Class
token
PrecedingContext
FollowingContext
totalDuration
TokenFrequency
Frequency Plot

Select a number for top frequent words (ex. 10 top frequent words)

10
20
30
40
50
60
70
80
Appendix 1: Density

Number of bins can have a disproportionate effect on visualization

Histogram of Dependent
**Density**: a non-parametric model of the distribution of points based on a smooth density estimate

Number of bins in histogram (approximate):

20

Show density estimate

Appendix 2 - Data Modification

Table  Summary  Data Structure  Cross Tabulation  Frequency

Adjust Data
Adjust Data

- **Retain**: Select data subset
- **Exclude**: Exclude variables from a factor group
- **Recode**: Combine and rename variables
- **Change class**: Numeric $\rightarrow$ factor; factor $\rightarrow$ numeric
- **Transform**: Apply log transformation to a specific column

**ADJUSTED DATASET**:

- **Run** - to apply all above changes
- **Reset** - to reset to the original dataset
Exclude: Emphatic Style

Select RUN to start excluding

- NO
- RUN

Select a factor group

- Style
- NULL
- R.Use
- Lexical.Item
- Store

Which value(s) to exclude from your group?

- emphatic
- NULL
- normal
Adjusted Dataset

Select RUN to make changes or RESET to revert the original dataset.

- R.Use: deletion: 322
- Lexical.Item: Floor: 223, Fourth: 236
- Style: normal: 459
- Store: Kleins: 130, Macys: 224, Saks: 105
To revert to the original data, select **RESET**:
Appendix 3 - Model Comparison

### Instructions

Running Stepwise Analysis

- **Select Apply**
  - none
  - apply

Choose the best model

Only for fixed models with binary and continuous dependent variables.

1. Run Stepwise regression
2. Select the best model (Loglikelihood, AIC or BIC)
3. Return to MODELLING and select factors suggested by the best model

Stepwise selection function stepAIC - both directions: up and down.

Calculation is based on stepAIC from the package MASS. Loglikelihood

Thank you!

obscrivn@indiana.edu

What features/analyses would you like to see in LVS?


[7] https://daniellestolt.files.wordpress.com/2013/01/are-you-ready1.gif