Interactive Visual Data Analysis
Part One
Language Variation Suite

Olga Scrivner

Indiana University

Workshop in Methods
1. Introduce a web application for quantitative analysis: LVS

2. Develop practical skills

3. Understand and interpret advanced statistical models
Language Variation Suite

It is a Shiny web application originally 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)
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**
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
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>
Terminology Review

a. **Categorical** - non-numerical data with **two** values
   - yes - no; male - female

b. **Continuous** - numerical data
   - duration, age, year

c. **Multinomial** - non-numerical data with **three or more** values
   - regions, nationalities

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

a. **Categorical** - non-numerical data with **two** values
   - yes - no; male - female

b. **Continuous** - numerical data
   - duration, age, year

c. **Multinomial** - non-numerical data with **three or more** values
   - regions, nationalities

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

ARE YOU READY?
http://ssrc.indiana.edu/seminars/wim.shtml

1. **movie_metadata.csv**
   Simplified set from https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

2. **LVS web site**: https://languagevariationsuite.com
Movie Data

- Budget
- Director
- Actor 1
- Director facebook likes
- Actor 1 facebook likes
- Genre
- Year
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

**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

Language Variation Suite (LVS)

About  Demo  Data  Visual Analysis  RBRUL  Inferential Statistics

Upload *movie_metadata.csv*

Step 1: Upload CSV File

Choose CSV File

Browse...  movie_metadata.csv

Upload complete
The data content is imported as a table and allows for sorting columns.

<table>
<thead>
<tr>
<th>director_name</th>
<th>director_facebook_likes</th>
<th>actor_1_facebook_likes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joel Schumacher</td>
<td>541</td>
<td>920</td>
</tr>
<tr>
<td>Tim Burton</td>
<td>13000</td>
<td>920</td>
</tr>
<tr>
<td>Michael Winnick</td>
<td>155</td>
<td>981</td>
</tr>
<tr>
<td>Alec Asten</td>
<td>5</td>
<td>472</td>
</tr>
<tr>
<td>Jon Hess</td>
<td>29</td>
<td>683</td>
</tr>
<tr>
<td>John Stockwell</td>
<td>134</td>
<td>260000</td>
</tr>
</tbody>
</table>
Summary provides a quantitative summary for each variable, e.g. frequency count, mean, median.
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 Plot

Select Dependent Variable (Rows)
Which column contains your dependent variable?

| NULL | actor_1_facebook_likes | genres | actor_1_name | movie_title | cast_total_facebook_likes | budget | title_year | movie_facebook_likes |

Select One Independent Variable (Columns)
Variable for Column

| NULL | NULL | director_name | director_facebook_likes | actor_1_facebook_likes | genres | actor_1_name | movie_title | cast_total_facebook_likes |

18 / 44
Cross Tabulation Plot

Select Dependent Variable (Rows)
Which column contains your dependent variable?
- NULL
- actor_1_facebook_likes
- genres
- actor_1_name
- movie_title
- cast_total_facebook_likes
- budget
- title_year
- movie_facebook_likes

Select One Independent Variable (Columns)
Variable for Column
- NULL
- director_name
- director_facebook_likes
- actor_1_facebook_likes
- genres
- actor_1_name
- movie_title
- cast_total_facebook_likes

Continuous Dependent Variable (n=23)

<table>
<thead>
<tr>
<th>Genre</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>65673657.30</td>
</tr>
<tr>
<td>Animation</td>
<td>82867225.43</td>
</tr>
<tr>
<td>Biography</td>
<td>26536984.86</td>
</tr>
<tr>
<td>Comedy</td>
<td>27161439.39</td>
</tr>
<tr>
<td>Drama</td>
<td>22394283.83</td>
</tr>
</tbody>
</table>
Shiny pages are fluid and reactive.

Continuous Dependent Variable (mean) at

To adjust plot display, place cursor at the right edge of browser and stretch it to the right
Continuous Dependent Variable (mean) and genres

- Action
- Animation
- Biography
- Comedy
- Drama
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

Language Variation Suite (LVS)
Visual Analytics: “The science of analytical reasoning facilitated by visual interactive interfaces” (Thomas et al. 2005)
One Variable Plot

Language Variation Suite (LVS)

About  Demo  Data  Visual Analysis  RBRUL  Inferential Statistics

- One Variable Plot
- Two Variables Plot
- Three Variables Plot
- Cluster Plot
- Frequency Plot
One Variable Plot

Select one variable

- NULL
- director_name
- director_facebook_likes
- actor_1_facebook_likes
- genres
- actor_1_name
- movie_title
- cast_total_facebook_likes
Customizing Plot

- Action
- Animation
- Biography
- Comedy
- Drama
Customizing Plot

Choose colour
- blue
- red
- green
- grey

Enter the title for your plot
- My plot

Name for your x-axis
- Genre
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
Classification of data into **sub-groups** is based on **pairwise similarities**

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

Variable must contain at least three values to be clustered.

Your dependent variable
- NULL
- budget
- genres
- actor_1_name
- movie_title
- cast_total_facebook_likes

One independent variable for cluster
- NULL
- genres
- director_name
- director_facebook_likes
- actor_1_facebook_likes
- actor_1_name
- movie_title
- cast_total_facebook_likes
Group 1 *Animation, Biography* and Group 2 *Action, Drama, Comedy*
Inferential Statistics
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
Select one dependent variable

Choose one column:

- NULL
- director_facebook_likes
- actor_1_facebook_likes
- genres
- actor_1_name
- movie_title
- cast_total_facebook_likes
- budget
- title_year

Select one or more independent variables

Choose columns:

- director_name
- director_facebook_likes
- actor_1_facebook_likes
- genres
- actor_1_name
- movie_title
- cast_total_facebook_likes
- budget
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
Call:
lm(formula = as.formula(paste(y, paste(listfactors, collapse = "+"),
   sep = "~")), data = plotData(), na.action = na.omit)

Residuals:
          Min       Q1  Median       Q3       Max
-82717225 -21661439 -7880755  12838561  234326343

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    65673657    1421800   46.19  < 2e-16 ***
genresAnimation 17193568    3262681    5.27  1.46e-07 ***
genresBiography -39136672    3650154   -10.72  < 2e-16 ***
genresComedy    -38512218    1791456   -21.50  < 2e-16 ***
genresDrama     -43279374    1979184   -21.87  < 2e-16 ***

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

Residual standard error: 38620000 on 3081 degrees of freedom
Multiple R-squared:  0.2165,  Adjusted R-squared:  0.2154
F-statistic: 212.8 on 4 and 3081 DF,  p-value: < 2.2e-16
Call:
\texttt{lm(formula = as.formula(paste(y, paste(listfactors, collapse = "+"),
sep = "~")), data = plotData(), na.action = na.omit)}

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>-82717225</td>
<td>-21661439</td>
<td>-7880755</td>
<td>12838561</td>
<td>234326343</td>
</tr>
</tbody>
</table>

Coefficients:

|                      | Estimate | Std. Error | t value | Pr(>|t|)  |
|----------------------|----------|------------|---------|-----------|
| (Intercept)          | 65673657 | 1421800    | 46.19   | < 2e-16   *** |
| genresAnimation      | 17193568 | 3262681    | 5.27    | 1.46e-07 *** |
| genresBiography      | -39136672 | 3650154   | -10.72  | < 2e-16   *** |
| genresComedy         | -38512218 | 1791456   | -21.50  | < 2e-16   *** |
| genresDrama          | -43279374 | 1979184   | -21.87  | < 2e-16   *** |

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

Residual standard error: 38620000 on 3081 degrees of freedom
Multiple R-squared: 0.2165, Adjusted R-squared: 0.2154
F-statistic: 212.8 on 4 and 3081 DF, p-value: < 2.2e-16
Genre **Action** is the reference value

- Positive coefficient - positive effect
- Negative coefficient - negative effect

http://www.free-online-calculator-use.com/scientific-notation-converter.html
Genre **Action** is the reference value

- Positive coefficient - positive effect
- Negative coefficient - negative effect

http://www.free-online-calculator-use.com/scientific-notation-converter.html
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)
Conditional Tree

Select Apply
- none
- apply

Conditional Tree

Select Apply
- none
- apply


Conditional Inference Tree
1. **Genre** is the significant factor for budget.

2. Budget distribution is split in two groups:
   - Action and Animation
   - Biography, Comedy and Drama

3. Budget is significantly higher for Animation and Action.
Random Forest

1. Variable importance for predictors
2. Robust technique with small $n$ large $p$ data
3. All predictors considered jointly (allows for inclusion of correlated factors)
Let’s add more factors!

- Return to **Modeling**

- Add independent factors: *director facebook likes, actor 1 facebook likes, title year*

```
Choose columns:
genres  director_facebook_likes
actor_1_facebook_likes  title_year
```

```
Select one dependent variable
Choose one column:
budget
```

The same dependent variable
Random Forest

Select Apply

- none
- apply
Random Forest

- Genre is the most important predictor for this model.
- Close to zero or red-dotted line (cut off values) - irrelevant for this model
Let’s Have a Short Break


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