Logistic Regression Software for Speech Perception Data

© Geoffrey Stewart Morrison http://geoff-morrison.net

Release 2009-03-13	– fixed problem with hyp_pairs when compareG2 set to false
Release 2009-03-07	– plot works over any range of values, default values are correctly assigned if input fields are empty
Release 2009-03-04	– first release

– This software is provided as-is without any guarantee that it will work. I'm willing to give some end-user support, but please read this document thoroughly and follow all the instructions before contacting me.

– This software is provided free-of-charge for academic not-for-profit research. Please include appropriate acknowledgments in published papers.

– Commercial use of this software in whole or in part is strictly prohibited except with the prior consent of the copyright holder.

– The packaged data were kindly provided by Maria V. Kondaurova © 2008.

– The data are provided free-of-charge for academic not-for-profit research. Please include appropriate acknowledgments in published papers.

– Commercial use of the data in whole or in part is strictly prohibited except with the prior consent of the copyright holder.

Description

This software runs some logistic regression modelling tasks on speech perception data, namely:

- – build simple models (but not diphone biassed)
- – ΔG^2 significance tests for pairs of nested models
- – probability surface plots

For an introduction to this type of logistic regression modelling see:

> Morrison, G. S. (2007). Logistic regression modelling for first- and second-language perception data. In M. J. Solé, P. Prieto, & J. Mascaró (Eds.), *Segmental and prosodic issues in Romance phonology* (pp. 219–236). Amsterdam: John Benjamins.

This software runs the logistic regression analyses reported in:

> Morrison, G. S., & Kondaurova, M. V. (2009). Analysis of categorical response data: Use logistic regression rather than endpoint-difference scores or discriminant analysis (L). *Journal of the Acoustical Society of America, 126* (5), xxx. DOI: 10.1121/1.3216917

Requirements

Compiled version

– 32 bit Windows

– Matlab Compiler Runtime version 7.9 (this can be downloaded from my ftp site; however, you will need to ask me for the username and password - the licencing agreement doesn't allow it to be posted to a generally accessible location on the internet).

Matlab code version

– Matlab installed and licenced

– Statistics Toolbox

Software has been successfully tested in Matlab R2008b running under Windows XP.

Installing and Running

Compiled version

– Install Matlab Compiler Runtime version 7.9

– Place Logistic_Regression_pkg.exe into a folder of your choice and double click to install.

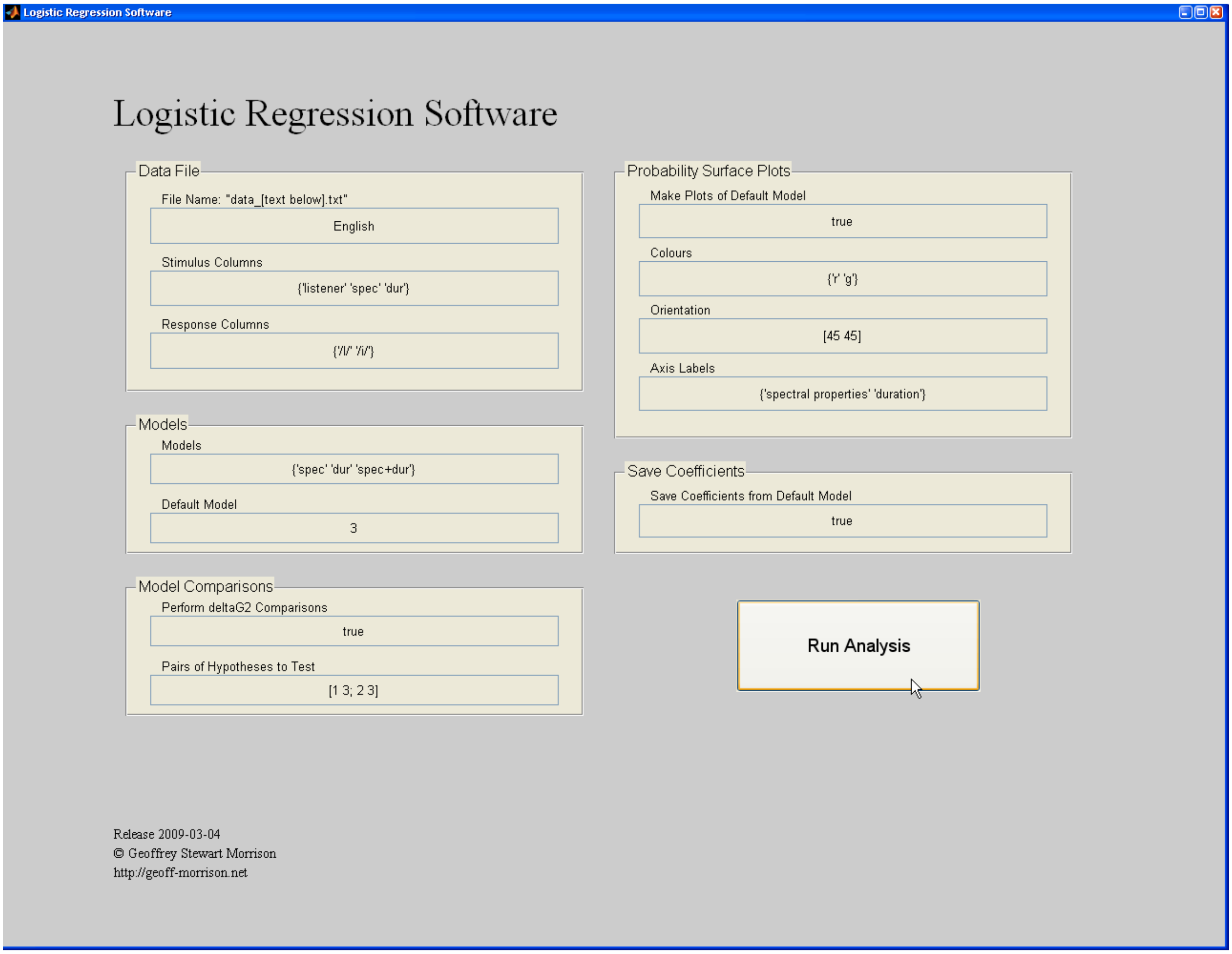
– To start the software, double click on Logistic_Regression.exe.

Matlab code version

– Unzip the archive Logistic_Regression.zip

– Run logistic_regression.m from Matlab

The following start screen will appear



Logistic Regression Software

Data File

File Name: "data_[text below].txt"

English

Stimulus Columns

{'listener' 'spec' 'dur'}

Response Columns

{'/I/' '/i/'}

Models

Models

{'spec' 'dur' 'spec+dur'}

Default Model

3

Model Comparisons

Perform deltaG2 Comparisons

true

Pairs of Hypotheses to Test

[1 3; 2 3]

Probability Surface Plots

Make Plots of Default Model

true

Colours

{'r' 'g'}

Orientation

[45 45]

Axis Labels

{'spectral properties' 'duration'}

Save Coefficients

Save Coefficients from Default Model

true

Release 2009-03-04
© Geoffrey Stewart Morrison
http://geoff-morrison.net

– Input arguments can be manually edited at the start screen, e.g., change “English” to “Spanish” and leave everything else the same to run the same analyses on the Spanish data set.

– A description of the input arguments is provided below under *Input Argument File*

– I haven’t implemented checks on the validity of the input, so if there are errors in your input arguments or your data file, then the software will crash or give inappropriate output.

Data Files

– The data to be analysed should be in a text file named data_XXX.txt where XXX is a label which distinguished this data set from other data sets, e.g., data_English.txt, data_Russian.txt, data_Spanish.txt contain English, Russian, and Spanish listeners’ responses to the same stimuli. Data files should be in the same folder as Logistic_Regression.exe / logistic_regression.m.

– The file should contain columns of **numbers only** with each column separated by a **tab**.

list	spec	dur	/I/	/i/	← this line is NOT part of the data file
1	0	0	2	8	
1	0	1	4	6	
1	0	2	1	9	
1	0	3	1	9	
1	0	4	2	8	
1	0	5	4	6	
1	0	6	3	7	
1	0	7	8	2	
1	0	8	5	5	
1	1	0	0	10	
1	1	1	1	9	
1	1	2	2	8	
...	...	...	...	...	
2	0	0	0	10	
2	0	1	0	10	
2	0	2	1	9	
2	0	3	0	10	
...	...	...	...	...	

– The first column must contain numeric ids for listeners (if you only have one listener, fill the whole of this column with the same number)

– The next columns should contain stimulus properties, e.g., spectral values in column 2 and duration values in column 3 (in the example data the stimulus values are integers but they can be any real number). There should be as many columns as there are stimulus dimensions.

– The final columns should contain counts of the listeners’ responses to the stimuli, e.g., counts of /ɪ/ responses in column 4 and counts of /i/ responses in column 5. There should be as many columns as there are response categories. An alternative to specifying response counts for each stimulus is to enter proportions. Another alternative is to repeat the listener and stimulus information for every individual response and code the category chosen as 1 and all the other categories as 0.

Input Argument File

– A text file Logistic_Regression_input_arguments.txt must be included in the same folder as Logistic_Regression.exe / logistic_regression.m. This file specifies the default options which will appear on the start screen.

– The file must have **two columns** separated by a **tab**.

```
lang_label      English
stimcols        {'listener' 'spec' 'dur'}
respcols        {'/I/' '/i/'}
models          {'spec' 'dur' 'spec+dur'}
default_model   3
save_coefs      true
compareG2       true
hyp_pairs       [1 3; 2 3]
make_plots      true
colours         {'r' 'g'}
orientation     [45 45]
axis_labels     {'spectral properties' 'duration'}
```

– Do **NOT** make any changes to the **first column**.

– Edit the **second column** according to the design of your logistic regression analysis:

lang_label The label of the second part of the data file you want to analyse

stimcols The labels for the listener and stimulus columns, e.g., {'listener' 'spec' 'dur'}. The first column must be 'listener'. There must be a label for every column corresponding to every stimulus dimension.

respcols The labels for the response columns, e.g., {'/I/' '/i/'}. There must be a label for every column corresponding to every response category.

models Specification of the models to fit, e.g., {'spec' 'dur' 'spec+dur' 'spec+dur+spec*dur'}. all models include a bias (intercept) coefficient, only the stimulus-tuned coefficients are specified. {'spec'} is a model which include a bias coefficient and a spectrally-tuned coefficient. {'spec+dur'} is a model which include a bias coefficient, a spectrally-tuned coefficient, and a duration-tuned coefficient. {'spec+dur+spec*dur'} is a model which include a bias coefficient, a spectrally-tuned coefficient, a duration-tuned coefficient, and a spectral-by-duration-interaction-tuned coefficient..

default_model Specifies the model to plot and the model for which coefficient values will be saved (if these options are selected). Must be a positive integer no greater than the number of models specified. If make_plots is set to true, the default model must have two stimulus-tuned coefficients.

save_coefs Must have the value true or false. If set to true, a file coef_XXX.txt will be saved containing deviation-from-mean logistic regression coefficient values. The first column indicates the listener, then there is one column for each response category. The first row for each listener contains the bias coefficient values, and the remaining columns contain the stimulus-tuned coefficient values in the order specified in the default model, e.g., if the default model is {'spec+dur'}, then row 1: bias, row 2: spec, row 3: dur.

compareG2 Must have the value true or false. If set to true, ΔG^2 significance tests for pairs of nested models are conducted for each listener. The pairs of models are specified by hyp_pairs.

hyp_pairs Specification of the models to compare in the ΔG^2 significance tests, e.g., [1 3; 2 3] sequentially compares model 1 spec against model 3 spec+dur, and model 2 dur against model 3 spec+dur. The first model must be nested within the second model, i.e., the first model must contain a subset of the terms included in the second model. A significant result may indicate that the extra stimulus term in the larger model has an effect on the listener’s response pattern, e.g., if a significant result of obtained for model 3 vs model 1, then this may indicate that the listener is responding to duration cues.

make_plots Must have the value true or false. If set to true, probability plots will be drawn. If set to true, the default_model must have two stimulus-tuned coefficients, e.g., {'spec+dur'}. As is, this software will not plot the model {'spec+dur+spec*dur'} (which is a model allowing for curved boundaries), but it would be possible to alter the Matlab code to do this.

colours

Specifies the colours to use in the probability surface plots, e.g., {'r' 'g'} will draw the /ɪ/ surface in red and the /i/ surface in green. Colours allowed are: **r**ed, **g**reen, **b**lue, **y**ellow, **m**agenta, **c**yan, **w**hite, blac**k**. There must be at least as many colours specified as there are response categories.

orientation

Specifies the azimuth and elevation from which the probability surface plots are viewed, e.g. [45 45]. Azimuth is horizontal rotation about the z-axis measured in degrees anticlockwise from the y-axis. Elevation is measured in degrees above the x-y-plane.

axis_labels

Labels for the x and y axes of the probability surface plots, e.g., {'spectral properties' 'duration'}.

Output Files

– The file Logistic_Regression_output_XXX.txt will contain the result of model fitting and, if requested, ΔG^2 significance tests for model comparisons. In the list of coefficient values, the first row for each listener contains the bias coefficient values, and the remaining columns contain the stimulus-tuned coefficient values in the order specified in the specified model, e.g., for model spec+dur: row 1: bias, row 2: spec, row 3: dur. Coefficient values are given using deviation-from-mean coding.

```
LOGISTIC REGRESSION ANALYSIS
------------------------
------------------------
LISTENER: 01
------------------------
MODELS:
------------------------
spec
G2: 264            df: 79
/I/       /i/
-0.385    0.385
-0.004    0.004
------------------------
dur
G2: 108            df: 79
/I/       /i/
-1.331    1.331
0.207     -0.207
------------------------
spec+dur
G2: 108            df: 78
/I/       /i/
-1.312    1.312
```

```
-0.005    0.005
0.207     -0.207
------------------------
MODEL COMPARISONS:
------------------------
spec v spec+dur
deltaG2:          156.72
delta_df:         1
p:                0.0000
------------------------
dur v spec+dur
deltaG2:          0.09
delta_df:         1
p:                0.7692
```

– If requested, the file coefs_XXX.txt will contain coefficients for the default_model. See save_coefs above.

Output Graphics

– If requested, see make_plots, a probability surface plot will be drawn for the model fitted to each listener. Several tools are available for manipulating the graphics, including a 3D rotation tool.

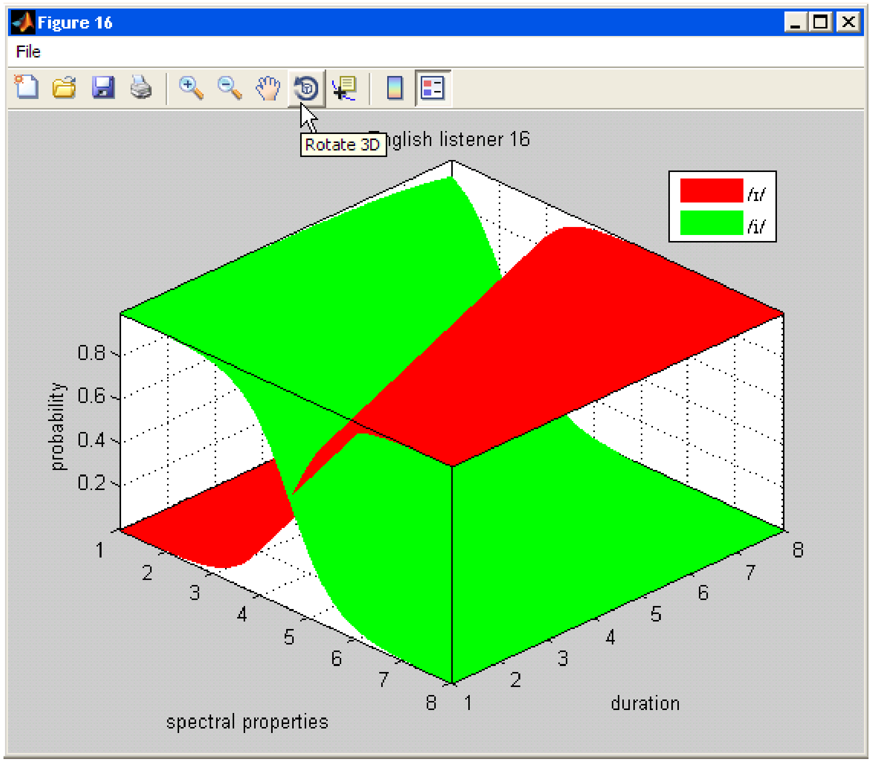


Figure 16
File
Rotate 3D
English listener 16
🟥 /ɪ/
🟩 /i/
probability
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
8
spectral properties
1
2
3
4
5
6
7
8
duration