Analysing linguistic variation with Rbrul – a step-by-step guide

Dr Agata Daleszynska
Contact email address: agata.daleszynska@gmail.com

1. Rbrul – what does it do?

The type of statistical analysis performed in Rbrul is called **multiple logistic regression.** In multiple logistic regression the probability of one outcome is modelled as a function of the linear combination of several explanatory variables (basically, it establishes the relationship between a dependent variable and multiple independent variables). Please consult the manual for more details: (http://www.danielezrajohnson.com/Rbrul_manual.html)

Advantages of Rbrul:

- Easy to use (no need to type commands, you always select options from the menu),
- Easy input from Excel
- Allows for mixed effects modelling
- Allows for modelling continuous variables
- Helpful in handling interactions

Disadvantages of Rbrul:

- I'll emphasise one which is crucial - It allows for only one type of analysis (logistic regression). There are plenty of other tests and types of analyses which you might apply in variation analysis through R or SPSS.

Below, I will conduct a step-by-step analysis of variation within /t,d/ deletion in Bequia Creole (Meyerhoff and Walker 2007; Daleszynska 2011). For a detailed discussion of the variable cf. Guy (1980; 1991), and for /t,d/ deletion in Caribbean English Creoles, cf. Patrick (1991; 1999) or (http://privatewww.essex.ac.uk/~patrickp/TDintro.htm)

2. Preparing your data for Rbrul

a) Download R (http://cran.r-project.org/)
b) Open R and type in (or paste) the following commands following the > arrow

> source(http://www.danielezrajohnson.com/Rbrul.R)

After a while, another > arrow should appear where you type in > rbrul()

Once you do this, your Rbrul 'main page' should look as follows:

```
Failed with error:  '(converted from warning) there is no package called 'lme4''
Installing lme4 package...

trying URL 'http://cran.ma.imperial.ac.uk/bin/windows/contrib/2.14/lme4_0.999375-42.zip'
Content type 'application/zip' length 1302757 bytes (1.2 Mb)
opened URL
downloaded 1.2 Mb

package 'lme4' successfully unpacked and MD5 sums checked

The downloaded packages are in
        C:\Users\Jon\AppData\Local\Temp\Rtmpem4aeb\downloaded_packages

No data loaded.

MAIN MENU
1-load/save data
9-reset 0-exit
1: 1
```

Tip: Before you load your data file, there are several things you should check:

- Make sure your data file has no empty cells (every row of your spreadsheet needs to be a token)
- One column needs to contain your dependent variable (a response)
- The other columns must contain the independent variables (predictors)
- Save your spreadsheet as .csv (comma separated value). **Rbrul won't load .xls files!**

3. Loading data into Rbrul

To load your data into Rbrul select 1 from the menu. Rbrul will then ask you what separates the columns in the file we want to load. Select c for commas (since our file is in a csv format).

```
MAIN MENU
1-load/save data
9-reset 0-exit
1: 1

No data loaded.

What separates the columns in the data file to open?
(c-commas s-semicolons t-tabs tf-token file)
Press Enter to exit, keeping current data file, if any.
1: c
```

Next, select the data file from the location on your computer. Once loaded the data should look as follows:

```
Current data file is: C:\Users\Jon\Downloads\Rbrul workshop data.csv

Current data structure:
Village (factor with 3 values): Mount Pleasant Paget Farm Hamilton
Speaker (factor with 30 values): 108 10 107 9 H12 ...
Sex (factor with 2 values): F M
Age (factor with 2 values): Older Younger
Word (factor with 157 values): left build dust told send ...
T.d.deleted.or.not (factor with 2 values): retained deleted
Preceding.phon.segment (factor with 3 values): fricative Lateral nasal
Following.phon.segment (factor with 9 values): glide nasal pause sibilant stop ...
Grammatical.class (factor with 5 values): Irregular devoicing verbs monomorph negative contraction regular nonsyllabic semi-weak
Token (factor with 998 values): , the old guy- that- those land was[ left], handed over to the family. So Uncle-Sallo buy it and
Total tokens: 999

MAIN MENU
1-load/save data 2-adjust data
4-crosstabs 5-modeling 6-plotting
8-restore data 9-reset 0-exit
1:
```

- Tip: Remember, once you type/select something into Rbrul, you can't untype or undo it! So think twice before you select it, otherwise you have to restart your model and start from scratch.

4. Main Rbrul Menu

Below your data, you will see several options within the main menu. Here's what they're for (partially adopted from the Rbrul manual):

Load/save data - allows you to save the current data to a file, and to load a data file into R
Adjust data – here's where you recode your data, e.g. collapse factors, combine predictors
Crosstabs – here's where you cross-tabulate your data (see below)
Modelling – here's where you select your model and run the analysis
Plotting – here's where the graphics are created
Restore data – resets your model
Reset – resets Rbrul
Exit – exits to R

5. Cross-tabulations

- Cross-tabulations should be an important part of your data analysis for several reasons. First, they allow you to observe unbalanced distributions of your data, and secondly they allow you to trace interactions.

- Cross-tabulations are often referred to also as contingency tables and report the frequency counts of two or more categorical variables in order to show a proportional relationship between them (Tagliamonte, 2006).

- In Rbrul these are reported through "counts" (a total distribution of tokens in each cell), or mean, which shows the average value of the response.

Let's try to cross-tabulate two categories: t/d deletion (our response) and the preceding phonological segment.

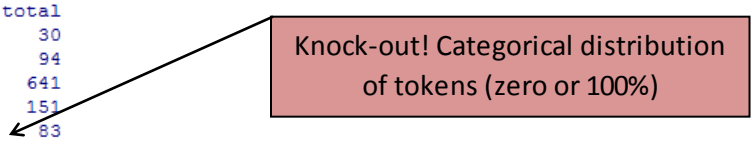
- From the main menu select 4 – crosstabs
- Factor for columns – I select 6 for t/d deletion
- Cross-tabs for rows – I select 7 for preceding segment

[We will skip 'pages' for now. (It's a useful function which allows you to cross-tab more than two independent variables at a time.)]

- Hit Enter
- Now we have to select whether we want to view the cross-tabulation as a mean or just raw data. I want to just see the raw data (counts) so I hit Enter.

```
MAIN MENU
1-load/save data 2-adjust data
4-crosstabs 5-modeling 6-plotting
8-restore data 9-reset 0-exit
1: 4
Cross-tab: factor for columns?  (1-Village 2-Speaker 3-Sex 4-Age 5-Word 6-T.d.deleted.or.not 7-Preceding.phor
1: 6
Cross-tab: factor for rows? (1-Village 2-Speaker 3-Sex 4-Age 5-Word 7-Preceding.phon.segment 8-Following.phor
1: 7
Cross-tab: factor for 'pages'? (1-Village 2-Speaker 3-Sex 4-Age 5-Word 8-Following.phon.segment 9-Grammatical
1:
Cross-tab: variable for cells? (1-response proportion/mean Enter-counts)
1:
counts
                      T.d.deleted.or.not
Preceding.phon.segment deleted retained total
             fricative      18       12    30
             Lateral        66       28    94
             nasal         405      236   641
             sibilant      132       19   151
             stop           83        0    83
             total         704      295   999

Current data file is: C:\Users\Jon\Downloads\Rbrul workshop data (1).csv
```



This cross-tabulation revealed important information, namely, that there are no tokens in my data, where t/d would be retained if preceded by a stop. This is extremely important for further qualitative and quantitative analysis:

- Qualitatively, it's telling me that in Bequia preceding stops are the most favourable environment for deletion to occur.
- Secondly, leaving out this knockout could potentially skew my overall result. I need to make a decision what to do with this portion of the data. Here're some options:

  - You could remove such tokens from the data set. This is perhaps the most straightforward scenario, although you need to remember that this way you're excluding data, which could be valuable! This is not ideal if your model is not "data-heavy".

- Also, remember that just because you decide to exclude these tokens, it doesn't mean they don't exist. It's an important result, and you could report on them in your final write-up!
- You could collapse the KO group with another predictor. BUT, you need to have a very good qualitative (conceptual, linguistic) reason for doing so. E.g. the predictors are somehow related. Then you could create a combined factor group.

5a. Cross-tabulating with 'pages'

This option allows for a more detailed insight into how your data is distributed across categories. Let's say I want to see how /t,d/ deletion is distributed across speakers of different age groups (older vs. younger). Here's what I do:

- From the main menu - 4 for Crosstabs
- Factor for columns - 6 for /t,d/ deletion
- Factor for rows - 1 for Village
- Factor for 'pages' - 4 for Age

```
MAIN MENU
1-load/save data 2-adjust data
4-crosstabs 5-modeling 6-plotting
8-restore data 9-reset 0-exit
1: 4
Cross-tab: factor for columns?  (1-Village 2-Speaker 3-Sex 4-Age 5-Word 6-T.d.deleted.or.not
1: 6
Cross-tab: factor for rows? (1-Village 2-Speaker 3-Sex 4-Age 5-Word 7-Preceding.phon.segment
1: 1
Cross-tab: factor for 'pages'? (2-Speaker 3-Sex 4-Age 5-Word 7-Preceding.phon.segment 8-Foll
1: 4
Cross-tab: variable for cells? (1-proportion/mean of T.d.deleted.or.not 2-change response va
1:
counts
, , Age = Older

               T.d.deleted.or.not
Village         retained deleted total
  Hamilton            10     143   153
  Mount Pleasant      33     166   199
  Paget Farm           7     177   184
  total               50     486   536

, , Age = Younger

               T.d.deleted.or.not
Village         retained deleted total
  Hamilton             1      24    25
  Mount Pleasant       6      45    51
  Paget Farm           3      56    59
  total               10     125   135

, , Age = total

               T.d.deleted.or.not
Village         retained deleted total
  Hamilton            11     167   178
  Mount Pleasant      39     211   250
  Paget Farm          10     233   243
  total               60     611   671
```

Distribution of /t,d/ deletion among older speakers in the 3 villages

Distribution of /t,d/ deletion among younger speakers in the 3 villages

6. Combining/removing factors

Through cross-tabulating grammatical class and /t,d/ deletion I have decided to remove preceding stops from the data set. Here's how I do it.

- From the main menu select 2 – adjust data, then 3 – exclude.
- Rbrul now asks which factor group I want to exclude from. I select 7 – grammatical class.
- Then I select 5 for preceding stops.

The bad thing is that I just got rid of 83 tokens, but the good thing is that my model is more likely to be statistically accurate.

> ➢ Tip: It's good practice to cross-tabulate all your factor groups together to account for knockouts, interactions and unevenly distributed data.

```
MAIN MENU
1-load/save data 2-adjust data
4-crosstabs 5-modeling 6-plotting
8-restore data 9-reset 0-exit
1: 2

ADJUSTING MENU
1-change class 2-rename 3-exclude 4-retain 5-recode
6-relevel 7-center/transform 8-count 9-main menu 0-exit
10-make interaction group
1: 3
Factor group to exclude from? (1-Village 2-Speaker 3-Sex 4-Age 5-Word 6-T.d.deleted.or.not 7-Preceding.phon.segment 8-Following.phon.segment 9-Grammatical.class 10-Toke
1: 9
Factors to exclude from Grammatical.class? (1-Irregular devoicing verbs 2-monomorph 3-negative contraction 4-regular nonsyllabic 5-semi-weak past verbs)
1: 3
2:
```

7. Modelling

We will come back to cross-tabulations soon…For now, let's try to build our model.

- From the main menu select 5 – modelling.

First, we want to choose which variables we want to include in the model.

- So select 1.

Now we get to choose our response (dependent variable).

- I select 6 for t/d deletion.
- My response is binary (deletion vs. retention of /t,d/), so I just hit enter, but if your response is continuous, select 1.

Now, Rbrul is asking which of the variants included in your envelope of variation you want to make your application value. If you have only 2 factors, the decision is easy. For example,

I select 1 for deleted /t,d/. This means Rbrul will be analysing deletion vs. retention. But I could also select 2 in which case Rbrul would analyse retention vs. deletion.

- Tip: If you have more than 2 factors it's an important qualitative and quantitative decision whether you want to analyse e.g. X vs. Y and Z, or X vs. Y, or X vs. Z. This is particularly important for variables beyond the level of phonology. Your application value and non-application value(s) have to fall within the same envelope of variation (they need to carry semantically and grammatically equivalent meanings). For more details, cf. Lavandera 1978; Torres-Cacoullos 2011)

After I selected deletion as my application value, I now get to choose the predictors. This is also an important step. You have to know your data well to know which predictors can be analysed together, and which ones shouldn't.

- Tip: Generally, predictors shouldn't be included in the same run if they: are collinear (that is, they are highly correlated), if they interact (they have a strong effect on each other), or if there is another conceptual reason for not analysing them together.

The (non-interactive) predictors I select are: Village, Age, Word, Preceding Segment, Following Segment, Grammatical class:

```
MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-trim 7-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1: 1
Choose response (dependent variable) by number (1-Village 2-Speaker 3-Sex 4-Age 5-Word 6-T.d.deleted.or.not 7-Preceding.phon.segment 8-Following.phon.segment 9-Grammatical.class 10-To$
1: 6
Type of response? (1-continuous Enter-binary)
1:
Choose application value(s) by number?  (1-deleted 2-retained)
1: 1
Choose predictors (independent variables) by number (1-Village 2-Speaker 3-Sex 4-Age 5-Word 7-Preceding.phon.segment 8-Following.phon.segment 9-Grammatical.class 10-Token)
1: 1
2: 4
3: 5
4: 7
5: 8
6: 9
7:
Are any predictors continuous? (1-Village 4-Age 5-Word 7-Preceding.phon.segment 8-Following.phon.segment 9-Grammatical.class Enter-none)
1:
Any grouping factors (random effects)? (1-Village 4-Age 5-Word 7-Preceding.phon.segment 8-Following.phon.segment 9-Grammatical.class Enter-none)
1: 5
2:
```

- In the end Rbrul asks if any of my predictors are continuous. I hit Enter for none but if any of yours are, select them here.
- Similarly, Rbrul asks if any of my predictors are random. Here, I select Word as a random effect. I'll explain random effects shortly…
- Next, you can select 2 categories which you think might interact. I leave out this option for now. If you know your data well and you've cross-tabulated your data, you don't really need this option. But if some of the predictors interact, and you want to include both of them in the model, Rbrul will ask you: Consider a pairwise interaction of fixed effects? Choose them from the list.

8. Mixed effects

Above, I have selected Word as a random effect. Why did I do that?

A few words about mixed-effects modelling:

Mixed models make a distinction between two types of factors that can affect a response:

- **Fixed effects** - predictors with a fairly small number of possible factors, which are usually replicable in a further study, and…
- **Random effects** - factors drawn from larger populations, unlikely to be replicable, such as individual speakers or words (Johnson, 2009: 365; Baayen, 2008: 241)

While fixed-effects factors are modelled by means of contrasts, random effects are modelled as random variables with a mean of zero and unknown variance (Baayen, 2008: 242). In my analysis of /t,d/ deletion, the individual word factor group consists of unbalanced tokens which are not exhaustively sampled across the dataset. For example, some words represent very high or low rates of inflection. However, this doesn't mean that this variability should be automatically excluded from the model. Rather, it should be controlled in the testing of fixed effects. Especially if it is OTHER factors which are at the centre of the analysis. Therefore, including the word class as a random effect provides a good opportunity to embrace and model this variation, while at the same time removing the individual level of variance from the outcome in testing for the effect of other independent variables.

Speaker is another predictor where mixed modelling could be applied. This doesn't mean that we are levelling out individual variation! Rather, it means that we are accounting for it and including it in the model. You can still see the rate of individual variation by tracing the intercept of each individual in the sample. This way you will recognise which speakers contribute most or least strongly to the variation in question.

9. Stepwise regression

Back to our analysis…
Now that we have built up the model, we can finally do some testing.

- From the Modelling Menu I select 5 for step-up/step-down. The program will start conducting stepwise regression. What exactly is stepwise regression?

Basically, the program evaluates null hypothesis (which assumes that there is no relationship between the factors included, and that variation is random, by trying to find the best model including the predictors. Two different, let's call them, 'mini-analyses' are conducted by Rbrul:

- *Step up* – Rbrul adds predictors one at a time, starting with the one that has the greatest effect on the response and repeating the process until no more significant variables can be added

- *Step down* – the program fits the full model and then removes those predictors which are not significant. If both "step up" and "step down" result in the same model – then the best model has been achieved and the two runs match

Here's the preliminary result of my analysis:

```
All remaining predictors are significant, best model from last step is Run 2

BEST STEP-DOWN MODEL IS WITH Word [random] and Village (3.68e-07) + Grammatical.class (1.71e-05) + Following.phon.segment (0.000251) + Preceding.phon.segment
[p-values dropping from full model]

$Village
        factor logodds tokens deleted/deleted+retained centered factor weight
      Hamilton   0.462    428                    0.79                 0.614
 Mount Pleasant  -0.462    488                    0.58                 0.386

$Preceding.phon.segment
    factor logodds tokens deleted/deleted+retained centered factor weight
  sibilant   1.191    151                   0.874                 0.767
   Lateral   0.014     94                   0.702                 0.504
     nasal  -0.104    641                   0.632                 0.474
 fricative  -1.101     30                   0.600                  0.25

$Following.phon.segment
    factor logodds tokens deleted/deleted+retained centered factor weight
      stop   0.812    162                   0.815                 0.693
     glide   0.716    165                   0.770                 0.672
 fricative   0.242     46                   0.696                  0.56
     pause   0.064     86                   0.640                 0.516
     nasal  -0.016     98                   0.714                 0.496
     vowel  -0.234    250                   0.580                 0.442
    rhotic  -0.440     15                   0.600                 0.392
  sibilant  -0.485     64                   0.547                 0.381
   lateral  -0.659     30                   0.533                 0.341

$Grammatical.class
                factor logodds tokens deleted/deleted+retained centered factor weight
       regular nonsyllabic   1.127    140                   0.879                 0.755
      semi-weak past verbs   0.812     19                   0.684                 0.692
      negative contraction   0.086    254                   0.740                 0.521
                monomorph  -0.551    442                   0.629                 0.366
 Irregular devoicing verbs  -1.474     61                   0.311                 0.186

$Word
       random logodds tokens deleted/deleted+retained centered factor weight std dev
       ground   1.201      8                   1.000                 0.771   0.845
       around   1.132     18                   0.889                 0.759   0.845
```

10. But what do these numbers mean?

As you can see, Rbrul reports a lot of information. Some of this data will be crucial for interpreting your results, other is less crucial but good to know. Let's focus on these result step by step.

a) First let's have a look at the top bar – BEST STEP UP STEP DOWN MODEL IS...

Here, Rbrul reports which factor groups are statistically significant, and in which order.

- In my data, Village membership is by far the most significant predictor. That is, it matters whether you're from Hamilton or Mount Pleasant for how high your rates of /t,d/ deletion are.

- Grammatical class is also strongly significant which goes in line with previous studies on /t,d/ deletion in Caribbean English Creoles (e.g. Patrick 1991). This makes me happy because it means that my results fit in the overall trend for this variable.

- Finally, the preceding and following segments are also important which again goes in line with previous studies on this variable.

b) Now let's have a look at the contribution of each of these categories in turn.

As an example, let's have a look at Grammatical class:

```
$Grammatical.class
                      factor logodds tokens deleted/deleted+retained centered factor weight
       regular nonsyllabic   1.127    140                      0.879                  0.755
      semi-weak past verbs   0.812     19                      0.684                  0.692
      negative contraction   0.086    254                      0.740                  0.521
                  monomorph  -0.551    442                      0.629                  0.366
 Irregular devoicing verbs  -1.474     61                      0.311                  0.186
```

Two different sets of numbers are of special interest here: logodds and factor weights

– **Log odds** – are a measure of the effect size. They reflect the strength of the relationship between a factor and dependent variable. If log-odds are negative, there is a negative correlation between the variables, if they are above 0, the correlation is positive. The higher the value the stronger the correlation.
– **Factor weight** – simply, it reports the same thing but within the range of 0 - 1.00. If the correlation is 1.00 it is a knock-out. Here we observe that regular verbs are most likely to occur with /t,d/ deletion. If the result is close to 0 for log odds or close to 0.50 for factor weights it is almost neutral. Here this means that it almost doesn't matter if a variant is a negative contraction or not (as in *can't*, *ain't*). On the other hand, in irregular verbs deletion is least likely to occur.

➢ **Although log odds probably show a more accurate fit of each category to the data, factor weights are useful for drawing overall comparisons across different sets of data.**

NOTE! For models with continuous variables, only logodds are reported. Continuous predictors do not have associated factor weights.

– **Uncentered input prob.** – is another number reported by Rbrul. Roughly speaking, this reports the overall prediction of the model. More precisely, the centered input probability is the inverse logit of the model intercept. But what does that mean? All these models make a prediction for the mean (in linear regression) or the proportion (in logistic regression) in each cell, where a cell is defined as a given setting of the independent variable(s). The input probability is the average of the predicted values for each cell (cf. the Rbrul manual).

c) Important information about the model:

```
$misc
 deviance df intercept grand mean centered input prob
  948.178 18     0.874      0.678              0.706


BEST STEP-UP MODEL IS WITH Word [random] and Village (2.93e-07) + Grammatical.class (4.58e-05) + Following.phon.segment (0.000107) + Precedin
[p-values building from null model]
```

Rbrul also reports some more general information about the model. This is important if we are modelling the data several times and want to compare the overall fit of the model.

- **Deviance** - a measure of how well the model fits the data, or how much the actual data deviates from the predictions of the model. The larger the deviance, the worse the fit. As we add predictors to the model, we will see this number decrease.

- **Degrees of freedom** - The df (degrees of freedom) is the number of parameters in the model, a measure of model complexity (the more factors we add the higher the df).

- **Intercept** - If the dependent variable is continuous: the intercept is the estimated value of the dependent variable if x=0. If the dependent variable is binary: the intercept is the log odds of the dependent variable if x=0.

- **Grand mean** – overall data proportion

- **Input probability** – See the above definition of the uncentered input prob. To put it non-technically, it is the overall probability that the dependent variable will occur in a given variable context

11. Comparing different runs

The analysis showed that in the following phonol. segment glides and rhotics disfavour /t,d/ deletion in a similar manner. This warrants combining these two factors into one category – e.g. approximants.

Why would I want to do this?

Above we said that the deviance goes up as we add predictors to the model. Therefore, one way of obtaining the neatest, most accurate model is by reducing the number of predictors. **However** (and this is crucial!), combining factors or factor groups needs to be conceptually validated. That is, we have to have a really good reason to do this. In this case combining rhotics and laterals is justified because: (i) they correlate with t/d deletion in a similar way, and (ii) they represent the same manner of articulation.

To combine factors, go back to Main Menu and follow the same steps as above.

After the factors are combined, the final result looks as follows:

```
BEST STEP-DOWN MODEL IS WITH Word [random] and Village (3.77e-07) + Grammatical.class (1.72e-05) + Following.phon.segment (0.000118) + Preceding.phon.segment (0.000591)
[p-values dropping from full model]

$Village
        factor logodds tokens deleted/deleted+retained centered factor weight
      Hamilton   0.462    428                    0.79                 0.613
 Mount Pleasant  -0.462    488                    0.58                 0.387

$Preceding.phon.segment
    factor logodds tokens deleted/deleted+retained centered factor weight
  sibilant   1.187    151                   0.874                 0.766
   Lateral   0.014     94                   0.702                 0.503
     nasal  -0.108    641                   0.632                 0.473
 fricative  -1.093     30                   0.600                 0.251

$Following.phon.segment
      factor logodds tokens deleted/deleted+retained centered factor weight
        stop   0.748    162                   0.815                 0.679
       glide   0.651    165                   0.770                 0.657
   fricative   0.184     46                   0.696                 0.546
       pause  -0.001     86                   0.640                   0.5
       nasal  -0.081     98                   0.714                  0.48
       vowel  -0.300    250                   0.580                 0.426
    sibilant  -0.550     64                   0.547                 0.366
 approximant  -0.651     45                   0.556                 0.343

$Grammatical.class
                  factor logodds tokens deleted/deleted+retained centered factor weight
       regular nonsyllabic   1.129    140                   0.879                 0.756
      semi-weak past verbs   0.812     19                   0.684                 0.692
      negative contraction   0.083    254                   0.740                 0.521
                monomorph  -0.550    442                   0.629                 0.366
 Irregular devoicing verbs  -1.474     61                   0.311                 0.186

$Word
       random logodds tokens deleted/deleted+retained centered factor weight std dev
       ground   1.201      8                   1.000                 0.771   0.845
       around   1.131     18                   0.889                 0.759   0.845
         cold   0.819      5                   1.000                 0.697   0.845
        round   0.722     15                   0.867                 0.676   0.845

$misc
 deviance df intercept grand mean centered input prob
  948.266 17     0.942      0.678               0.719


BEST STEP-UP MODEL IS WITH Word [random] and Village (2.93e-07) + Grammatical.class (4.58e-05) + Following.phon.segment (4.76e-05) + Preceding.phon.segment (
[p-values building from null model]

BEST STEP-DOWN MODEL IS WITH Word [random] and Village (3.77e-07) + Grammatical.class (1.72e-05) + Following.phon.segment (0.000118) + Preceding.phon.segment
[p-values dropping from full model]

STEP-UP AND STEP-DOWN MATCH!
```

As we can see, the result is not drastically different, but the significance of the Following phon. segment has slightly changed, and so has the overall deviance. Because we have reduced the number of predictors, the number of degrees of freedom has decreased.

How do we know then which model offers a better fit to the data (the ones where glides and rhotics are combined, or the one where they are separated)?

To test this we will apply the log likelihood-ratio test.

11a. Log-likelihood ratio test

- First, we need to calculate the log likelihood from both runs we've conducted. To do this, you just need to divide the deviance value by -2.

- So: Run 1 Log likel. = -474.089 and Run 2 Log likel. =-474.133

- Now we subtract Ll 1 rom Ll2 and multiply by 2. The subtracted value is 0.08 (we can round it up to 0.1).

- Now we calculate the difference in degrees of freedom between the two runs. We had 18 df in Run 1 and 17 in Run 2, so the number of df we will be looking at is 1.

- Now, all we need to do is to see if 0.1 is significant at 1 df. To do this we will look at this chi-square table. The are available online:

E.g. (http://people.richland.edu/james/lecture/m170/tbl-chi.html)

The areas given across the top are the areas to the right of the critical value. To look up an area on the left, subtract it from one, and then look it up (ie: 0.05 c

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156

The top bar in the table shows the p value. The left column shows degrees of freedom. We are looking at row one since we've calculated one degree of freedom. Our value 0.8 is between the 5th and the 6th cell in the first row. The p value for this cell is 0.90 and 0.10. These values are greater than the usual $p<0.05$ which means that 0.8 is not significant at 1df.

This means overall that Run 2 is NOT significantly worse than Run 1 so we could happily accept it as our best run.

12. Saving your file.

- Use the Load/Save Model in the Main Menu

  With this option, Rbrul allows you to save models to disk and retrieve them. It first prompts you to save the current model in the current directory using a filename of your choice. Then, regardless of whether you have saved the model, it prompts you to load a saved model from any directory. This means that in theory, if you load a saved model, you will be able to make plots without having to reload the data separately.

- You can also save the data to a txt file through File>Save to File command

13. Reporting your results – what to report?

Here's an example of a table output of your results. I chose to report my results in factor weights but log odds are also fine. It's up to you! (Unless you've included continuous variables in which case you can only report log odds):

	T/d deletion		
Input prob.	0.719		
Total N	883		
Deviance		948.226	
	F.w	%	N
Village	p.<3.77e-07		
Hamilton	0.61	79	428
Mount Pleasant	0.38	58	488
Gramm. class	p.<1.72e-05		
Reg. nonsyllabic	0.75	87	140
Semi-weak	0.69	68	19
Negative contr.	0.52	74	254
Monomorphemic	0.36	63	442
Irreg. devoicing	0.18	31	61
Following Seg.	p.<0.0001		
Stop	0.67	80	162
Glide	0.65	77	165
Fricative	0.54	69	46
Pause	0.50	64	86
Nasal	0.48	71	98
Vowel	0.42	58	250
Sibilant	0.36	55	64
Approximant	0.34	55	45
Preceding seg.	p.<0.0005		
Sibilant	0.76	87	151
Lateral	0.50	70	94
Nasal	0.47	63	641
Fricative	0.25	60	30
Age	Not significant		
Older	[0.47]	66	845
Younger	[0.52]	81	71
Lexeme	Random		

- **Total N** – total number of raw tokens in the data set

- **Percentages of the variant per cell**. How much % of the variant which is your application value constitutes the total N in a given category. This allows us to spot any remaining interactions. Basically, the greater the factor weight, the greater the % should be. In this case, we can observe that the % values of /t,d/ deletion is 79%. There's an interaction in my Semi-weak verb category, perhaps due to a small number of tokens. These should be removed.
- The proportion of factors in each cell can be established through cross-tabs

Again, notice the interaction between Pause and Nasal following phon. Context (the % value for nasals is higher than for pause).

It's good practice to report predictors, which were not selected as significant. Remember, just because they are not significant, doesn't mean they don't matter! To indicate such predictors we put the values in parentheses. You can find the values of these factors in your Rbrul run.

Of course, there are other ways in which the output of your analysis can be presented. Here's another example (adopted from Scleef et al. 2011):

https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxseW5uY2xhcmtsaW5nfGd4OjQwYWVhNGQzYWQwZTFiZg&pli=1

14. Data interpretation.

You have now obtained a set of preliminary results (at least my results are preliminary because there're still some areas which should be improved, e.g. interactions). It's up to you now to interpret these results and complete the story. Here're a few questions you could consider in this process:

- How do these results compare to the previous studies on this variable/variable context?
- Is there anything expected/unexpected?
- Do these results match your initial hypotheses?
- Do these results answer the questions raised in the beginning of the study?
- Are there any limitations of this analysis? (You need to report these too)
- How do the social variables match what you know about the speaker/community you analyse?
- Do these results corroborate/contradict any of the socio-linguistic patterns found in other studies focussing on similar issues?
- Do the linguistic constraints make sense considering what you know about the variable? Can you support the results with relevant examples from the dataset, e.g. from discourse?