Some Statistical Models to Analyze Language Production Data

(in strict alphabetical order)

F.-Xavier Alario  
francois-xavier.alario@univ-amu.fr

Matt Goldrick  
matt-goldrick@northwestern.edu

IWOLP 2014 – Genève
DON’T PANIC
Objective n°1

Going from this . . .

. . . to this.
Grasping statements like... 

\[ RT = I + \beta_{\text{group}} \times \left[ \begin{array}{c} -0.5_{\text{mono}} \\ 0.5_{\text{bil}} \end{array} \right] + \mathcal{N}_{\text{part}}(\mu_p, \sigma_p) + \]

\[ \beta_{\text{freq}} \times \left[ \begin{array}{c} 1.79 \\ 2.08 \\ 2.48 \\ \vdots \\ 7.67 \\ 7.77 \end{array} \right] + \mathcal{N}_{\text{item}}(\mu_i, \sigma_i) + \mathcal{N}_{\text{freq}}(\mu_f, \sigma_f) \times \left[ \begin{array}{c} 1.79 \\ 2.08 \\ 2.48 \\ \vdots \\ 7.67 \\ 7.77 \end{array} \right] + \epsilon(\mu_n, \sigma_n) \]
Structure of this session

Work through slides, with a partner, at your own pace.

- There are an accompanying R files referenced in the text

You will encounter problems to work on, to check your understanding of the material.

- Questions in red have answers on the following slide.
  - Don’t cheat!! Solve the problem before moving on.

Let us know if you have questions!

- Things will probably be confusing at some points.
Structure of this session

Things you will be doing

- Inspect a dataset in detail
- Examine relationship of regression models to data (using “artificial” data)
- Explore your own datasets

When you’re ready, get started...

- Ask us questions!
- Ask each other questions!!
Inspecting data

Open the R file inspect.R

- Code chunk 1 loads in a set of simulated data.
- This is based on the lexdec dataset from languageR.
- It is based on the lexdec dataset from languageR.
The columns of the data dataset are defined below

<table>
<thead>
<tr>
<th>Column</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>Individual ID number</td>
</tr>
<tr>
<td>Sex</td>
<td>Gender of simulated individual</td>
</tr>
<tr>
<td>NativeLanguage</td>
<td>Language background of simulated individual</td>
</tr>
<tr>
<td>Word</td>
<td>Word participants were asked to make lexical decision on</td>
</tr>
<tr>
<td>Frequency</td>
<td>Log-transformed frequency per million of words</td>
</tr>
<tr>
<td>centerFrequency</td>
<td>Transformed “Frequency” variable : Deviation from average</td>
</tr>
<tr>
<td>RT</td>
<td>“Simulated” response time</td>
</tr>
</tbody>
</table>
Inspecting data

Code chunk 2

- Step through the commands in code chunk 2.
- These provide an initial overview of the data.
- What sort of output does each command yield?

Question: What does the output of the `table()` commands tell you about the design of the experiment?

(please, answer before advancing to next slide)
Question: What does the output of the `table()` commands tell you about the design of the experiment?
(please, answer before advancing to next slide)

Answer

- Native language background is (by definition) between subjects; `table()` command shows the data respect this
- All subjects see every item exactly once.
Inspecting data

Code chunk 3

- Step through the commands in code chunk 3.
- Loads in a graphing library.
- Creates a scatter plot (with least squares regression line)

Looking at the graph, what is the relationship between frequency and response time?
Inspecting data

Code chunk 4

- Step through the commands in code chunk 4.
- Breaks up the data by native language background:
  1. is a repeat of the scatter plot.
  2. is a histogram (empirical density distribution)
  3. is a density plot, a smoothed histogram

Looking at these graphs, what is the relationship between native language background and response time?
Inspecting data

Code chunk 5

- Step through the commands in code chunk 5.
- Graphs the data only for non-native speakers.
- To see how speakers differ from one another:
  1. Density plot. Do all speakers have the same mean response time?
  2. Scatter plot. Does this show the same effect as observed in the density plot?
  3. Smoothers. Do all speakers show the same effect of frequency?

Looking at these graphs, can you tell the relationship between native language background and response time?
Inspecting data

Question: Modify code chunk 5 to confirm variation in mean response time and frequency effects for native speakers.

(please, answer before advancing to next slide)
Inspecting data

Question: Modify code chunk 5 to confirm variation in mean response time and frequency effects for native speakers.

Answer

- `nativeOnly = dat[dat$NativeLanguage=="English",]`
- Replace ‘otherOnly’ with ‘nativeOnly’ in each of the ggplot commands
Summary: Data inspection

- High freq. words have faster response times than low freq. words.
- Non-native speakers are slower than native speakers.
- There is variation across speakers:
  - Some speakers are slower than others.
  - Not everyone shows the same sensitivity to frequency.
The source of these data

Artificial data

- Unlike real data, these “observations” we made up . . .
  ...so we actually know how they were generated.

- The file `simulationCode.R` illustrates how this was done...
  ...using the `lexdec` dataset in `languageR` as a basis.

Take several minutes to carefully step through the code in
`simulationCode.R`.
[This includes actually running the code]
Question: How do the “fixed” and “random” factors in the code quantify the observations you made when inspecting the data? (please, answer before advancing to next slide)
The source of these data

Question: How do the “fixed” and “random” factors in the code quantify the observations you made when inspecting the data?

Fixed factors:
1. Non-native speakers are 70 milliseconds slower than native speakers.
2. Word frequency decreases response time at a rate of 20 ms/sec per unit of log frequency

Random factors:
1. Participant mean response times differ from one another
2. These differences drawn from a normal distribution ($\mu = 0$, $sd = 50$)
3. Each participants’ frequency effect differs from overall freq. effect
4. These differences drawn from a normal distribution ($\mu = 0$, $sd = 5$)
The source of these data

Re-run code chunk 4 in file inspect.R.
Make sure you use `Mydat` and not `dat`.

1. Notice at 0 (center of the graph) the difference between the two regression lines is about 70 milliseconds.

2. Notice the slope of the regression lines. As you move about one unit in log frequency (light lines on the graph), notice the line decreases by about 20 milliseconds.
The source of these data

Question: What random effect is present in the data, but was not revealed by the graphs we constructed while inspecting the data?

(please, answer before advancing to next slide)
The source of these data

Question: What random effect is present in the data, but was not revealed by the graphs we constructed while inspecting the data?

Answer

Words’ mean response times aren’t just due to their frequency; there is some random variation (drawn from a normal distribution with a mean of 0 and standard deviation of 25).

To see this, enter command:

```r
ggplot(otherOnly, aes(x = centerFrequency, y = RT,color=Word)) + geom_point(size = 2, position=position_jitter(width=.05, height=0)) + geom_smooth(method = "lm",color="red") + labs(y = "Synthetic Reaction Time (RT in msec)", x = "Centered Log Frequency",title="")```
Mixed-effect regression models

Each simulated data simultaneously reflects the contribution of...

1. fixed properties of...
   - . . . individuals,
   - . . . words,

2. as well as random variation in...
   - speakers,
   - words,
   - our measurement

This was shown in the last line of `simulationCode.R`:

```r
dat$RT = dat$Intercept + dat$langEff +
        dat$centerFrequency * dat$freqEff +
        dat$subEff + dat$subFreqEff * dat$centerFrequency +
        dat$itemEff + dat$Noise
```

Compare to slide 4
Mixed-effect regression models attempt to infer the properties of these multiple sources.

Specifically...

Given a set of observations, along with some known factors, the models estimate numbers for the terms in the equation that produces response times.

- What is the value of the intercept?
- How large is the language effect?
- How do subjects, items randomly vary?
- etc.
Mixed-effect regression models

Building a model of our simulated data using model.R

- Code chunk A repeats the first plot we saw – the effect of frequency, ignoring all other factors.
- We begin with a model that adopts a similarly simplistic structure.
- This model assumes that each observation reflects this equation:

  \[ RT = \text{Intercept} + \]
  \[ \text{Frequency Effect} \times \text{Word Frequency} + \]
  \[ \text{Deviation from intercept for each participant} + \]
  \[ \text{Noise} \]

Compare to slide 4
Mixed-effect regression models

\[ RT = \text{Intercept} + \]
\[ \text{Frequency Effect} \times \text{Word Frequency} + \]
\[ \text{Deviation from intercept for each participant} + \]
\[ \text{Noise} \]

Corresponding `lmer` (linear mixed-effects regression) code

- Run code chunk A
- `lmer(RT \sim \text{centerFrequency}+(1|\text{Subject}),\text{data=dat})`

This says: Every response time reflects a constant value, some variation due to the frequency of the word, random variation in average RT for subjects, and noise. You don’t have to specify the intercept or noise contributions, as they are constant.

- `lmer` will estimate a coefficient for the `centerFrequency` predictor (i.e. the frequency effect)
- `(1|\text{Subject})` requires that subject specific deviations from the overall intercept are modeled as a random normal distribution
Mixed-effect regression models: Coding

Code chunk B examines language background

- Recall our earlier graph: Language background makes a difference.
- But how do you incorporate language background into an equation?
- “English” isn’t a number, it’s a word (isn’t it?)
- Note this wasn’t a problem with frequency, as it’s already a number!

The solution is to provide a coding scheme that represents distinctions between groups as numbers: R can do that for you. Recall slide 4 and check 2 next slides.
Mixed-effect regression models : Coding

Execute code Chunk B to build this new model

• Compare the summary of this new model to the old one.
• There's a new fixed effect! The estimate of the language background effect
• The model’s estimate of the by-speaker random effect has changed. This is because we’ve taken into account additional sources of variation in speakers’ response times.
Mixed-effect regression models : Coding

Behind the scenes : More on coding schemes

• If you have two levels to a variable, you can represent one contrast level with a positive number and the other with a negative number.

• If you use \([ +0.5, -0.5 ]\), then the coefficient will be equal to the (average) difference \(D\) between the two levels.

• In the equation, Group 1 will be \(+.5D\) different from the overall mean, and Group 2 will be \(-.5D\) different from the mean; the difference between them will \(+.5D - (-.5D) = D\)!

Comparing two contrast coding schemes

Code chuck B allows comparing

• \texttt{frequencyLanguage.lmer1} and

• \texttt{frequencyLanguage.lmer2}
Mixed-effect regression models: Coding

For additional information, Maureen Gillespie has a nice tutorial on coding available here:


This deals with more complicated coding schemes (with more than 2 levels, etc.)
Mixed-effect regression models: Random slopes

The magnitude of an effect may vary across participants.

Code chunk C deals with this:

- Recall the graph showing variation in the frequency effect across participants.
- This calls for yet another extension of our equation.

\[
RT = \text{Intercept} + \text{Frequency Effect} * \text{Word Frequency} + \text{Language Background Effect} * \text{Language Code} + \text{Deviation from intercept for each participant} + \text{Deviation of frequency effect for each participant} + \text{Noise}
\]
Mixed-effect regression models: Random slopes

Code chunk C incorporates a new term to our regression code.

- `(0+centerFrequency|Subject)`
- 0 explicitly excludes the intercept from this random effect term
- Recall we already have that term in our model: `(1|Subject)`

If you put both the intercept and the frequency effect into a single term (i.e. `(1+centerFrequency|Subject)`) you are specifying a model in which the correlation between these two random deviations is estimated. In our case, we know the correlation is not present (see simulation code). In real datasets, there is debate about what to do in these situations.
Mixed-effect regression models:
Random slopes

Inspect the output for this regression model
New random effects term!

Question: What’s missing from the model? What do we know is in our simulation code but is not in the model specification? How do we add it to the model?
(Please, answer before moving to new slide)
Mixed-effect regression models:
Random slopes

Question: What’s missing from the model? What do we know is in our simulation code but is not in the model specification? How do we add it to the model?

Answer

- Random variation in overall response times for each word.
- Extend model by adding in one more random effects term $(1 | \text{Word})$
Mixed-effect regression models: Assessment

Uncovering the actual “effects”?  

• Compare the output for full model specified in the last slide of simulationCode.R  
• Does lme4 successfully recover the parameter values from the simulation?  
• Where are the parameter values in the summary output?  

Exercise  

1 Modify the simulation code by increasing the frequency effect (or other effects).  
2 MYdat will be modified  
3 Verify with graphs that the frequency effect is stronger.  
4 Use lme4 to try to recover the equation parameters for your new simulated dataset.
Analyzing real data

At this point, you can either:

1. Analyze the (real) data in the lexdec dataset, whose structure is similar to that of the simulated data
   or
2. Load your own data, select one or two factors, graph and analyze it
   or
3. both...!
Done with all of the above?

“You have reached your final destination!”
Thank you for your attention