mixed-effects models in R (and Rbrul)

9:30 – 10:45	1. Intro to R and regression
	tea interval
11:15 – 12:30	2. Why use mixed models?
	lunch break
1:30 – 2:15	3. Mixed models in Rbrul
	tea interval
2:45 – 5:00	4. Mixed models in R

Daniel Ezra Johnson, Lancaster University
danielezrajohnson@gmail.com
www.danielezrajohnson.com/york_workshop2.pdf

1. Intro to R and regression

statistics: methods and software

- in theory there is independence between them
- in practice there is some association
- with software, more a matter of taste (budget)
- with methods, more substance to the discussion

- you can (now) do mixed models in SPSS and SAS
- I have never used these programs
- I enjoy using R, you may not
- Rbrul is another option though relatively limited

introduction to R

- programming language for statistics, graphics
- command-line interface or scripts Command-Enter or Ctrl-R
- variables (objects) and functions
  - variable <- > *height <- c(120, 140, 160)*
  - function() > *sum(height); mean(height)*
    - common functions in *base environment* > *t.test(); boxplot()*
    - others are in *packages* > *install.packages("lme4")*
      > *library(lme4)*
- data frames
  - columns are variables, rows are observations
  - to import .csv files: > *my.data <- read.csv("data.csv")*

introduction to R

- difficult: learning the syntax (quotes, commas, parentheses)
- difficult: learning the functions (Google, "R help")
- powerful: write your own functions, packages
- powerful: customize graphics, simulations
- http://tryr.codeschool.com/ (R simulator)
- go through Chapter 1
- then open R itself and try similar commands
- then open a new script (Command-N or Ctrl-N) and practice
  - entering commands in the script
  - running commands from the script (Command-Return or Ctrl-R)

some R functions/operators

() [] {} + - * / ^ ! & | %in% %% : = <- == # ? ??

abline
abs
anova
as.character
as.factor
as.numeric
c
cat
cbind
class
coef
cor
data.frame
else
exp
head
if
ifelse
fisher.test

fixef
for
function
getwd
glm
glmer
grep
head
image
install.packages
is.na
ks.test
length
library
log
logLik
max
mean
median

min
mosaicplot
names
paste
pchisq
pf
plogis
plot
print
qlogis
ranef
range
rbind
read.csv
rep
repeat
rnorm
round
runif

sample
seq
setwd
set.seed
shapiro.test
signif
sqrt
str
summary
table
tail
t.test
vector
which
wilcox.test
write.csv
xtabs
xyplot
lm

introduction to regression

- model that uses “independent variables” to predict the value of a “dependent variable”
- not the only type of statistics, has assumptions
- simplest type: linear regression (continuous DV)
- test script 1: http://www.danielezrajohnson.com/york_regression.R
- also common: logistic regression (binary DV)
- linear > *model <- lm(y ~ x1 [+ x2 + x3…], data)*
- logistic > *model <- glm(y ~ x…, data, family = binomial)*
- compare nested models to get significance
  > *anova(model.1, model.2, test = “Chisq”)* LRT gives p-value!

introduction to regression

- let's load the Labov department store data:

  > <u>*ds*</u> *<- read.csv("http://www.danielezrajohnson.com/ds.csv")*

  > *head(ds)* or > *str(ds)* 1 DV (response) and 3 IVs (predictors)
- DV is called *r*, IVs are called *store, word, emphasis*

  store Saks, Macy's, Klein's *word* fourth, floor *emphasis* normal, emphatic
- fit the null model: > <u>*m0*</u> *<- glm(r ~ 1, ds, family = binomial)*
- which IVs make a sig. difference when added to *m0*?
- fit the full model: > <u>*mf*</u> *<- glm(r ~ store + word + emphasis, ds, family = binomial)*
- which IVs make a difference when removed from *mf*?
- what do these models say? print a model (type its name)
- the output can be challenging, let's look through it…

output of a *glm* model in R

```
> ds <- read.csv("http://www.danielezrajohnson.com/ds.csv")
> mf <- glm(r ~ store + word + emphasis, ds, family = binomial)
> mf

Call:  glm(formula = r ~ store + word + emphasis, family = binomial,
    data = ds)

Coefficients:
   (Intercept)     storeMacy's       storeSaks      wordfouRth
       -1.6192          1.8028          2.2428         -1.0013
emphasisnormal
       -0.3291

Degrees of Freedom: 728 Total (i.e. Null);  724 Residual
Null Deviance:     909
Residual Deviance: 789.6
  AIC: 799.6
```

p = probability   p/(1-p) = odds   ln(p/(1-p)) = log-odds   1 log-odds = 50% to 73% = 73% to 88% = 88% to 95%

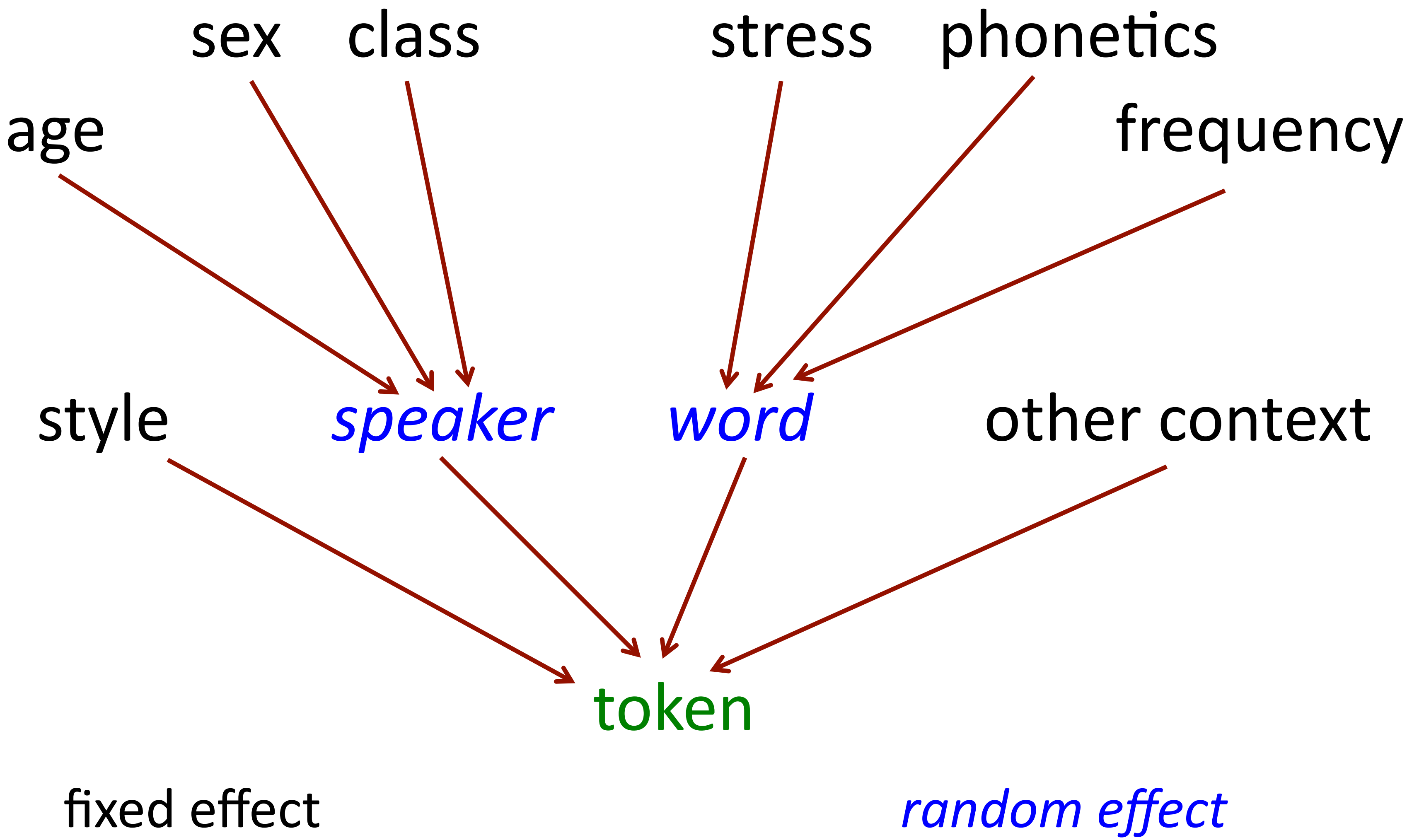
2. Why use mixed models?

why use mixed models?

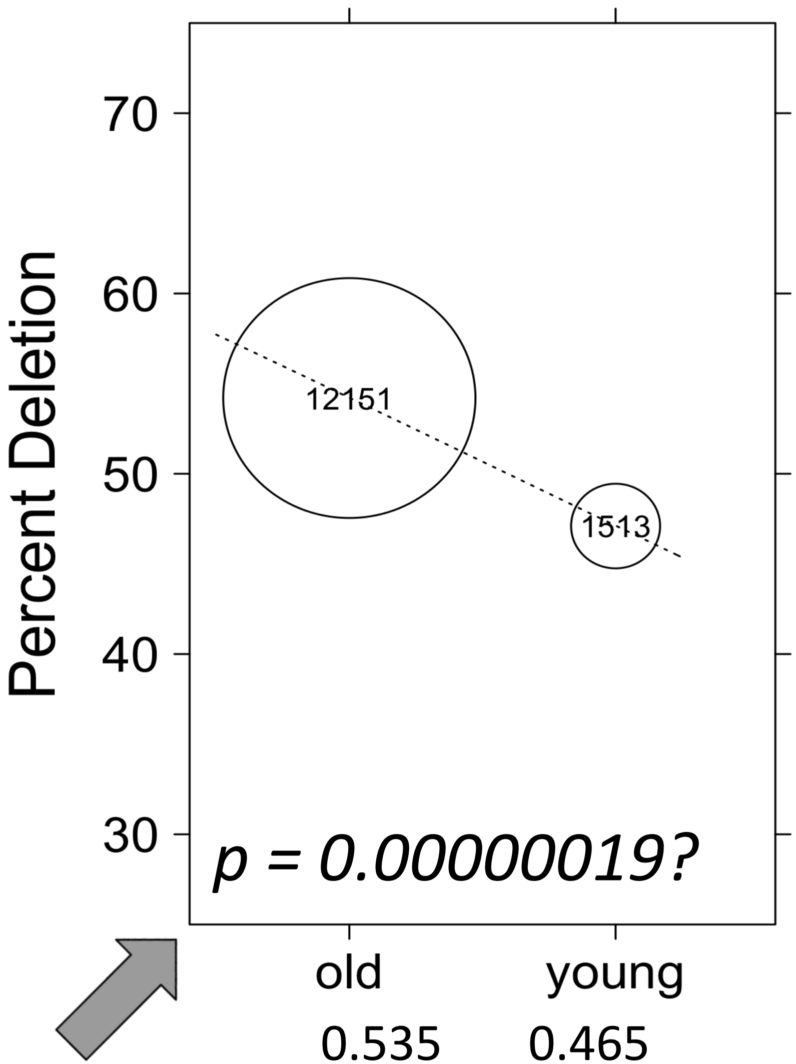
- what is special about the department store study?
- accounting for store, word, emphasis: 1 obs. per speaker
- in most sociolinguistic data sets, many obs. per speaker
- with binary data, even more observations per speaker
- regression assumes uncorrelated errors
- if speakers differ, each speaker's errors will be correlated
- significance of between-speaker predictors exaggerated

height	F0 (forensic)	rhoticity
2 genders		2 genders
50 people	2 people	50 people
	50 observations	5000 observations

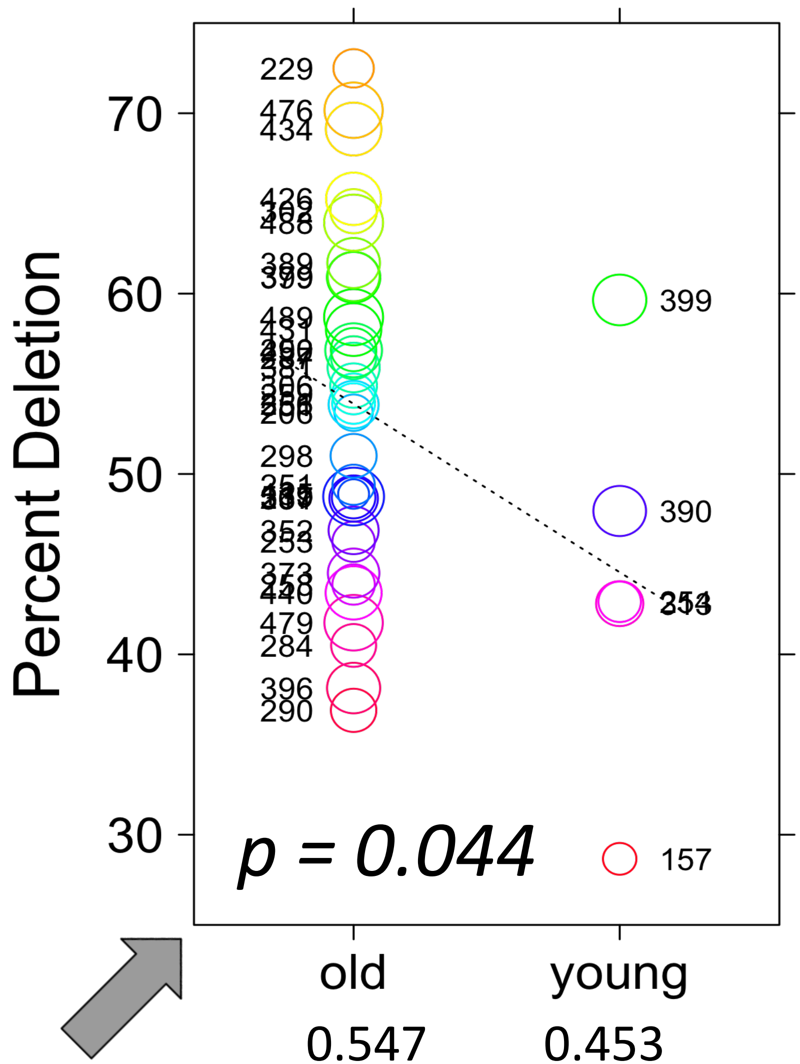
architecture of variables



test script 2: http://www.danielezrajohnson.com/york_four.R

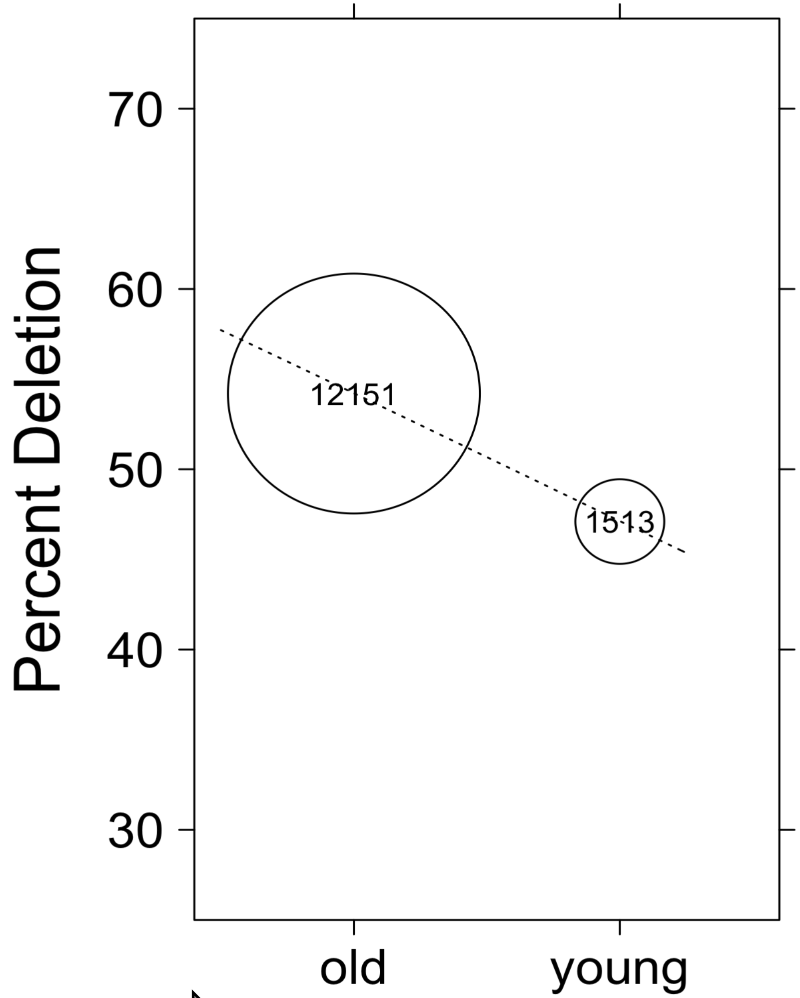


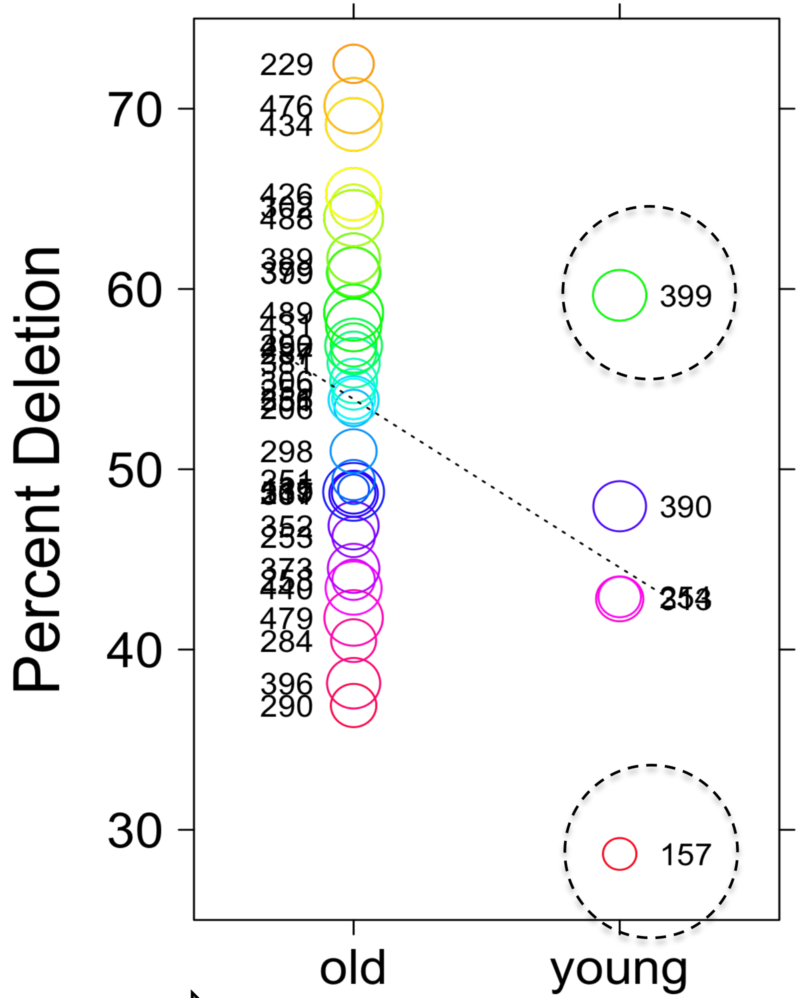
age w/ no random effect



age + random intercept: speaker

Significance of between-speaker predictor





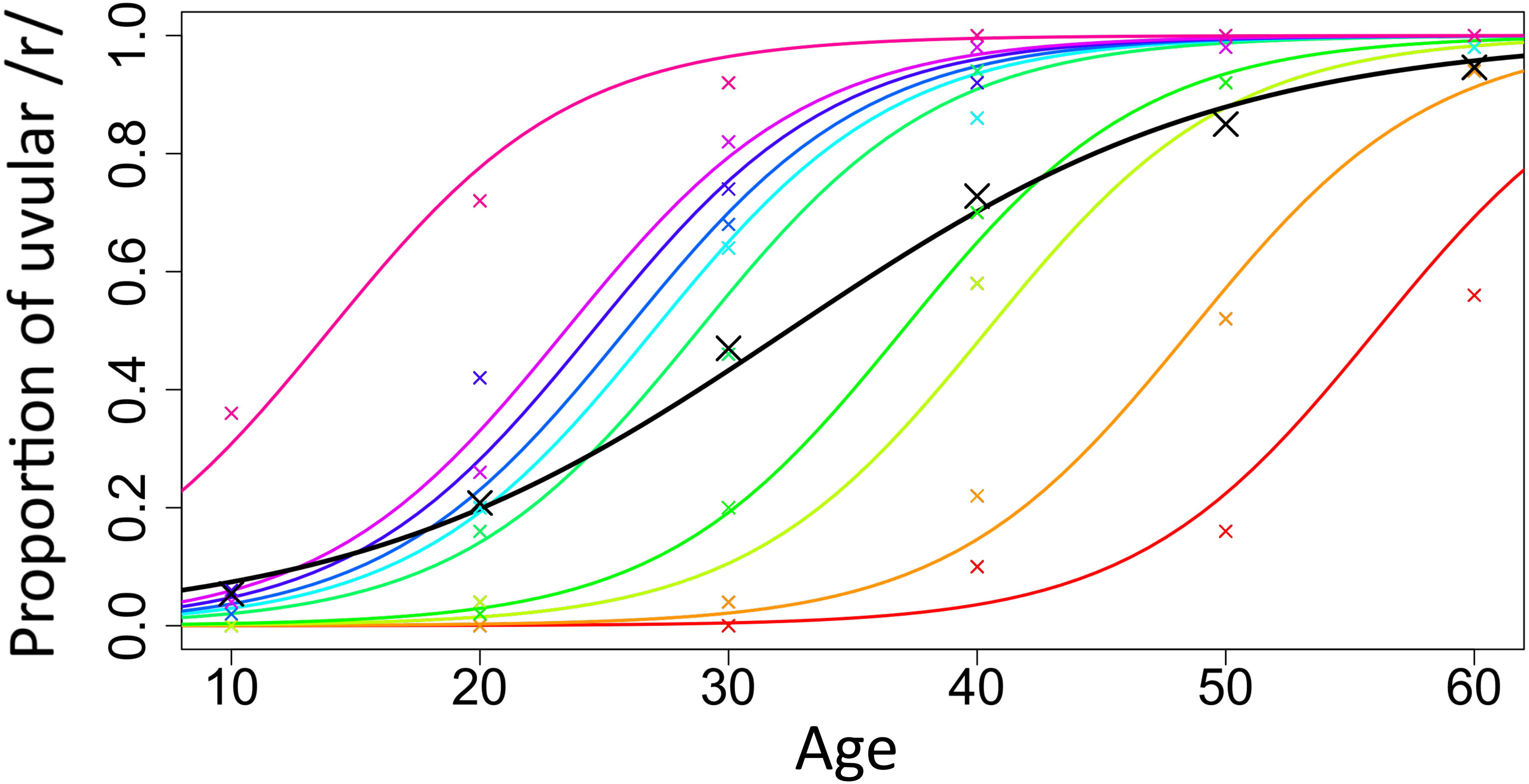
0.535 0.465

age w/ no random effect

0.547 0.453

age + random intercept: speaker

Effect size of between-speaker predictor



age coefficient w/ no random effect: 0.113 log-odds/year

age coeff. w/ speaker random effect: 0.205 log-odds/year

Effect size of within-speaker predictor (logistic regression only)

fitting mixed models with *lmer()*

- let's load some Gretna rhoticity data:
  gr <- read.csv("http://www.danielezrajohnson.com/gretna.csv")
- *head(gr)* or *str(gr)*: 2 DVs (response) and many IVs (predictors)
- DV is *r2*, IV is *gender,* grouping factor is *speaker*
- does *gender* make a significant difference to *r2*?
- fixed-effects approach ignores speaker variation
  > *f.0 <- glm(r2 ~ 1, gr, family = binomial)*
  > *f.g <- glm(r2 ~ gender, gr, family = binomial)*
  > *f.g* [size in log-odds of gender effect]
  > *anova(f.0, f.g, test = "Chisq")* [p = significance of gender effect]
- mixed-effects approach accounts for speaker variation:
  > *g.0 <- lmer(r2 ~ (1|speaker), gr, family = binomial)*
  > *g.g <- lmer(r2 ~ gender + (1|speaker), gr, family = binomial)*
  > *g.g* [size in log-odds of gender effect]
  > *anova(m.0, m.g)* [p = significance of gender effect]

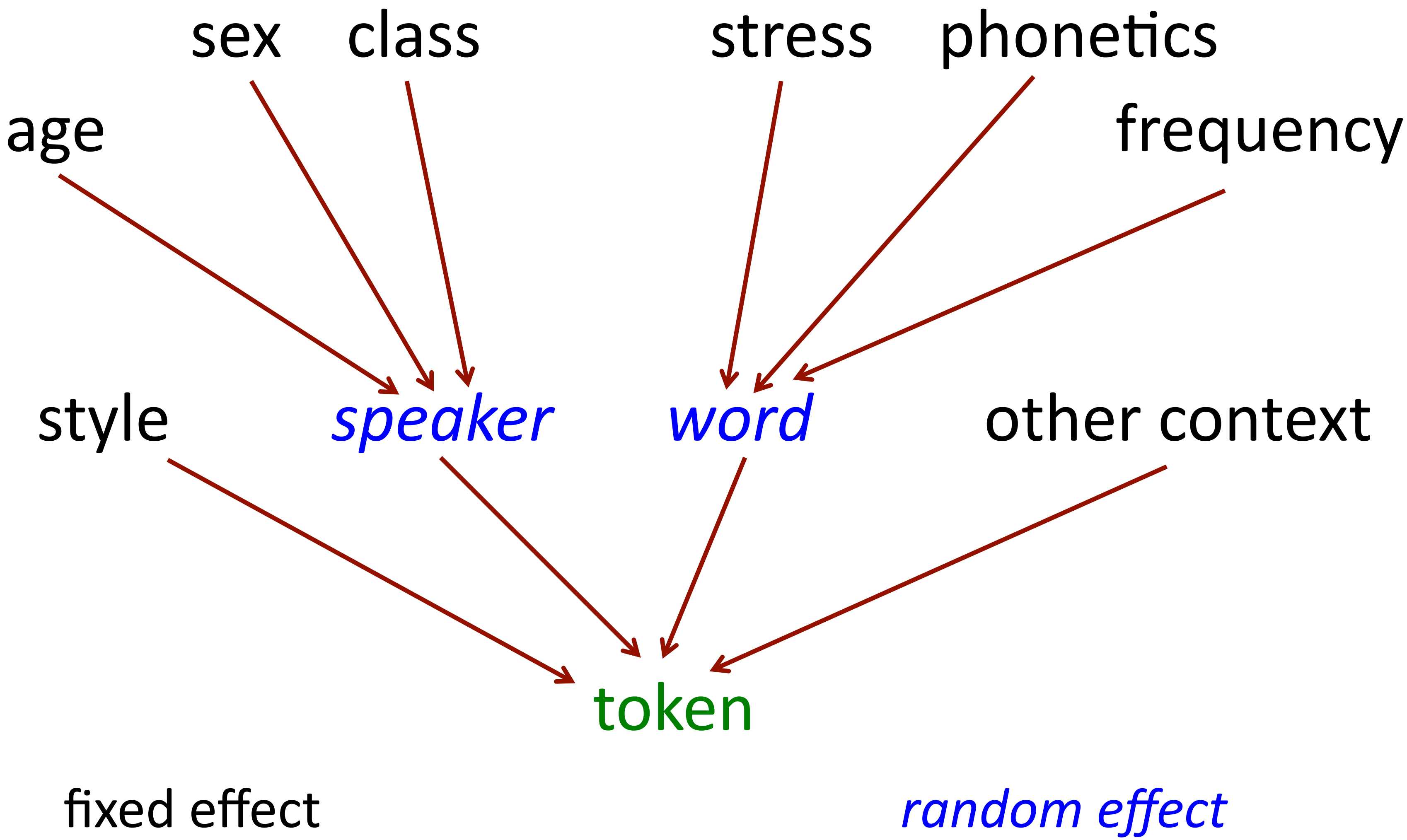
output of a *lmer* model in R

```
> g.g <- lmer(r2 ~ gender + (1|speaker), gr, family = binomial)
> g.g
Generalized linear mixed model fit by the Laplace approximation
Formula: r2 ~ gender + (1 | speaker)
   Data: gr
  AIC  BIC logLik deviance
 1432 1448 -713.1     1426
Random effects:
 Groups  Name        Variance Std.Dev.
 speaker (Intercept) 6.523    2.554
Number of obs: 1597, groups: speaker, 40

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3345     0.5917   0.565    0.572
gendermale    0.3841     0.8342   0.460    0.645

Correlation of Fixed Effects:
           (Intr)
gendermale -0.709
```

random intercepts, random slopes

- random intercepts
  - e.g. (1|speaker) or (1|text) or (1|word)
  - allows the DV to vary according to that factor
- random slopes
  - e.g. (stress|speaker) or (gender|word)
  - allows the IV's effect to vary according to that factor
  - e.g. (gender|speaker) does not make sense
  - (stress|speaker) = (1 + stress|speaker) *intercept and slope*
  - maximal random effects preferred, may not fit
- to test fixed effect significance
  - leave random effects alone, remove fixed effect only

architecture of variables



Gretna /r/, fixed vs. mixed models

- fit a fixed-effects model using *glm()*
  - dependent variable *r2*
  - independent variables *gender* and *position* (*far* vs. *farm*)
- fit smaller models and use *anova()* to determine
  - the p-value associated with dropping *gender*
  - the effect estimated for male gender
  - the p-value associated with dropping *position*
  - the effect estimated for coda-internal position
- fit a similar mixed-effects model using *lmer()*
  - use random intercepts for *speaker* and *word*
  - how do the p-values and effect estimates change?
  - add a random slope: *gender* by *word*? *position* by *speaker*?

3. Mixed models in Rbrul

what is Rbrul?

- a front-end interface to R that is an improvement on the GoldVarb (VARBRUL) program from c. 1980
- allows many things that GoldVarb does not
- does not allow / operator to partially exclude tokens
- does not (yet) allow random slopes, only intercepts
- makes stepwise regression easy – maybe a bad thing
- > *source("http://www.danielezrajohnson.com/Rbrul.R")*
- compared to GoldVarb, Rbrul almost certainly better
- compared to R, advantages and disadvantages
- <u>default</u> output from Rbrul is different from R
- R: treatment contrasts; Rbrul/GoldVarb: sum contrasts

VARBRUL / GoldVarb	other
dependent variable (DV)	DV, response, y
factor group, independent variable (IV)	IV, factor (categorical), predictor, x
factor	level
factor weight	coefficient, effect, estimate, β
factor weight range	similar to ‘effect size’
input probability	intercept
applications / total	(response) proportion

`lme4`	other
mixed model	mixed-effects, hierarchical, or multilevel model
fixed effect	main effect
(all) fixed-effects model	flat model
conditional modes of random effects	random effect estimates, random effect BLUPs

Do you speak VARBRUL?

Advantages of Rbrul

- flexible in terms of type of response and predictors
- handles typical nested data structure (with mixed models)
- accepts data in non-annoying format(s)
- output in factor weights, log-odds… backward compatible
- some support for interactions between predictors
- runs much faster than GoldVarb
- user may accidentally learn to use R
- “A tool such as Rbrul offers a compromise of the old and the new that I believe will be widely used in the near future.” (Baayen, 2009)
- worth learning, if you like it – it does nothing that special!

Concerns with Rbrul

- how to address problem of multiple comparisons
- how best (not) to use stepwise regression (tomorrow AM)
- cannot test assumptions of models we are building
- does not report standard errors of coefficients
- issues with p-values for mixed models
- how to test for and resolve multicollinearity
- error handling is not good (but customer service is good!)
- programming is not good, difficult to adapt/improve
- like GV, black box, can use w/o understanding what doing
- can get “answers” without thinking about problems
- no pain, no gain? better to learn R?

treatment vs. sum contrasts

- two equivalent ways of reporting effects of a factor
- treatment contrasts *> options(contrasts = c("contr.treatment", "contr.poly"))*
  - baseline level: effect is 0
  - other levels: difference from the baseline
  - in R: baseline level is omitted
  - in R: other levels are named
  - in Rbrul: all levels are named
- sum contrasts *> options(contrasts = c("contr.sum", "contr.poly"))*
  - all levels: difference from the mean
  - in R: levels are numbered (first alphabetically is 1, etc.)
  - in R: last level (alphabetically) is omitted, is 0 - sum(others)
  - in Rbrul: all levels are named

treatment vs. sum contrasts (R)

```
> ds <- read.csv("http://www.danielezrajohnson.com/ds.csv")
> mf <- glm(r ~ store + word + emphasis, ds, family = binomial)
> mf

Call:  glm(formula = r ~ store + word + emphasis, family = binomial,
    data = ds)

Coefficients:
   (Intercept)     storeMacy's       storeSaks      wordfouRth
       -1.6192          1.8028          2.2428         -1.0013
emphasisnormal
       -0.3291
```

```
> ds <- read.csv("http://www.danielezrajohnson.com/ds.csv")
> options(contrasts = c("contr.sum", "contr.poly"))
> mf.sum <- glm(r ~ store + word + emphasis, ds, family = binomial)
> mf.sum

Call:  glm(formula = r ~ store + word + emphasis, family = binomial,
    data = ds)

Coefficients:
(Intercept)       store1       store2        word1    emphasis1
    -0.9359      -1.3485       0.4542       0.5007       0.1646
```

sum vs. treatment contrasts (Rbrul)

```
ONE-LEVEL ANALYSIS WITH store (7.57e-20) + word (5.62e-09) + emphasis (0.0661)
$store
  factor logodds tokens 1/1+0 centered factor weight
    Saks   0.894    177 0.475                   0.71
  Macy's   0.454    336 0.372                  0.612
 Klein's  -1.349    216 0.097                  0.206

$emphasis
   factor logodds tokens 1/1+0 centered factor weight
 emphatic   0.165    271 0.347                  0.541
   normal  -0.165    458 0.297                  0.459

$word
 factor logodds tokens 1/1+0 centered factor weight
  flooR   0.501    347 0.412                  0.623
 fouRth  -0.501    382 0.228                  0.377
```

```
ONE-LEVEL ANALYSIS WITH store (7.57e-20) + word (5.62e-09) + emphasis (0.0661)
$store
  factor logodds tokens 1/1+0 centered factor weight
    Saks   2.243    177 0.475                   0.71
  Macy's   1.803    336 0.372                  0.612
 Klein's   0.000    216 0.097                  0.206
```

Basic Rbrul - New York City /r/

Start R.
(Some Windows users may need to right-click and Run As Administrator. Mac users should use R64.)

Load Rbrul.
The usual way to load Rbrul is:

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

Or you can use the GUI to do this, under the File menu.
A browser window will open to track Rbrul use. Close it.

Start Rbrul.

```
rbrul()
```

The first time, there will be various messages.
If you get to the MAIN MENU, then everything is working.
If you don't, then there was a problem loading packages.

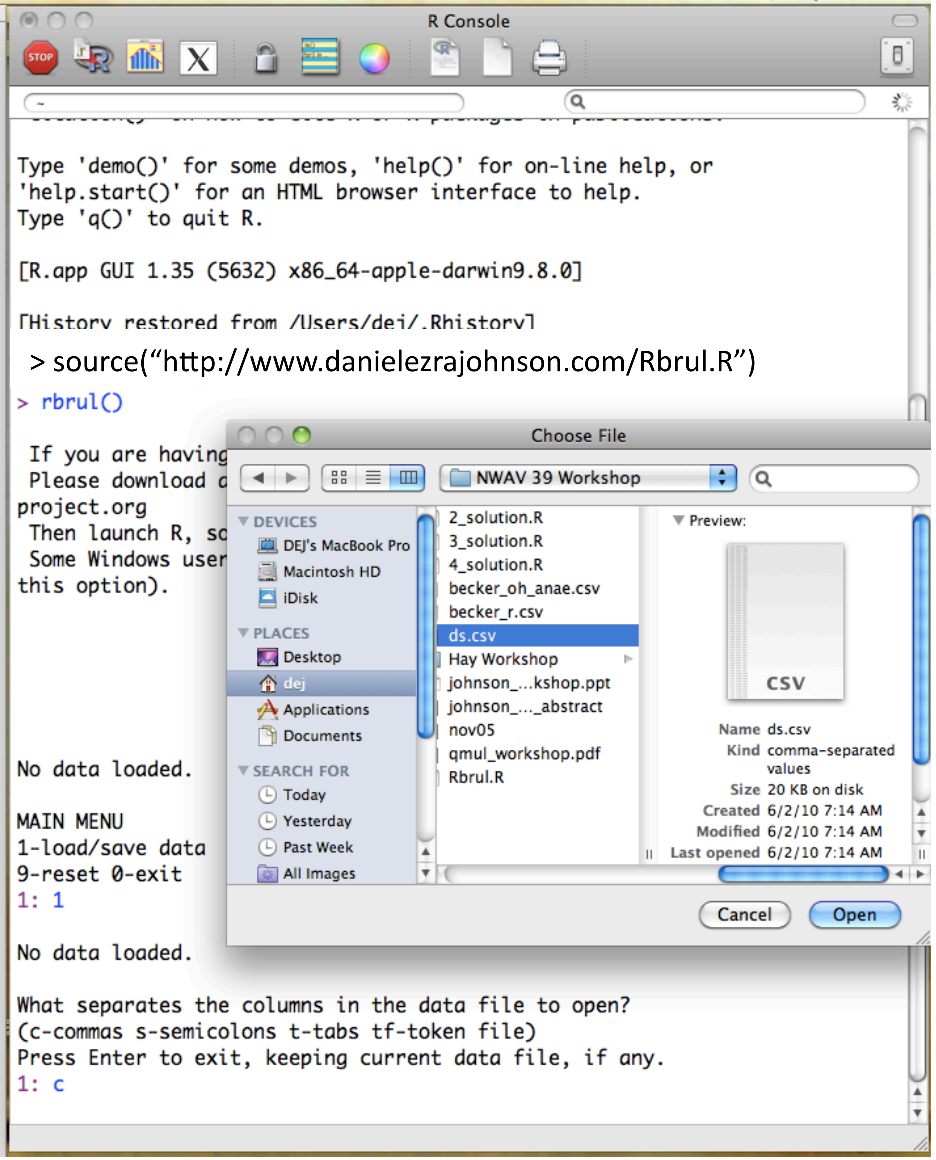
From the MAIN MENU, choose "1" to "load/save data".

```
1
```

If prompted to save current data, press "Enter" for "No".
It will ask "What separates the columns...?"
Enter "c" for commas, because we want to open a .csv file.

```
c
```

A file-choosing window appears. Locate and open "ds.csv".



The MAIN MENU should appear again, after giving the "Current data structure" as follows:

Current data structure:
r (integer with 2 values): 1 0
store (factor with 3 values): Saks Macy's Klein's
emphasis (factor with 2 values): normal emphatic
word (factor with 2 values): fouRth flooR

This is the Labov department store data, collected in 1962. The response variable is "r" (presence or absence of /r/). There are three potential predictor variables:
"store" (Saks, Macy's, or Klein's),
"emphasis" (normal or emphatic),
"word" ("fourth" or "floor").

Now we will run a logistic regression with all predictors. Enter "5-modeling" and the MODELING MENU will open.
5
Enter "1" to "choose variables".
1
Enter "1" to choose "r" as the response, or dependent variable.
1
Just press "Enter" as "r" is a binary response.

Enter "2" to choose "1" (presence of /r/) as the application value. We will be see what favors and disfavors the presence of /r/. If we used "0" as the application value instead, the model coefficients would be reversed.)
2

R Console

```
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

Current data file is: /Users/dej/Linguistics/NWAV 39/NWAV 39 Workshop/
ds.csv

Current data structure:
r (integer with 2 values): 1 0
store (factor with 3 values): Saks Macy's Klein's
emphasis (factor with 2 values): normal emphatic
word (factor with 2 values): fouRth flooR

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

No variables chosen.

MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1: 1
Choose response (dependent variable) by number (1-r 2-store 3-emphasis 4-
word)
1: 1
Type of response? (1-continuous Enter-binary)
1:
Choose application value(s) by number?  (1-0 2-1)
1: 2
```

R File Edit Format Workspace Packages & Data Misc Window Help

Enter "2", "Enter", "3", "Enter", "4", "Enter" to choose "store", "emphasis", and "word" as the predictors.
2
3
4

"Are any predictors continuous?" The answer is no, they are all categorical predictors, so just press "Enter".
[Enter]
"Any grouping factors (random effects)?" The answer is no, we will treat these predictors as fixed effects. Unlike many, this data set does not really require a random effect for speaker; it is not coded for it anyway. Press "Enter".
[Enter]
"Consider a(nother) pairwise interaction between predictors?" For now, we will not look at interactions. So just press "Enter".
[Enter]

The MODELING MENU should now show, after we see that
Current variables are:
response.binary: r (1 vs. 0)
fixed.factor: store emphasis word

The one-level analysis, similar to the first step of a step-down analysis, is recommended. Enter "2".
2

R Console

~/Linguistics/NWAV 39/NWAV 39 Workshop

```
No variables chosen.

MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1: 1
Choose response (dependent variable) by number (1-r 2-store 3-emphasis 4-
word)
1: 1
Type of response? (1-continuous Enter-binary)
1:
Choose application value(s) by number?  (1-0 2-1)
1: 2
Choose predictors (independent variables) by number (2-store 3-emphasis 4-
word)
1: 2
2: 3
3: 4
Are any predictors continuous? (2-store 3-emphasis 4-word Enter-none)
1:
Any grouping factors (random effects)? (2-store 3-emphasis 4-word Enter-
none)
1:
Consider a(nother) pairwise interaction between predictors? Choose two at
a time. (2-store 3-emphasis 4-word Enter-done)
1:

Current variables are:
response.binary: r (1 vs. 0)
fixed.factor: store emphasis word

MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1: 2
```

R File Edit Format Workspace Packages & Data Misc Window Help

The first line of the output says:

ONE-LEVEL ANALYSIS WITH store (7.57e-20) + word (5.62e-09) + emphasis (0.0661)

The numbers in parentheses are the p-values associated with dropping each predictor from a full model with all of them. These are in order from most to least significant. The notation such as 7.57e-20 is shorthand for 7.57×10^{-20}.

We see that both "store" and "word" are unquestionably statistically significant, while "emphasis" is in a marginal range. In theory, we might collect more data to see if the degree of emphasis really affects /r/ or not.

Next there is a section for each predictor. They are in the order entered, not in order of significance or effect size.

The rightmost column gives factor weights as in GoldVarb. (There is one difference: GoldVarb uses "uncentered" factor weights by default, while Rbrul's default is "centered" factor weights. Uncentered weights depend on the relative numbers of tokens within the factor group. They are highly idiosyncratic and are not recommended.)

The leftmost column has the same coefficients in log-odds units. When the log-odds coefficient x is expressed in "sum contrasts", as here, it relates to the factor weight p by:

$$x = \ln(p/(1-p)) \qquad p = \exp(x)/(1+\exp(x))$$

R Console

~/Linguistics/NWAV 39/NWAV 39 Workshop

```
MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1: 2

ONE-LEVEL ANALYSIS WITH store (7.57e-20) + word (5.62e-09) + emphasis
(0.0661)
$store
  factor logodds tokens 1/1+0 centered factor weight
    Saks   0.894    177 0.475                   0.71
  Macy's   0.454    336 0.372                  0.612
 Klein's  -1.349    216 0.097                  0.206

$emphasis
    factor logodds tokens 1/1+0 centered factor weight
 emphatic   0.165    271 0.347                  0.541
   normal  -0.165    458 0.297                  0.459

$word
 factor logodds tokens 1/1+0 centered factor weight
  flooR   0.501    347 0.412                  0.623
 fouRth  -0.501    382 0.228                  0.377

$misc
 deviance df intercept grand mean centered input prob Nagelkerke R2
  789.624  5    -0.936      0.316                0.282         0.212


Current variables are:
response.binary: r (1 vs. 0)
fixed.factor: store emphasis word

MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1:
```

Stickies File Edit Font Note Color Window Help

The output also gives the number of tokens for each level of the predictor, as well as the raw proportion of the application value for that level - here, labeled "1/1+0".

```
$store
  factor logodds tokens 1/1+0 centered factor weight
    Saks   0.894    177 0.475                  0.71
  Macy's   0.454    336 0.372                 0.612
 Klein's  -1.349    216 0.097                 0.206
```

The employees of the highest-end store, Saks Fifth Avenue, favor post-vocalic /r/ the most (f.w. .710, log-odds .894), and Klein's do the least (.206, -1.349), with Macy's in between (.454, 0.612). (Despite some common practice, we cannot say that factor weights above .500 favor and those below disfavor an alternant. These numbers can only be interpreted relative to the others in the group, or perhaps compared to other studies where the same levels were used.)

For the other predictors, we see that the word "floor" - which bears more stress than "fourth" - is associated with more post-vocalic /r/, and that there is a small effect of emphasis favoring /r/ within repetitions of the same word.

R Console

~/Linguistics/NWAV 39/NWAV 39 Workshop

```
MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1: 2

ONE-LEVEL ANALYSIS WITH store (7.57e-20) + word (5.62e-09) + emphasis
(0.0661)
$store
  factor logodds tokens 1/1+0 centered factor weight
    Saks   0.894    177 0.475                  0.71
  Macy's   0.454    336 0.372                 0.612
 Klein's  -1.349    216 0.097                 0.206

$emphasis
   factor logodds tokens 1/1+0 centered factor weight
 emphatic   0.165    271 0.347                 0.541
   normal  -0.165    458 0.297                 0.459

$word
 factor logodds tokens 1/1+0 centered factor weight
  flooR   0.501    347 0.412                 0.623
 fouRth  -0.501    382 0.228                 0.377

$misc
 deviance df intercept grand mean centered input prob Nagelkerke R2
  789.624  5    -0.936      0.316                0.282         0.212


Current variables are:
response.binary: r (1 vs. 0)
fixed.factor: store emphasis word

MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1:
```

R File Edit Format Workspace Packages & Data Misc Window Help

```
$misc
 deviance  df  intercept  grand mean  centered input prob
  789.624   5     -0.936       0.316                0.282
 Nagelkerke R2
         0.212
```

For a simple logistic model, the "misc" section gives: deviance (measure of model non-fit, remembering that observed 0's and 1's never match the model prediction); degrees of freedom (1 + k-1 for each predictor); intercept i (mean predicted log-odds across cells); grand mean (proportion of response = application value); centered input probability (exp(i)/(1+exp(i)); Nagelkerke pseudo-R-squared value.

We will now look at another New York City /r/ data set.
Enter "9" to return to the main menu.
9

From the MAIN MENU, choose "1" to "load/save data".
1
If prompted to save current data, press "Enter" for "No".
It will ask "What separates the columns...?"
Enter "c" for commas, because we want to open a .csv file.
c

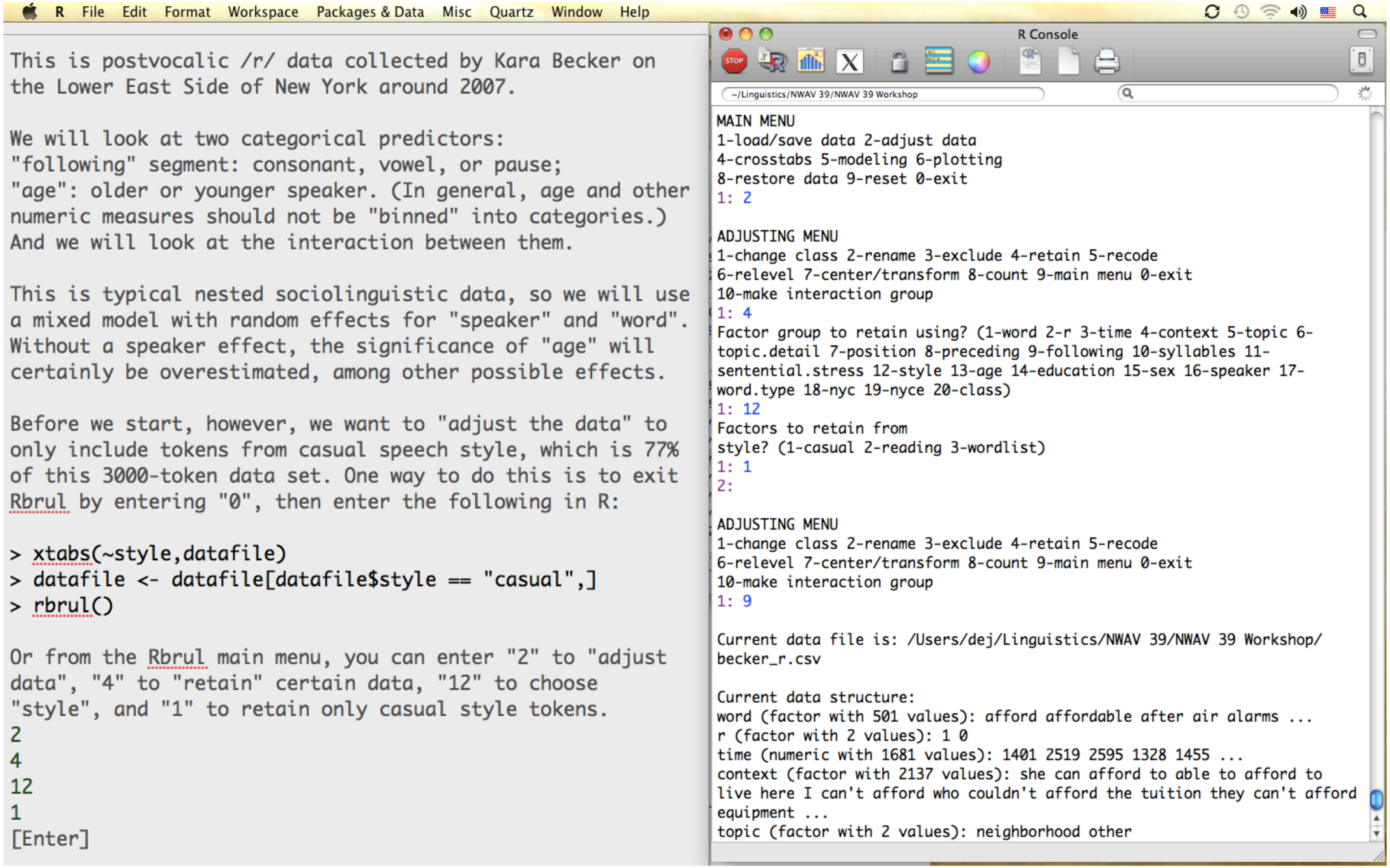
A file-choosing window appears. Open "becker_r.csv".

The MAIN MENU should appear again, with a large number of variables under "Current data structure".

R Console

~/Linguistics/NWAV 39/NWAV 39 Workshop

```
ONE-LEVEL ANALYSIS WITH store (7.57e-20) + word (5.62e-09) + emphasis
(0.0661)
$store
  factor logodds tokens 1/1+0 centered factor weight
    Saks   0.894    177 0.475                  0.71
  Macy's   0.454    336 0.372                 0.612
 Klein's  -1.349    216 0.097                 0.206

$emphasis
   factor logodds tokens 1/1+0 centered factor weight
 emphatic   0.165    271 0.347                 0.541
   normal  -0.165    458 0.297                 0.459

$word
 factor logodds tokens 1/1+0 centered factor weight
  flooR   0.501    347 0.412                 0.623
 fouRth  -0.501    382 0.228                 0.377

$misc
 deviance df intercept grand mean centered input prob Nagelkerke R2
  789.624  5    -0.936      0.316               0.282         0.212


Current variables are:
response.binary: r (1 vs. 0)
fixed.factor: store emphasis word

MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1: 9

Current data file is: /Users/dej/Linguistics/NWAV 39/NWAV 39 Workshop/
ds.csv

Current data structure:
r (factor with 2 values): 1 0
```


R File Edit Format Workspace Packages & Data Misc Quartz Window Help
This is postvocalic /r/ data collected by Kara Becker on the Lower East Side of New York around 2007.
We will look at two categorical predictors:
"following" segment: consonant, vowel, or pause;
"age": older or younger speaker. (In general, age and other numeric measures should not be "binned" into categories.)
And we will look at the interaction between them.
This is typical nested sociolinguistic data, so we will use a mixed model with random effects for "speaker" and "word". Without a speaker effect, the significance of "age" will certainly be overestimated, among other possible effects.
Before we start, however, we want to "adjust the data" to only include tokens from casual speech style, which is 77% of this 3000-token data set. One way to do this is to exit Rbrul by entering "0", then enter the following in R:
> xtabs(~style,datafile)
> datafile <- datafile[datafile$style == "casual",]
> rbrul()
Or from the Rbrul main menu, you can enter "2" to "adjust data", "4" to "retain" certain data, "12" to choose "style", and "1" to retain only casual style tokens.
2
4
12
1
[Enter]
R Console
~/Linguistics/NWAV 39/NWAV 39 Workshop
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: 4
Factor group to retain using? (1-word 2-r 3-time 4-context 5-topic 6-topic.detail 7-position 8-preceding 9-following 10-syllables 11-sentential.stress 12-style 13-age 14-education 15-sex 16-speaker 17-word.type 18-nyc 19-nyce 20-class)
1: 12
Factors to retain from
style? (1-casual 2-reading 3-wordlist)
1: 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: 9
Current data file is: /Users/dej/Linguistics/NWAV 39/NWAV 39 Workshop/becker_r.csv
Current data structure:
word (factor with 501 values): afford affordable after air alarms ...
r (factor with 2 values): 1 0
time (numeric with 1681 values): 1401 2519 2595 1328 1455 ...
context (factor with 2137 values): she can afford to able to afford to live here I can't afford who couldn't afford the tuition they can't afford equipment ...
topic (factor with 2 values): neighborhood other

R File Edit Format Workspace Packages & Data Misc Quartz Window Help

The "Current data structure" should reflect the change.
We now fit the model by entering "9" to return to the main menu, then "5" so that the MODELING MENU opens.
9
5
Enter "1" to "choose variables".
1
Enter "2" to choose "r" as the response, or dependent variable.
2
Press "Enter" as "r" is a binary response.
[Enter]
Enter "2" to choose "1" (presence of /r/) as the application value.
2
Enter "9", "13", "16", "1" to choose "following", "age", "speaker", and "word" as the predictors.
9
13
16
1
[Enter]
"Are any of these predictors continuous?" The answer is no, they are all categorical, so press "Enter".
[Enter]
"Any grouping factors (random effects)?" The answer is yes, "speaker" and "word" are random effects, so type "16", "Enter", "1", "Enter".
16
1
[Enter]
"Consider the interaction between any predictors?" Let's

R Console

~/Linguistics/NWAV 39/NWAV 39 Workshop

```
response.binary: r (1 vs. 0)
fixed.factor: following age sentential.stress
random.intercept: speaker

MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1: 1
Choose response (dependent variable) by number (1-word 2-r 3-time 4-
context 5-topic 6-topic.detail 7-position 8-preceding 9-following 10-
syllables 11-sentential.stress 13-age 14-education 15-sex 16-speaker 17-
word.type 18-nyc 19-nyce 20-class)
1: 2
Type of response? (1-continuous Enter-binary)
1:
Choose application value(s) by number?  (1-0 2-1)
1: 2
Choose predictors (independent variables) by number (1-word 3-time 4-
context 5-topic 6-topic.detail 7-position 8-preceding 9-following 10-
syllables 11-sentential.stress 13-age 14-education 15-sex 16-speaker 17-
word.type 18-nyc 19-nyce 20-class)
1: 9
2: 13
3: 16
4: 1
5:
Are any predictors continuous? (9-following 13-age 16-speaker 1-word
Enter-none)
1:
Any grouping factors (random effects)? (9-following 13-age 16-speaker 1-
word Enter-none)
1: 16
2: 1
3:
Consider a(nother) pairwise interaction between predictors? Choose two at
a time. (9-following 13-age 16-speaker 1-word Enter-done)
1:
```

[Enter]
"Consider a(nother) pairwise interaction between predictors?" To include the "age:following" interaction, type "9", "13".
9
13
[Enter]
The MODELING MENU should now show "Current variables are:"
response.binary: r (1 vs. 0)
fixed.factor: following age
fixed.interaction: following:age
random.intercept: speaker word

From the MODELING MENU, enter "2" for "one-level".
2

A very long output results, because the default setting is to display all the levels of the random effects. Here we note that the standard deviation for words is estimated at 0.765 log-odds and that for speakers is estimated at 0.452.

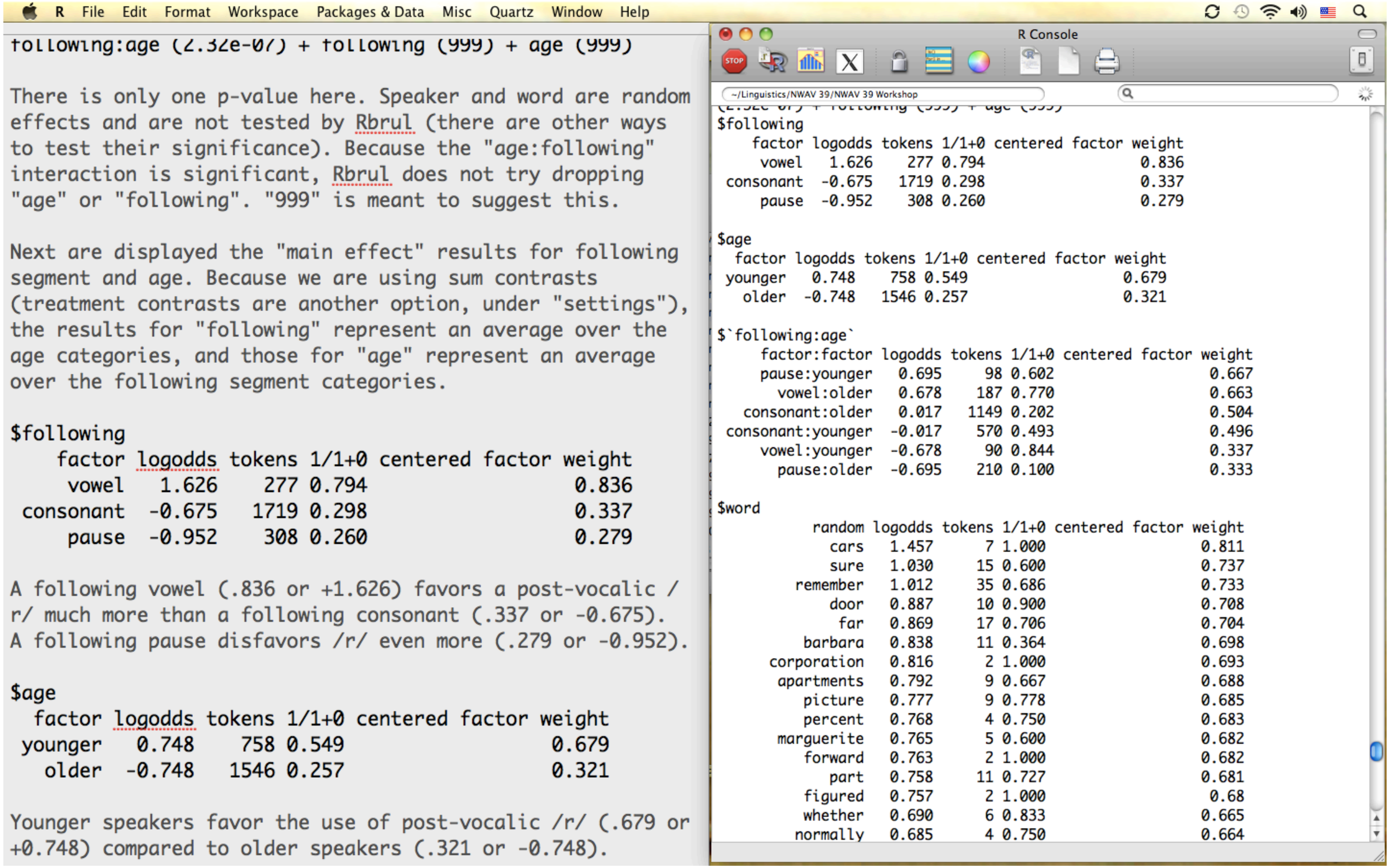
The first line of the output is:

ONE-LEVEL ANALYSIS WITH speaker (random) + word (random) + following:age (2.32e-07) + following (999) + age (999)

There is only one p-value here. Speaker and word are random effects and are not tested by Rbrul (there are other ways to test their significance). Because the "age:following" interaction is significant, Rbrul does not try dropping "age" or "following". "999" is meant to suggest this.

R Console

```
context 5-topic 6-topic.detail 7-position 8-preceding 9-following 10-
syllables 11-sentential.stress 13-age 14-education 15-sex 16-speaker 17-
word.type 18-nyc 19-nyce 20-class)
1: 9
2: 13
3: 16
4: 1
5:
Are any predictors continuous? (9-following 13-age 16-speaker 1-word
Enter-none)
1:
Any grouping factors (random effects)? (9-following 13-age 16-speaker 1-
word Enter-none)
1: 16
2: 1
3:
Consider a(nother) pairwise interaction between predictors? Choose two at
a time. (9-following 13-age 16-speaker 1-word Enter-done)
1: 9
2: 13
Consider a(nother) pairwise interaction between predictors? Choose two at
a time. (9-following 13-age 16-speaker 1-word Enter-done)
1:

Current variables are:
response.binary: r (1 vs. 0)
fixed.factor: following age
fixed.interaction: following:age
random.intercept: speaker word

MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1: 2

ONE-LEVEL ANALYSIS WITH speaker (random) + word (random) + following:age
(2.32e-07) + following (999) + age (999)
```



following:age (2.32e-07) + following (999) + age (999)

There is only one p-value here. Speaker and word are random effects and are not tested by Rbrul (there are other ways to test their significance). Because the "age:following" interaction is significant, Rbrul does not try dropping "age" or "following". "999" is meant to suggest this.

Next are displayed the "main effect" results for following segment and age. Because we are using sum contrasts (treatment contrasts are another option, under "settings"), the results for "following" represent an average over the age categories, and those for "age" represent an average over the following segment categories.

```
$following
    factor logodds tokens 1/1+0 centered factor weight
     vowel   1.626    277 0.794                  0.836
 consonant  -0.675   1719 0.298                  0.337
     pause  -0.952    308 0.260                  0.279
```

A following vowel (.836 or +1.626) favors a post-vocalic /r/ much more than a following consonant (.337 or -0.675). A following pause disfavors /r/ even more (.279 or -0.952).

```
$age
  factor logodds tokens 1/1+0 centered factor weight
 younger   0.748    758 0.549                  0.679
   older  -0.748   1546 0.257                  0.321
```

Younger speakers favor the use of post-vocalic /r/ (.679 or +0.748) compared to older speakers (.321 or -0.748).

```
(2.32e-07) + following (999) + age (999)
$following
    factor logodds tokens 1/1+0 centered factor weight
     vowel   1.626    277 0.794                  0.836
 consonant  -0.675   1719 0.298                  0.337
     pause  -0.952    308 0.260                  0.279

$age
  factor logodds tokens 1/1+0 centered factor weight
 younger   0.748    758 0.549                  0.679
   older  -0.748   1546 0.257                  0.321

$`following:age`
     factor:factor logodds tokens 1/1+0 centered factor weight
     pause:younger   0.695     98 0.602                  0.667
       vowel:older   0.678    187 0.770                  0.663
   consonant:older   0.017   1149 0.202                  0.504
 consonant:younger  -0.017    570 0.493                  0.496
     vowel:younger  -0.678     90 0.844                  0.337
       pause:older  -0.695    210 0.100                  0.333

$word
          random logodds tokens 1/1+0 centered factor weight
            cars   1.457      7 1.000                  0.811
            sure   1.030     15 0.600                  0.737
        remember   1.012     35 0.686                  0.733
            door   0.887     10 0.900                  0.708
             far   0.869     17 0.706                  0.704
         barbara   0.838     11 0.364                  0.698
     corporation   0.816      2 1.000                  0.693
      apartments   0.792      9 0.667                  0.688
         picture   0.777      9 0.778                  0.685
         percent   0.768      4 0.750                  0.683
      marguerite   0.765      5 0.600                  0.682
         forward   0.763      2 1.000                  0.682
            part   0.758     11 0.727                  0.681
         figured   0.757      2 1.000                   0.68
         whether   0.690      6 0.833                  0.665
        normally   0.685      4 0.750                  0.664
```

Stickies File Edit Font Note Color Window Help

The results for the interaction show a coefficient for all six combinations of the interacting variables. This can be misleading because there are only three parameters in a 3x2 interaction such as we have here. The redundancy is seen by each coefficient's having a "mirror image". Using treatment instead of sum contrasts, the numbers would look different.

```
$`following:age`
     factor:factor logodds tokens 1/1+0 centered f.w.
     pause:younger   0.695     98 0.602          0.667
       vowel:older   0.678    187 0.770          0.663
   consonant:older   0.017   1149 0.202          0.504
 consonant:younger  -0.017    570 0.493          0.496
     vowel:younger  -0.678     90 0.844          0.337
       pause:older  -0.695    210 0.100          0.333
```

We could parse this by saying that while following consonants behave similarly for the two age groups, the /r/-favoring effect of a following vowel and the /r/-disfavoring effect of a following pause are both two or three times larger for the older speakers.

We get this result by adding the interaction (log-odds) coefficients to those for the main effects. For example, the average following-vowel effect is +1.626, but it is 1.626 + 0.678 = 2.304 for the older speakers compared to 1.626 - 0.678 = 0.948 for the younger speakers.

These slides have introduced most of the features of Rbrul. If a continuous predictor is used, its coefficient(s) are for a one-unit increase in the predictor. With a continuous response, coefficients are in the units of the response.

R Console

~/Linguistics/NWAV 39/NWAV 39 Workshop

```
$following
    factor logodds tokens 1/1+0 centered factor weight
     vowel   1.626    277 0.794                  0.836
 consonant  -0.675   1719 0.298                  0.337
     pause  -0.952    308 0.260                  0.279

$age
  factor logodds tokens 1/1+0 centered factor weight
 younger   0.748    758 0.549                  0.679
   older  -0.748   1546 0.257                  0.321

$`following:age`
     factor:factor logodds tokens 1/1+0 centered factor weight
     pause:younger   0.695     98 0.602                  0.667
       vowel:older   0.678    187 0.770                  0.663
   consonant:older   0.017   1149 0.202                  0.504
 consonant:younger  -0.017    570 0.493                  0.496
     vowel:younger  -0.678     90 0.844                  0.337
       pause:older  -0.695    210 0.100                  0.333

$word
          random logodds tokens 1/1+0 centered factor weight
            cars   1.457      7 1.000                  0.811
            sure   1.030     15 0.600                  0.737
        remember   1.012     35 0.686                  0.733
            door   0.887     10 0.900                  0.708
             far   0.869     17 0.706                  0.704
         barbara   0.838     11 0.364                  0.698
     corporation   0.816      2 1.000                  0.693
      apartments   0.792      9 0.667                  0.688
         picture   0.777      9 0.778                  0.685
         percent   0.768      4 0.750                  0.683
      marguerite   0.765      5 0.600                  0.682
         forward   0.763      2 1.000                  0.682
            part   0.758     11 0.727                  0.681
         figured   0.757      2 1.000                   0.68
         whether   0.690      6 0.833                  0.665
        normally   0.685      4 0.750                  0.664
```

try using Rbrul with your own data

- must be saved as .csv file (for example from Excel)
- the first row must be the names of the variables
- there must not be any extra cells or partial rows

- try fitting an ordinary model (no random effects)
- try fitting a mixed model (one or more random effects)
- what difference did it make?

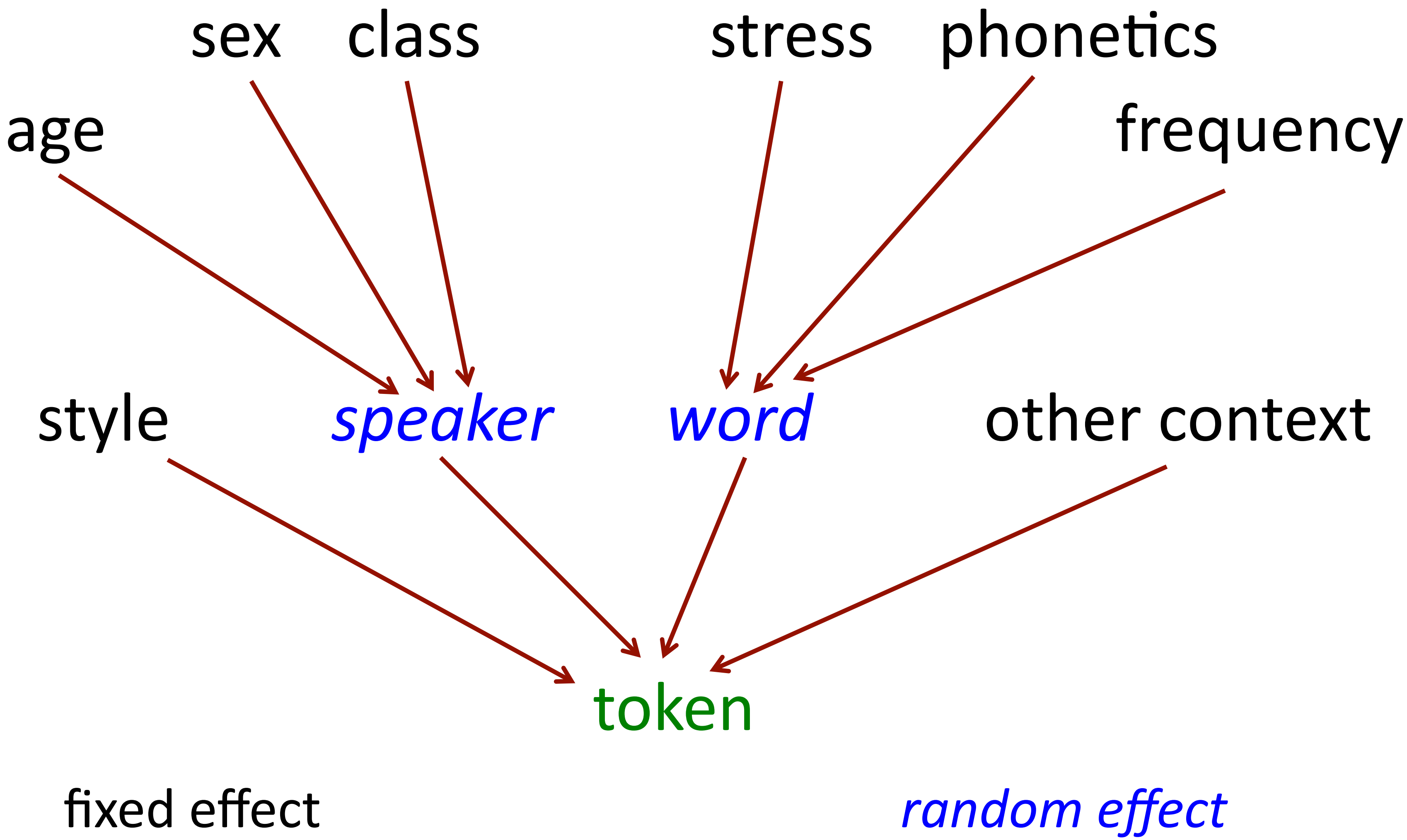
- resist stepwise regression as much as possible
- use "one-level" option to fit the model you specify

loading data into Rbrul

- Rbrul uses a file picker that can’t access websites
- you can circumvent file picker as follows:
- datafile <- read.csv(“~/Desktop/my_data.csv”)
- datafile <- read.csv(“http://www.web.com/data.csv”)

- some secret data if you don’t have your own today:
  - Labov department store coda /r/ (not good for mixed model) http://www.danielezrajohnson.com/ds.csv
  - Becker New York coda /r/ http://www.danielezrajohnson.com/becker_r.csv
  - AISEB Gretna coda /r/ http://www.danielezrajohnson.com/gretna.csv

architecture of variables



4. Mixed models in R

try using R with your own data

- don’t use the command line, use a script, and save it
- try a cross-tabulation:
  > xtabs(~ x + y, data)
- try a plot:
  > plot(y ~ x, data)
- try a model:
  > glm(y ~ x, data, family = …)
  > lmer(y ~ x + (1|speaker), data, family = …) random intercept
  > lmer(y ~ x + (x|speaker), data, family = …) random slope
- try some other statistic – look it up or ask me
- exploratory: graphing preferred (origins of R)
- confirmatory (specific hypothesis testing): models good

syntax for lm(), glm(), (g)lmer()

- no random effects? use lm() or glm()
  - m.linear <- lm(y ~ x1 *+ x2 + x1:x2…*, data)
  - m.linear <- glm(y ~ x1 *+ x2 + x1:x2…*, data, *family = gaussian*)
  - m.logistic <- glm(y ~ x1 *+ x2 + x1:x2…*, data, family = binomial)
- random effects? use lmer() aka glmer()
  - install.packages("lme4"); library(lme4)
  - m.mixed <- lmer(y ~ x1 *+ x2 + x1:x2* + (1|speaker) + (1|word), data*, family = binomial)*
  - m.maximal <- lmer(y ~ x1 + x2 + (x1|speaker) + (x2|word), data, *family = binomial*)
- to test fixed effect significance, e.g. of x1
  - m.maximal.no.x1 <-lmer(y ~ x2 + (x1|speaker) + (x2|word), data, *family = binomial*)
  - anova(m.maximal.no.x1, m.maximal) *note: don't overly trust p-values from mixed models*

some references

R-sig-ME listserv: stat.ethz.ch/mailman/listinfo/r-sig-mixed-models

FAQ for the above: http://glmm.wikidot.com/faq

R-Lang listserv: http://pidgin.ucsd.edu/mailman/listinfo/r-lang

Journal of Memory and Language 59 (issue on Emerging Data Analysis)

Barr et al. 2013. “Random effects structure… keep it maximal.”
    http://idiom.ucsd.edu/~rlevy/papers/barr-etal-2013-jml.pdf

Bates lme4 bible (draft): http://lme4.R-forge.R-project.org/book/

older testament:

    Pinheiro & Bates 2000. Mixed-Effects Models in S and S-PLUS.

other books:

    Harrell, Frank E. 2001. Regression Modeling Strategies.

    Baayen, R. Harald. 2008. Analyzing Linguistic Data.

    Field, Andy. 2012. Discovering Statistics Using R.

    Fruehwald, Josef. Forthcoming. Starting Quantitative Analysis with R.