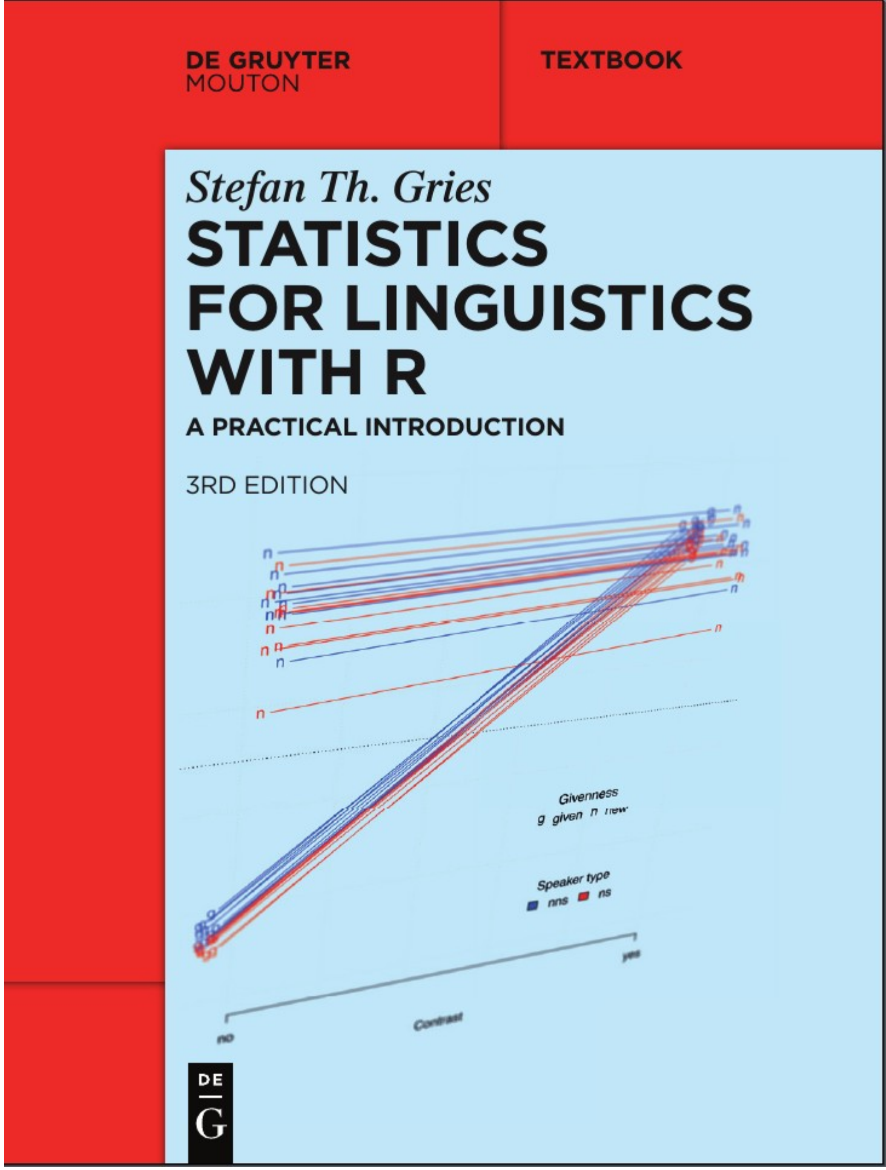
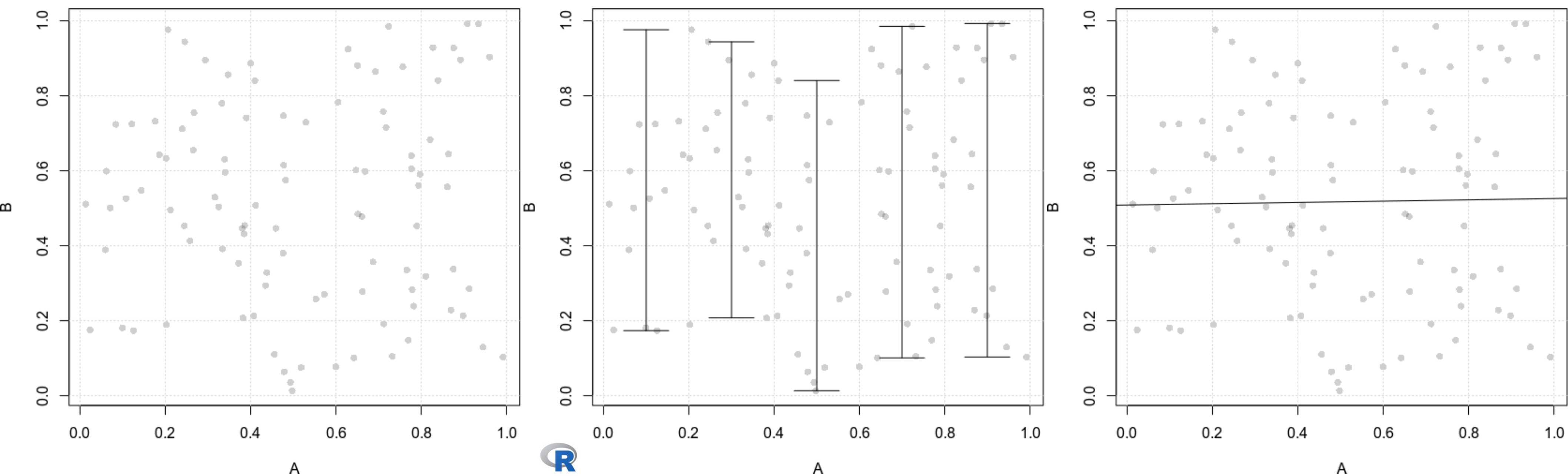
Predictive modeling in linguistics with R



Stefan Th. Gries
UC Santa Barbara & JLU Giessen
http://www.stgries.info

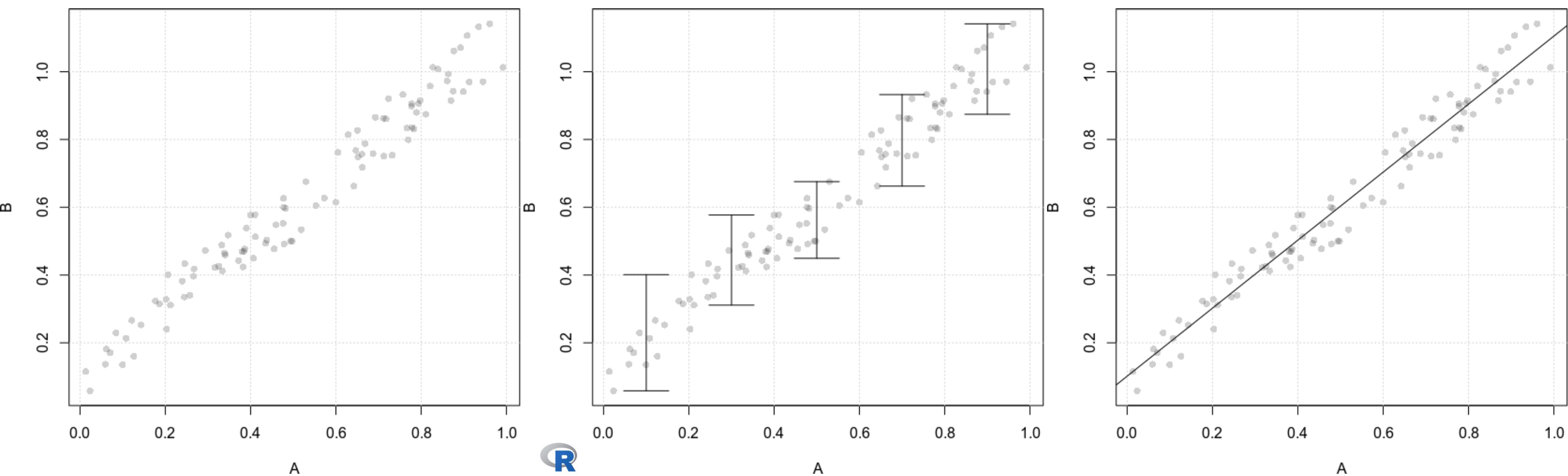
A brief excursus on the notion of correlation (part 1)

- What does it mean if one says that 'variables A and B are correlated'?
- variables (A & B) are correlated if knowing the value (range) of A makes it easier to 'predict' the value (range) of B than not knowing A



A brief excursus on the notion of correlation (part 1)

- What does it mean if one says that 'variables A and B are correlated'?
- variables (A & B) are correlated if knowing the value (range) of A makes it easier to 'predict' the value (range) of B than not knowing A



A brief excursus on the notion of correlation (part 2)

- What kinds of variables can be correlated with each other?
- all kinds of variables can be correlated with each other
  - null/no correlation
    - knowing the level of the categorical or ordinal variable A (*A1* vs *A2* vs *A3*) doesn't help you 'predict' the level of the categorical or ordinal variable B (*B1* vs *B2* vs *B3*)
  - strong correlation
    - knowing the level of the categorical or ordinal variable A (*A1* vs *A2* vs *A3*) does help you 'predict' the level of the categorical or ordinal variable B (*B1* vs *B2* vs *B3*)

	B1	B2	B3	Sum
A1				300
A2				300
A3				300
Sum	300	300	300	900

	B1	B2	B3	Sum
A1	100	100	100	300
A2	100	100	100	300
A3	100	100	100	300
Sum	300	300	300	900

	B1	B2	B3	Sum
A1	250	25	25	300
A2	25	250	25	300
A3	25	25	250	300
Sum	300	300	300	900

From monofactorial to multifactorial

- Monofactorial tests involve
  - 1 dependent/response variable
  - 1 independent/predictor variable
- there are two big howevers here, though
  - however$_1$, there is probably no linguistic phenomenon that is monofactorial – they're probably all multifactorial
  - however$_2$, there is probably hardly any situation where you should really be doing a monofactorial test
- monofactorial studies have probably nothing to contribute to most linguistic work (there, I said it!)
- for the sake of simplicity, let's explore this with a very mundane non-linguistic example, the efficiency of cars measured in mpg

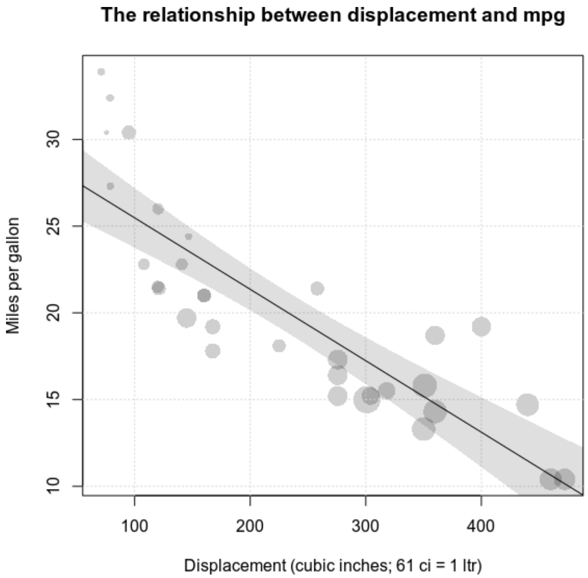
What affects the efficiency of cars (measured in mpg)?

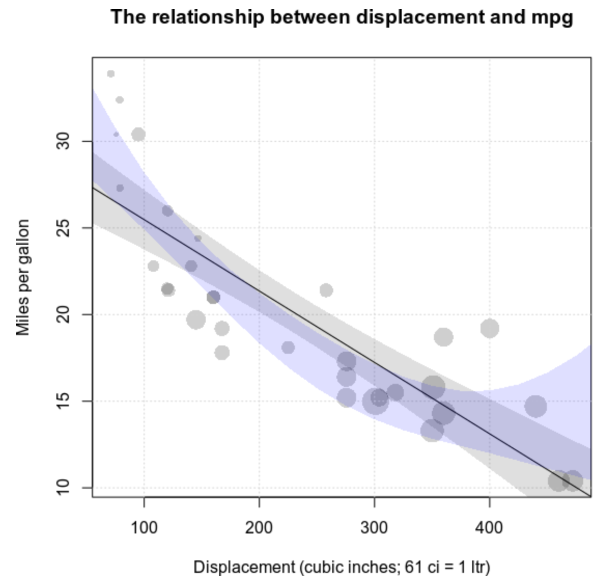
- You read up on this and think about it and
  - find a 1992 study that shows that cylinder does
  - find a 1996 study that shows that horsepower does
  - consider it reasonable physics that weight does
- and then you think that displacement should have an effect and collect data to test this correlationally

```
> summary(test.of.new.hyp <- lm(mpg ~ disp, data=mtcars))
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.599855   1.229720  24.070  < 2e-16 ***
disp        -0.041215   0.004712  -8.747 9.38e-10 ***
Multiple R-squared:  0.7183,  Adjusted R-squared:  0.709
F-statistic: 76.51 on 1 and 30 DF,  p-value: 9.38e-10
```

- but this test is ridiculously anti-conservative
  - you're testing the H1 that disp is related to mpg against the H0 that it is not
  - you're not testing the H1 that disp is related to mpg against the H0 of everything else we already know

What affects the efficiency of cars (measured in mpg)?





What affects the efficiency of cars (measured in mpg)?

- In other words,
  - you're pretending we know nothing about mpg already
  - you're leaving all mpg variability up for grabs by disp
- but that's delusional/too generous: you already know that cylinder, horsepower, & weight affect mpg

```
> summary(prior.knowl <- lm(mpg ~ cyl+hp+wt, data=mtcars))
[…]
Multiple R-squared:  0.843,   Adjusted R-squared:  0.8263 ***
```

- option 1: you need to test whether disp **adds to** what we already know

```
> summary(real.test.of.new.hyp <- lm(mpg ~ disp+cyl+hp+wt, data=mtcars))
Multiple R-squared:  0.8486,  Adjusted R-squared:  0.8262 ***

> anova(prior.knowl, real.test.of.new.hyp, test="F")
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
2     27 170.44  1    6.1762 0.9784 0.3314
```

What affects the efficiency of cars (measured in mpg)?

- In other words,
  - you're pretending we know nothing about mpg already
  - you're leaving all mpg variability up for grabs by disp
- but that's delusional/too generous: you already know that cylinder, horsepower, & weight affect mpg

```
> summary(prior.knowl <- lm(mpg ~ cyl+hp+wt, data=mtcars))
[…]
Multiple R-squared:  0.843,   Adjusted R-squared:  0.8263 ***
```

- option 2: you need to test whether disp **replaces** what we already know

```
> exp((MuMIn::AICc(test.of.new.hyp) - MuMIn::AICc(prior.knowl))/2)
[1] 765.9413 # prior knowledge is this much more likely to be the right model
```

- it does neither …
- why?
- disp is >90% predictable from cyl, hp, wt, …

From monofactorial to multifactorial

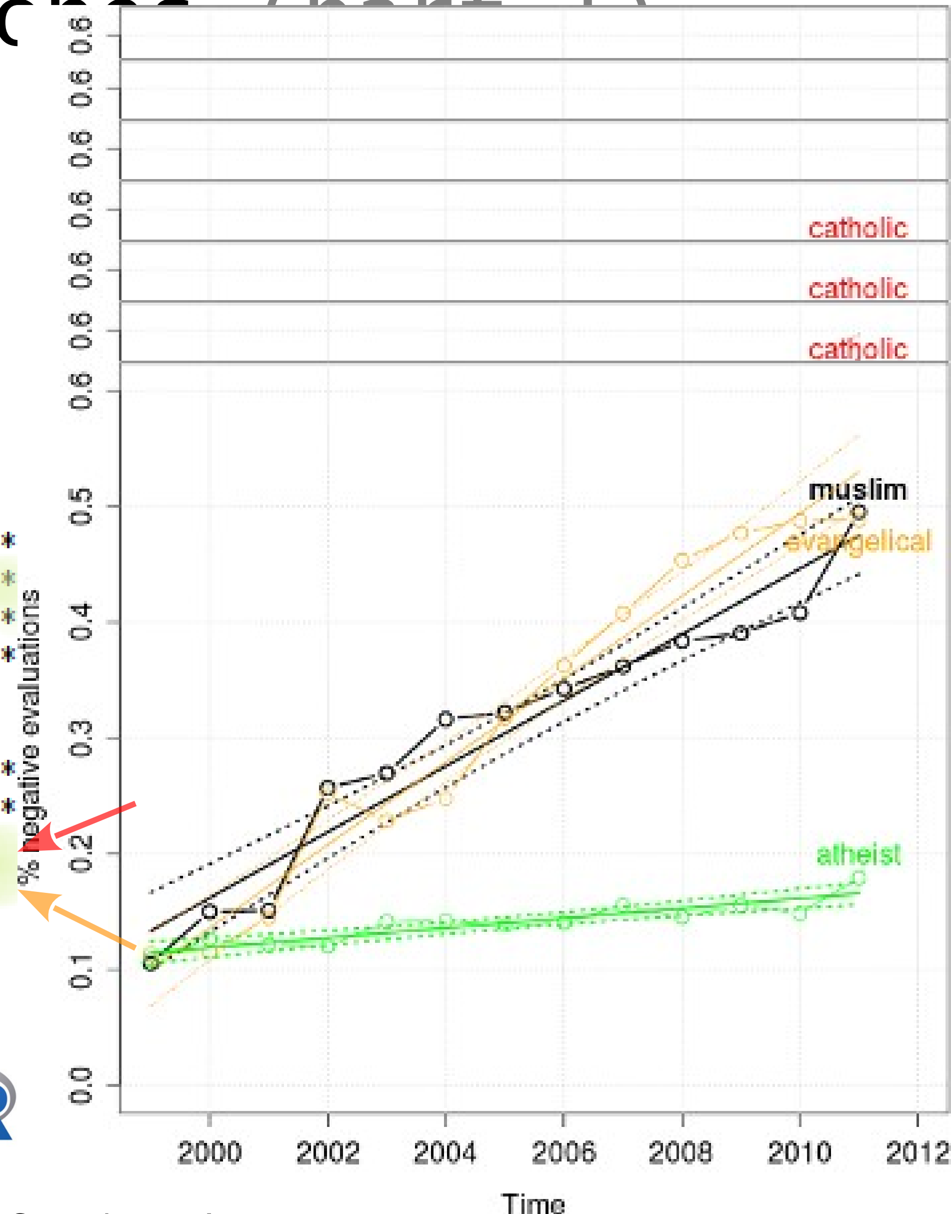
- We now said that we need multifactorial methods
- however, once you have 2+ independent/predictor variables (as in `lm(mpg ~ cyl+hp+wt)`), they can behave in two different ways together
  - additively, which is what we just saw
  - interactively, and the concept of interaction is one of the easiest yet also probably one of the most underestimated, underutilized, and misunderstood notions
  - interaction is related to (effect) modification and a special type of conditional dependence: when the association between a predictor & the response is not constant across another characteristic
  - interaction: X doesn't do the same everywhere/always

What interactions can reveal: differences in slopes (part 1)

	SOURCE: guardian			SOURCE: the sun			…
YEAR	TERM	FREQUENCY	VALUE	TERM	FREQUENCY	VALUE	…
2002	muslim	10	negative	muslim	…	negative	…
2003	muslim	16	negative	muslim	…	negative	…
2004	muslim	23	negative	muslim	…	negative	…
2005	muslim	30	negative	muslim	…	negative	…
2002	muslim	100	neutral	muslim	…	neutral	…
2003	muslim	158	neutral	muslim	…	neutral	…
2004	muslim	225	neutral	muslim	…	neutral	…
2005	muslim	270	neutral	muslim	…	neutral	…
2002	muslim	30	positive	muslim	…	positive	…
2003	muslim	54	positive	muslim	…	positive	…
2004	muslim	88	positive	muslim	…	positive	…
2005	muslim	115	positive	muslim	…	positive	…
2002	evangelical	…	negative	evangelical	…	negative	…
2003	evangelical	…	negative	evangelical	…	negative	…
2004	evangelical	…	negative	evangelical	…	negative	…
2005	evangelical	…	negative	evangelical	…	negative	…
2002	evangelical	…	neutral	evangelical	…	neutral	…
2003	evangelical	…	neutral	evangelical	…	neutral	…
2004	evangelical	…	neutral	evangelical	…	neutral	…
2005	evangelical	…	neutral	evangelical	…	neutral	…
2002	evangelical	…	positive	evangelical	…	positive	…
2003	evangelical	…	positive	evangelical	…	positive	…
2004	evangelical	…	positive	evangelical	…	positive	…
2005	evangelical	…	positive	evangelical	…	positive	…
2002	catholic	…	negative	catholic	…	negative	…

What interactions can reveal: differences in slopes (part 1)

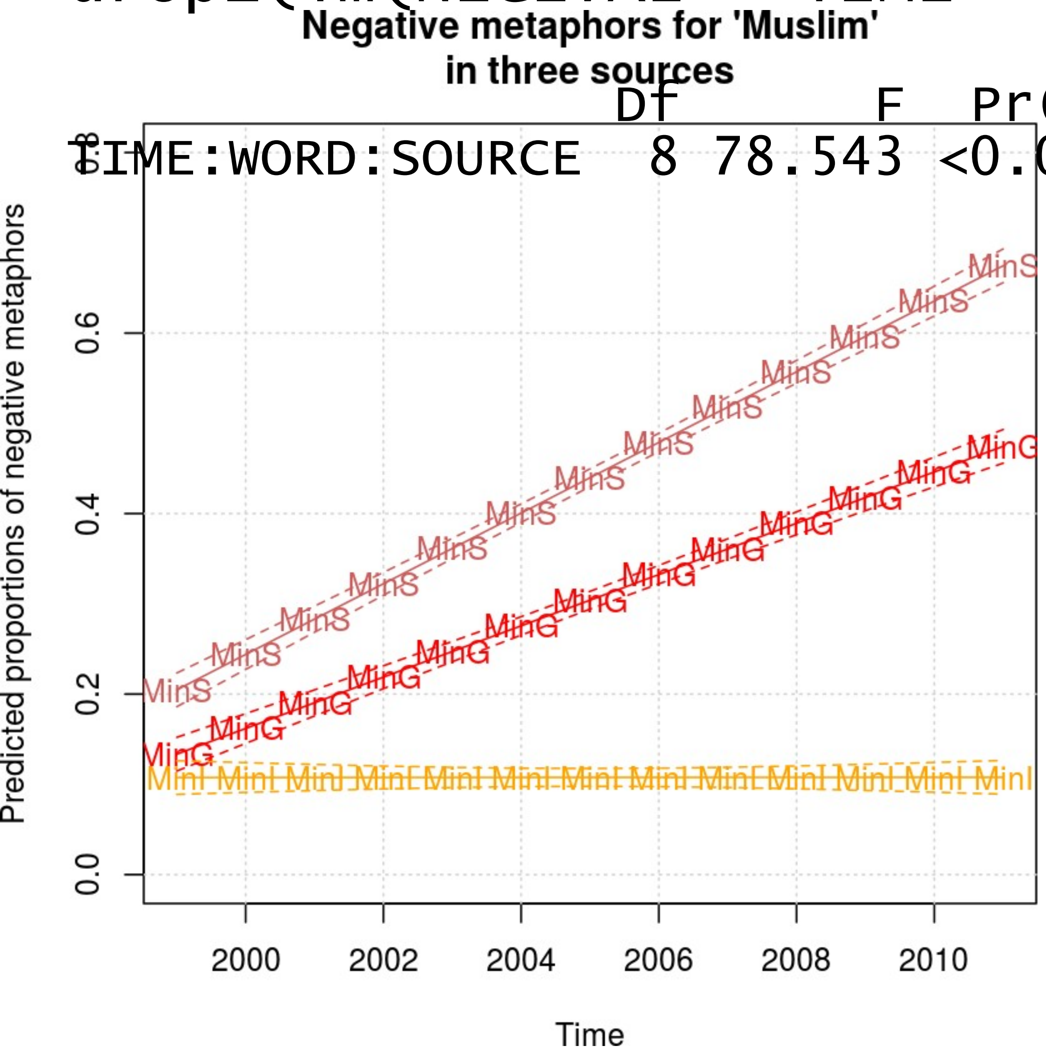
```
> summary(model.01)

Call:
lm(formula = NEGEVAL ~ TIME * WORD)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.041519 -0.008056 -0.000604  0.011717  0.044168 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        -56.692127   3.043619 -18.627  < 2e-16 ***
TIME                 0.028427   0.001518  18.726  < 2e-16 ***
WORDatheist         48.328480   4.304327  11.228 7.49e-16 ***
WORDbuddhist        28.394695   4.304327   6.597 1.72e-08 ***
WORDcatholic        -5.994329   4.304327  -1.393  0.16934    
WORDevangelical    -14.686952   4.304327  -3.412  0.00122 ** 
TIME:WORDatheist    -0.024186   0.002147 -11.266 6.58e-16 ***
TIME:WORDbuddhist   -0.014194   0.002147  -6.612 1.63e-08 ***
TIME:WORDcatholic    0.003030   0.002147   1.412  0.16370    
TIME:WORDevangelical 0.007331   0.002147   3.415  0.00121 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02048 on 55 degrees of freedom
Multiple R-squared: 0.9783, Adjusted R-squared: 0.9748
F-statistic:   276 on 9 and 55 DF,  p-value: < 2.2e-16
```

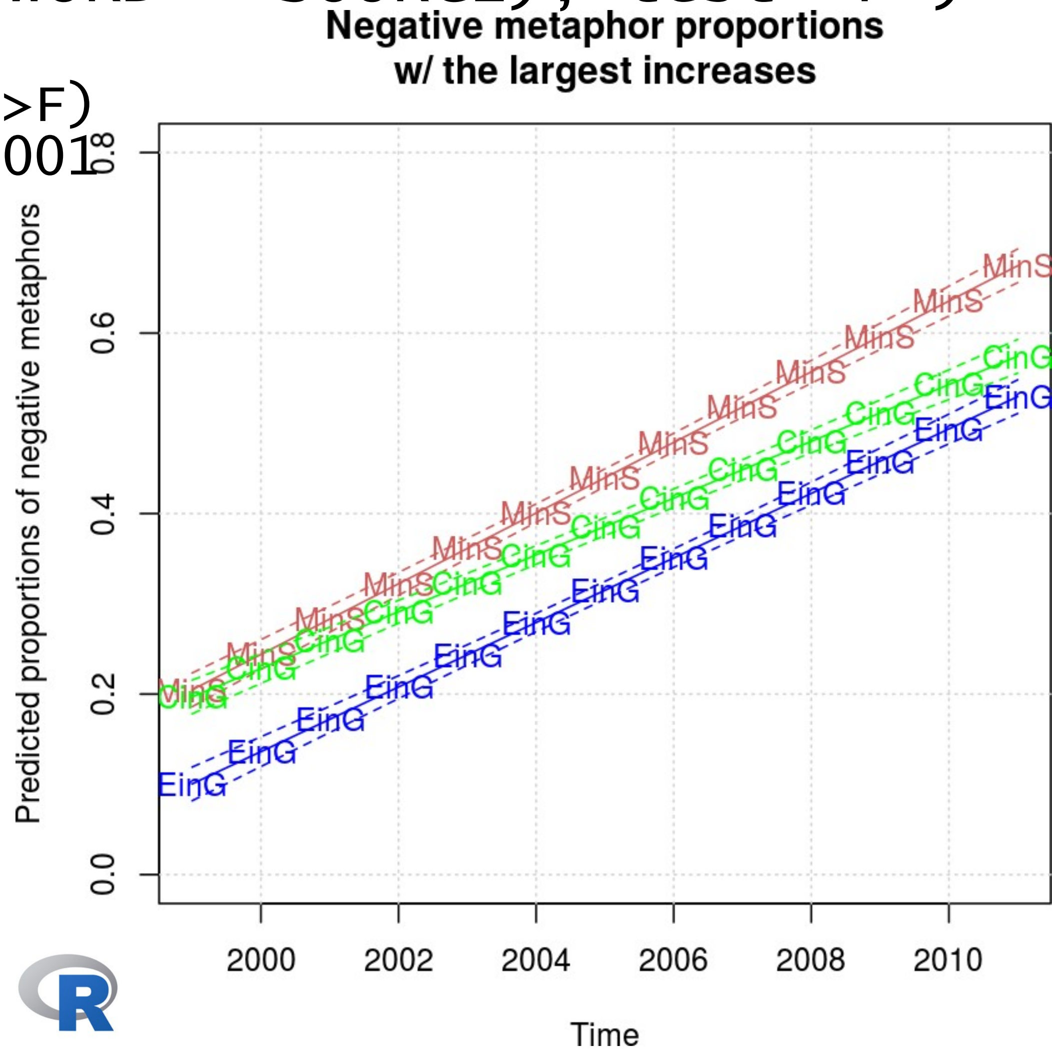


What interactions can reveal: differences in slopes (part 1)

- drop1(lm(NEGEVAL ~ TIME * WORD * SOURCE), test="F")

	Df	F	Pr(>F)
TIME:WORD:SOURCE	8	78.543	<0.0001

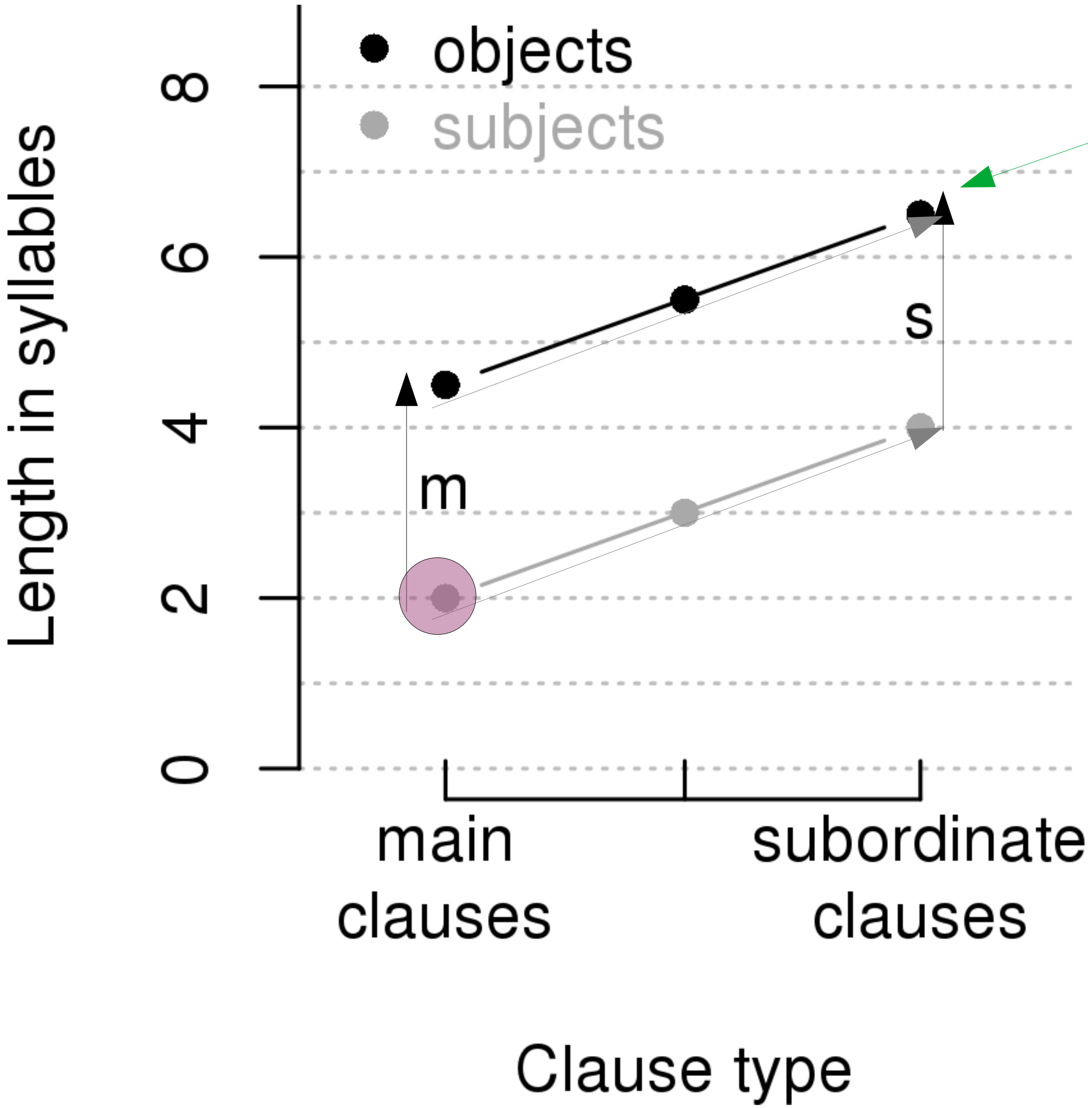




What often happens in multifactorial approaches: an example

- Subjects and direct objects in 60 main and 60 subordinate clauses are studied
- half of the subjects and objects are in main clauses, the other half in subordinate clauses
- the dependent variable is the length of the subjects/objects in syllables …
- that is, we are dealing with a multifactorial design
  - independent variable 1: clause type (main vs. subord.)
  - independent variable 2: grm relation (subj. vs. obj.)
- example results
  - monofactorial finding 1: mean $_{\text{length main}}$ < mean $_{\text{length subord}}$
  - monofactorial finding 2: mean $_{\text{length subj}}$ < mean $_{\text{length obj}}$
- given these monofactorial findings,
  - which of the four combinations will exhibit the longest constituents?
  - which of the four combinations will exhibit the shortest constituents?

```
> summary(lm(LENGTH ~ 1 + GRAMREL + CLAUSE + GRAMREL:CLAUSE, data=s1))
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.01   0.016297 123.396   <2e-16 ***     the intercept +
GRAMREL: subj→obj     2.49   0.023048 108.349   <2e-16 ***     this          +
CLAUSE: main→subord   2.00   0.023048  86.789   <2e-16 ***     this          (i.e. additive behavior)
GRAMREL:CLAUSE        0.00   0.032595   0.145    0.885         predicts 6.5 (as it should w/ no iact)
```



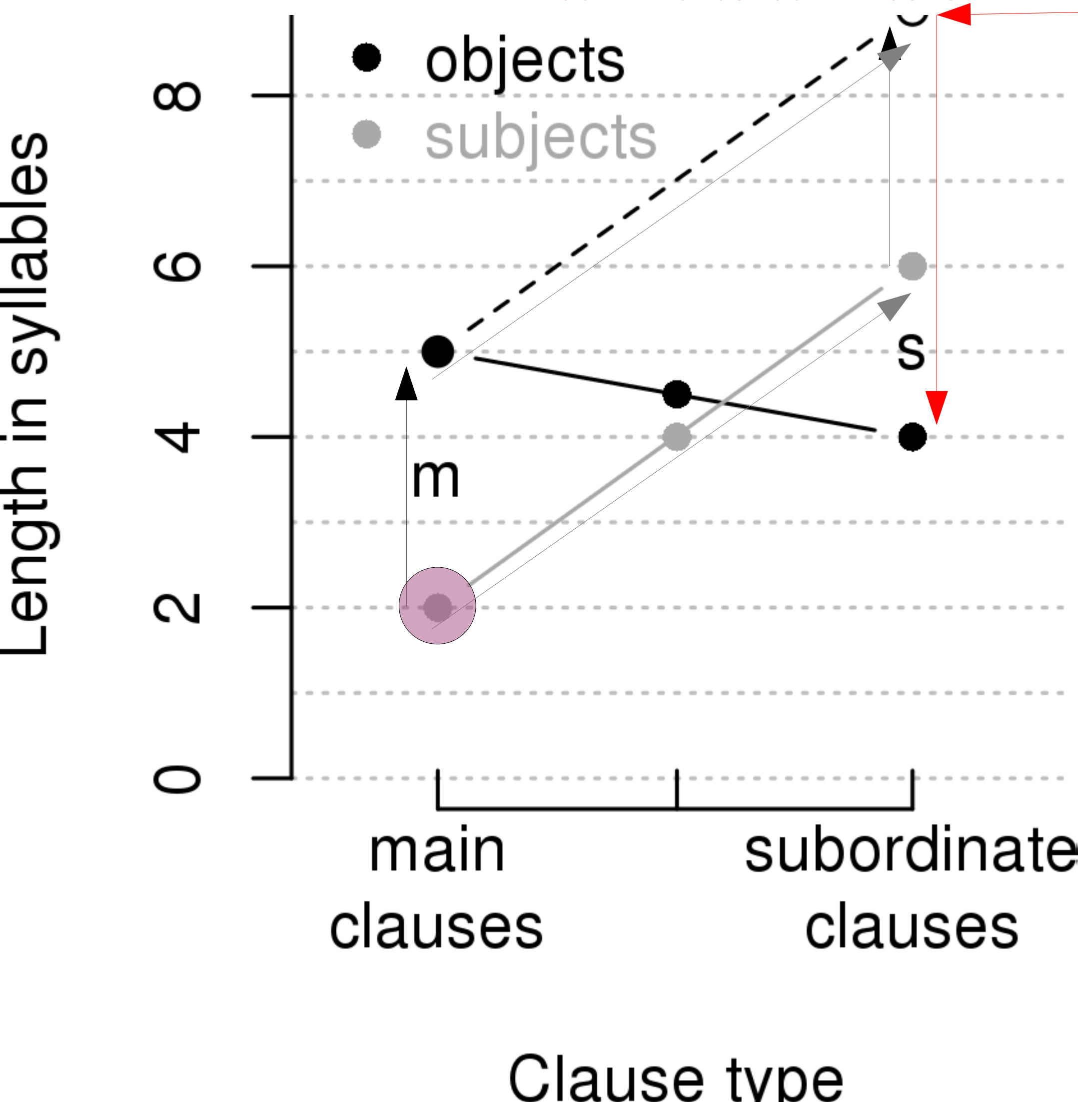
mean (subject) $<_{2.5}$ mean (object)

mean (main) $<_{2}$ mean (subordinate)

clause type and grammatical relation influence length additively

	obj	subj	diff
main	4.5	2	2.5
subord	6.5	4	2.5
diff	2	2	

```
> summary(lm(LENGTH ~ 1 + GRAMREL + CLAUSE + GRAMREL:CLAUSE, data=s2))
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.01    0.01630   123.4   <2e-16 ***    the intercept +
GRAMREL: subj→obj          2.99    0.02305   130.0   <2e-16 ***    this           +
CLAUSE: main→subord        4.00    0.02305   173.6   <2e-16 ***    this           (i.e. additive behavior)
GRAMREL:CLAUSE            -4.99    0.03259  -153.3   <2e-16 ***    predicts 9 but we need to predict 4
```



mean (subject) $<_{0.5}$
mean (object)

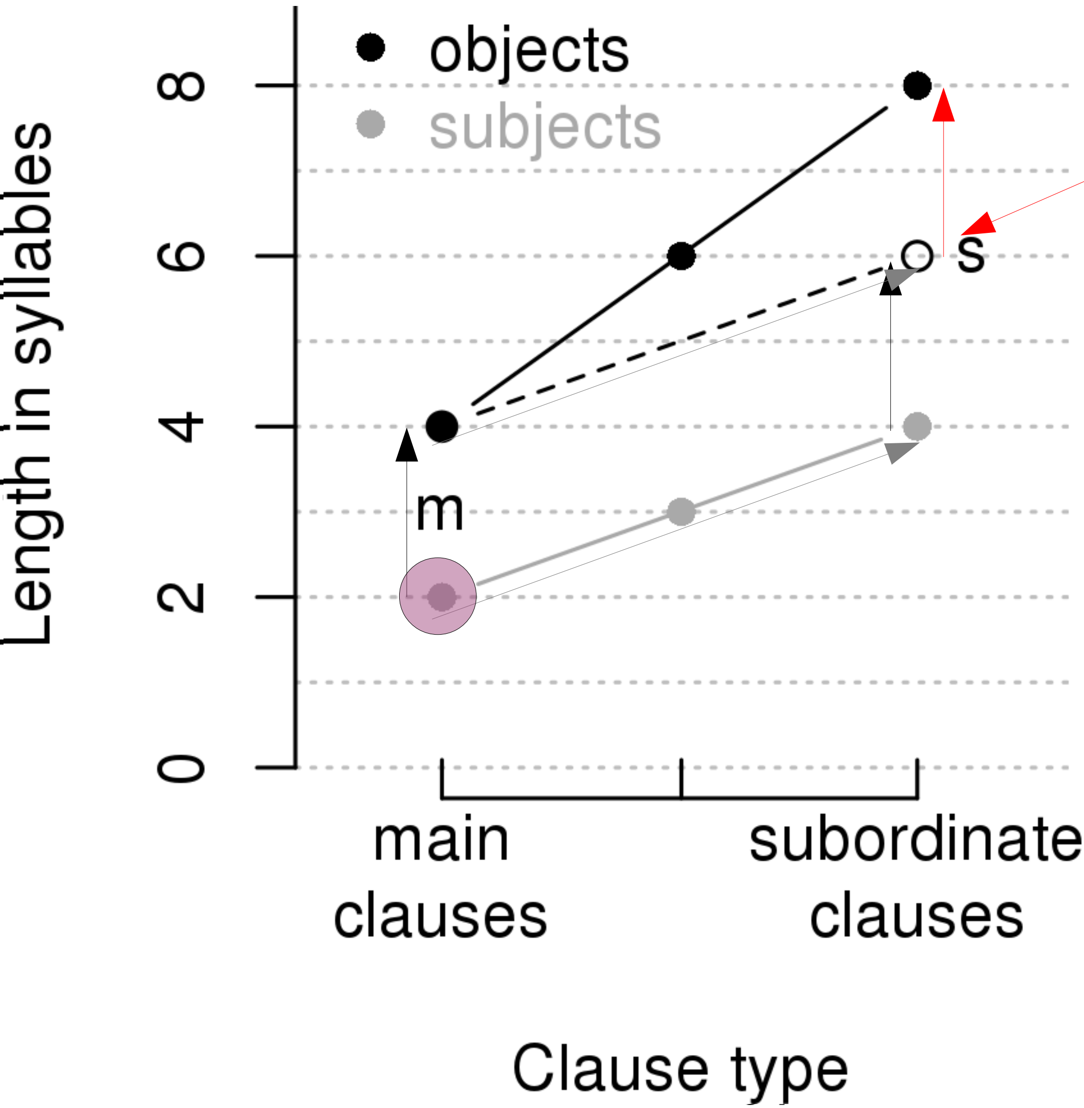
mean (main) $<_{1.5}$
mean (subordinate)

clause type and grammatical relation influence length interactively

	obj	subj	diff
main	5	2	3
subord	4	6	-2
diff	-1	4	

```
> summary(lm(LENGTH ~ 1 + GRAMREL + CLAUSE + GRAMREL:CLAUSE, data=s3))
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.01    0.01630   123.4   <2e-16 ***
GRAMREL: subj→obj     1.99    0.02305   86.66   <2e-16 ***
CLAUSE: main→subord   2.00    0.02305   86.79   <2e-16 ***
GRAMREL:CLAUSE        2.00    0.03259   61.51   <2e-16 ***
```

the intercept +
this +
this (i.e. additive behavior)
predicts 6, but we need to predict 8



mean (subject) $<_3$ mean (object)

mean (main) $<_3$ mean (subordinate)

clause type and grammatical relation influence length interactively

	obj	subj	diff
main	4	2	2
subord	8	4	4
diff	4	2	

What interactions can reveal: mean vs. slope

- Example: predicting mistakes in L2-English dictation
  - indep. vars: mistakes in L1-German dictation and class
- model 1: ENGL ~ GER + CLASS + GER:CLASS

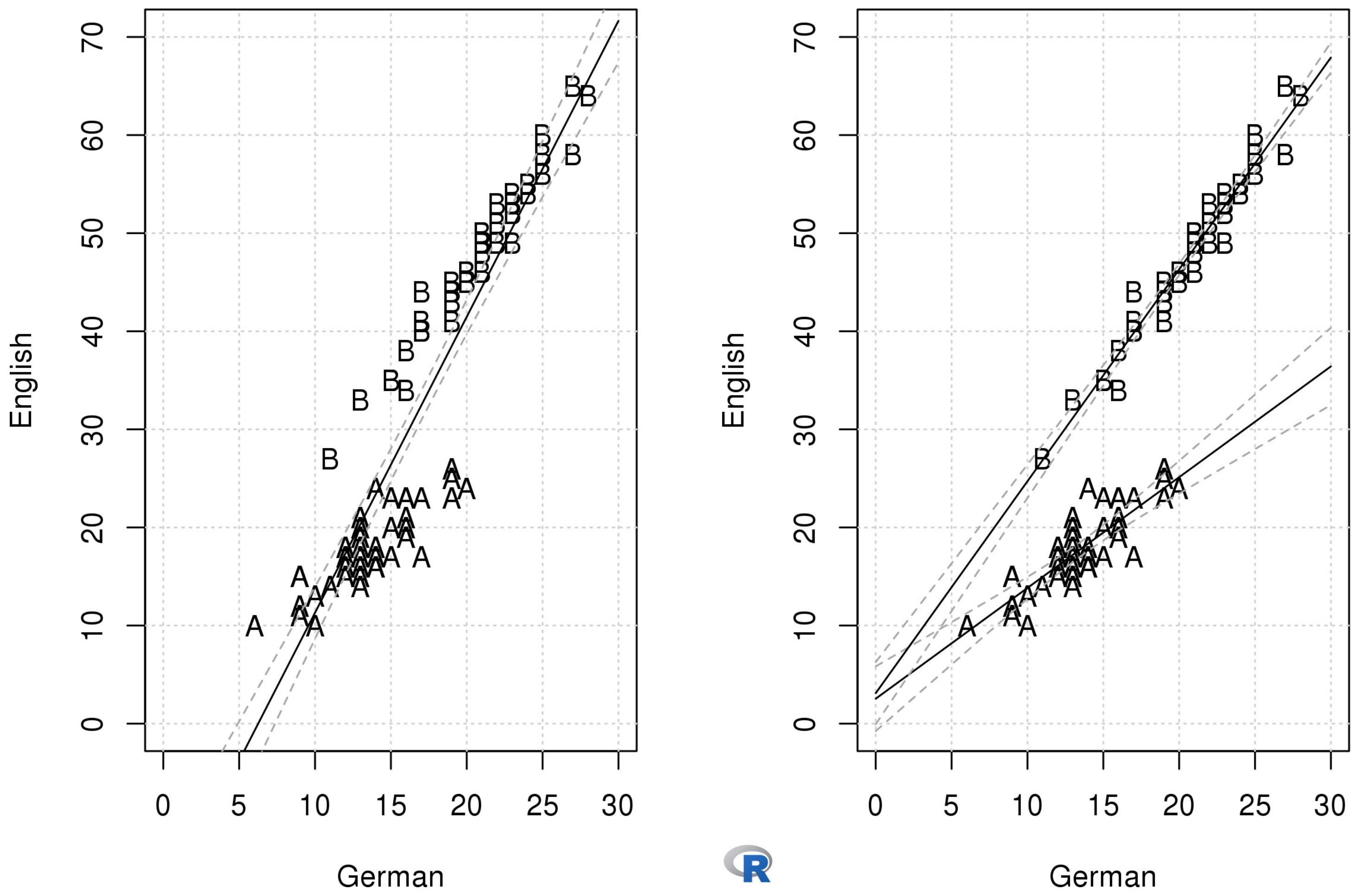
model 1	Sum Sq	Estimate	Std. error	*t*	*p*
GER	2931.69	1.1292	0.1054	10.713	<0.0001
CLASS *A→B*	3010.3	0.565	2.3098	0.245	0.8074
GER:CLASS *A→B*	241.73	1.0308	0.1354	7.613	<0.0001
Residual var.	316.95				

- model 2: ENGL ~ GER + CLASS

model 2	Sum Sq	Estimate	Std. error	*t*	*p*
GER	2931.69	1.75395	0.08726	20.101	<0.0001
CLASS *A→B*	3010.3	17.44117	0.85627	20.369	<0.0001
Residual var.	558.68				

- 1: better variance explanation ($p<10^{-10}$)
- 2: different (more accurate) coefficients
  - model 1's estimate for a student in class A who made 17 mistakes in German is off by 8.7% - model 2: 19.2% off!
- 3: different (more accurate) *p*-values

What interactions can reveal: mean vs. slope



What interactions can reveal: mean vs. slope

- A way out? If CLASS plays a role, we test the effect of GER separately in each class …
- model 3: ENGL[CLASS=="A"] ~ GER[CLASS=="A"]
  - estimate for GER: 1.13, $p<0.0001$
- model 4: ENGL[CLASS=="B"] ~ GER[CLASS=="B"]
  - estimate for GER: 2.16, $p<0.0001$
- the coefficients are different, which suggests an interaction, but …
- 4: *separate* tests of ENGL~GER per class never contrast the *separate* coefficients for GER in the two classes:
  - the interaction does not show up in either model
  - thus, 1.13 is never explicitly compared to 2.16
  - thus, the interaction does not get a *p*-value
  - thus, one does not know whether the difference between the two slopes of 1.03 (1.13-2.16) is significant or not
  - model 1 is not the only, but the best, way to do this

What interactions can reveal: differences in slopes (part 2)

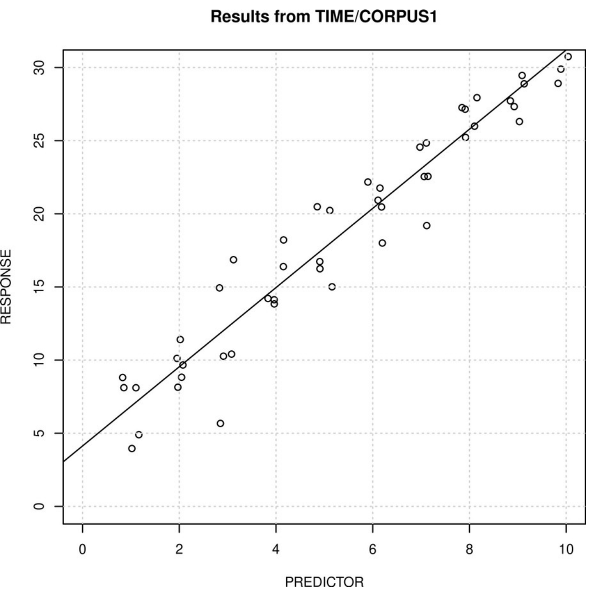
- Sometimes, interactions are the whole point, even if authors don't notice that …
- I once saw a conference presentation where someone wanted to discuss how a response was affected by a predictor differently over 3 time periods (each represented by a different corpus representative of one time period) …
- why is this useless?
  - you see the slopes are different across the corpus: 2.7>2.3>1.9
  - you see each slow is * different from 0 (*p*-values)
  - you do not see whether they are * different from each other!
- we need 1 big regression model where the slope of PREDICTOR can be different in each corpus
- the interaction PREDICTOR:CORPUS

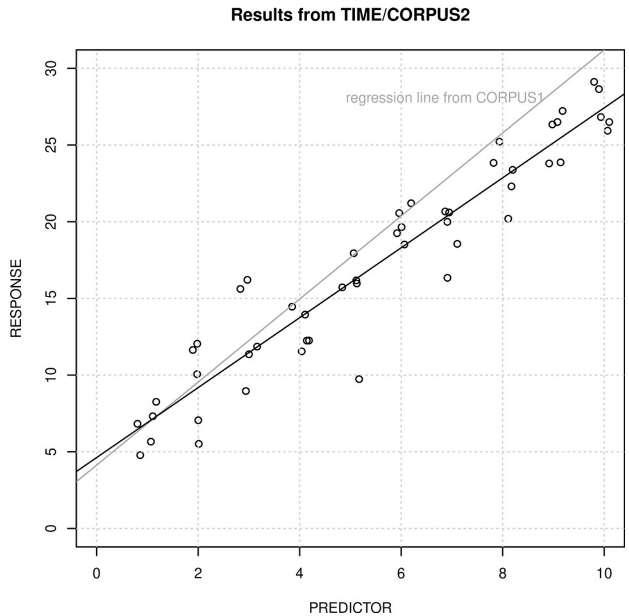
What interactions differences in slop

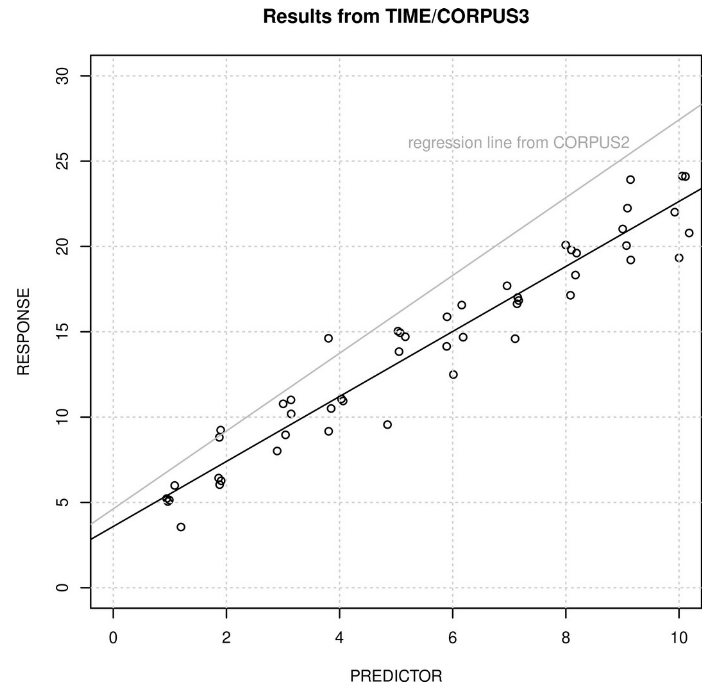
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.134      0.636   6.505        0
## PREDICTOR        2.706      0.102  26.423        0
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.625      0.642   7.202        0
## PREDICTOR        2.280      0.103  22.025        0
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3.592      0.478   7.508        0
## PREDICTOR        1.905      0.077  24.701        0
```



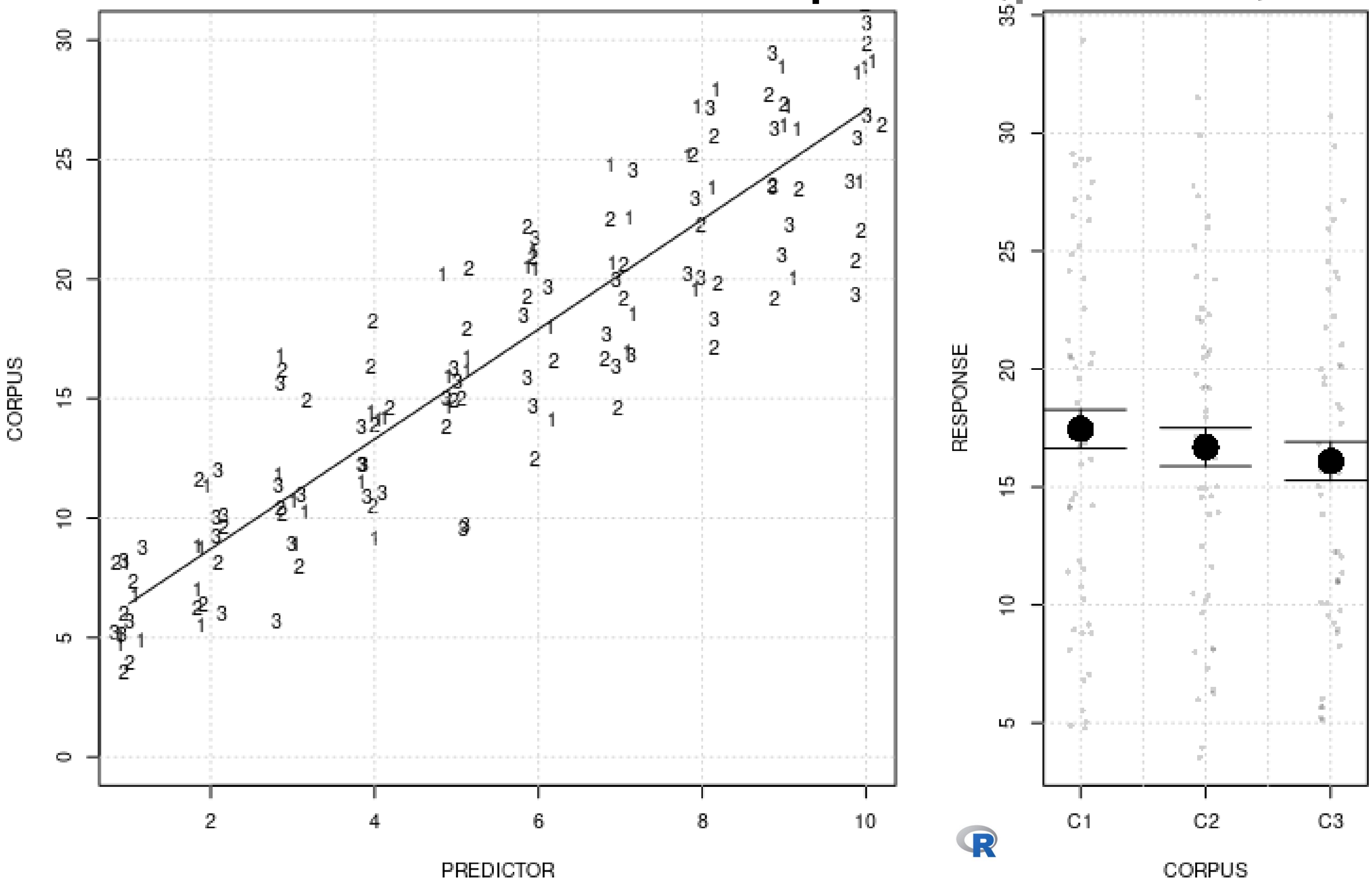




What interactions can reveal: differences in slopes (part 2)

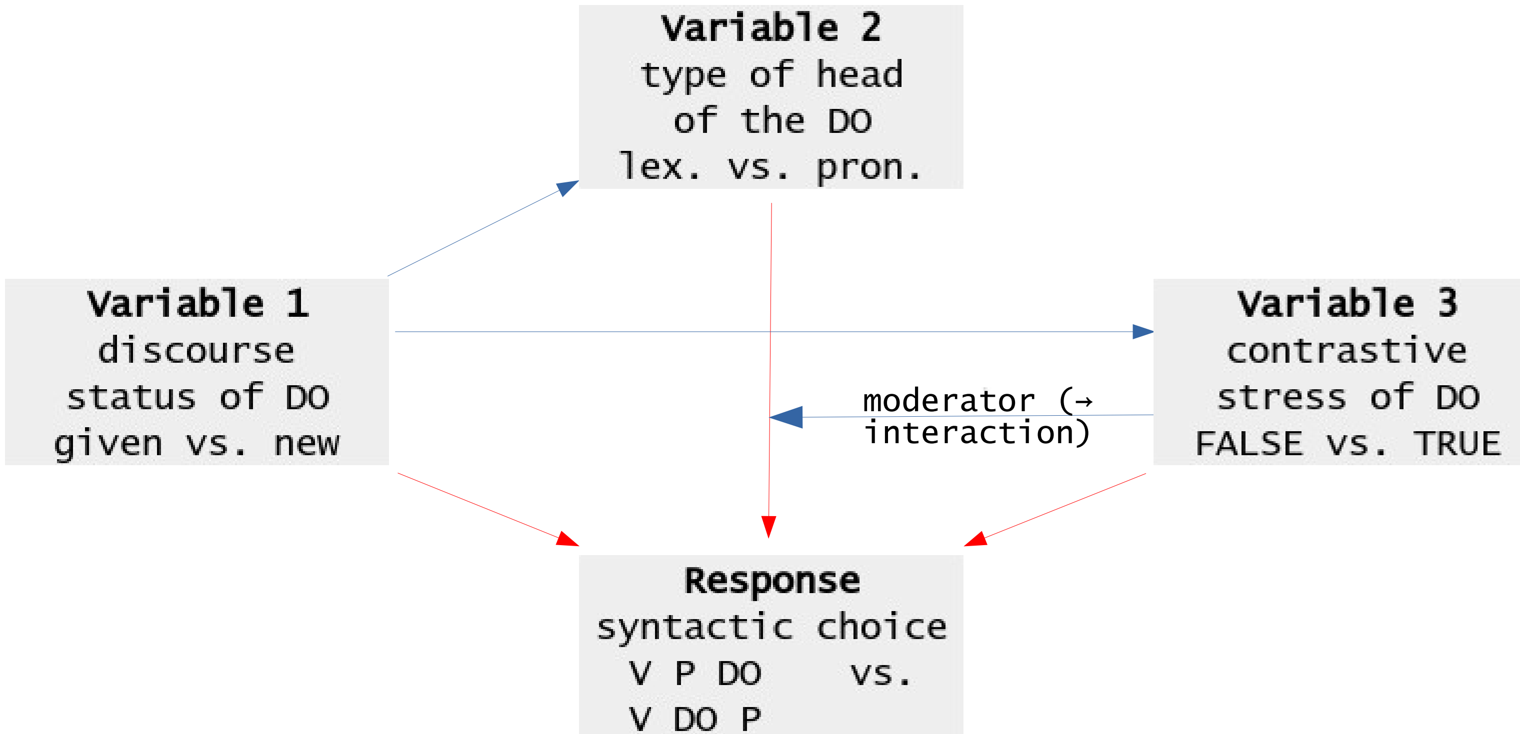
- Instead, we need 1 big regression model where the slope of PREDICTOR can be different in each corpus
- i.e. where the effect of PREDICTOR is not the same everywhere …,
- i.e. we need the interaction PREDICTOR:CORPUS
- if one does that here,
  - the interaction is not significant (*p*=0.09833 ns)
  - none of the differences between the slopes of the 3 corpora is significant
- thus, if the effects of PREDICTOR and CORPUS are significant, the results would be this
- in that case, there is a diachronic effect – RESPONSE is decreasing over time/CORPUS – but the author wanted the effect of PREDICTOR to change of time/CORPUS!

What interactions can reveal: differences in slopes (part 2)



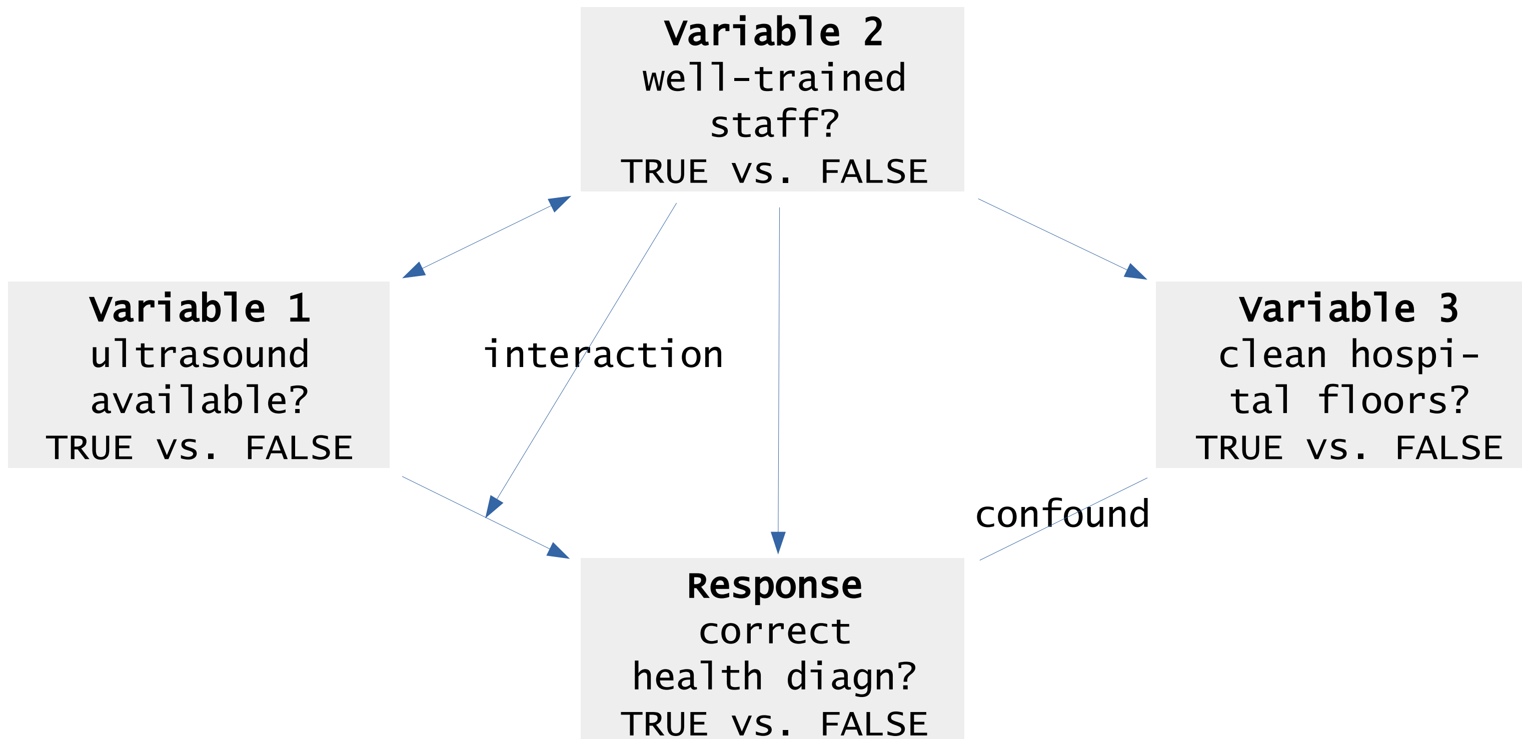
Another (linguistic) example/reminder

- What affects the probability of putting the particle of a trans. phras. verb before/after the DO?
  - *picked up* N
  - *picked* N *up*



Another (non-linguistic) example/reminder

- What affects the probability of correct diagnoses of fetal health during pregnancy?
  - predictors
  - interactions between them
  - confounds



So we just add predictors/interactions until we're blue in the face?

- No, because of … Occam's razor
  - prefer models with fewer parameters over models with more parameters
    - i.e., prefer models with fewer predictors over models with more predictors
    - i.e., prefer predictors with fewer levels over predictors with more levels
    - i.e., prefer linear models to non-linear models
    - i.e., prefer additive relationships to interactions
- what does "prefer" mean?
  - typically, it means 'if two models that try to account for data don't differ (enough), use the simpler one'
    - enough = according to p, or
    - enough = according to *AIC*, …

How are the effects of (multiple) predictors explored?

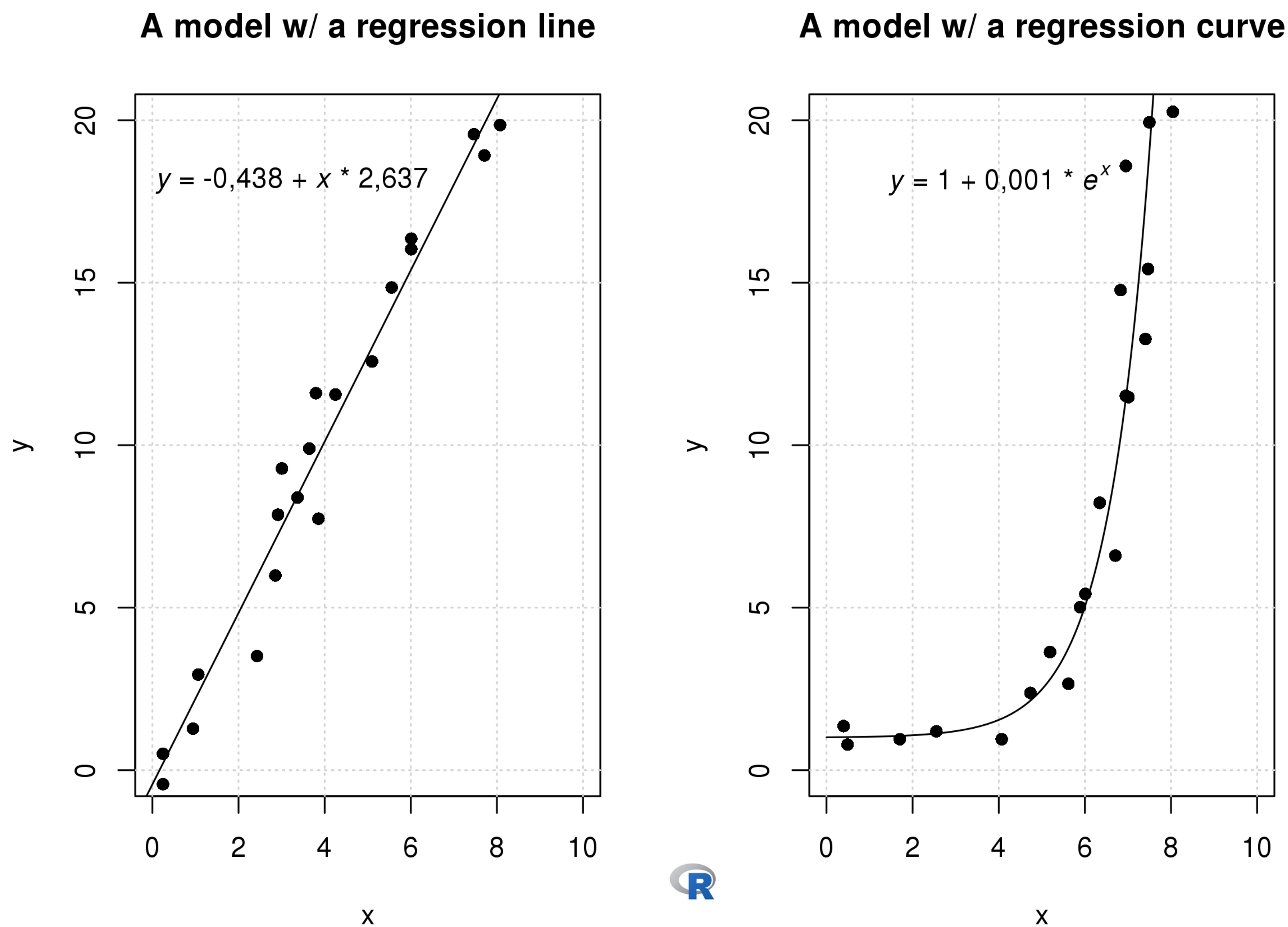
- Models and their selection
  - model = formal characterization of the relation between
    - predictors
      - independent variables
      - their interactions
      - (sometimes even levels of predictors)
    - dependent variables, or responses
    - usually in the form of a regression equation
    - note: many tests you already know are actually the simplest cases of regression modeling: r, t-test, X^2, …
  - model selection = the process of developing the most appropriate model for a given data set
    - direction of model selection
      - backwards selection
      - forward selection
      - bidirectional
    - criterion of model selection
      - p-values (of different kinds)
      - AIC (or AIC_c or BIC or …)
  - model amalgamation

How are the effects of (multiple) predictors explored?

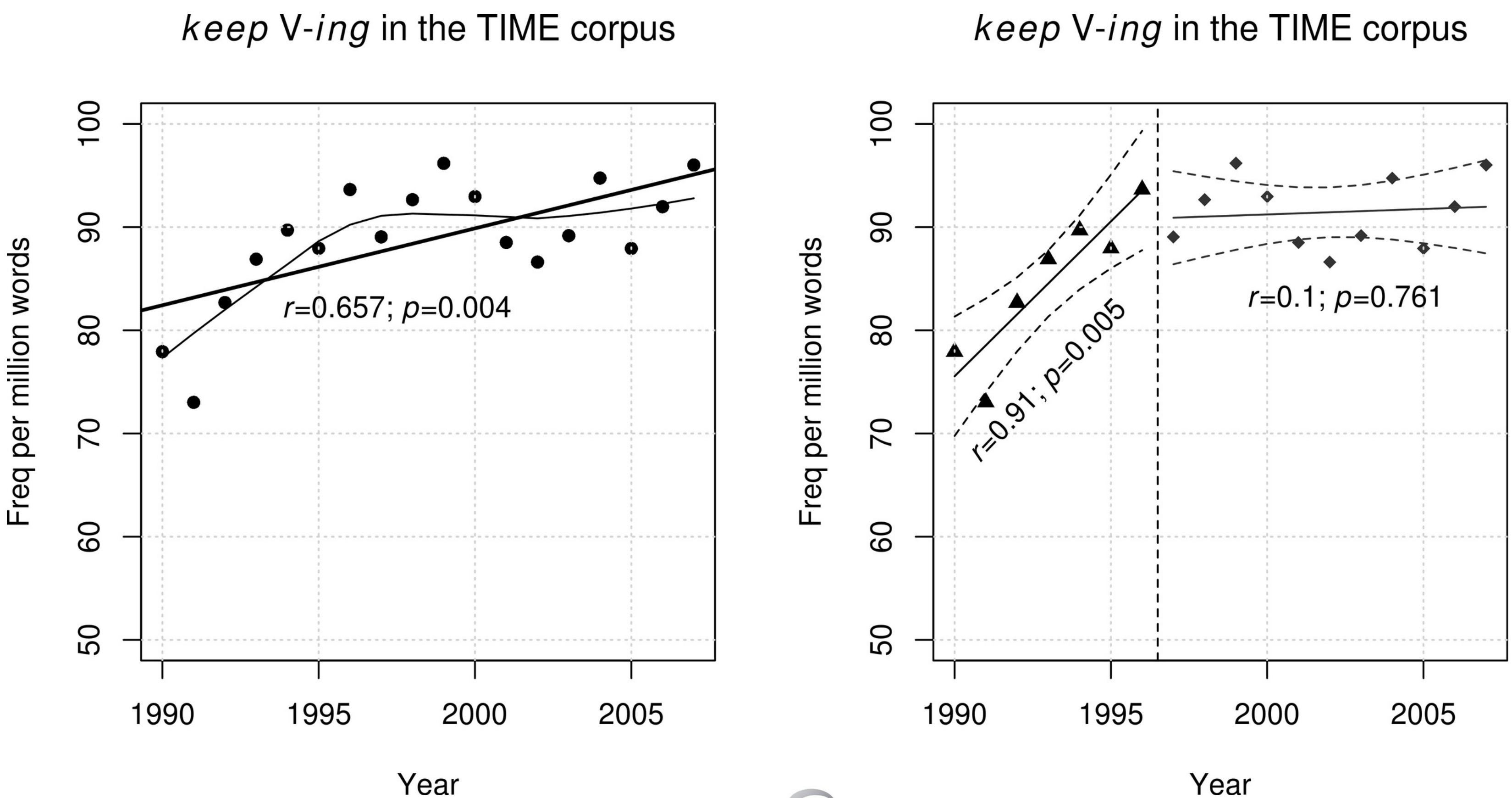
- Formulating the first model
  - what is the nature of the response?
    - numeric? → linear regression (often)
    - binary? → binary logistic regression
    - ordinal? → ordinal logistic regression
    - categorical? → multinomial regression
    - frequencies → Poisson regression

    and of course others …
  - which scales for the predictors are most useful?
    - raw values? logged? roots? centered? standardized? other?
  - what type of regression line is predicted?
    - straight line? curve? polynomial? w/ breakpoints? other?
  - which predictors and interactions to include/explore?

How are the effects of (multiple) predictors explored?



How are the effects of (multiple) predictors explored?



When the model selection process has been completed …

- Is there a significant correlation between the predictor(s) and the response?
  - typical answers: *yes* or *no*
- what is the nature of the significant correlation?
  - how high/strong is the overall correlation? how well does the model explain the data? (NB: *explain* = 'predict' or 'account for variability')
    - typical answer: some kind of R^2-value(s) or an accuracy score or a similar value
  - what are the effects of the individual predictors?
    - typical answer: coefficients from the regression equation
      - intercepts
      - (differences between) means
      - (differences between) slopes
  - often easier: what values does the model predict?
    - typical 'answer': plots of predicted values (usually better than plots of observed values, but sometimes you want both)

Additional considerations

- Validation
  - validity: does variable x measure what it's supposed to measure?
  - validation: does a model based on data set x also work well (enough) on data set y? the issue of overfitting …
  - frequent approaches
    - cross-validation (often *k*-fold with *k*=10, i.e. with 10% samples)
    - leave-one-out method
    - sampling/permutation methods
- model assumptions/diagnostics
  - randomness and normality of residuals
  - no collinearity
  - special data points are considered
    - outliers and/or points with high influence (dffits/dfbetas)
  - missing data are considered
    - exploration or imputation of missing data
- non-independence of data points → multilevel models