Predicting is not explaining

targeted learning of the dative alternation in English

Guillaume Desagulier[1] Antoine Chambaz[2]

[1]MoDyCo (UMR 7114)
Paris 8, CNRS, Paris Ouest Nanterre La Défense
gdesagulier@univ-paris8.fr

[2]Modal'X (EA 3454)
Paris Ouest Nanterre La Défense
achambaz@u-paris10.fr

LiC, Paris, December 4, 2015

outline

a mathematician and a linguist walk into a lab

a sample from a real conversation:

a mathematician and a linguist walk into a lab

a sample from a real conversation:

- math.: as a linguist working in the Construction Grammar framework, what do you do?

a mathematician and a linguist walk into a lab

a sample from a real conversation:

- math.: as a linguist working in the Construction Grammar framework, what do you do?
- linguist: take the dative alternation for instance...

a mathematician and a linguist walk into a lab

a sample from a real conversation:

- math.: as a linguist working in the Construction Grammar framework, what do you do?
- linguist: take the dative alternation for instance...

(1) John gave the book to Mary. (PD)
S_{AGENT} V O_{THEME} Prep $O_{RECIPIENT}$

(2) John gave Mary the book. (DO)
S_{AGENT} V $O_{RECIPIENT}$ O_{THEME}

a mathematician and a linguist walk into a lab

a sample from a real conversation:

- math.: as a linguist working in the Construction Grammar framework, what do you do?
- linguist: take the dative alternation for instance...

(1) John gave the book to Mary. (PD)
S_{AGENT} V O_{THEME} Prep $O_{RECIPIENT}$

(2) John gave Mary the book. (DO)
S_{AGENT} V $O_{RECIPIENT}$ O_{THEME}

. . . typically, alternations are handled with predictive methods [Bre+07; Baa11]

outline of a typical parametric-model-based, predictive approach

- select a response variable (*e.g.*, PD vs. DO)

outline of a typical parametric-model-based, predictive approach

- select a response variable (*e.g.*, PD vs. DO)
- extract observations of the variable from a corpus (*e.g.*, Switchboard, 3263 observations)

outline of a typical parametric-model-based, predictive approach

- select a response variable (*e.g.*, PD vs. DO)
- extract observations of the variable from a corpus (*e.g.*, Switchboard, 3263 observations)
- annotate for variables (*e.g.*, 15: speaker, modality, verb meaning, semantic class of verb, animacy/pronominality/length/definiteness/accessibility of theme/recipient, and PD vs. DO)

outline of a typical parametric-model-based, predictive approach

- select a response variable (*e.g.*, PD vs. DO)
- extract observations of the variable from a corpus (*e.g.*, Switchboard, 3263 observations)
- annotate for variables (*e.g.*, 15: speaker, modality, verb meaning, semantic class of verb, animacy/pronominality/length/definiteness/accessibility of theme/recipient, and PD vs. DO)
- fit parametric statistical models and interpret each coefficient as an effect of the corresponding variable

outline of a typical parametric-model-based, predictive approach

- select a response variable (*e.g.*, PD vs. DO)
- extract observations of the variable from a corpus (*e.g.*, Switchboard, 3263 observations)
- annotate for variables (*e.g.*, 15: speaker, modality, verb meaning, semantic class of verb, animacy/pronominality/length/definiteness/accessibility of theme/recipient, and PD vs. DO)
- fit parametric statistical models and interpret each coefficient as an effect of the corresponding variable
- compare the models

example: logistic regression/prediction [Bre+07]

(26) Model B: Relative magnitudes of significant effects

	Coefficient	Odds Ratio PP	95% C.I.
nonpronominality of recipient	1.73	5.67	3.25–9.89
inanimacy of recipient	1.53	5.62	2.08–10.29
nongivenness of recipient	1.45	4.28	2.42–7.59
indefiniteness of recipient	0.72	2.05	1.20–3.5
plural number of theme	0.72	2.06	1.37–3.11
structural parallelism in dialogue	–1.13	0.32	0.23–0.46
nongivenness of theme	–1.17	0.31	0.18–0.54
length difference (log scale)	–1.16	0.31	0.25–0.4
indefiniteness of theme	–1.74	0.18	0.11–0.28
nonpronominality of theme	–2.17	0.11	0.07–0.19

problems

- math.: interesting... but then does your job as a linguist consist in making predictions?

problems

- math.: interesting... but then does your job as a linguist consist in making predictions?
- linguist: uh... not really

problems

- math.: interesting... but then does your job as a linguist consist in making predictions?
- linguist: uh... not really
- math: and if I were to give you the true, unknown law of the data, could you tell me what feature of the law you are targeting with your parametric statistical models?

problems

- math.: interesting... but then does your job as a linguist consist in making predictions?
- linguist: uh... not really
- math: and if I were to give you the true, unknown law of the data, could you tell me what feature of the law you are targeting with your parametric statistical models?
- linguist: ...

problems

- math.: interesting... but then does your job as a linguist consist in making predictions?
- linguist: uh... not really
- math: and if I were to give you the true, unknown law of the data, could you tell me what feature of the law you are targeting with your parametric statistical models?
- linguist: . . .
- linguist: (sobs)

predicting *vs.* explaining the dative alternation

- predicting:

predicting *vs.* explaining the dative alternation

- predicting:
  - building an algorithm that poses as a native speaker of English when she formulates a construction involving a dative alternation

predicting *vs.* explaining the dative alternation

- predicting:
  - building an algorithm that poses as a native speaker of English when she formulates a construction involving a dative alternation
  - the algorithm does not need to tell us how the dative alternation works

predicting *vs.* explaining the dative alternation

- predicting:
  - building an algorithm that poses as a native speaker of English when she formulates a construction involving a dative alternation
  - the algorithm does not need to tell us how the dative alternation works
- explaining:

predicting *vs.* explaining the dative alternation

- predicting:
  - building an algorithm that poses as a native speaker of English when she formulates a construction involving a dative alternation
  - the algorithm does not need to tell us how the dative alternation works
- explaining:
  - uncovering what drives the choice of one dative form over the other

predicting *vs.* explaining the dative alternation

- predicting:
  - building an algorithm that poses as a native speaker of English when she formulates a construction involving a dative alternation
  - the algorithm does not need to tell us how the dative alternation works
- explaining:
  - uncovering what drives the choice of one dative form over the other
  - by building upon/targeting the above algorithm

method

targeted learning [LR11, monograph]

- Chambaz and Desagulier [CD15]

method

targeted learning [LR11, monograph]

- Chambaz and Desagulier [CD15]
- through causal analysis, we operationalize the set of scientific questions that we wish to address regarding the dative alternation

method

targeted learning [LR11, monograph]

- Chambaz and Desagulier [CD15]
- through causal analysis, we operationalize the set of scientific questions that we wish to address regarding the dative alternation
- we answer these questions by targeting some versatile machine learners borrowing from the latest advances in semi-parametric statistics

method

targeted learning [LR11, monograph]

- Chambaz and Desagulier [CD15]
- through causal analysis, we operationalize the set of scientific questions that we wish to address regarding the dative alternation
- we answer these questions by targeting some versatile machine learners borrowing from the latest advances in semi-parametric statistics
- we derive estimates, confidence regions and p-values for well-defined parameters that can be interpreted as the influence of each contextual variable on the outcome PD *vs.* DO

data

the `dative` dataset [Bre+07]

available from the `languageR` package [Baa09]

categorical contextual information variables

variable	*vs.*	estimate	CI	*p*-value
Modality	written%spoken	0.0277	[-0.0031,0.0585]	0.0776
AnimacyOfRec	inanimate%animate	0.0938	[0.0549,0.1327]	0.0000
DefinOfRec	indefinite%definite	0.0395	[0.0102,0.0688]	0.0083
PronomOfRec	pronominal%nonpronominal	-0.1398	[-0.2171,-0.0624]	0.0004
AnimacyOfTheme	inanimate%animate	0.0843	[0.0337,0.1348]	0.0011
DefinOfTheme	indefinite%definite	-0.0568	[-0.0865,-0.0272]	0.0002
PronomOfTheme	pronominal%nonpronominal	-0.1168	[-0.1377,-0.0959]	0.0000
AccessOfRec	new%accessible	-0.3824	[-0.5458,-0.2189]	0.0000
	given%accessible	0.0411	[-0.0149,0.0971]	0.1506
AccessOfTheme	new%accessible	-0.0782	[-0.1100,-0.0463]	0.0000
	given%accessible	-0.0415	[-0.0673,-0.0157]	0.0016
SemanticClass	t%a	0.1152	[0.0548,0.1755]	0.0002
	p%a	-0.0928	[-0.1532,-0.0324]	0.0026
	f%a	-0.1471	[-0.1946,-0.0997]	0.0000
	c%a	0.1657	[0.1238,0.2077]	0.0000

categorical contextual information variables – PD, decrease

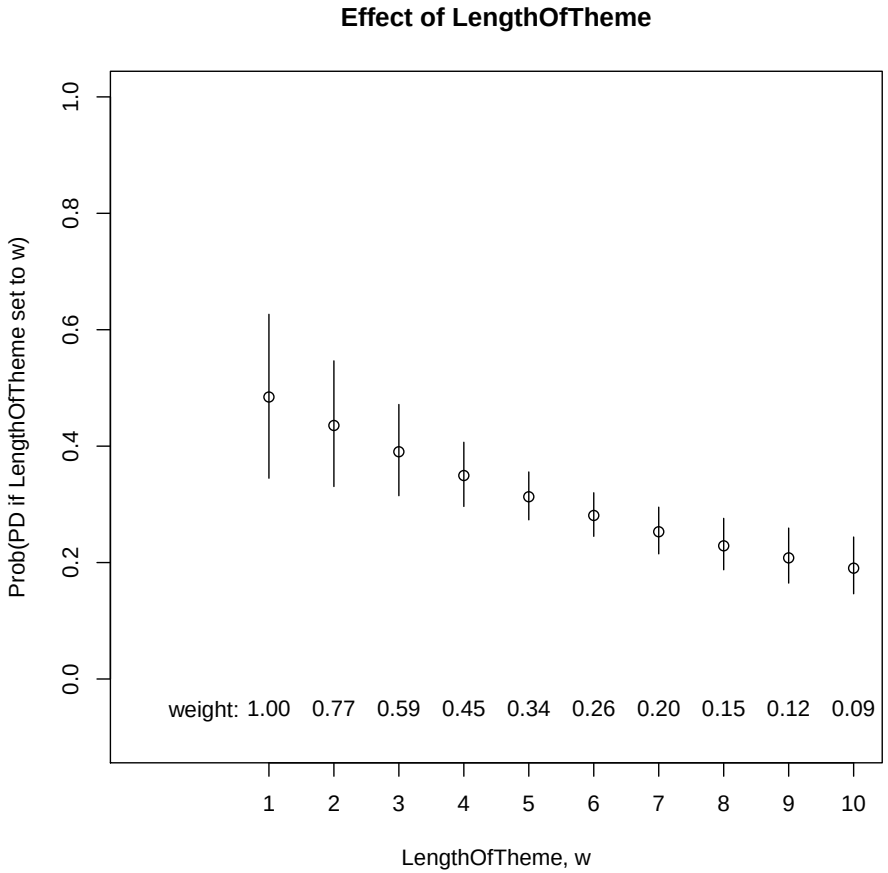
variable	*vs.*	estimate	CI	*p*-value
Modality	written%spoken	0.0277	[-0.0031,0.0585]	0.0776
AnimacyOfRec	inanimate%animate	0.0938	[0.0549,0.1327]	0.0000
DefinOfRec	indefinite%definite	0.0395	[0.0102,0.0688]	0.0083
PronomOfRec	pronominal%nonpronominal	-0.1398	[-0.2171,-0.0624]	0.0004
AnimacyOfTheme	inanimate%animate	0.0843	[0.0337,0.1348]	0.0011
DefinOfTheme	indefinite%definite	-0.0568	[-0.0865,-0.0272]	0.0002
PronomOfTheme	pronominal%nonpronominal	-0.1168	[-0.1377,-0.0959]	0.0000
AccessOfRec	new%accessible	-0.3824	[-0.5458,-0.2189]	0.0000
	given%accessible	0.0411	[-0.0149,0.0971]	0.1506
AccessOfTheme	new%accessible	-0.0782	[-0.1100,-0.0463]	0.0000
	given%accessible	-0.0415	[-0.0673,-0.0157]	0.0016
SemanticClass	t%a	0.1152	[0.0548,0.1755]	0.0002
	p%a	-0.0928	[-0.1532,-0.0324]	0.0026
	f%a	-0.1471	[-0.1946,-0.0997]	0.0000
	c%a	0.1657	[0.1238,0.2077]	0.0000

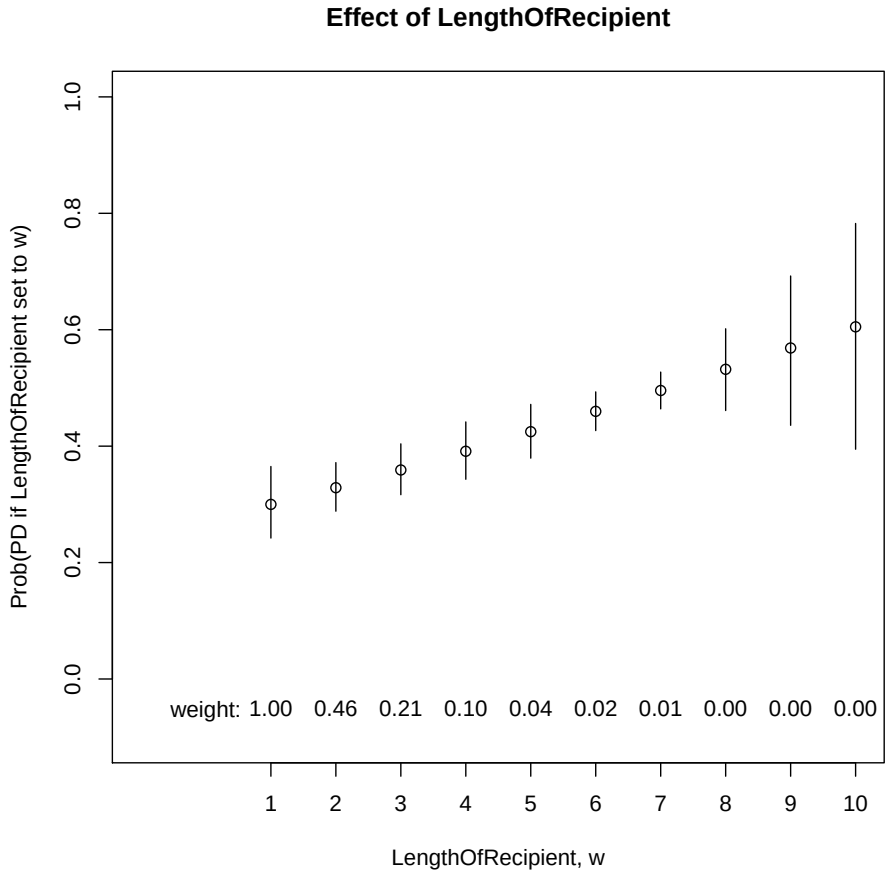
categorical contextual information variables – PD, increase

variable	*vs.*	estimate	CI	*p*-value
Modality	written%spoken	0.0277	[-0.0031,0.0585]	0.0776
AnimacyOfRec	inanimate%animate	0.0938	[0.0549,0.1327]	0.0000
DefinOfRec	indefinite%definite	0.0395	[0.0102,0.0688]	0.0083
PronomOfRec	pronominal%nonpronominal	-0.1398	[-0.2171,-0.0624]	0.0004
AnimacyOfTheme	inanimate%animate	0.0843	[0.0337,0.1348]	0.0011
DefinOfTheme	indefinite%definite	-0.0568	[-0.0865,-0.0272]	0.0002
PronomOfTheme	pronominal%nonpronominal	-0.1168	[-0.1377,-0.0959]	0.0000
AccessOfRec	new%accessible	-0.3824	[-0.5458,-0.2189]	0.0000
	given%accessible	0.0411	[-0.0149,0.0971]	0.1506
AccessOfTheme	new%accessible	-0.0782	[-0.1100,-0.0463]	0.0000
	given%accessible	-0.0415	[-0.0673,-0.0157]	0.0016
SemanticClass	t%a	0.1152	[0.0548,0.1755]	0.0002
	p%a	-0.0928	[-0.1532,-0.0324]	0.0026
	f%a	-0.1471	[-0.1946,-0.0997]	0.0000
	c%a	0.1657	[0.1238,0.2077]	0.0000

integer valued contextual information variables

(using a working model)





surprising results

- *e.g.*, linguists know that PD is preferred when the theme is pronominal:

  Anthony sent it to you. (PD)

  ??Anthony sent you it. (DO)

surprising results

- *e.g.*, linguists know that PD is preferred when the theme is pronominal:

  Anthony sent it to you. (PD)

  ??Anthony sent you it. (DO)

- averaging out the context yields a 11% decrease:

  "all other things being equal, switching theme from nonpronominal to pronominal yields an 11% decrease of the probability of PD"

surprising results

- *e.g.*, linguists know that PD is preferred when the theme is pronominal:

  Anthony sent it to you. (PD)

  ??Anthony sent you it. (DO)

- averaging out the context yields a 11% decrease:

  "all other things being equal, switching theme from nonpronominal to pronominal yields an 11% decrease of the probability of PD"

- this is an example of Simpson's paradox

surprising results

- *e.g.*, linguists know that PD is preferred when the theme is pronominal:

  Anthony sent it to you. (PD)

  ??Anthony sent you it. (DO)

- averaging out the context yields a 11% decrease:

  "all other things being equal, switching theme from nonpronominal to pronominal yields an 11% decrease of the probability of PD"

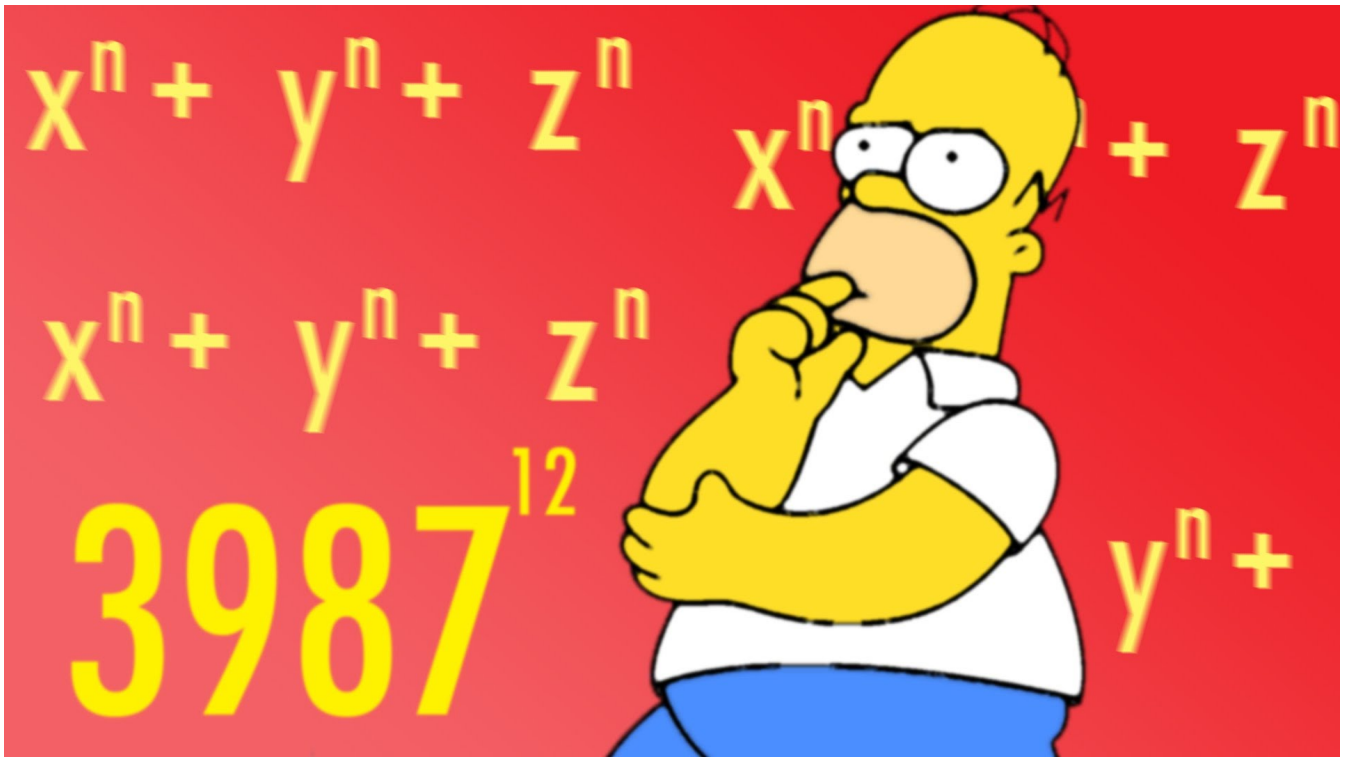
- this is an example of Simpson's paradox
- further illustrates that the parameter matching pronominality of theme in a logistic regression model is an awkward function of the law of data

Simpson's paradox

"a trend appearing in different data groups may reverse once these groups are combined"

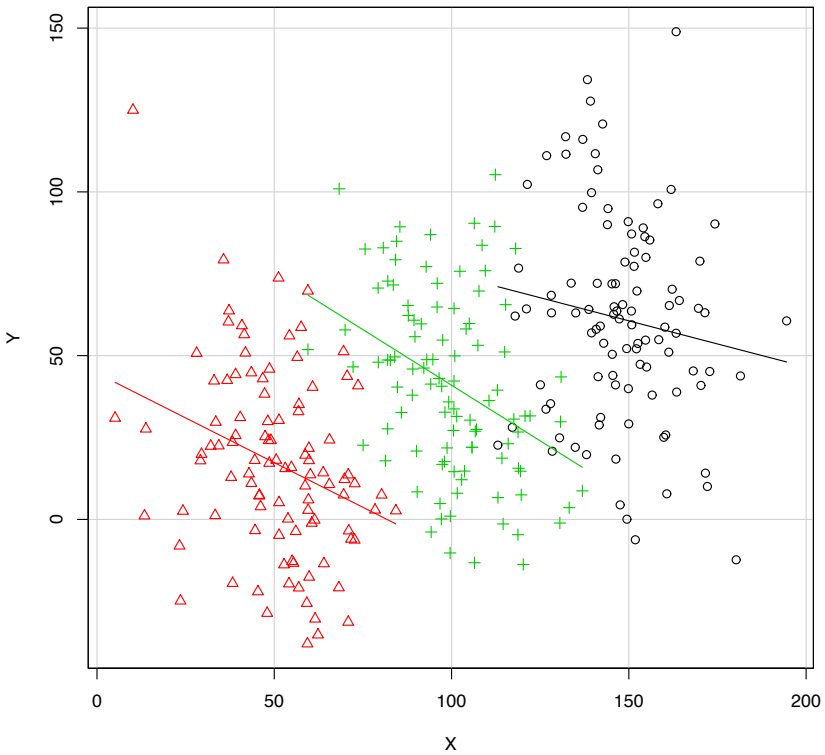
Simpson's paradox

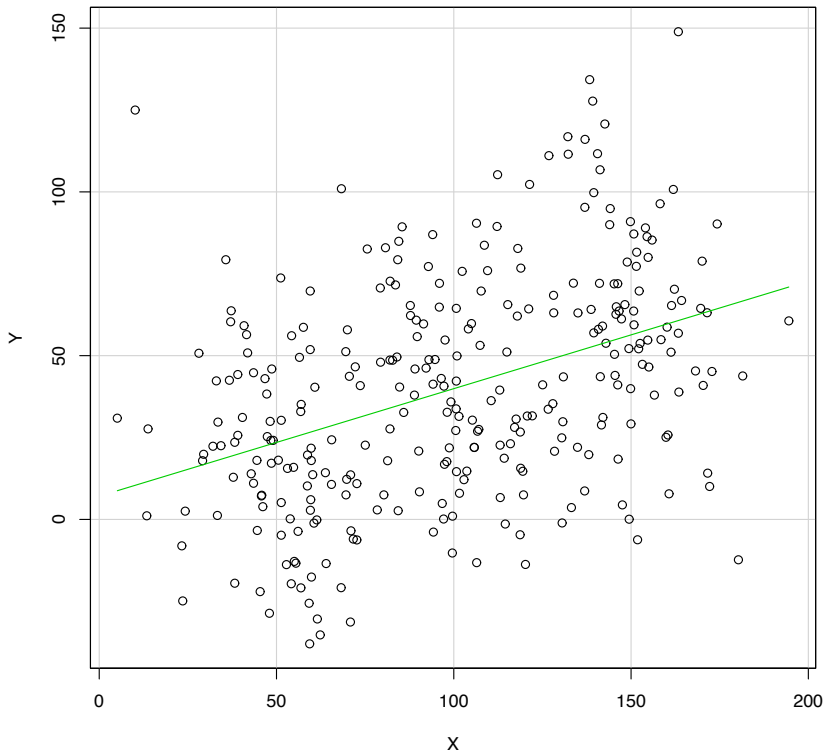
"a trend appearing in different data groups may reverse once these groups are combined"



Simpson's paradox

"a trend appearing in different data groups may reverse once these groups are combined"





Simpson's paradox

"a trend appearing in different data groups may reverse once these groups are combined"

numerical example: "all other things being equal, switching theme from definite to indefinite yields a 5% decrease of the probability of PD"

theme	definite	indefinite
PD	63	378
DO	28	858

uncontextualized data

- $\mathrm{ER}(P_n) = \frac{63}{63+28} - \frac{378}{378+858} \approx 38\%$
- $\Psi(P_n^*) \approx -5\%$ (significant difference)

when explanation is sought, prediction is only a means to an end

- the take-home message on the dative alternation cannot be provided in the form of a fitted prediction model

when explanation is sought, prediction is only a means to an end

- the take-home message on the dative alternation cannot be provided in the form of a fitted prediction model
- *e.g.*,
  - we observed a significant decrease of the probability of obtaining PD when, all other things being equal, the theme is switched from nonpronominal to pronominal
  - a crude measure of statistical association such as the excess risk would have indicated a significant increase
  - this is an illustration of Simpson's paradox

what we did

what we provide instead is two-fold:

- we framed our account of the dative alternation in a causal model
  $\neq$ prediction model
- we investigated the effect of each available, contextual information variable on the choice of PD over DO, resulting in a table of estimates, confidence intervals, and *p*-values

our approach is based on causal inference, machine learning, and semi-parametric statistics

- we operationalized the effect of any given element of context on the dative alternation as a functional evaluated at the true, unknown law of the data
- we also showed how to estimate this effect in a targeted way, under the form of that functional evaluated at an empirical law built specifically to estimate the corresponding effect

future work

our method can be applied to case-studies involving contrasts or alternations, such as

- the choice of the predeterminer *vs.* preadjectival position of intensifiers (*e.g.*, *quite* and *rather*),
- the choice of one word over a near-synonym (*e.g.*, *almost/nearly*, *big/large*, *broad/wide*, *freedom/liberty*, . . . the sky is the limit)

thanks for your attention!

[Baa09] Rolf Harald Baayen. *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"*. 2009. URL: http://CRAN.R-project.org/package=languageR.

[Baa11] Rolf Harald Baayen. "Corpus linguistics and naive discriminative learning." In: *Revista Brasileira de Linguística Aplicada* 11.2 (2011), pp. 295–328.

[Bre+07] Joan Bresnan et al. "Predicting the dative alternation." In: *Cognitive foundations of interpretation* (2007), pp. 69–94.

[CD15] Antoine Chambaz and Guillaume Desagulier. "Predicting is not explaining: Targeted learning of the dative alternation." In: *Journal of Causal Inference* (2015).

[LR11] Mark J. van der Laan and Sherri Rose. *Targeted learning*. New York: Springer, 2011, pp. lxxi+626. ISBN: 978-1-4419-9781-4. DOI: 10.1007/978-1-4419-9782-1.