Inflection class and semantic analogies

Matías Guzmán Naranjo - Université de Paris

02.2020

Inflection classes

The literature on inflection classes suggests that inflection classes are morphomic:

- no interaction with syntax
- no interaction with semantics

We do not expect that the NOM.SG in inflection class A to be different from the NOM.SG in inflection class B.

Although intuitive, this hasn't been explored in a systematic manner.

Proportional analogy

The basic idea of proportional analogy can be expressed as:

- A:B=C:D

Although this is usually intended to express form relations, semantic relations can also be captured this way.

Guiding question:

- Are semantic analogies sensitive to inflection class?

Data-sets

For now I am focusing on Russian. I used the following data-sets:

- Semantics: Wiki Word Vectors as trained by Bojanowski et al. (2017).
- Word forms: Zalis Dictionary Zaliznyak (1977).

A data example

We can visually compare some vectors:

	gloss	X1	X2	X3	X4	...	X296	X297	X298	X299	X300
фирма	'firm'	0.293	0.618	-0.0996	0.106	...	0.0235	0.223	0.0188	0.210	0.164
мужчина	'man'	0.216	0.105	-0.221	0.161	...	-0.800	0.119	-0.0168	-0.0434	0.361
женщина	'woman'	0.268	0.0444	-0.389	0.150	...	-0.617	0.255	0.0221	0.139	0.520

The vectors for *женщина* and *мужчина* seem closer together than the vector for *фирма*.

Class extraction

1. find the non-continuous phonological sub-sequence common to all cells in the paradigm of a lexeme (from now on the *stem*),
2. remove this sub-sequence from each cell. In cases of discontinuous sub-sequences add a separation mark (-),
3. the result in each cell is the *marker* for that cell,
4. the inflection class of the lexeme is the set of markers for all cells.

Class extraction

Because this method makes no assumptions about underlying representations, it is very conservative and thus it produces the maximum possible number of inflection classes.

	singular		plural	
Cell	form	marker	form	marker
NOM	fʲirma	-a	fʲirmɨ	-ɨ
GEN	fʲirmɨ	-ɨ	fʲirm	-ø
DAT	fʲirm’e	-’e	fʲirmam	-am
ACC	fʲirmu	-u	fʲirmɨ	-ɨ
INS	fʲirmoj; firmou	-oj; -ou	fʲirmami	-am’i
PRE	fʲirm’e	-’e	fʲirmax	-ax

Table: Markers for *фирма* (‘fimr’).

Model

We predict the semantics of a (group of) cell(s) from another (group of) cell(s):

- group lexemes into inflection classes
- fit a model: `cell-1 ~ cell-2 + (1|sem-dimension)` for each inflection class
- get cross-validated R2 (withing class)
- fit the same model in class A and predict class B
- this this the across-class R2

We make sure that the cells were are considering in class A and B have the same marker syncretisms.

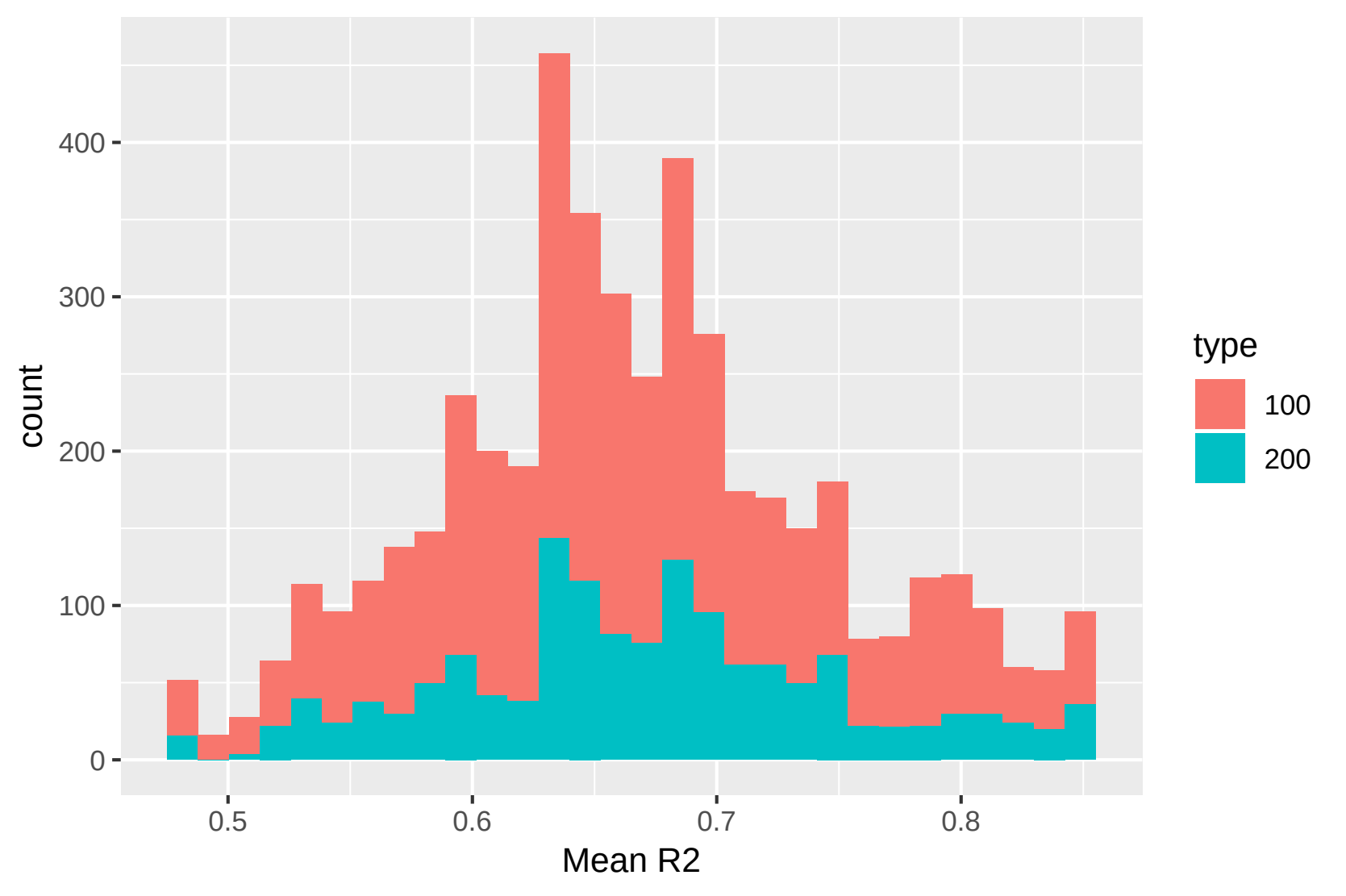
Example

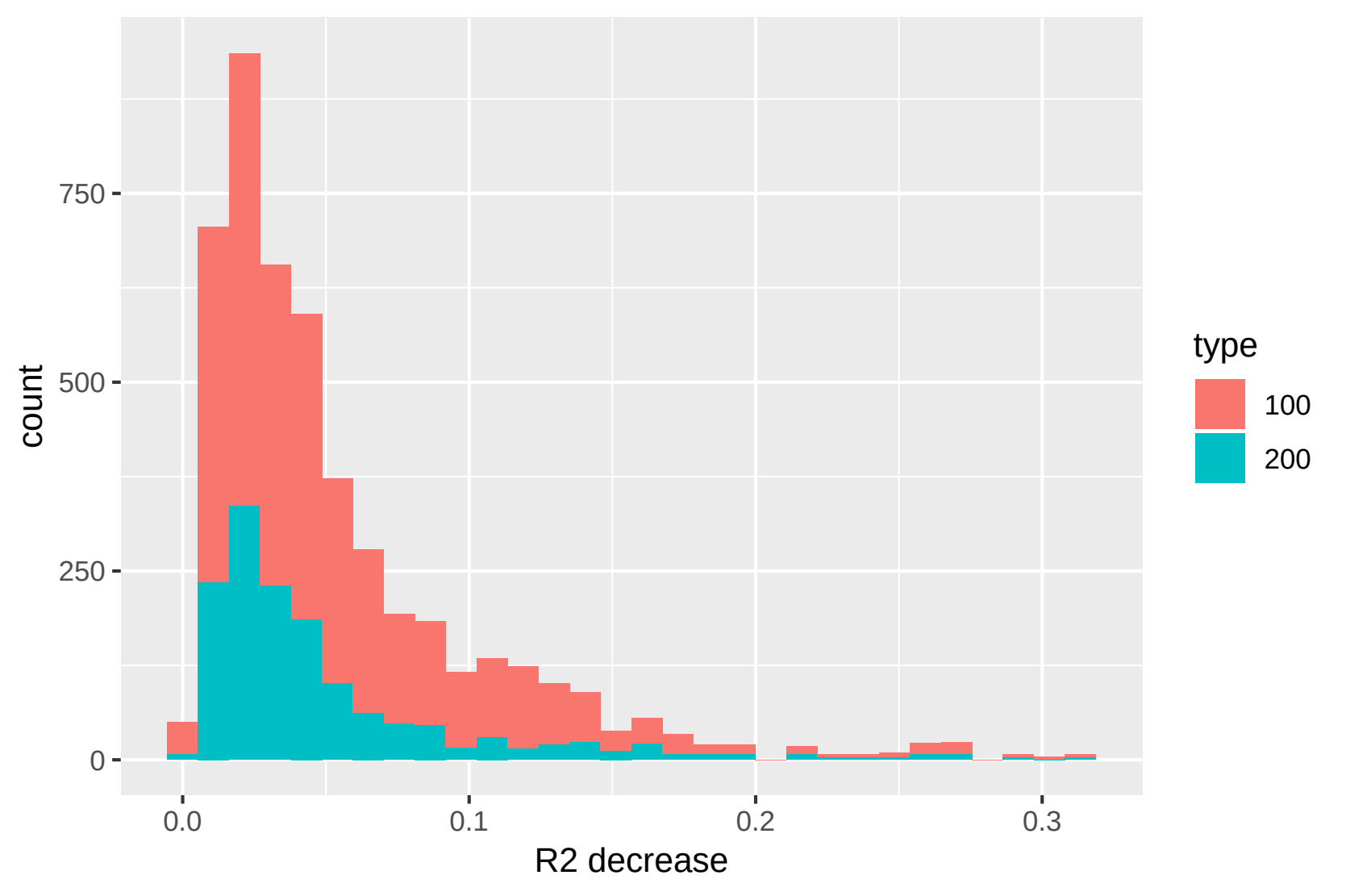
NOM.S + ACC.S ↔ GEN.S:

NOM.S	GEN.S	DAT.S	ACC.S	INS.S	PRE.S	NOM.P	GEN.P	DAT.P	ACC.P	INS.P	PRE.P
-ø	**-a**	-u	**-ø**	-om	-^{j}e	-~~i~~	-ov	-am	-~~i~~	-am^{j}i	-ax
-ø	**-a**	-u	**-ø**	-om	-^{j}e	-^{j}i	-ov	-am	-^{j}i	-am^{j}i	-ax
-ø	**-a**	-u	**-ø**	-om	-e	-~~i~~	-ov	-am	-~~i~~	-am^{j}i	-ax
o-	**-a**	-u	**o-**	-om	-^{j}e	-^{j}i	-ov	-am	-^{j}i	-am^{j}i	-ax

Example

cell-1	cell-2	sem-dimension	lexeme
0.1	0.3	X1	lexeme-1
0.2	0.9	X2	lexeme-1
0.1	0.012	X3	lexeme-1
...	...	...	lexeme-1
1.2	-0.24	X300	lexeme-1
-0.1	0.69	X1	lexeme-2
...	...	...	...


400
300
200
100
0
count
0.5
0.6
0.7
0.8
Mean R2
type
100
200

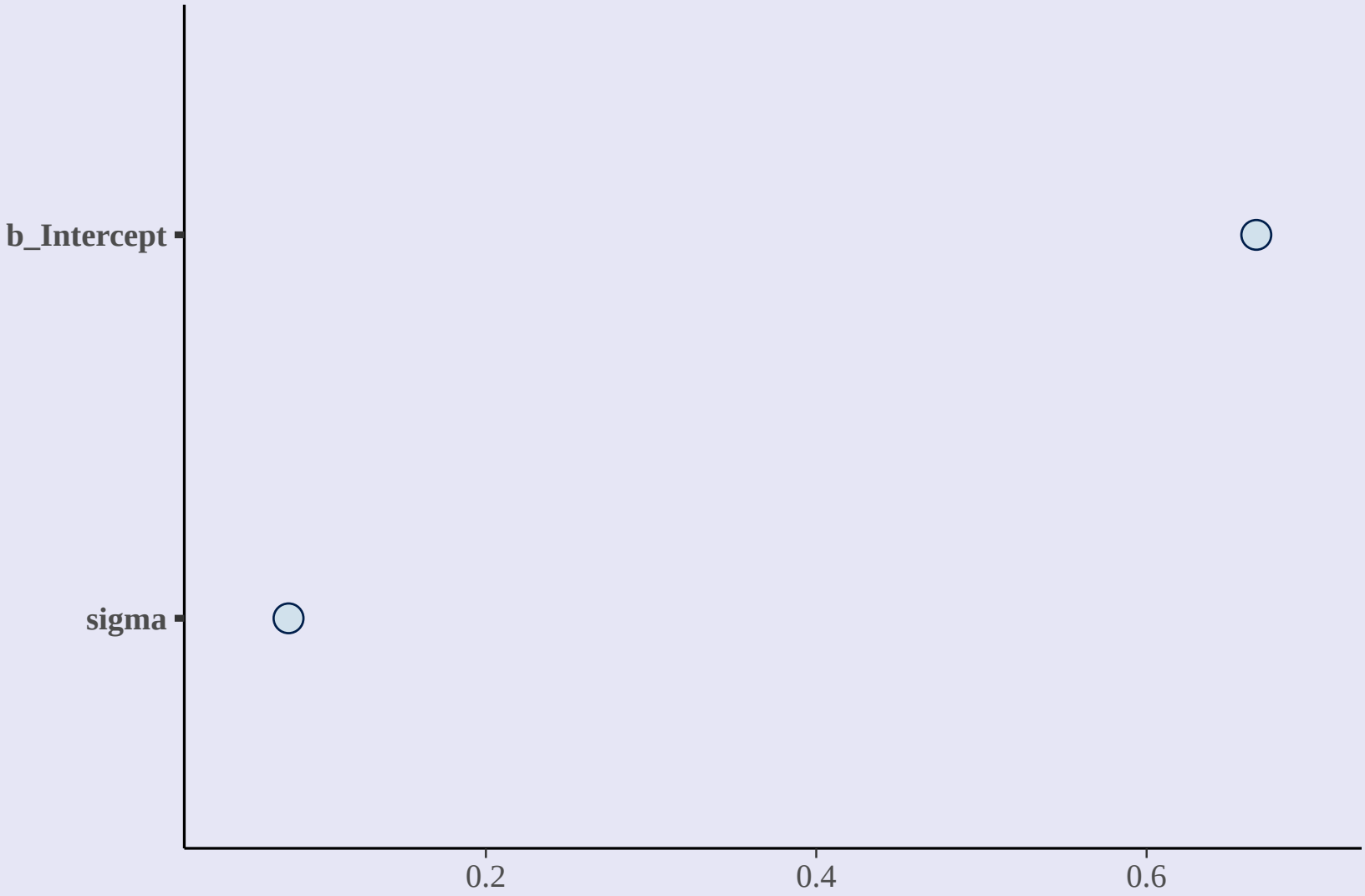

750
500
250
0
count
0.0
0.1
0.2
0.3
R2 decrease
type
100
200

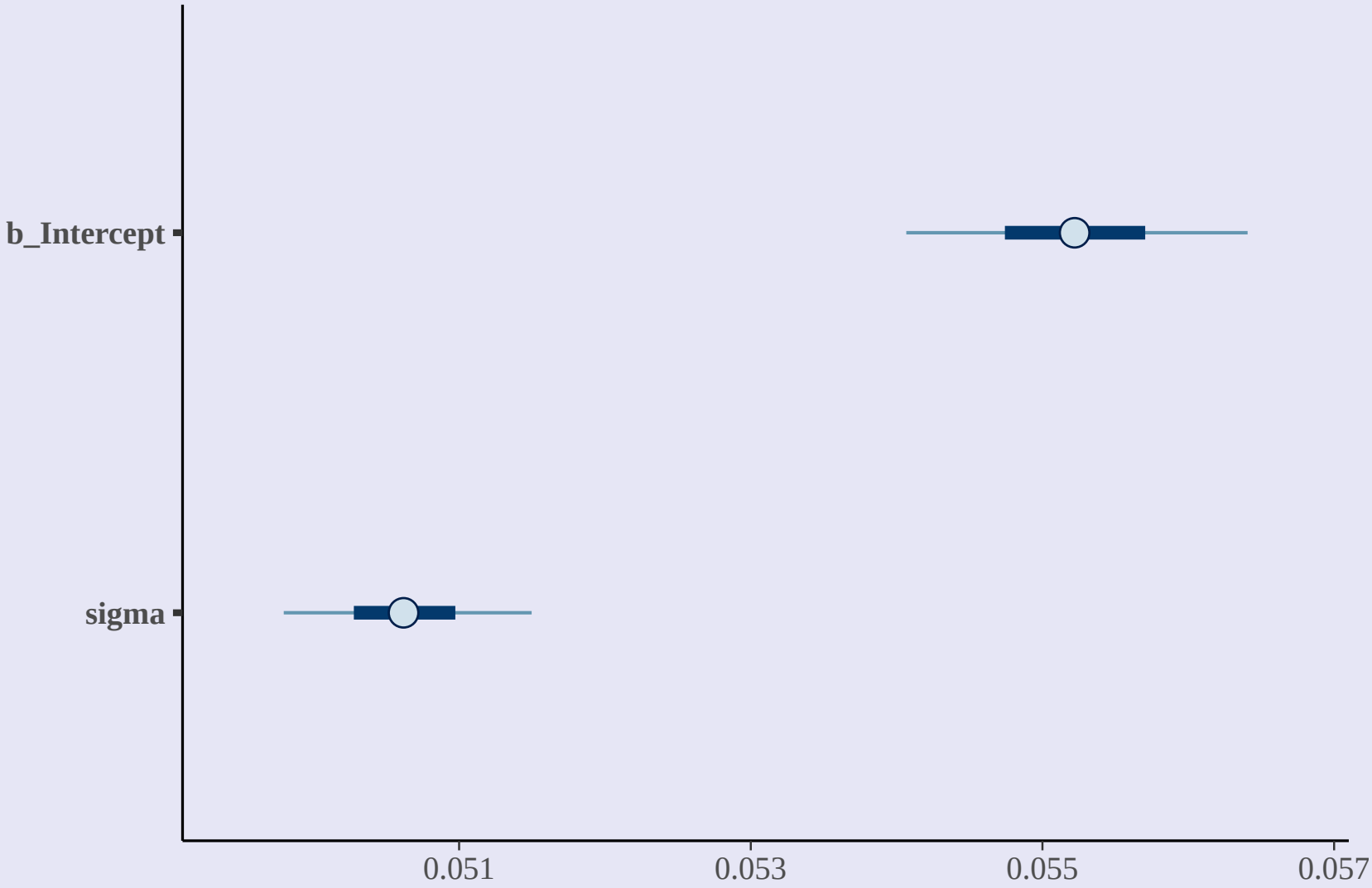
Modelling

We fit two simple model as follows:

$$\text{Diff-R2}_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha$$
$$\alpha \sim \text{Normal}(0, 1)$$
$$\sigma \sim \text{Uniform}(10, 10)$$

$$\text{Mean-R2}_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha$$
$$\alpha \sim \text{Normal}(0.5, 1)$$
$$\sigma \sim \text{Uniform}(10, 10)$$


b_Intercept
sigma
0.2
0.4
0.6


b_Intercept
sigma
0.051
0.053
0.055
0.057

Example

NOM.S ↔ INS.S (mean diff: 0.3):

NOM.S	GEN.S	DAT.S	ACC.S	INS.S	PRE.S	NOM.P	GEN.P	DAT.P	ACC.P	INS.P	PRE.P
-ø	-i	-i	-ø	**-u**	-i	-i	-ej	-am	-i	-am^{j}i	-ax
-ø	-a	-u	-ø	**-om**	-e	-ɨ	-ov	-am	-ɨ	-am^{j}i	-ax