Towards a de-ranged study of variation

Vsevolod Kapatsinski

University of Oregon

vkapatsi@uoregon.edu

The issue

- In variationist research, we often have a number of categorical predictors that influence a categorical linguistic variable.
  - Kapatsinski (2009): choice of adversative conjunction in Russian
    - types of constituents that are conjoined,
    - whether or not the event described by the first conjunct is interrupted by the event described by the second conjunct,
    - length of the second conjunct,
    - register.
  - Travis (2007): whether a pronominal subject is expressed overtly in Spanish,
    - semantic verb class,
    - distance from previous mention of the subject,
    - previous realization of the subject,
    - ambiguity of tense/aspect/modality marking.
  - Smith, Durham & Fortune (2007): monophthongization of /au/ in Scottish English
    - identity of the lexical item containing /au/
    - situational context
- Tool of choice: logistic regression

Effect size and range

- Which of the influential predictors have stronger effects?
- When we are dealing with categorical predictors with multiple values, effect size comparisons become tricky because every value of every predictor is associated with its own regression coefficient.
- To estimate the size of the effect of a predictor we need to somehow aggregate all regression coefficients associated with it.
- Standard solution: subtract the smallest coefficient associated with each predictor from the largest coefficient associated with the same predictor → the *range* of coefficients associated with it.
- Ranges of predictors are then compared to each other numerically, and predictors with larger ranges are said to be more important / have a stronger effect.
- Pervasive practice: 40/44 papers in LVC reporting logistic regressions report range comparisons

Example 1: Subject expression in Spanish

TABLE 2. *Two independent variable rule analyses of the contribution of factors selected as significant to the probability of expressed subjects in NM and Colombian data*

	NMCOSS			Colombia		
Total *N*	853			878		
% expressed S	33%			48%		
Corrected mean	.31			.48		
	Weight	%	% of data	Weight	%	% of data
Verb class						
psychological	.70	55	18	.68	67	20
copula	.55	39	7	.63	60	6
speech	.53	35	20	.53	51	16
other	.43	25	41	.42	39	47
motion	.35	20	12	.36	31	9
Range	35			32		
Distance						
5^+ clauses	.63	46	23	.58	55	38
2–4 clauses	.58	39	16	.51	49	17
1 clause	.53	34	16	.45	44	13
0 (subject continuity)	.39	23	44	.42	38	30
Range	24			16		
Previous realization						
expressed	.67	50	35	.57	55	49
unexpressed	.41	23	64	.43	40	50
Range	26			14		
TAM						
ambiguous TAM	.62	37	22	.62	60	10
unambiguous TAM	.47	32	77	.48	46	89
Range	15			14		

"The magnitude of effect of the factor groups is determined. This is captured in the range, which represents the difference between the factor that most favors realization of the variant (with the weight closest to 1), and that which least favors its realization (with the weight closest to 0). "

Travis, C. E. 2007. Genre effects on subject expression in Spanish: Priming in narrative and conversation. *Language Variation & Change*, 19(2), 101-135.

Example 2: Monophthongization of /au/ in Scotland

TABLE 4. *Variable rule analysis of the contribution of lexical item and situational context to the use of the monophthong in caregiver and child speech*

	Caregivers	Children
Corrected mean	.52	.41
Log likelihood	−484.100	−367.455
Lexical item		
Now	.23	.56
Down	.72	.71
Out	.77	.76
House	.63	.74
About	.98	.77
Round	.53	.55
How	.03	.02
Other	.52	.58
Range	*95*	*75*
Situational context		
Play	.76	.66
Routine	.67	.57
Discipline	.30	N/A
Teaching	.18	.19
Range	*58*	*47*
	Total N = 1048	Total N = 714

Smith, J., M. Durham, & L. Fortune. 2007. “Mam, my trousers is fa’in doon!”: Community, caregiver, and child in the acquisition of variation in a Scottish dialect. *Language Variation & Change*, 19(1), 63-99.

TABLE 14. *Social factors influencing CSD (Goldvarb analysis)*

Input: .667	Chi-square/cell = 1.5281		
Total *N*: 9554	Log likelihood = −4341.043		
	FW	%	*N*
Ethnicity			
African American	.71	72%	820
European American	.48	59%	8734
Range	*23*		
Age			
Group 4	.57	66%	3012
Group 2	.54	58%	3337
Group 3	.39	57%	3205
Range	18		
Social class			
Lower middle	.53	60%	4331
Working	.51	60%	2648
Upper middle	.44	61%	2575
Range	9		
Sex			
Female	.53	63%	4577
Male	.47	57%	4977
Range	6		
Region			
North	{NS}	60%	5167
South	{NS}	60%	4387
College experience			
Some college	{NS}	61%	6183
No college	{NS}	59%	3371

Note: Other factor groups included in analysis: following phonological environment, preceding phonological environment, and morphological type (see Table 11).

Hazen, K. 2011. Flying high above the social radar: Coronal stop deletion in modern Appalachia. *Language Variation & Change*, 23(1), 105-137.

Why this is wrong

- The predictor with more levels is very likely to come out as more important simply by chance
- Range comparisons may ignore most of the data
  - They don't answer the question we are really interested in: How important is it to know the value of predictor X for predicting the value of Y?

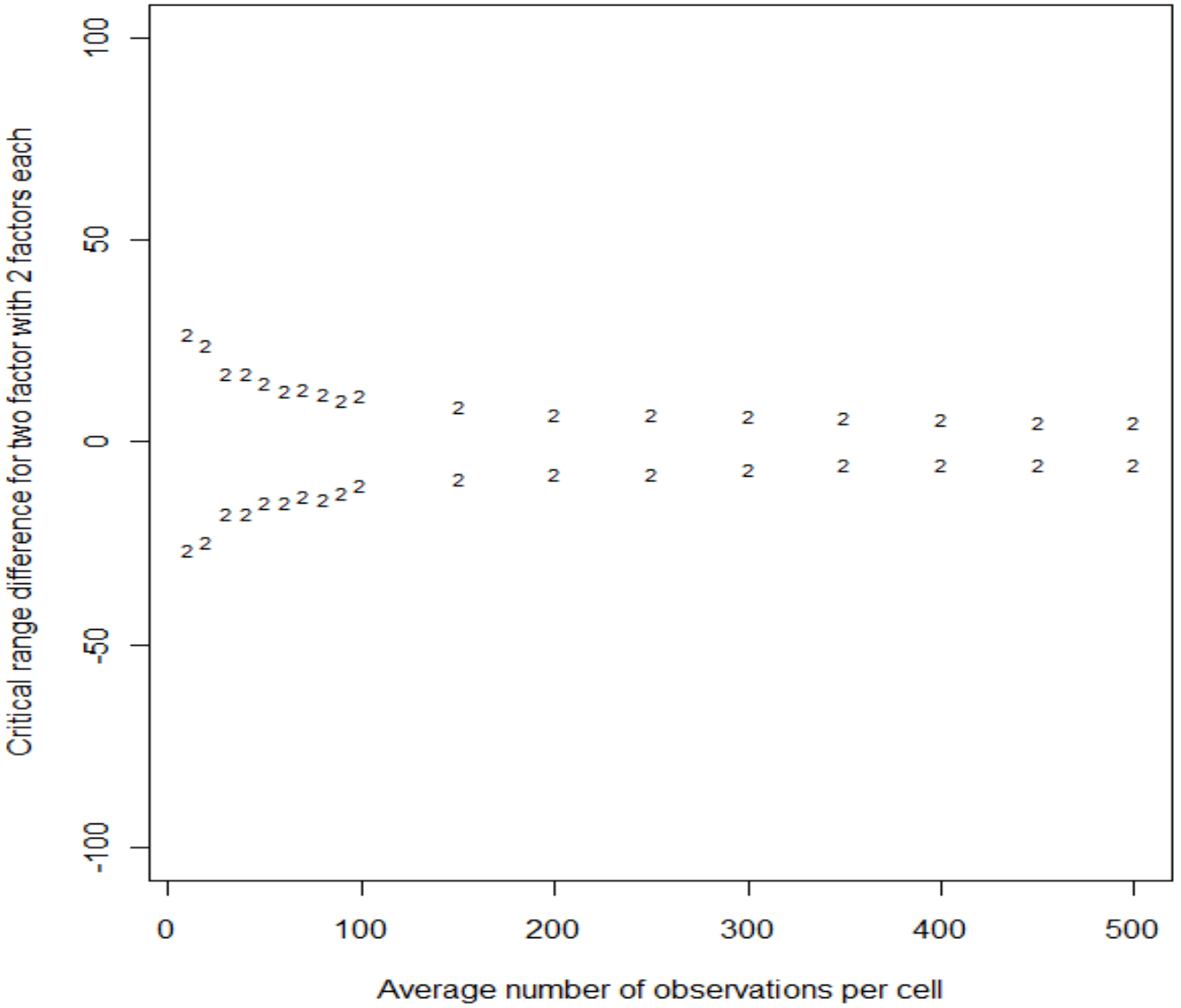
Design of simulation study

X and Z are predictors, with i and j values respectively.
There are observations for every combination of values of X and Z.
N is the number of observations, which can vary independently across values of X and Z.
Y is the number of times the binary dependent variable takes the value coded as "1".

	X_1	X_2	$X\ldots$	X_i
Z_1	Y_{11}/N_{11}	Y_{12}/N_{12}	$Y_{1\ldots}/N_{1\ldots}$	Y_{1i}/N_{1i}
Z_2	Y_{21}/N_{21}	Y_{22}/N_{22}	$Y_{2\ldots}/N_{2\ldots}$	Y_{2i}/N_{2i}
$Z\ldots$	$Y_{\ldots1}/N_{\ldots1}$	$Y_{\ldots2}/N_{\ldots2}$	$Y_{\ldots}/N_{\ldots}$	$Y_{\ldots i}/N_{\ldots i}$
Z_j	Y_{j1}/N_{j1}	Y_{j1}/N_{j1}	$Y_{j\ldots}/N_{j\ldots}$	Y_{ji}/N_{ji}

Create a thousand of samples of size mean(N)*i*j for each combination of mean(N),
i, j, and the distribution of N across i and j.
Run a logistic regression analysis on each sample.
Calculate the range of each of the predictors based on the regression analysis.
Subtract the range of the predictor with more values from the range of the predictor with fewer values.
How often is range of X greater than range of Z depending on i and j?

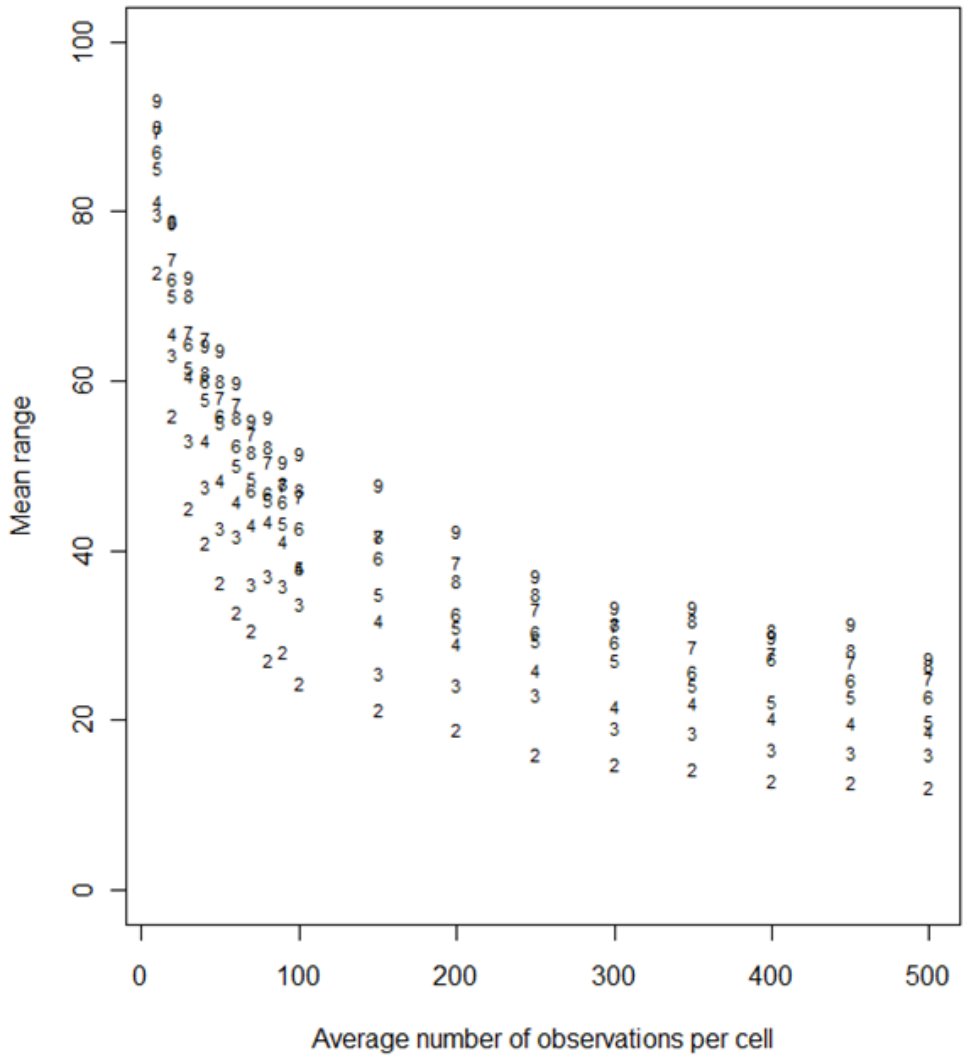
Critical range differences

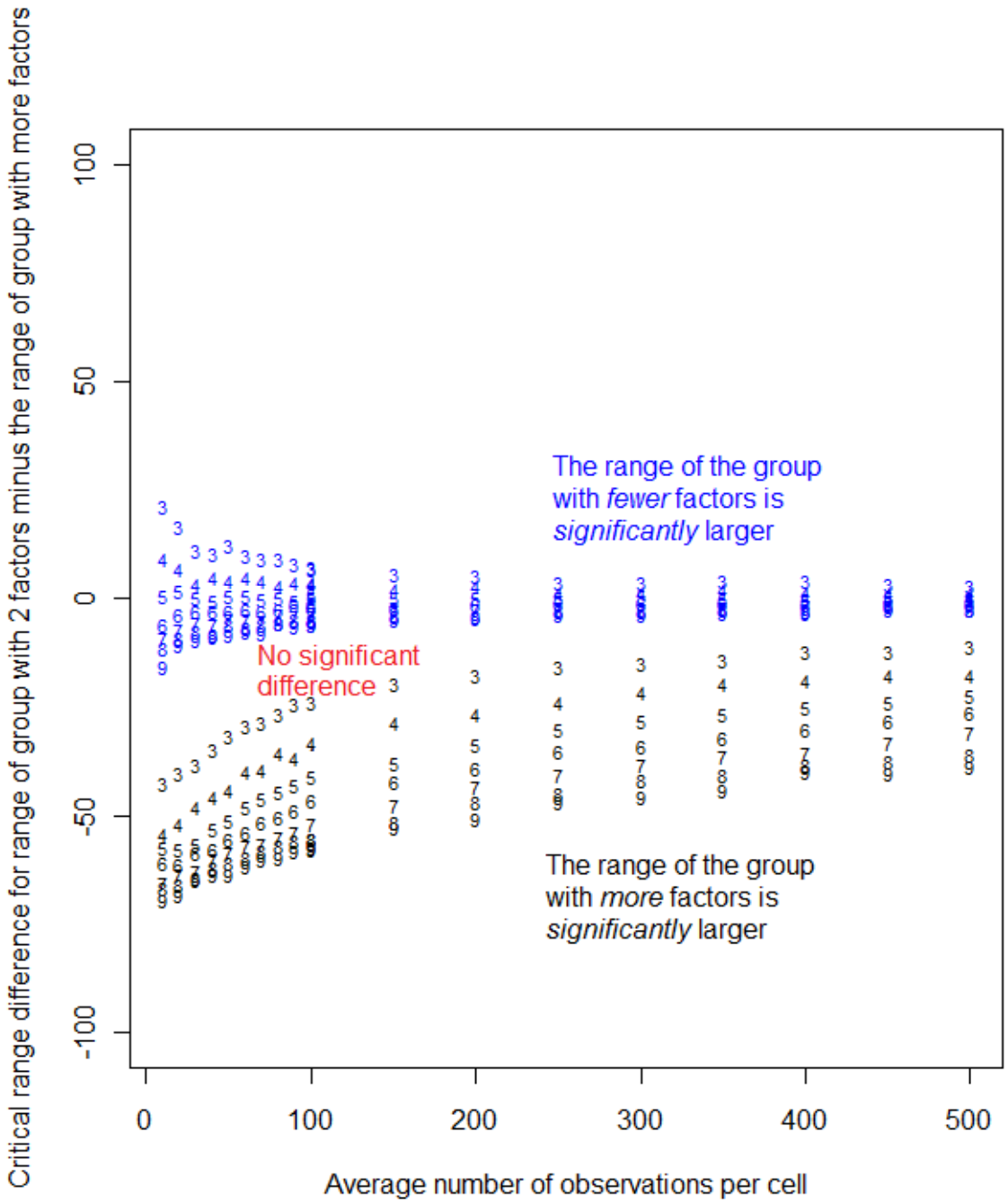


Not every numerical range difference is significant!

Bias in favor of predictors with more levels

Mean range of a factor group that came out as significant simply by chance depending on how many factors the group has and how many observations there are per factor.
Each point represents the mean from a sample of 1000 samples in which the predictor came out as significant by chance.




Critical range difference for range of group with 2 factors minus the range of group with more factors
100
50
0
-50
-100
0
100
200
300
400
500
Average number of observations per cell
The range of the group with *fewer* factors is *significantly* larger
No significant difference
The range of the group with *more* factors is *significantly* larger

Testing significance of range differences in Travis (2007)

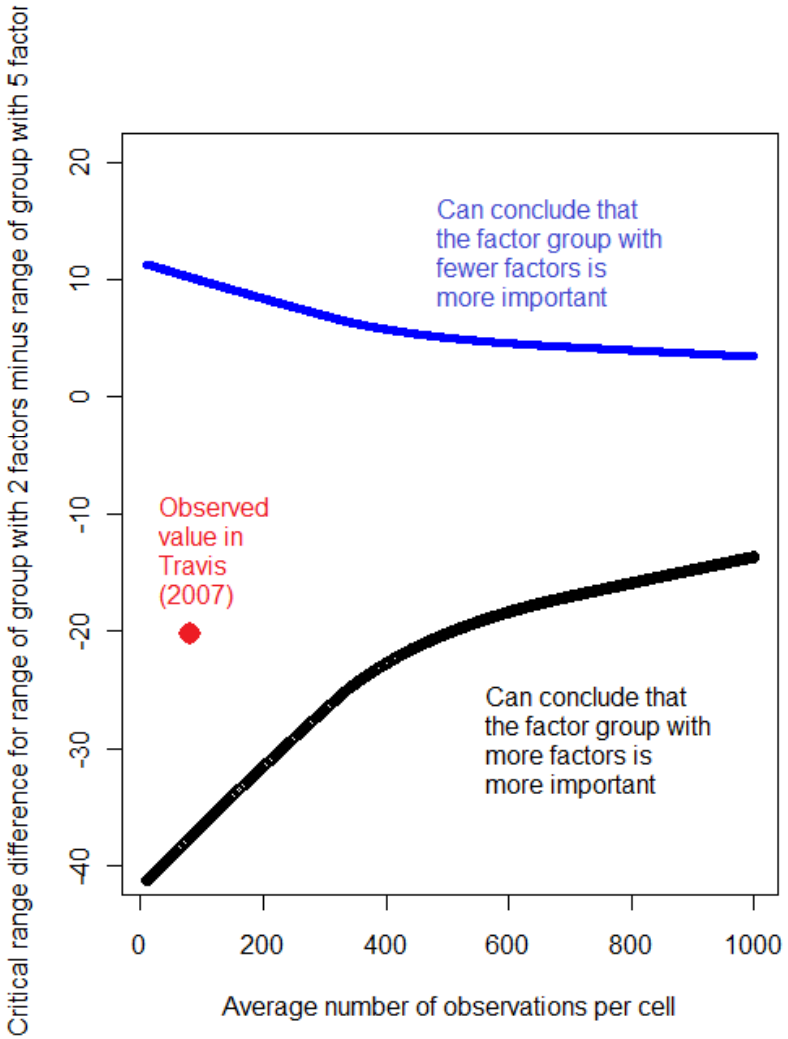


TABLE 2. *Two independent variable rule analyses of the contribution of factors selected as significant to the probability of expressed subjects in NM and Colombian data*

	NMCOSS			Colo	
Total *N*	853			87	
% expressed S	33%			48	
Corrected mean	.31			.4	
	Weight	%	% of data	Weight	%
Verb class					
psychological	.70	55	18	.68	67
copula	.55	39	7	.63	60
speech	.53	35	20	.53	51
other	.43	25	41	.42	39
motion	.35	20	12	.36	31
Range	35			32	
Distance					
5+ clauses	.63	46	23	.58	55
2–4 clauses	.58	39	16	.51	49
1 clause	.53	34	16	.45	44
0 (subject continuity)	.39	23	44	.42	38
Range	24			16	
Previous realization					
expressed	.67	50	35	.57	55
unexpressed	.41	23	64	.43	40
Range	26			14	
TAM					
ambiguous TAM	.62	37	22	.62	60
unambiguous TAM	.47	32	77	.48	46
Range	15			14	

Observed range difference: -20

Validity of range

- Consider the range of a predictor that has three values.
- Unlike in the case of a binary predictor, a range of 1 does not mean that the factor group is a perfect predictor.
- Suppose that we observe one case in which the predictor has value X, one case in which it has value Y, and 100 cases in which it has value Z.
- Suppose further, that the one time we observe the predictor having the value of X the dependent variable has value A, whereas the one time the predictor has value Y the dependent has value B.
- Then, the predictor has the maximum possible range of 1 regardless of what happens with the 100 observations, in which it has value Z.

→ A predictor can have maximum range without being at all predictive.

But we only look at significant predictors

- One might object that we only compare ranges of factor groups that are selected as being significant predictors by the logistic regression model.
- This is highly unlikely in the case of a factor group that has only one observation for two of its values.
- However, the general point holds: range only takes into account the two factors in a group that are exhibiting the most extreme behavior, paying no attention to whether the predictor is of any use in accounting for the choice when it does not take those values, essentially throwing out a significant portion, potentially the vast majority, of the data.
- Further, based on the Central Limit Theorem, the values of the predictor most likely to show extreme behavior are precisely the ones that are observed rarely,

→ It's quite likely that one would base one's conclusion about the importance of a predictor based on a minority of the data.

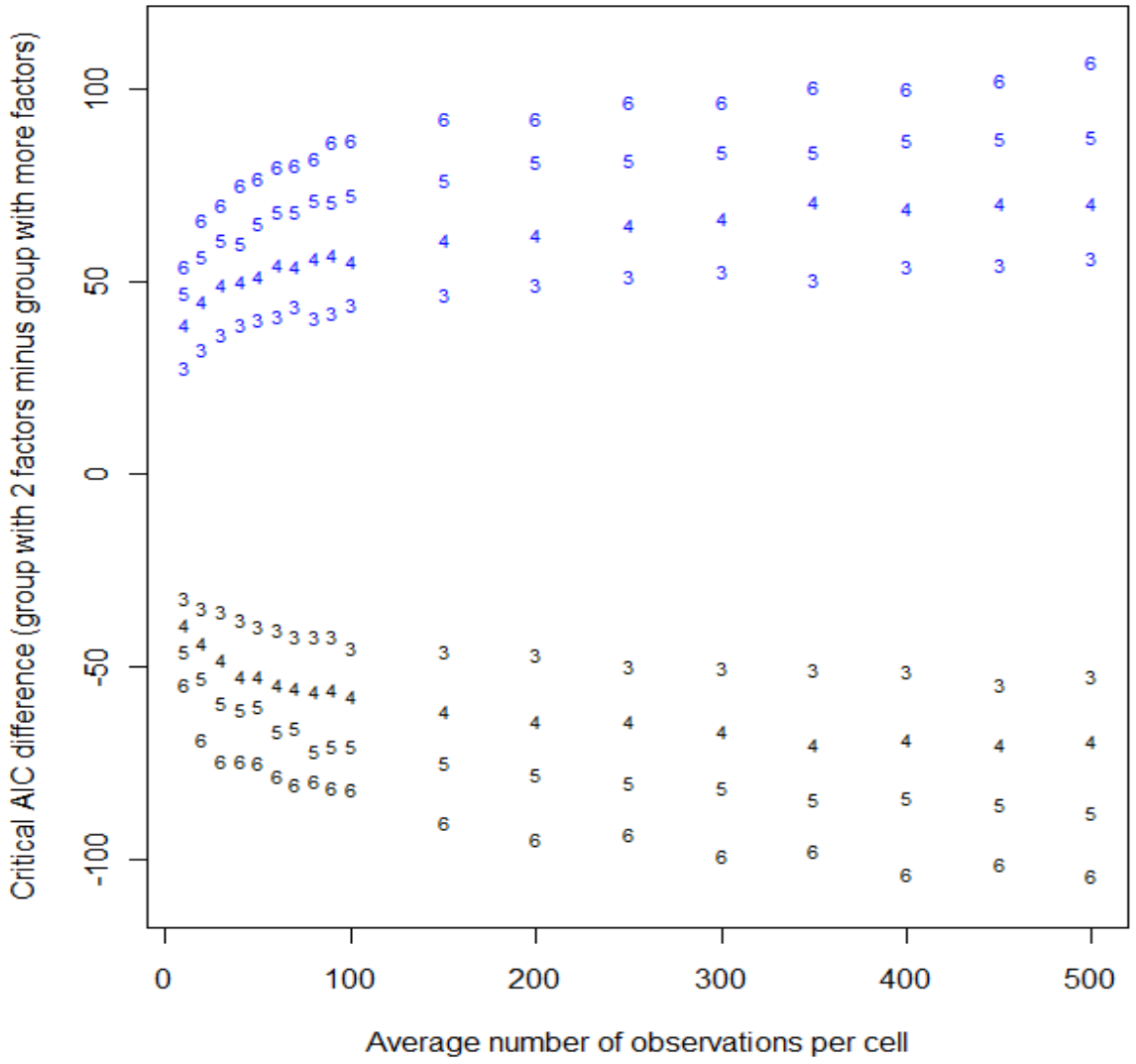
Solution: Akaike Information Criterion

(Akaike 1974)

- Compare two logistic regression models:
  - One lacks predictor X
  - The other has predictor X
- For each model, compute AIC:

  -2(log_likelihood - number_of_model_parameters)
- AIC captures how much the predictor improves predictiveness of the model (AIC is valid)

AIC is unbiased



Conclusion

- It is time to start using AIC instead of range to evaluate predictor importance.

- Whatever you use, you need to evaluate whether the predictors are *significantly different* in predictiveness.

References:

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control,* 19(6): 716–723.

Hazen, K. 2011. Flying high above the social radar: Coronal stop deletion in modern Appalachia. *Language Variation & Change*, 23(1), 105-137.

Kapatsinski, V. 2009. Adversative conjunction choice in Russian (*no*, *da*, *odnako*): Semantic and syntactic influences on lexical selection. *Language Variation & Change*, 21(2), 157-173.

Smith, J., M. Durham, & L. Fortune. 2007. “Mam, my trousers is fa’in doon!”: Community, caregiver, and child in the acquisition of variation in a Scottish dialect. *Language Variation & Change*, 19(1), 63-99.

Travis, C. E. 2007. Genre effects on subject expression in Spanish: Priming in narrative and conversation. *Language Variation & Change*, 19(2), 101-135.