Statistics Seminar, Spring 2009

Binomial (or Binary) Logistic Regression

Anja Schüppert

a.schueppert@rug.nl

Linear regression: Univariate

One independent variable, one (continuous) dependent variable.

$$\text{Outcome}_i = \text{Model}_i + \text{Error}_i$$
$$Y_i = b_0 + b_1 X_1 + \varepsilon i$$

b_0: interception at y-axis

b_1: line gradient

X_1: predictor variable

ε : Error

X_1 predicts Y.

Linear regression: Multivariate

Several independent variables, one (continuous) dependent variable.

$$Y_i = b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n + \varepsilon_i$$

b_0: interception at y-axis

b_1: line gradient

b_n: regression coefficient of X_n

X_1: predictor variable

ε: Error

X_1 predicts Y.

Assumption

- Linear regression assumes linear relationships between variables.
- This assumption is usually violated when the dependent variable is categorical.
- The logistic regression equation expresses the multiple linear regression equation in logarithmic terms and thereby overcomes the problem of violating the linearity assumption.

Assumption cont.

$\log_{base}[\text{number}]$

$\log_2 16 \quad = 4 \qquad => \qquad 2^4 = 2 \text{ x } 2 \text{ x } 2 \text{ x } 2 = 16$

'natural logarithm': ln

$\ln = \log_e[\text{number}]$ | e = Eulers constant $\approx$ 2,7182818284…

ln[odds] => 'logit'

$$\text{logit}(p) = \ln \frac{p}{(1-p)}$$

$$e^{\text{logit}(p)} = \frac{p}{1-p}$$

$$\begin{aligned} e^{\text{logit(p)}}(1-p) = \quad p \quad &= e^{\text{logit(p)}} - pe^{\text{logit(p)}} \\ p + pe^{\text{logit(p)}} &= e^{\text{logit(p)}} \\ p(1+e^{\text{logit(p)}}) &= e^{\text{logit(p)}} \\ p &= \frac{1}{1+e^{-\text{logit}(p)}} \end{aligned}$$

Binary logistic regression:
Univariate

One independent variable, one *categorical* dependent variable.

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 x_1)}}$$

P: probability of Y occuring

e: natural logarithm base (= 2,7182818284…)

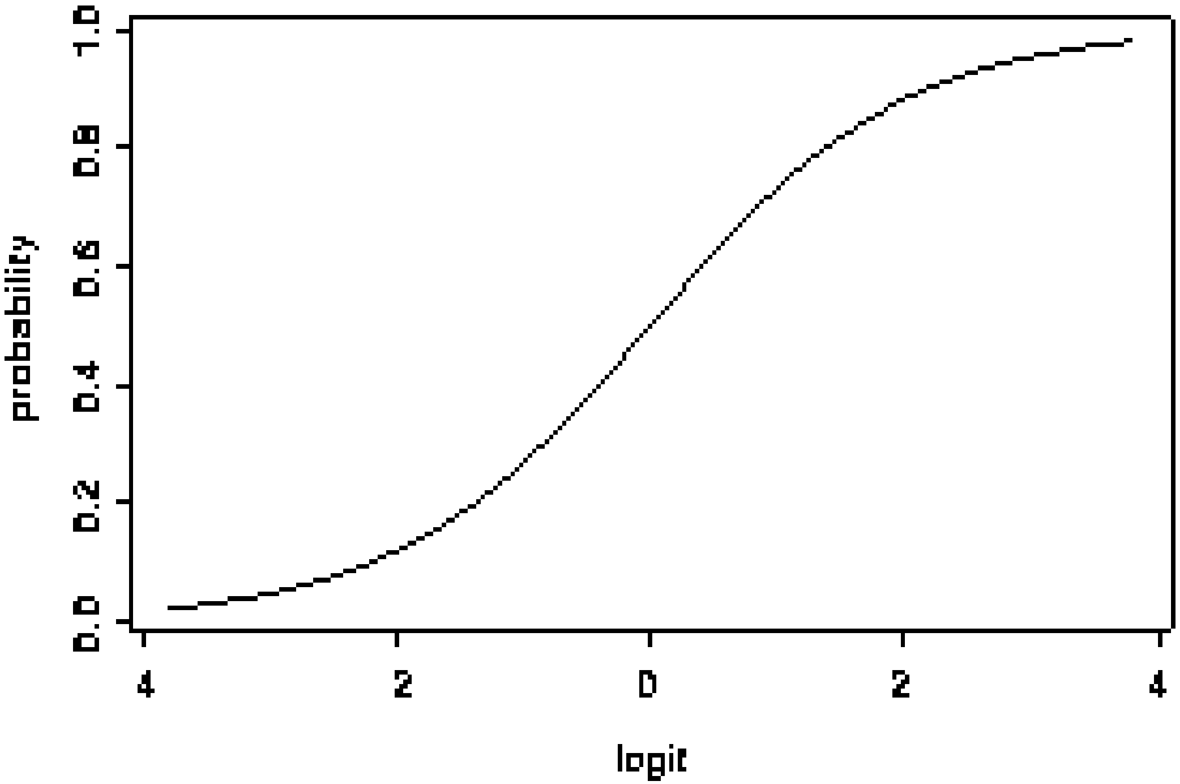
b_0: interception at y-axis

b_1: line gradient

X_1 predicts the probability of Y.

Binary logistic regression: Univariate cont.

As P(Y) ranges from 0 to 1, the logit ranges from -∞ to +∞.



http://data.princeton.edu/wws509/notes/c3s1.html

Binary logistic regression: Multivariate

Several independent variables, one *categorical* dependent variable.

$$P(Y) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n}}$$

P: probability of Y occuring

e: natural logarithm base

b_0: interception at y-axis

b_1: line gradient

b_n: regression coefficient of X_n

X_1: predictor variable

X_1 predicts the probability of Y.

Binary logistic regression: Multivariate cont.

=> Linear regression predicts the *value* that Y takes.

Instead, in logistic regression, the *frequencies* of values 0 and 1 are used to predict a value:

=> Logistic regression predicts the *probability* of Y taking a specific value.

Research question

- Broad: How intelligible is Danish to Swedish listeners without previous exposure?

  Here: Which factors predict whether a Danish word is easily decoded by Swedish pre-schoolers or not?
- Dependent variable: Word intelligibility

  Every word can be
  - Decoded (1), or
  - Not decoded (0)

  Binary variable
- Independent variables:

– Phonetic distance	Continuous variable (0.00 - 1.00)
– Toneme	Binary variable (0,1)
– Number of 'difficult' sounds for the listener	Categorical variable (0,1,2,3,…)

Experiment

- 50 Danish word were auditorily presented to 12 Swedish children via headphones
- Similarly, 200 pictures (i.e. 4 pictures per sound) were presented visually on a touch screen
- The children were instructed to point to the corresponding picture
- Resulting data: Intelligibility scores per word per subject

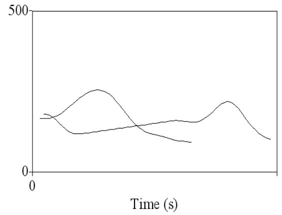
50 x 12 = 600 scores

Independent variables: Examples

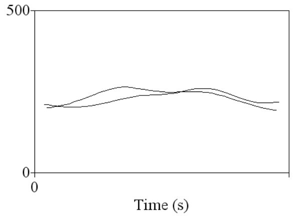
- Phonetic distance:

	måne/måne	hund/hund	apa/abe
Sw.	[mo:nə]	[hɵnd]	[ɑ:pa]
Da.	[mo:nə]	[hunʔ]	[ɛ:bə]
	0%	50%	100%

- Swedish tonemes:
  (NB: Tonemes not found in test language, only in listeners' native language!)

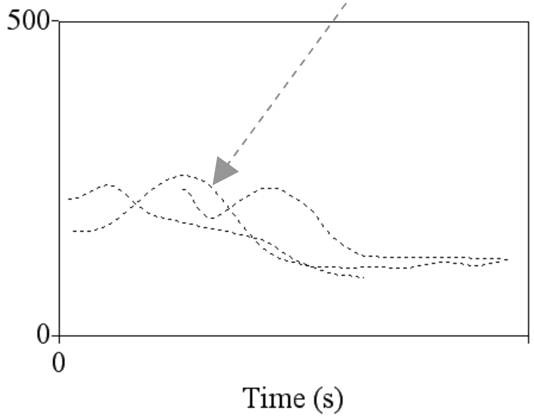


Swedish
bäbis vs *äpple*

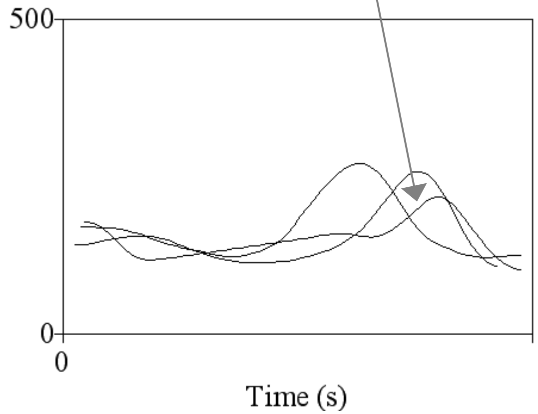


Danish
baby vs *æble*

Toneme 1 (e.g. *bäbis*)



Toneme 2 (e.g. *äpple*)



- 'Difficult sounds': Danish sounds that have been shown to be significantly more difficult to decode for Swedes (Schüppert & Gooskens, in prep.): [ɛ], [ð], [j], [œ]

Data

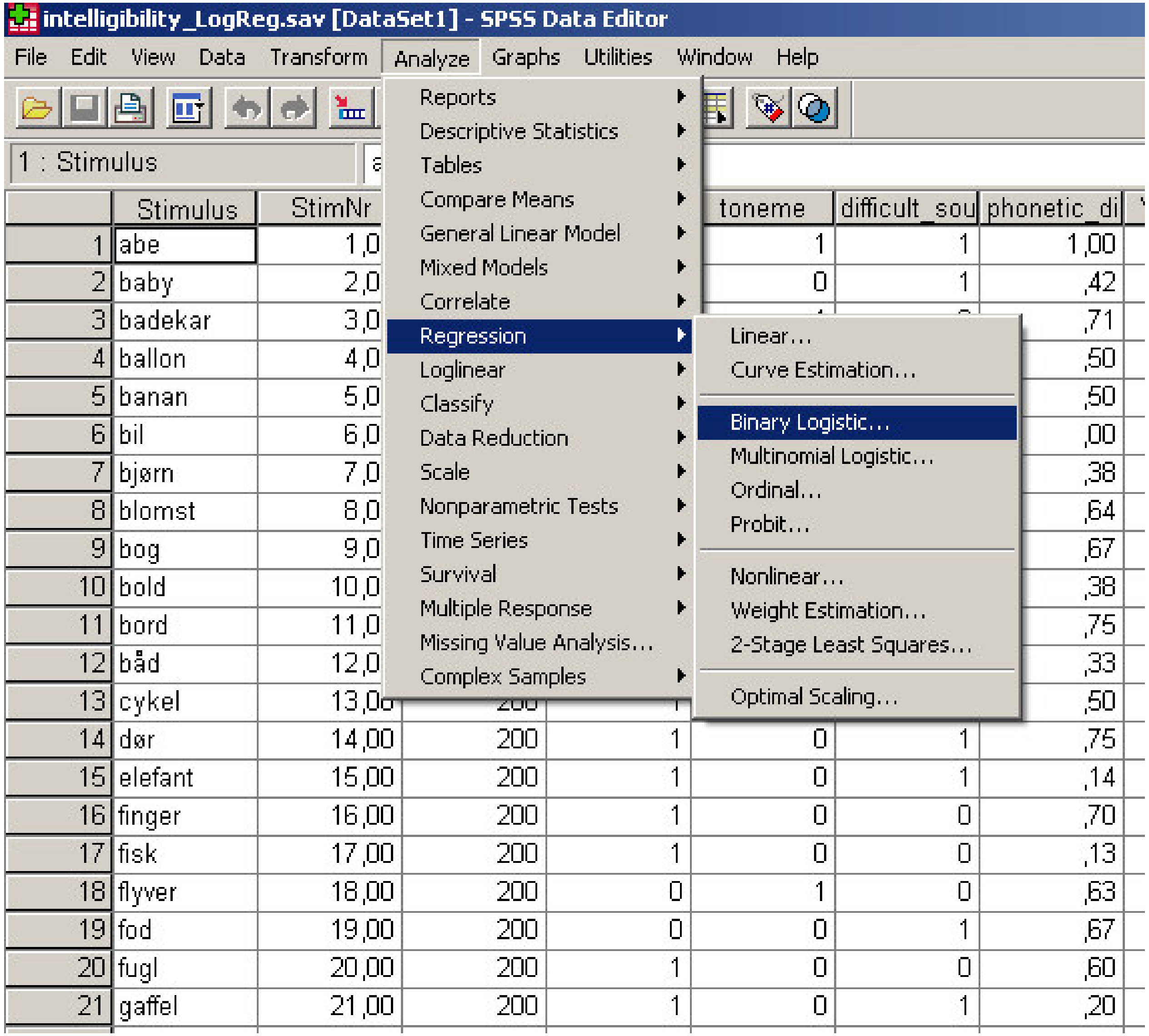
intelligibility_LogReg.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

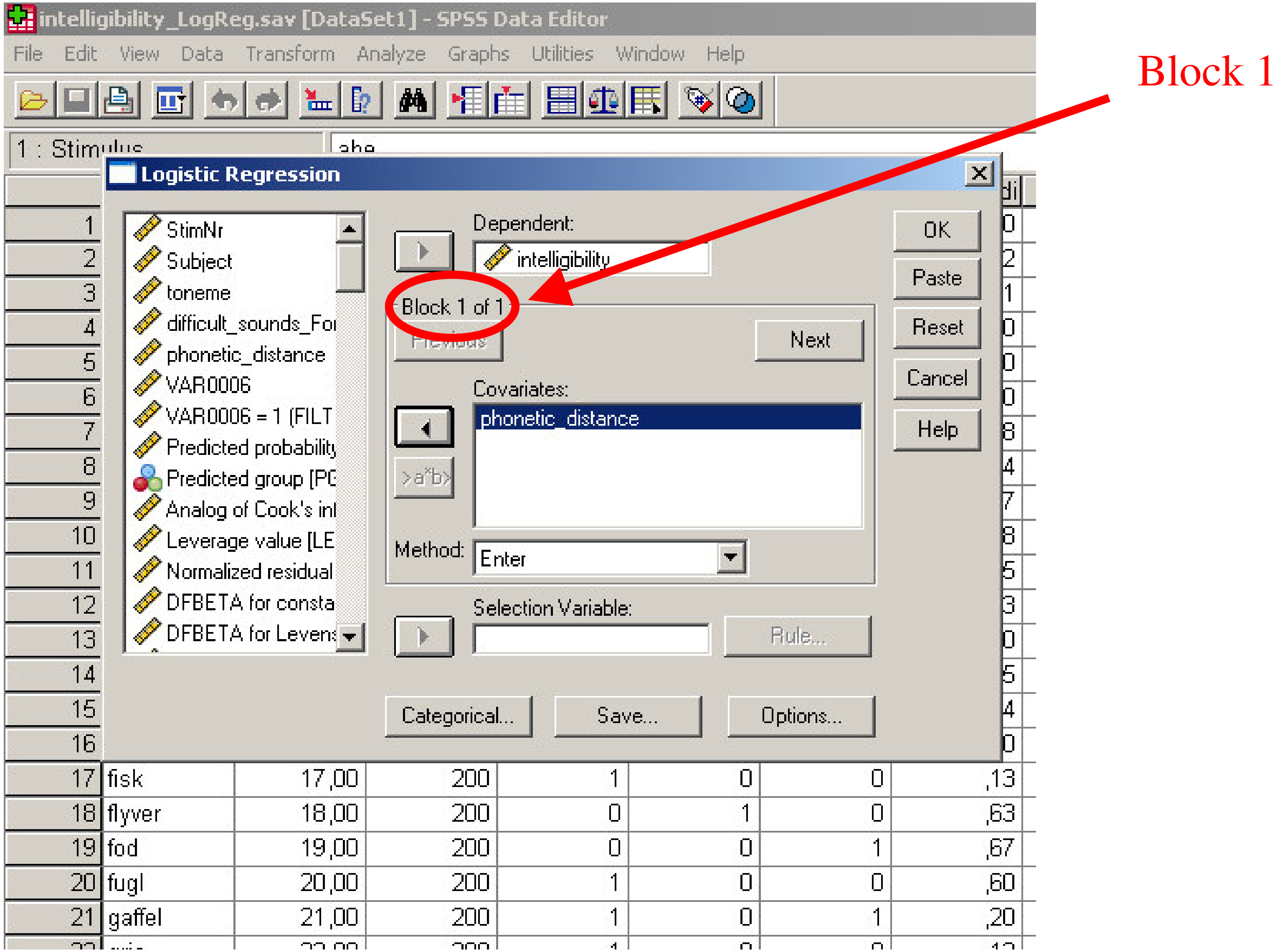
1 : Stimulus abe

	Stimulus	StimNr	Subject	intelligibility	toneme	difficult_sou	phonetic_di
1	abe	1,00	200	0	1	1	1,00
2	baby	2,00	200	1	0	1	,42
3	badekar	3,00	200	1	1	2	,71
4	ballon	4,00	200	1	0	0	,50
5	banan	5,00	200	0	0	1	,50
6	bil	6,00	200	1	0	0	,00
7	bjørn	7,00	200	1	0	1	,38
8	blomst	8,00	200	0	1	0	,64
9	bog	9,00	200	1	0	0	,67
10	bold	10,00	200	1	0	0	,38
11	bord	11,00	200	0	0	0	,75
12	båd	12,00	200	1	0	1	,33
13	cykel	13,00	200	1	0	0	,50
14	dør	14,00	200	1	0	1	,75
15	elefant	15,00	200	1	0	1	,14
16	finger	16,00	200	1	0	0	,70
17	fisk	17,00	200	1	0	0	,13
18	flyver	18,00	200	0	1	0	,63
19	fod	19,00	200	0	0	1	,67
20	fugl	20,00	200	1	0	0	,60
21	gaffel	21,00	200	1	0	1	,20

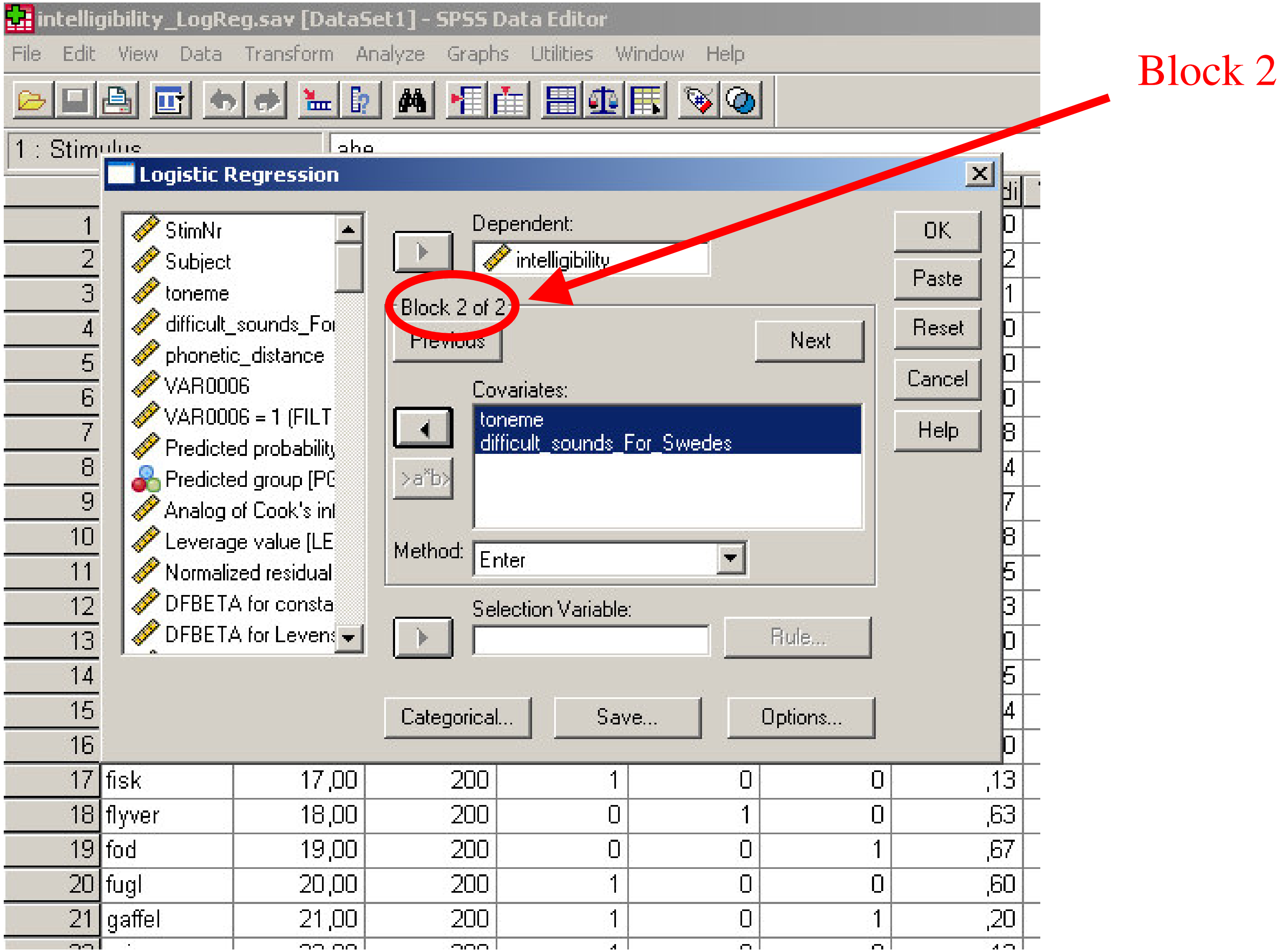
Data Entry


intelligibility_LogReg.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : Stimulus
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Multiple Response
Missing Value Analysis...
Complex Samples
Linear...
Curve Estimation...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling...
Stimulus
StimNr
toneme
difficult_sou
phonetic_di
abe
baby
badekar
ballon
banan
bil
bjørn
blomst
bog
bold
bord
båd
cykel
dør
elefant
finger
fisk
flyver
fod
fugl
gaffel

Data Entry

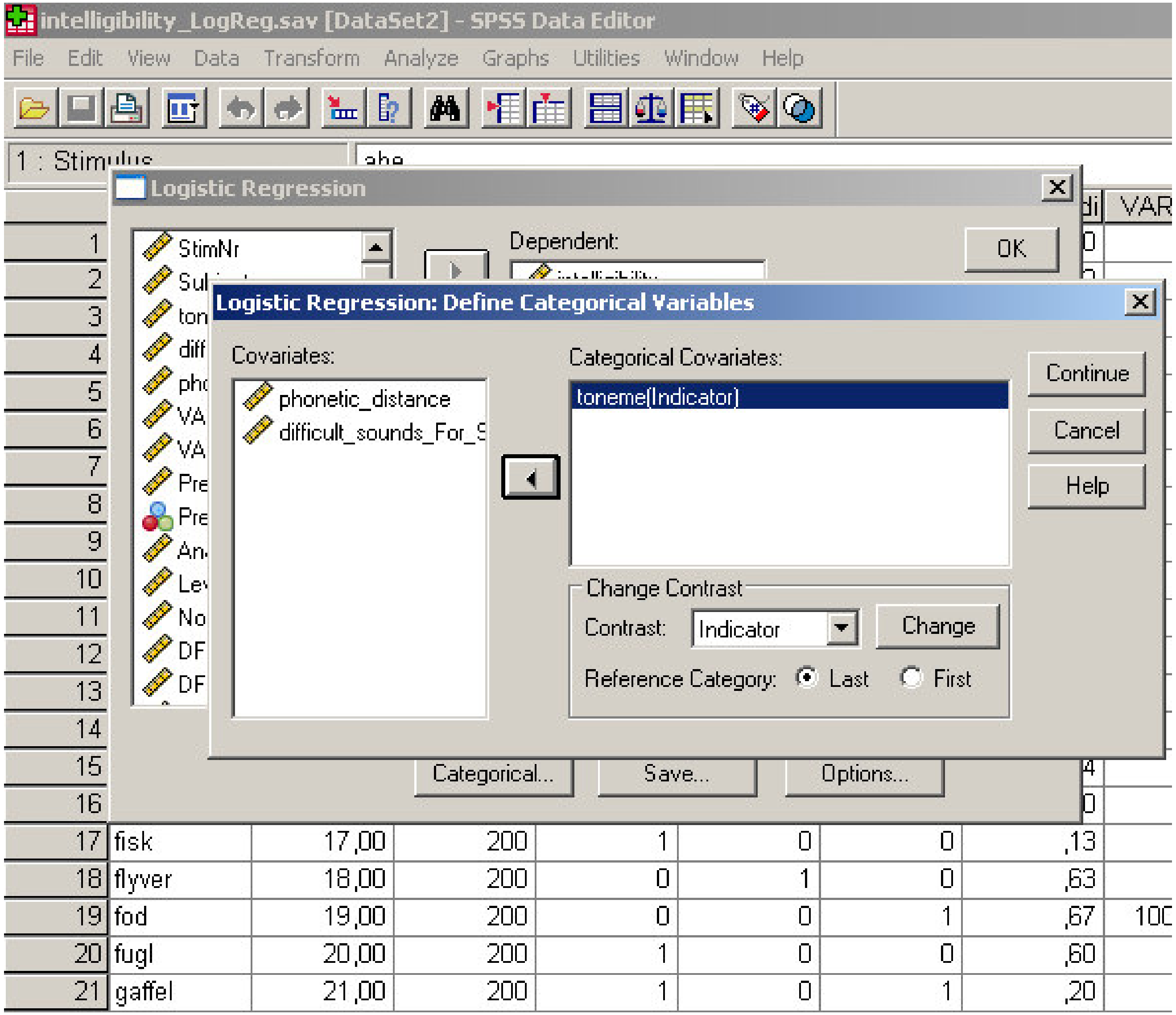

intelligibility_LogReg.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Logistic Regression
StimNr
Subject
toneme
difficult_sounds_Fo
phonetic_distance
VAR0006
VAR0006 = 1 (FILT
Predicted probability
Predicted group [PG
Analog of Cook's inf
Leverage value [LE
Normalized residual
DFBETA for consta
DFBETA for Levens
Dependent:
intelligibility
Block 1 of 1
Next
Covariates:
phonetic_distance
Method:
Enter
Selection Variable:
Rule...
Categorical...
Save...
Options...
OK
Paste
Reset
Cancel
Help
17 fisk 17,00 200 1 0 0 ,13
18 flyver 18,00 200 0 1 0 ,63
19 fod 19,00 200 0 0 1 ,67
20 fugl 20,00 200 1 0 0 ,60
21 gaffel 21,00 200 1 0 1 ,20
Block 1

Data Entry

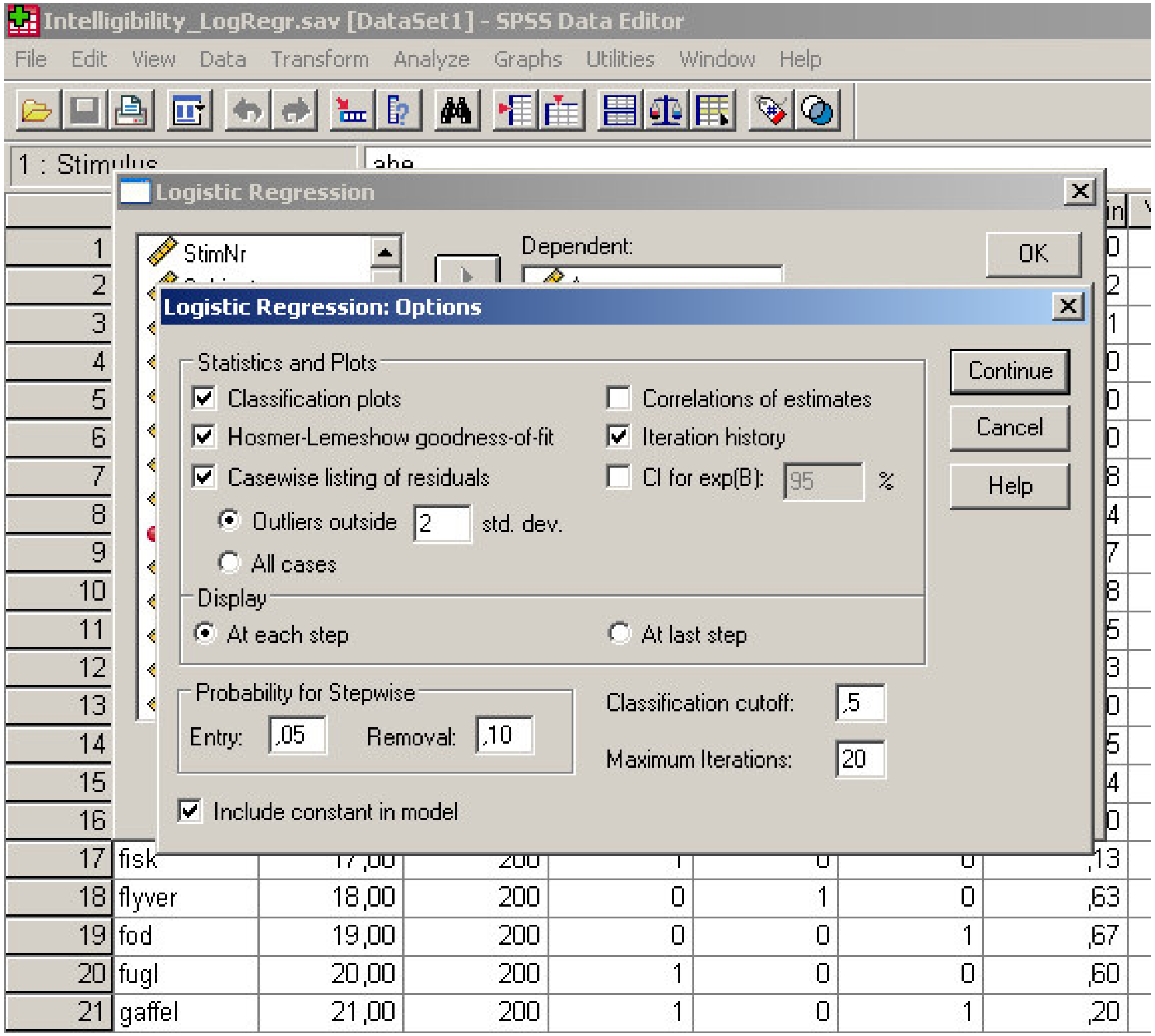


intelligibility_LogReg.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Logistic Regression
StimNr
Subject
toneme
difficult_sounds_For
phonetic_distance
VAR0006
VAR0006 = 1 (FILT
Predicted probability
Predicted group [PG
Analog of Cook's inf
Leverage value [LE
Normalized residual
DFBETA for consta
DFBETA for Levens
Dependent:
intelligibility
Block 2 of 2
Previous
Next
Covariates:
toneme
difficult_sounds_For_Swedes
Method:
Enter
Selection Variable:
Rule...
Categorical...
Save...
Options...
OK
Paste
Reset
Cancel
Help
17 fisk 17,00 200 1 0 0 ,13
18 flyver 18,00 200 0 1 0 ,63
19 fod 19,00 200 0 0 1 ,67
20 fugl 20,00 200 1 0 0 ,60
21 gaffel 21,00 200 1 0 1 ,20
Block 2

Data Entry


intelligibility_LogReg.sav [DataSet2] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Logistic Regression
Dependent:
OK
Logistic Regression: Define Categorical Variables
Covariates:
phonetic_distance
difficult_sounds_For_S
Categorical Covariates:
toneme(Indicator)
Continue
Cancel
Help
Change Contrast
Contrast:
Indicator
Change
Reference Category:
Last
First
Categorical...
Save...
Options...
fisk
flyver
fod
fugl
gaffel

Data Entry


Intelligibility_LogRegr.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 : Stimulus
Logistic Regression
StimNr
Dependent:
OK
Logistic Regression: Options
Statistics and Plots
Classification plots
Hosmer-Lemeshow goodness-of-fit
Casewise listing of residuals
Outliers outside 2 std. dev.
All cases
Correlations of estimates
Iteration history
CI for exp(B): 95 %
Continue
Cancel
Help
Display
At each step
At last step
Probability for Stepwise
Entry: ,05
Removal: ,10
Classification cutoff: ,5
Maximum Iterations: 20
Include constant in model
18 flyver 18,00 200 0 1 0 ,63
19 fod 19,00 200 0 0 1 ,67
20 fugl 20,00 200 1 0 0 ,60
21 gaffel 21,00 200 1 0 1 ,20

Output: Block 1 (Phonetic distance)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	62,049	1	,000
	Block	62,049	1	,000
	Model	62,049	1	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	513,089[a]	,117	,171

Improvement through added variable 'phonetic distance'

Improvement is significant: predictor 'phonetic distance' contributes to the model

-2LL: Amount of unexplained variance

$R^2_{CS} = 0.12$

$R^2_{N} = 0.17$

Output: Block 1 (Phonetic distance)

Classification Table[a]

	Observed		Predicted: intelligibility 0	Predicted: intelligibility 1	Percentage Correct
Step 1	intelligibility	0	18	113	13,7
		1	12	357	96,7
	Overall Percentage				75,0

Exp(B) < 1

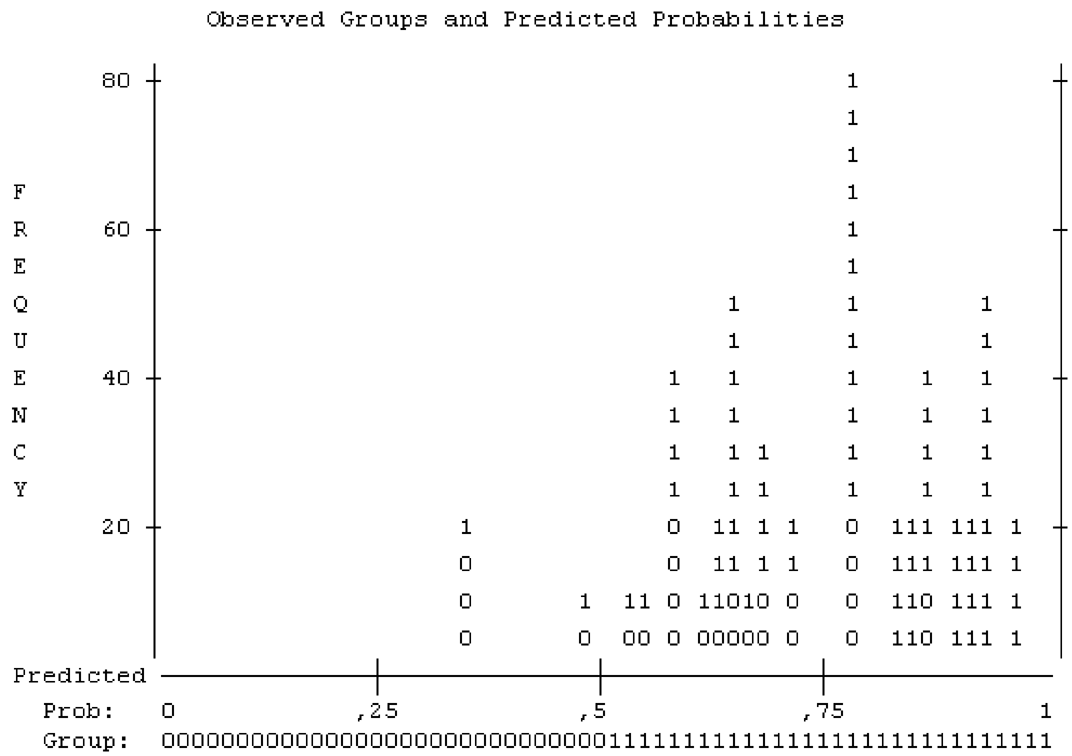
Indicates that phonetic distance correlates *negatively* with intelligibility.

Model predicts correct value in 75% of the cases.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B) Lower	Upper
Step 1[a]	phonetic_distance	-3,704	,526	49,658	1	,000	,025	,009	,069
	Constant	3,084	,331	86,654	1	,000	21,849		

Output: Block 1 (Phonetic distance)



Nondecoded stimuli seem to be difficult to predict (the zeroes should be concentrated further to left).

Decoded stimuli are more correctly predicted by the model (note the 1-columns on the right hand side of the plot).

Output: Block 2
(Phonetic distance, Toneme, Difficult Sounds)

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	14,754	3	,002
	Block	14,754	3	,002
	Model	76,804	4	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	498,335[a]	,142	,208

Model has further improved through added variable(s)

Significant value: indicates that one or both of the new predictors improve the new model.

-2LL: Amount of unexplained variance is reduced from 513,09 to 498,34

$R^2_{CS} = 0.14$ $\quad$ $R^2_N = 0.21$

(Block 1: $R^2_{CS} = 0.12$ $\quad$ $R^2_N = .17$)

Output: Block 2
(Phonetic distance, Toneme, Difficult Sounds)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B) Lower	Upper
Step 1[a]	phonetic_distance	-2,950	,572	26,555	1	,000	,052	,017	,161
	toneme(1)	,272	,286	,905	1	,341	1,313	,749	2,302
	difficult_sounds_For_Swedes	-,363	,171	4,506	1	,034	,695	,497	,973
	Constant	2,746	,456	36,209	1	,000	15,573		

Significant value indicates that variable 'difficult sounds' *improves* the model. Exp(B) < 1 indicates a negative correlation.

Non-significant value indicates that variable 'toneme' *does not* improve the model.

Results and Conclusion

Phonetic distance correlates negatively with intelligibility and contributes significantly to the model.

Tonemes seem not to be contributing to the model. This phenomenon, that listeners are familiar with from their native language but that is missing in the test language, does not seem to puzzle the listeners.

The number of difficult sounds correlate negatively with intelligibility and contribute significantly to the model.

Together, phonetic distance and number of strange sounds account for 14% to 21% of the variance.

References

- Cohen, J., P. Cohen, S.G. West, L.S. Aiken. 2003. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. London: Lawrence Erlbaum.
- Field, A. 2005. *Discovering Statistics Using SPSS*. London: Sage.
- Rietveld, T., R. van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behaviour*. Berlin: Mouton de Gruyter.
- Schüppert, A. & Gooskens, Ch. In prep. *Easy sounds, difficult sounds. Evidence from a word comprehension task with Swedish children listening to Danish.*
- Sieben, I. Logistische regressie analyse: een handleiding. http://scm.ou.nl/Manual/logisticregreskun.html
- Tabachnick, B.G., L.S. Fidell. 2007. *Using Multivariate Statistics*. Boston: Pearson.