Ordinal Regression

Seid Tvica

Overview

- Regression Models
- Logistic Regression
- Ordinal regression model
- Experiment
- Results
- Conclusion

Regression Models

- Linear Regression
  - Interval dependent variable
  - Independent variable predicts the value of the dependent
- Binomial Logistic Regression
  - Binary dependent variable (2 categories)
  - Independent variable predict the *probability* of the value of the dependent

Logistic Regression Model

$$\ln\left(\frac{prob(event)}{(1-prob(event))}\right)=\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_k X_k$$

- Log of the odds that an event occurs (logit)
- Multinomial regression can be used for the dependents with more than 2 categories.
- What about the *ranked* categorical variables?

Ordinal Regression Model

- Defining an event as a particular score or less
  - e.g. For an ordinal variable with four categories the following odds are modeled.

$$\theta_1 = \text{prob(score of 1)}/\text{prob(score greater than 1)}$$

$$\theta_2 = \text{prob(score of 1 or 2)}/\text{prob(score greater than 2)}$$

$$\theta_3 = \text{prob(score of 1,2 or 3)}/\text{prob(score greater than 3)}$$

  - No odds associated for the last category as the probability of scoring up to and the last score is 1.

Ordinal Regression Model

$$\theta_j = \text{prob}(\text{score} \le j)/(1 - \text{prob}(\text{score} \le j))$$

- The ordinal logistic model for a single independent variable is then:

$$\ln(\theta_j) = \alpha_j - \beta X$$

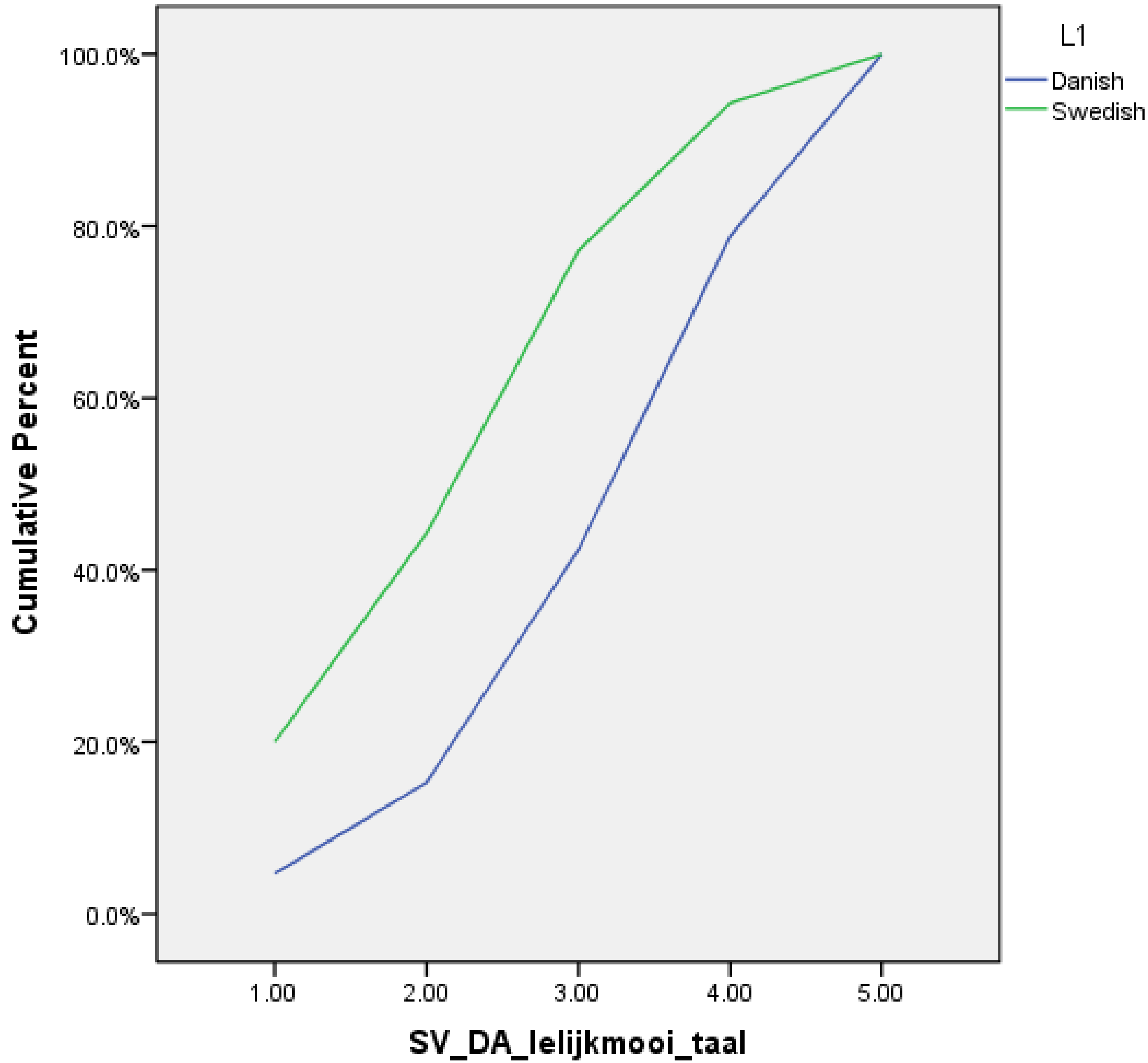
- j = number of categories of the dependent minus 1
- α_j = threshold values.

Experiment

- 160 Danish and Swedish children (ages 7–16) rated each other's languages on a 5-point scale, 1 being ugly, 5 beautiful.

- Dependent ordinal variable *ugly/beautiful*
- Independent variables *L1* and *age groups*

Data


L1
Danish
Swedish
100.0%
80.0%
60.0%
40.0%
20.0%
0.0%
Cumulative Percent
1.00
2.00
3.00
4.00
5.00
SV_DA_lelijkmooi_taal

Results – SPSS output

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[SV_DA_lelijkmooi_taal = 1.00]	-1.411	.278	25.8	1	.000	-1.955	-.867
	[SV_DA_lelijkmooi_taal = 2.00]	-.215	.229	.879	1	.349	-.664	.234
	[SV_DA_lelijkmooi_taal = 3.00]	1.210	.253	22.7	1	.000	.713	1.707
	[SV_DA_lelijkmooi_taal = 4.00]	2.824	.325	75.2	1	.000	2.186	3.462
Location	[L1=1.00]	1.514	.312	23.5	1	.000	.903	2.125
	[L1=2.00]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

“goodness of fit”

- The Pearson goodness-of-fit statistic

$$x^2 = \sum\sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

- The deviance measure

$$D = 2\sum\sum O_{ij} \ln\left(\frac{O_{ij}}{E_{ij}}\right)$$

Results – SPSS output

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	.091	3	.993
Deviance	.091	3	.993

Link function: Logit.

Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	31.787			
General	31.696	.091	3	.993

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

Results – SPSS output

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[SV_DA_lelijkmooi_taal = 1.00]	-.955	.321	8.843	1	.003	-1.585	-.326
	[SV_DA_lelijkmooi_taal = 2.00]	.287	.291	.971	1	.325	-.284	.857
	[SV_DA_lelijkmooi_taal = 3.00]	1.772	.326	29.62	1	.000	1.134	2.411
	[SV_DA_lelijkmooi_taal = 4.00]	3.461	.401	74.44	1	.000	2.675	4.248
Location	[L1=1.00]	1.633	.322	25.74	1	.000	1.002	2.264
	[L1=2.00]	0[a]	.	.	0	.	.	.
	[age_group=1.00]	1.252	.373	11.26	1	.001	.521	1.984
	[age_group=2.00]	.496	.347	2.038	1	.153	-.185	1.177
	[age_group=3.00]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

Results – SPSS output

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	16.442	17	.493
Deviance	17.878	17	.397

Link function: Logit.

Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	80.891			
General	73.207	7.683	9	.566

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

Conclusion

- For any rating level Danish participants who listened Swedish scored higher on the “ugly/beautiful” scale than the Swedish who listened to the Danish. This difference was significant (p=0.000). For the Danish participants the expected log odds ratio increases by 1.63 with each category of the dependent variable.

- For any rating level the first age group of the Scandinavian speakers scored higher on the dependent variable scale than the second and the third. This difference was significant. There were no significant effects from the age groups 2 an 3.

References

- Introduction to SAS. UCLA: Academic Technology Services, Statistical Consulting Group. http://www.ats.ucla.edu/stat/spss/dae/ologit.htm
- Norusis, J. M. (2005). *SPSS 14.0 Advanced statistical procedures companion*. Chicago: Prentice Hall Inc..
- Vedder, E. (2010) MA thesis.