Faculty of Computer Science | Data Analysis | Moscow 2025

Lecture 8
Logistic Regression

Lecturer: Alisa Melikyan, amelikyan@hse.ru, PhD,
Associate Professor of the School of Software Engineering

Logistic Regression

Type of regression model	Dependent Variable	Predictors
Binary	Binary (yes, no)	• Interval or ratio scale variables • Binary variables • Categorical variables (series of dummy variables)
Multinomial	More than two values (Computer Science, Mathematics, Business, Sociology)	
Ordinal	More than two values which are ordered (acceptable, good, excellent)	

Multinomial regression

Multinomial logistic regression is used to model nominal outcome variables. The analysis breaks the outcome variable down into a series of comparisons between two categories.

If we have three outcome categories (A, B and C), then the analysis will consist of two comparisons. The form that these comparisons take depends on how we have selected a baseline category. The important parts of the analysis and output are much the same as binary logistic regression.

We will calculate the probability that the case belongs to each of the groups, that are formed based on the values of the dependent variable. We will predict that according to the model a case belongs to the group with maximum probability.

Multinomial logistic regression

Dependent Variable: recommended specialization for a student
Values of the Dependent Variable:

- Software Engineering,
- Data Science,
- Business Analytics,
- **Computer Science (baseline category)**,
- Applied Mathematics.

We will test 4 binary models:

- Software Engineering vs. Computer Science
- Data Science vs. Computer Science
- Business Analytics vs. Computer Science
- Applied Mathematics vs. Computer Science

Preliminary Data Analysis

Frequency and contingency tables could help to reveal the completeness of information representation in the dataset and identify some trends and relationships.

Interpreting Results

Gender : 1 – male, 0 – female (reference group)
Dependent variable has 3 values: Breakfast bar, Oatmeal, Cereal (baseline category).
Question: How the gender influences the breakfast choice?

(-,147) – males compared to females are less likely to go for Breakfast bar rather than Cereal.
(,053) – males compared to females are more likely to go for Oatmeal rather than Cereal.

bfast	Coef.	Std. Err.
Breakfast_Bar		
gender		
Male	-.1466864	.1711641
_cons	-.3148682	.1167096
Oatmeal		
gender		
Male	.0531098	.1571972
_cons	-.1156302	.1104479
Cereal	(base outcome)	

Positive B coefficient indicates that as the predictor increases the chance to be in the current category increases.

Results Interpretation

OR = ODDSm/ODDSf = 0,864 т.е. odds for males is about 14% lower than odds for females to get Breakfast Bar instead of Cereal.

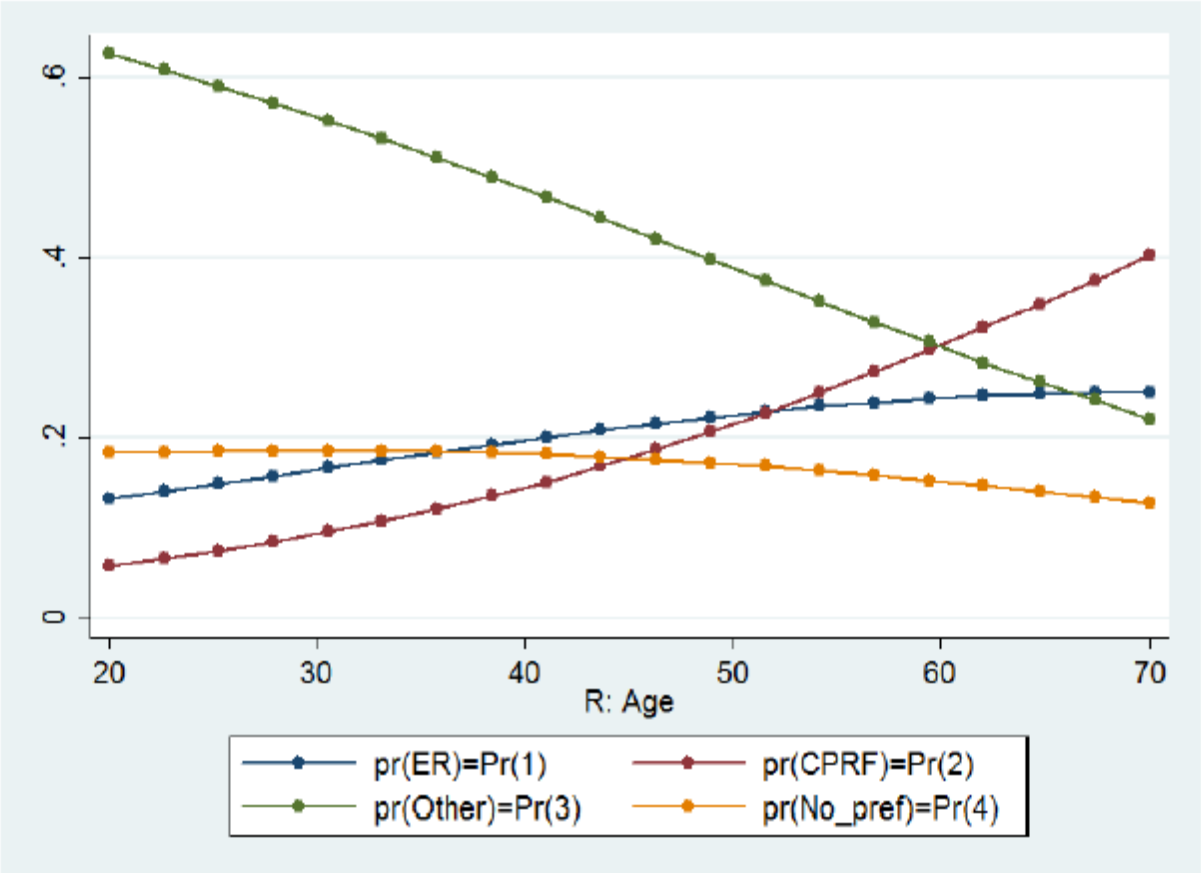
OR = ODDSm/ODDSf = 1,05 т.е. odds for males is about 5% greater than odds for females to get Breakfast Bar instead of Cereal.

bfast	OR	Std. Err.
Breakfast_Bar		
gender		
Male	.8635648	.1478112
_cons	.7298851	.0851846
Oatmeal		
gender		
Male	1.054545	.1657716
_cons	.8908046	.0983875
Cereal	(base outcome)	

Results Interpretation

Coef (B)	RRR=exp(b)	Interpretation
> 0	> 1	Males compared to females (reference group) are more likely to be in the current group than in the reference group. Exp(b)=1,6 -> odds for males is 60% higher than odds for females
= 0	= 1	Males and females are equally likely to be in the current and the reference groups.
< 0	from 0 to 1	Males compared to females (reference group) are less likely to be in the current group than in the reference group. Exp(b)=0,6 -> Odds for males is 40% lower than odds for females

Graphical representation



Ordinal regression

Ordinal regression is used to model ordinal outcome variables. We can rank the values of the ordinal variable, but the real distance between categories is unknown.

Instead of considering the probability of an individual event, we consider the probability of that event and all events that are ordered before it.

Dependent Variable

The ordinal dependent variable ‘Level of Education’ has the following values:

1 – no education

2 – secondary education

3 – higher education

4 – PhD

As a result of analysis, we will calculate the probability to achieve particular educational level and all the levels, which are below it.

Interpreting the results

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[plan = 1]	-1,589	,345	21,254	1	,000	-2,265	-,914
	[plan = 2]	-,640	,306	4,370	1	,037	-1,241	-,040
	[plan = 3]	,328	,300	1,197	1	,274	-,259	,915
	[plan = 4]	1,672	,385	18,882	1	,000	,918	2,426
Location	[g=1]	-1,106	,401	7,607	1	,006	-1,892	-,320
	[g=2]	0[a]	.	.	0	.	.	.

Predictor is gender: 1 – male, 2 – female (reference group). As the coefficient is negative (-1,106), males are less likely to get higher results than females.

Testing Parallel Lines

When you fit an ordinal regression, you assume that the relationships between the independent variables and the logits are the same for all the logits. That means that the results are a set of parallel lines or planes – one for each category of the outcome variable. You can check this assumption by allowing the coefficients to vary, estimating them, and then testing whether they are all equal.

Useful links

Multinomial Regression:

- http://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XL-8/193/2014/isprsarchives-XL-8-193-2014.pdf
- https://www.datasklr.com/logistic-regression/multinomial-logistic-regression

Ordinal Regression:

- www.norusis.com/pdf/ASPC_v13.pdf
- https://www.statsmodels.org/devel/examples/notebooks/generated/ordinal_regression.html?highlight=ordinal
- https://www.stata.com/meeting/germany08/GSUG2008.pdf

Thank you for your attention!