4 Multicategory Logistic Regression

4.1 Baseline Model for nominal response

- Response variable Y has $J > 2$ categories, $i = 1, , J$
- $\pi_1, \ldots, \pi_J$ are the probabilities that observations fall into the categories
  Question: What effect do certain predictors have on the $\pi_1, \ldots, \pi_J$?
- Simultaneously model for pairs of categories the odds of falling within one category instead of another.
- Pair a baseline category (usually last or "control" category) with all remaining categories.
- Model with one predictor x:

$$Log(\pi_j/\pi_J) = \alpha_j + \beta_j x, j = 1, ., J - 1,$$

  Note: For each category j compared, a new set of parameters (α_j, β_j) is introduced.

- Baseline category logit model allows comparison of any categories a and b

$$log(\pi_a/\pi_b) = log((\pi_a/\pi_J)/(\pi_b/\pi_J)) = log(\pi_a/\pi_J) - log(\pi_b/\pi_J) = (\alpha_a - \alpha_b) - (\beta_a - \beta_b)x$$

Example 1
Contraceptive Use in Dependency on Age
Data is in contraceptive.sav 3165 currently married women from El Salvador (1985)

Variables:	contra	(s(sterilization), o(other), n(none))
	age	(midpoints of 5 year intervals, viz: 17.5, 22.547.5)
	a2	age squared
	freq	frequency in each age midpoint/contraception level cell

To start create a contingency table, for visual acuity. Also test if age and contraception method are independent.

SPSS (PSPP) commands:

1. Data>Weight Cases
   Weight cases by Frequency Variable: freq
   OK
2. Analyze>Descriptive Statistics>Crosstabs
   Row(s): contra
   Column(s): age
   Statistics button: check chi square
   Continue
   OK
3. OK

PSPP output:

contra * age [count, row %, column %, total %].

contra	age							Total
	17.50	22.50	27.50	32.50	37.50	42.50	47.50	
n	232.00	400.00	301.00	203.00	188.00	164.00	183.00	1671.00
	13.88%	23.94%	18.01%	12.15%	11.25%	9.81%	10.95%	100.00%
	78.38%	64.83%	46.45%	37.11%	43.22%	48.52%	64.44%	52.80%
	7.33%	12.64%	9.51%	6.41%	5.94%	5.18%	5.78%	52.80%
o	61.00	137.00	131.00	76.00	50.00	24.00	10.00	489.00
	12.47%	28.02%	26.79%	15.54%	10.22%	4.91%	2.04%	100.00%
	20.61%	22.20%	20.22%	13.89%	11.49%	7.10%	3.52%	15.45%
	1.93%	4.33%	4.14%	2.40%	1.58%	.76%	.32%	15.45%
s	3.00	80.00	216.00	268.00	197.00	150.00	91.00	1005.00
	.30%	7.96%	21.49%	26.67%	19.60%	14.93%	9.05%	100.00%
	1.01%	12.97%	33.33%	48.99%	45.29%	44.38%	32.04%	31.75%
	.09%	2.53%	6.82%	8.47%	6.22%	4.74%	2.88%	31.75%
Total	296.00	617.00	648.00	547.00	435.00	338.00	284.00	3165.00
	9.35%	19.49%	20.47%	17.28%	13.74%	10.68%	8.97%	100.00%
	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	9.35%	19.49%	20.47%	17.28%	13.74%	10.68%	8.97%	100.00%

Chi-square tests.

Statistic	Value	df	Asymp. Sig. (2-tailed)
Pearson Chi-Square	430.03	12.00	.000
Likelihood Ratio	521.10	12.00	.000
N of Valid Cases	3165.00		

According to the output at significance level of 5% we can reject that the choice of contraceptive is independent from age ($\chi^2(12) = 430.03, p < 0.001$, LR$\chi^2(12) = 521.10, p < 0.001$).

To determine the nature of the relationship between the two variables a baseline logit model can be fit with "no contraceptive" as the baseline category.
It is therefore of interest to graph age versus $log(\pi_s/\pi_n)$ and age versus $log(\pi_o/\pi_n)$. These log values can be obtained for the age midpoints by calculating for each age category

$$Log(\hat{\pi}_s/\hat{\pi}_n) = Log(s/n) \quad \text{and} \quad Log(\hat{\pi}_o/\hat{\pi}_n) = Log(o/n)$$

SPSS commands:

1. Transform>Compute Variable
   Target Variable lsn
   Numeric Expression: LN(s/n) OK

2. Transform>Compute Variable
   Target Variable:lon
   Numeric Expression: LN(o/n) OK

3. Graph>Legacy Dialogs>Scatter/Dot>Overlay Scatter
   Define
   X-Y pairs

Pair	Y Variable	X Variable
1	lsn	age
2	lon	age

4. OK

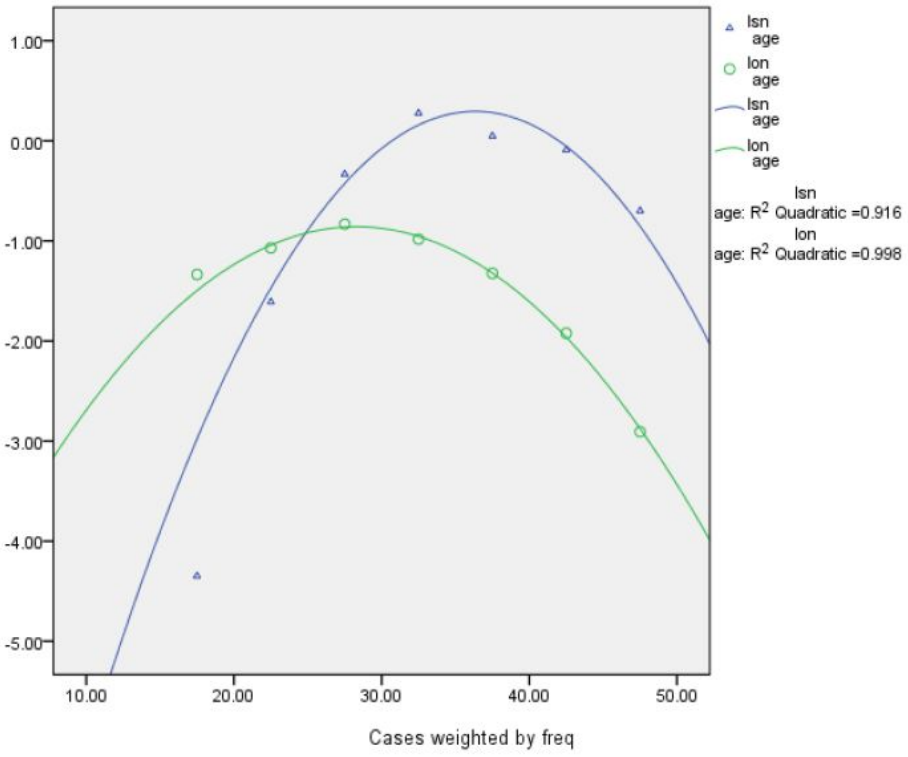
5. OK

6. Double click to activate graph in output

7. Click on lsn on legend, and change type to a triangle in the popup
   apply
   close

8. Right click on an lsn triangle (or a lon circle)
   select "add fit line" at total from the dropdown menu
   select quadratic in the popup
   apply

9. Close the activated graph window to return it to the output file.



From the graph one can conclude that a quadratic model on age seems to fit the log odds well.
The model:

$$Log(\pi_o/\pi_n) = \alpha_o + \beta_{o,1}age + \beta_{o,2}age^2$$
$$Log(\pi_s/\pi_n) = \alpha_s + \beta_{s,1}age + \beta_{s,2}age^2$$

Test if age has a significant effect on the odds for using one contraceptive method over another. Also test if the quadratic model provides a significant better fit than the linear model:
SPSS commands for quadratic model:

1. Analyze>Regression>Multinomial Logistic
   Dependent: contra
   Reference Category: first *(choosing n as baseline category)*
   Covariate(s):
   age
   a2

2. Model button:
   Choose Custom
   Build Terms: Main effects
   Select age and a2 from Factors & Covariates window and bring them over to Forced Entry Terms window.
   Continue

3. Statistics button:
   leave defaults
   Also check Goodness of Fit under model.

4. Save Window:
   select estimated response probabilities

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	605.377	617.497	601.377			
Final	112.749	149.108	100.749	500.628	4	.000

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	18.869	8	.016
Deviance	20.475	8	.009

Pseudo R-Square

Cox and Snell	.146
Nagelkerke	.170
McFadden	.080

From this table we obtain the information for testing if age has a significant effect on the odds for using one contraceptive over another. Testing the null hypothesis that all four slopes are zero, can be tested by comparing the quadratic model with the intercept only model. The LR-test statistic $\chi^2(4) = 500.628$, p-value< 0.001, indicates that the data provide sufficient evidence at significance level of 5% that at least one of the slopes is different from 0, therefore age has an effect on the use of contraceptive in El Salvador at the time of data collection.

The table "LR tests" from the SPSS output gives us the information we need to compare the two models:

Likelihood Ratio Tests

Effect	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC of Reduced Model	BIC of Reduced Model	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	492.264	516.504	484.264	383.515	2	.000
age	419.670	443.910	411.670	310.922	2	.000
a2	382.252	406.491	374.252	273.503	2	.000

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

According to the subscription of the table, to test if both slopes for age^2 equal zero in the model the test statistics gives $\chi^2(2) = 273.503$ with a p-value< 0.001, therefore reject at significance level of 5% that the slopes are both 0 and support the quadratic over the linear model.
Also note that AIC and BIC support the quadratic model, which is not surprising given the graph above.
The tables on Model fit indicate that the model fit for the quadratic model is acceptable, but the pseudo R^2 values indicate that predictions from these model will not be very accurate.

Use the table below to obtain the model equations to estimate odds and probabilities within the different age groups

Parameter Estimates

contra[a]		B	Std. Error	Wald	df	Sig.	Exp(B)
o	Intercept	-4.550	.694	42.998	1	.000	
	age	.264	.047	31.472	1	.000	1.302
	a2	-.005	.001	39.233	1	.000	.995
s	Intercept	-12.618	.757	277.545	1	.000	
	age	.710	.046	243.225	1	.000	2.033
	a2	-.010	.001	218.250	1	.000	.990

Instead one can have SPSS calculate the probabilities, by checking "estimate response probabilities" after clicking the Save button.
In the data spreadsheet three extra columns will be added for example giving for age= 22.5

EST1_1	EST2_2	EST3_3
.64	.23	.13

telling that at age 22.5 $\hat{\pi}_n = .64$, $\hat{\pi}_0 = .23$, and $\hat{\pi}_s = .13$.

4.2 Ordinal Logistic Regression

If the response is ordinal it is meaningful to take this into account by modelling an ordinal logistic regression.

- Y is an ordinal variable with categories $1, 2, \ldots, J$
- Consider cumulative probability for category $j = P(Y \leq J) = \pi_1 + \cdots + \pi_j, j = 1, 2, \ldots, J$
- It is $P(Y \leq 1) \leq P(Y \leq 2) \leq \cdots \leq P(Y \leq J) = 1$

- Model the cumulative log odds
  Logit $(P(Y \leq j)) = log(\frac{P(Y \leq j)}{1-P(Y \leq j)})$
- For one predictor variable x, **proportional odds model** is , for $j = 1, 2, \ldots, J-1$, Logit $(P(Y \leq j)) = \alpha_j + \beta_x$
  NOTE: the slope β is the same for all cumulative logits
- β describes the effect of x on the odds of falling into categories 1 to j versus higher categories. The model assumes that this effect is the same for all cumulative odds.
- NOTE: Property of this model is that, for $\beta > 0$, as x increases, the probability of falling in a lower category increases.
  NOTE: SPSS models Logit$(Y \leq j) = \alpha_j - \beta_j x$ to "fix" this property.

Example 2
HSB Data:
Use SES (low-1, middle-2, high-3) as response
Predictors:
Math
Writing
Sex (1-Male, 2-Female)
Race (1-Hispanic, 2-Asian, 3-Black, 4-White)

The two model equations fir by SPSS will be for j = 1, 2.

$$Logit(P(Y \leq j) = \alpha_j - \beta_1 math - \beta_2 writing - \beta_3 sex - \beta_{41} race_1 - \beta_{42} race_2 - \beta_{43} race_3$$

We first check to be sure we have no ses/sex cells or ses/race cells with very few observations, as this would cause problems in analyzing our model.
SPSS commands:

1. Analyze>Descriptive Statistics > Crosstabs
   Row(s): ses
   Column(s): sex
   Ok
2. Analyze>Descriptive Statistics>Crosstabs
   Row(s): ses
   Column(s): race
   Ok

ses * sex Crosstabulation

Count

		sex		Total
		1.00	2.00	
ses	1.00	50	89	139
	2.00	144	155	299
	3.00	79	83	162
Total		273	327	600

ses * race Crosstabulation

Count

		race				Total
		1.00	2.00	3.00	4.00	
ses	1.00	23	8	24	84	139
	2.00	34	14	24	227	299
	3.00	14	12	10	126	162
Total		71	34	58	437	600

The smallest cell count is 8, the model can be fit.

SPSS commands to fit the proportional cumulative odds model

1. Analyze>Regression>Ordinal
   Dependent: ses
   Factor(s):
   sex
   race
   Covariate(s):
   writing
   math

2. Output button:
   Check Test of parallel lines
   Check Predicted category
   Continue
   Ok

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1221.075			
Final	1148.593	72.482	6	.000

Link function: Logit.

At significance level of 5% the considered model fits significantly better than the intercept only model, with $\chi^2(6) = 74.482, p < 0.001$.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	1148.222	1138	.410
Deviance	1125.438	1138	.599

Link function: Logit.

Pseudo R-Square

Cox and Snell	.114
Nagelkerke	.130
McFadden	.058

Link function: Logit.

deviance/df= 0.989 indicating acceptable model fit and the pseudo R^2 all indicate that the predictors are relevant but will not result in reliable predictions.

Before interpreting the model parameters one more important test should be conducted. As mentioned above it is assumed that the slope parameters are the same for modelling all cumulative log odds. In the analysis it is important to confirm that this is an acceptable assumption.

Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	1148.593			
General	1142.142	6.450	6	.375

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.
a. Link function: Logit.

At significance level of 5% the data do not provide sufficient evidence that the slopes are not all the same for the two cumulative log odds. The model can be deemed appropriate.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[ses = 1.00]	2.509	.555	20.422	1	.000	1.421	3.597
	[ses = 2.00]	4.927	.587	70.497	1	.000	3.777	6.077
Location	math	.043	.012	14.121	1	.000	.021	.066
	writing	.027	.011	5.833	1	.016	.005	.049
	[sex=1.00]	.456	.170	7.210	1	.007	.123	.788
	[sex=2.00]	0[a]	.	.	0	.	.	.
	[race=1.00]	-.143	.255	.314	1	.575	-.644	.357
	[race=2.00]	-.151	.345	.193	1	.660	-.827	.524
	[race=3.00]	-.476	.279	2.910	1	.088	-1.024	.071
	[race=4.00]	0[a]	.	.	0	.	.	.

Link function: Logit.
a. This parameter is set to zero because it is redundant.

To interpret the sign of the slopes it is important to remember that SPSS inverts the sign of the slopes such that a positive slope indicates a positive relationship between the predictor and the probabilities to fall into higher categories.
Therefore: The higher a student scores on the writing and math exam the more likely they fall into higher SES categories correcting for sex and race. Since $e^{.043} = 1.044$, a one point increase in math increases a students odds of being in a higher SES category by 4.4% and with $e^{.027} = 1.027$, a one point increase in math increases your chance of being in a higher SES category by 2.7%, all other variables in the model held constant.
Since sex=1 indicates male students the positive sign for the sex=1 dummy indicates that male students were more likely to come from higher SES categories than female students after correcting for math and writing scores and race.
All race dummies were not significant, therefore the data do not indicate a significant effect of race on the probabilities of belonging into higher or lower SES categories. The negative slopes would have told that Asian, Hispanic, and Black students have lower probabilities for falling into higher SES categories than white students after correcting for sex and math and writing score.