1 Multicategory Logit Models

In this section we will introduce how the binary logistic regression model can be generalized to allow modeling a response variable with more than two categories.
We will distinguish between models with nominal and ordinal response variables.

1.1 Nominal Response

Let J be the number of response categories for variable Y and $\pi_1, \ldots, \pi_J$ be the probabilities for a randomly chosen individual to fall into categories $1, \ldots, J$, respectively. Then $\sum \pi_i = 1$.
Then the number of observations out of n independent observations falling into the different categories have a multinomial distribution. (review first section).

When modeling a nominal response variable we are interested in finding if certain predictors have an effect on the probabilities $\pi_1, \ldots, \pi_J$.
In a binary logistic regression model we model the odds for success, i.e. the probability of success in relationship to the probability for failure. In multicategory logit models we model simultaneously all relationships between probabilities for pairs of categories. This is done by modeling the odds of falling within one category instead of another.

1.1.1 Baseline logits

In a similar way to which a multicategory predictor with a GLM uses dummy variables to compare the first $J-1$ categories with the last, the baseline category, multicategory logit models pair a baseline category with all remaining categories.
Assume the last category (J) is the baseline category, then the baseline logits are

$$log(\pi_i/\pi_J), \quad i = 1, 2, \ldots, J-1$$

The baseline category logit model with one predictor x is then

$$log(\pi_i/\pi_J) = \alpha_i + \beta_i x, \quad i = 1, 2, \ldots, J-1$$

(for only two categories this is the binary logistic regression model.)
Observe that for each category i compared with the baseline category, a new set of parameters is introduced.
The baseline category logit model permits the comparison of any two categories since, for categories a and b

$$log(\pi_a/\pi_b) = log\left(\frac{\pi_a/\pi_J}{\pi_b/\pi_J}\right) = log(\pi_a/\pi_J) - log(\pi_b/\pi_J) = (\alpha_a - \alpha_b) + (\beta_a - \beta_b)x$$

Software like SPSS fits these $J-1$ model equations simultaneously, which results in smaller standard errors for the parameter estimates than when fitting them separately.
The choice of baseline category has no effect on the parameter estimates for comparing two categories a and b.

All the other formerly discussed concepts hold for this model.

Example 1

Contraceptive use in dependency on age (from G. Rodriguez, 2007, online notes)

The data has been taken of the report on the Demographic and Health Survey conducted in El Salvador in 1985. The table shows 3165 currently married women classified by age, grouped in five-year intervals, and current use of contraception, classified as sterilization, other methods, and no method.

age * contra Crosstabulation

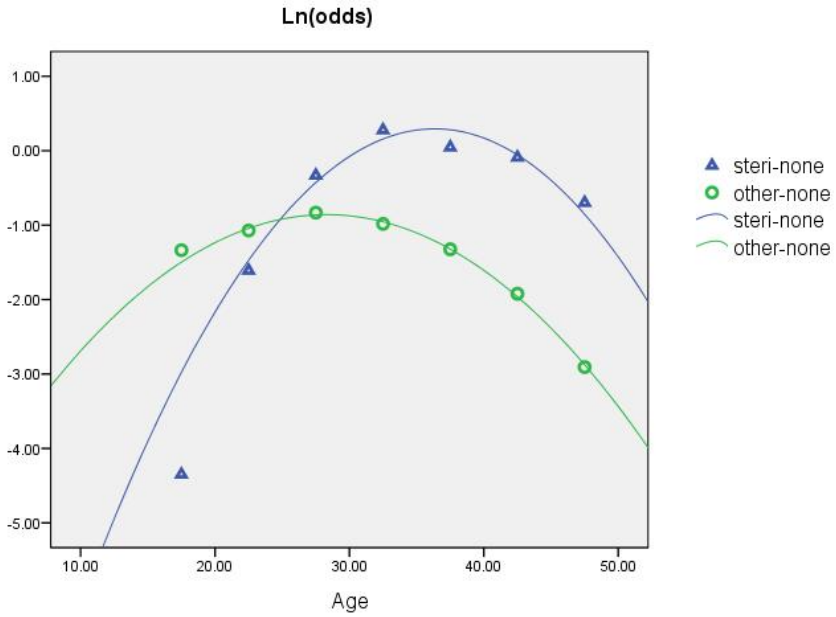
Count

		contra			Total
		n	o	s	
age	17.50	232	61	3	296
	22.50	400	137	80	617
	27.50	301	131	216	648
	32.50	203	76	268	547
	37.50	188	50	197	435
	42.50	164	24	150	338
	47.50	183	10	91	284
Total		1671	489	1005	3165

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	430.028[a]	12	.000
Likelihood Ratio	521.103	12	.000
N of Valid Cases	3165		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 43.88.



The graph indicates that there is a quadratic (non linear) relationship between the log odds (for sterilization versus none and other versus none) and age.

First we will fit a linear model for the logits and age, a multicategory logistic regression model with predictor age. The model equations are:

$$\begin{aligned} log(\pi_s/\pi_n) &= \alpha_s + \beta_s age, \\ log(\pi_o/\pi_n) &= \alpha_o + \beta_o age, \end{aligned}$$

Second we will fit a quadratic model for the logits and age, a multicategory logistic regression model with predictor age and age^2. The model equations are:

$$\begin{aligned} log(\pi_s/\pi_n) &= \alpha_s + \beta_{1s} age + \beta_{2s} age^2, \\ log(\pi_o/\pi_n) &= \alpha_o + \beta_{1o} age + \beta_{2o} age^2, \end{aligned}$$

The SPSS output can be found in file "CONTRACEPTIVE.PDF"
As expected the linear model shows an unacceptable fit with Pearson $\chi^2 = 268.17, df = 10$ and Deviance $= 293.98, df = 10$. Test for an effect are significant, but are meaningless because of the bad fit by the model.

The quadratic model shows a marginally acceptable model fit with Pearson $\chi^2 = 18.869, df = 8$ and Deviance $= 20.475, df = 8$. But the analysis of the standardized residuals shows an acceptable fit for almost all cells and supports this model.
The model can be improved by including age^3 with the model, but this makes the model less parsimonious.
According to the Wald χ^2 test both age and age^2 have a significant effect on the log odds for sterilization versus none, other versus none, and sterilization versus other.

odds-ratio	variable	χ^2	*df*	*P*
sterilization - none	age	243.22	1	< 0.001
	age^2	218.25	1	< 0.001
other - none	age	31.47	1	< 0.001
	age^2	39.23	1	< 0.001
sterilization - other	age	54.79	1	< 0.001
	age^2	28.24	1	< 0.001

For getting the comparison between sterilization I reran the analysis with baseline category "other".
The estimated model equations are:

$$\begin{aligned} log(\pi_s/\pi_n) &= -12.6 + .710\ age - 0.010\ age^2, \\ log(\pi_o/\pi_n) &= -4.5 + .264\ age - 0.005\ age^2, \end{aligned}$$

These are the curves shown in the graph above.
Then we can also find the following estimated model equation for sterilization versus other:

$$\begin{aligned} log(\pi_s/\pi_o) &= (-12.6 - (-4.5)) + (.710 - .264)\ age + (-0.010 - (-0.005))\ age^2, \\ &= -8.1 + .446\ age - 0.005\ age^2, \end{aligned}$$

Interpretation: Since this a quadratic model the interpretation of the slopes is not as easy, because the effect of an extra year in age depends on the age.

For example:
Given that the chosen contraceptive is either sterilization or none the odds that a woman used sterilization at 26 is $e^{\beta_{1s}(26-25)+\beta_{2s}(26^2-25^2)}$ times the odds of a woman at age 25.

Because: Given that the chosen contraceptive is either sterilization or none

$$\begin{aligned} odds(25) &= e^{\alpha_s+\beta_{1s}25+\beta_{2s}25^2} \\ odds(26) &= e^{\alpha_s+\beta_{1s}26+\beta_{2s}26^2} \end{aligned}$$

Thus

$$\begin{aligned} odds(26)/odds(25) &= e^{\alpha_s+\beta_{1s}26+\beta_{2s}26^2-(\alpha_s+\beta_{1s}25+\beta_{2s}25^2)} \\ &= e^{\beta_{1s}(26-25)+\beta_{2s}(26^2-25^2)} \\ &= e^{\beta_{1s}+\beta_{2s}(51)} \end{aligned}$$

$e^{\beta_{1s}(26-25)+\beta_{2s}(26^2-25^2)} = 1.22$, 22% increase from 25 to 26.

To test if a predictor x has an effect on the probabilities to fall with in the different categories, a test based on the comparison of two models is applied with the χ^2 statistics being the difference between the log likelihood for the model without the predictor and the log likelihood for the model including the predictor.

1. Hypotheses: $H_0 : x$ has no effect on $\pi_1 \dots \pi_J$ versus $H_a : H_0$ is not true
2. Assumptions:
3. Test statistic: $\chi^2 = -2(L_0 - L_1)$, $df = J - 1$

Example 2
For the example this looks a little different because the model is quadratic. To test if age has an effect on the choice of contraceptive we need to test:

1. $H_0 : \beta_{1s} = \beta_{2s} = \beta_{1o} = \beta_{2o} = 0$ versus H_a : at least one is different. $\alpha = 0.05$
2.
3. $\chi^2_0 = -2(L_0 - L_1) = 500.628$, $df = 4$ (page 5 from the output).
4. P-value< 0.001
5. Reject H_0, at significance level of 5% the data provide sufficient evidence that the age has an effect on the choice of contraceptive.

1.1.2 Estimating response probabilities

The multicategory logistic regression model can also be rewritten to state the probabilities for the J categories:

$$\pi_j = \frac{e^{\alpha_j+\beta_j x}}{\sum e^{\alpha_h+\beta_h x}}, \qquad j = 1, \dots, J$$

The denominator is always the same and when adding the probabilities we get $\sum_j \pi_j = 1$.

Let b be the baseline category. We set the parameters α_b and β_b for the baseline category to 0. We get:

$$\text{odds ratio}(b/b) = e^{\alpha_b + \beta_b x} = e^0 = 1$$

the only value that makes sense.
These equations can then be used to estimate the probabilities for the J categories in dependency on the predictors, by replacing the parameters with their estimates.

Example 3

$$\hat{\pi}_s = \frac{e^{-12.6+.710\ age-0.010\ age^2}}{e^{-12.6+.710\ age-0.010\ age^2} + e^{-4.5+.264\ age+-0.005\ age^2} + 1}$$

$$\hat{\pi}_o = \frac{e^{-4.5+.264\ age+-0.005\ age^2}}{e^{-12.6+.710\ age-0.010\ age^2} + e^{-4.5+.264\ age+-0.005\ age^2} + 1}$$

and

$$\hat{\pi}_n = \frac{1}{e^{-12.6+.710\ age-0.010\ age^2} + e^{-4.5+.264\ age+-0.005\ age^2} + 1}$$

Using these equations for age = 22.5:

$$\hat{\pi}_s = .13,\ \hat{\pi}_o = .23,\ \hat{\pi}_n = .64$$

To Based on the contingency tables we would estimate
for s: 80/617=0.129, o: 137/617=0.222, and n: 400/617=0.648.
This model has 6 parameters, the contingency table has $(3-1)\times(7-1) = 12$ degrees of freedom. Therefore the model needs a much smaller number of parameters to almost perfectly replicate the estimated probabilities.

1.2 Ordinal Response

If the multicategory response is ordinal a model should be used which reflects the order of the categories. A model taking the ordinal nature of the response variable into account should be easier to interpret and tests have greater power.
Definition:
Assume that Y is an ordinal variable with categories $1, 2, \ldots, J$, then the cumulative probability for category j is the probability to fall at most into category j

$$P(Y \le j) = \pi_1 + \cdots + \pi_j, \quad j = 1, 2, \ldots, J$$

It is

$$P(Y \le 1) \le P(Y \le 2) \le \cdots \le P(Y \le J) = 1$$

To model an ordinal response variable one models the cumulative response probabilities or cumulative odds. In the model cumulative odds for the last category do not have to be modeled since the cumulative probability for the highest category is always one (no category falls above).

Definition:
The cumulative logits are the logits for the cumulative probabilities

$$logit(P(Y \le j)) = log\left(\frac{P(Y \le j)}{1 - P(Y \le j)}\right) = log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right)$$

1.2.1 Cumulative Logit Models

A model for cumulative logit j is equivalent to a binary logistic regression model for combined categories 1 to j (I) versus the combined category $j+1$ to J (II)

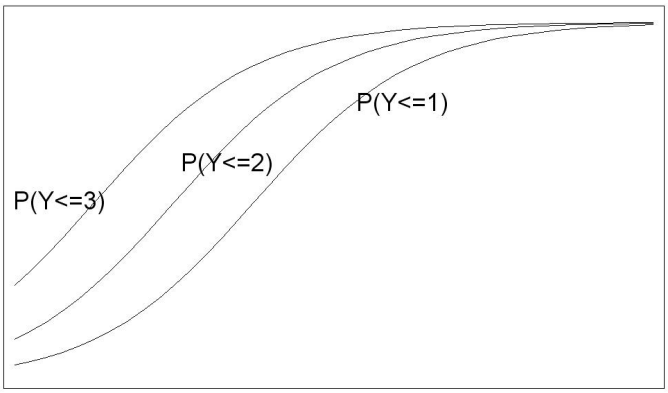
$$\text{categories: } \underbrace{1 \;\; 2 \;\; \ldots \;\; j}_{I} \;\; \underbrace{j+1 \;\; \ldots \;\; J}_{II}$$

For one predictor variable x the proportional odds model becomes for $j = 1, 2 \ldots, J-1$:

$$logit(P(Y \leq j)) = \alpha_j + \beta x$$

The slope β is the same for all cumulative logits, and therefore this model has a single slope parameter instead of $J-1$ in the multicategory logistic regression model.
The parameter β describes the effect of x on the odds for falling into categories 1 to j. The effect is assumed to be the same for all cumulative odds.

The cumulative probabilities in dependency on predictor variable x if $\beta > 0$.



Using that

$$P(Y = j) = P(Y \leq j) - P(Y \leq j-1)$$

We can obtain from the curves shown above the probabilities to fall within category j in dependency on x.
DIAGRAM!!(pg.181 in the text book)
If $\beta > 0$ then the probability to fall into a lower category increases with increasing x.
This is against the usual interpretation, positive slope is associated with a positive correlation.
For this reason SPSS models

$$logit(P(Y \leq j)) = \alpha_j - \beta x$$

and reports the negative of the slope. Be very careful!!

Example 4
High school and beyond

The ordinal response we are considering is the SES. We include gender, math, writing, and race as predictors.
The model equation for $j = 1, 2$:

$$logit(P(Y \leq j)) = \alpha_j + \beta_1 \text{math} + \beta_2 \text{writing} + \beta_3 \text{gender} + \beta_{41} \text{race}_1 + \beta_{42} \text{race}_2 + \beta_{43} \text{race}_3$$

First check if there are no empty cell when cross tabulating the predictors with the response.

ses * sex Crosstabulation

Count

		sex		
		1.00	2.00	Total
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				
		1.00	2.00	3.00	4.00	Total
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

Based on these contingency tables we do not have to be concerned because of empty or "small" cells.
Running the Ordinal Logistic Regression provides the following output:

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.

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.

From the first table we can see that the model including the predictors fits significantly better than the model omitting all predictors ($\chi^2 = 72.482, df = 6, p < 0.001$).

From the second table we find that the Pearson $\chi^2 = 1148.222, df = 1138$ and deviance= $1125.438, df = 1138$ both indicating good fit.

The third table report pseudo R^2 values. Using Nagelkerke=0.13, we get the impression that this model fits well, but we should not rely on it for predictions.

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.

The threshold values are usually not interpreted, they represent the cutpoints in the linear predictor for the different categories when the predictors are all 0. For the example the value of 2.509 is to differentiate low SES values from middle and high.

From the estimates table we learn that the results on the math test and writing test are significant predictors for the SES of a student (math: $\chi^2 = 14.121, df = 1, p < 0.001$, writing: $\chi^2 = 5.833, df = 1, p < 0.016$).
The interpretation of the parameter estimates give: A one point increase in the math test results in an increase in the ordered log-odds of being in a higher ses category by 0.043 while the other variables in the model are held constant. $e^{0.043} = 1.044$, therefore a one point increase in the math test increase your chance being in a higher SES category by 4.4%.

We also learn that race has no significant effect, but gender does on the odds being of a higher SES.

The results in this table hold if it is reasonable to assume that the slope for the cumulative log odds are all the same, i.e. the S-curves are truly parallel which means that changes for falling into a higher SES is equally affected by the predictor variables for all categories.
If this is a reasonable assumption can be tested:

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.

The test is based on the likelihood ratio statistic for comparing the loglikelihoods for the model with equal effect and the more general model without this requirement.
Here the null hypothesis that the lines are parallel and the effect is the same for all categories can not be rejected ($\chi^{=}6.450, df = 6, p = .375$) which supports the model above model, but remember that we do not know the error probability for this decision.

ses * Predicted Response Category Crosstabulation

Count

		Predicted Response Category			
		1.00	2.00	3.00	Total
ses	1.00	18	117	4	139
	2.00	10	259	30	299
	3.00	7	131	24	162
Total		35	507	58	600

The prediction table would be perfect if all measurement would fall onto the diagonal of the prediction table. This table supports what has been said before, that this model is not very powerful for predicting the SES of an individual.