Multiple Logistic Regression

- So far we have looked at the case where our dependent variable is binary, i.e. it has just two categories. However there are many nominal variables with more than two categories.

- We therefore need methods that can model this type of dependent variable as well. Luckily we can extend the logistic regression model to do this. We call this extended method ***multinomial logistic regression***, and refer to logistic regression for dichotomous dependent variables as ***binary logistic regression***.

- The basic principle of multinomial logistic regression is similar to that for binomial logistic regression, in that it is based on the probability of membership of each category of the dependent variable.

- **(Important)** The way multinomial logistic regression deals with the variables in this case is somewhat similar to the concept of dummy variables, in that it compares the probability of being in each of n-1 categories compared to a baseline or reference category.

- **(Important)** In a way we can say that we are fitting n-1 separate binary logistic models, where we compare category 1 to the baseline category, then category 2 to the baseline and so on.

- In practice software algorithms allow us to model the comparisons to the baseline simultaneously using maximum likelihood estimation, which is better as doing it sequentially could lead to misestimation of the standard errors.

- Therefore, multinomial logistic regression is basically an extension of binary logistic regression for nominal variables with more than two categories.

Examples of Multinomial Logistic Regression

Example 1. People's occupational choices might be influenced by their parents' occupations and their own education level. We can study the relationship of one's occupation choice with education level and parent's occupation. The occupational choices will be the outcome variable which consists of categories of occupations.

Example 2. A biologist may be interested in food choices that alligators make. Adult alligators might have difference preference than young ones. The outcome variable here will be the types of food, and the predictor variables might be the length of the alligators and other environmental variables.

Example 3. Entering high school students make program choices among general program, vocational program and academic program. Their choice might be modeled using their writing score and their social economic status.

British Voting Preferences

- A researcher wanted to understand whether the political party that a person votes for can be predicted from a belief in whether tax is too high and a person's income (i.e., salary).
- Therefore, the political party the participants last voted for was recorded in the politics variable and had three options: "Conservatives", "Labour" and "Liberal Democrats". When presented with the statement, "tax is too high in this country", participants had four options of how to respond: "Strongly Disagree", "Disagree", "Agree" or "Strongly Agree" and stored in the variable, `tax_too_high`.
- The researcher also asked participants their annual income which was recorded in the income variable. As such, in variable terms, a multinomial logistic regression was run to predict politics from `tax_too_high` and income.
- For the British political system, we are taking a stereotypical approach to the three major political parties, whereby the Liberal Democrats and Labour are parties in favour of high taxes and the Conservatives are a party favouring lower taxes.

Implementation

We created three variables:

(1) the independent variable, `tax_too_high`, which has four ordered categories: "Strongly Disagree", "Disagree", "Agree" and "Strongly Agree";

(2) the independent variable, `income`; and

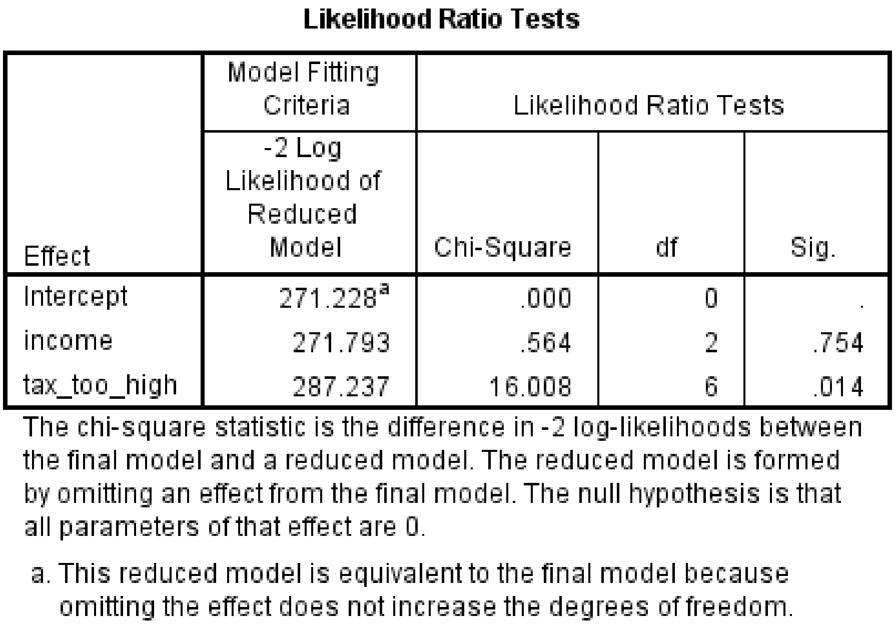
(3) the dependent variable, politics, which has three categories: "Con", "Lab" and "Lib" (i.e., to reflect the Conservatives, Labour and Liberal Democrats).

Factors and Covariate

- With SPSS You need to separate the variables into covariates and factors. For these particular procedures, SPSS Statistics classifies continuous independent variables as covariates and nominal independent variables as factors.

- Therefore, the continuous independent variable, income, is considered a covariate. However, where you have an ordinal independent variable, such as in our example (i.e., `tax_too_high`), you must choose whether to consider this as a covariate or a factor.

- This table shows which of your independent variables are statistically significant. You can see that income (the "income" row) was not statistically significant because p = 0.754 (the "Sig." column). On the other hand, the `tax_too_high` variable (the "`tax_too_high`" row) was statistically significant because $p = 0.014$. There is not usually any interest in the model intercept (i.e., the "Intercept" row).



Likelihood Ratio Tests

	Model Fitting Criteria	Likelihood Ratio Tests		
Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	271.228[a]	.000	0	.
income	271.793	.564	2	.754
tax_too_high	287.237	16.008	6	.014

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.

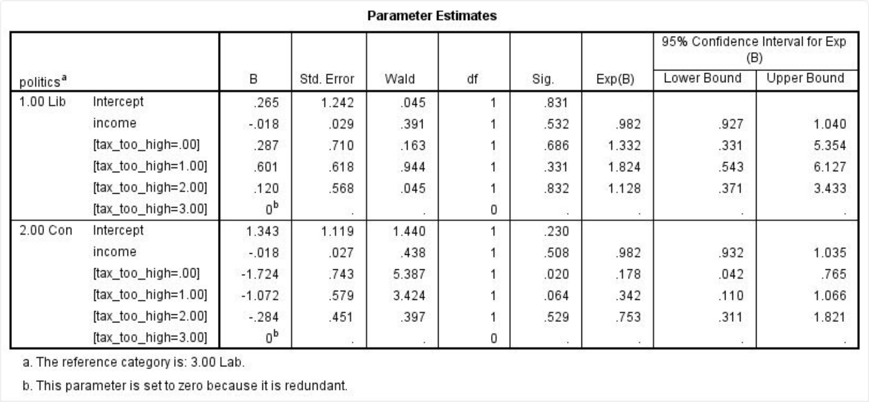
a. This reduced model is equivalent to the final model because omitting the effect does not increase the degrees of freedom.

Figure 1:

This table is mostly useful for nominal independent variables because it is the only table that considers the overall effect of a nominal variable, unlike the Parameter Estimates table, as shown below:

- This table presents the parameter estimates (also known as the coefficients of the model). As you can see, each dummy variable has a coefficient for the `tax_too_high` variable. However, there is no overall statistical significance value. This was presented in the previous table (i.e., the Likelihood Ratio Tests table).

- As there were three categories of the dependent variable, you can see that there are two sets of logistic regression coefficients.



Parameter Estimates

								95% Confidence Interval for Exp (B)	
politics^a		B	Std. Error	Wald	df	Sig.	Exp(B)	Lower Bound	Upper Bound
1.00 Lib	Intercept	.265	1.242	.045	1	.831			
	income	-.018	.029	.391	1	.532	.982	.927	1.040
	[tax_too_high=.00]	.287	.710	.163	1	.686	1.332	.331	5.354
	[tax_too_high=1.00]	.601	.618	.944	1	.331	1.824	.543	6.127
	[tax_too_high=2.00]	.120	.568	.045	1	.832	1.128	.371	3.433
	[tax_too_high=3.00]	0^b	.	.	0	.	.	.	.
2.00 Con	Intercept	1.343	1.119	1.440	1	.230			
	income	-.018	.027	.438	1	.508	.982	.932	1.035
	[tax_too_high=.00]	-1.724	.743	5.387	1	.020	.178	.042	.765
	[tax_too_high=1.00]	-1.072	.579	3.424	1	.064	.342	.110	1.066
	[tax_too_high=2.00]	-.284	.451	.397	1	.529	.753	.311	1.821
	[tax_too_high=3.00]	0^b	.	.	0	.	.	.	.

a. The reference category is: 3.00 Lab.

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

- The first set of coefficients are found in the "Lib" row (representing the comparison of the Liberal Democrats category to the reference category, Labour). The second set of coefficients are found in the "Con" row (this time representing the comparison of the Conservatives category to the reference category, Labour). You can see that "income" for both sets of coefficients is not statistically significant `(p = 0.532 and p = 0.508, respectively; the "Sig." column)`.
- The only coefficient (the "B" column) that is statistically significant is for the second set of coefficients. It is `[tax_too_high=.00] (p = 0.020)`, which is a dummy variable representing the comparison between "Strongly Disagree" and "Strongly Agree" to tax being too high.
- The sign is negative, indicating that if you "strongly agree" compared to "strongly disagree" that tax is too high, you are more likely to be Conservative than Labour. However, because the coefficient does not have a simple interpretation, the exponentiated values of the coefficients (the "`Exp(B)`" column) are normally considered instead.

Summary

Multinomial Logistic Regression is useful for situations in which you want to be able to classify subjects based on values of a set of predictor variables. This type of regression is similar to logistic regression, but it is more general because the dependent variable is not restricted to two categories.