Multinomial Logistic Regression

1) Introduction

Multinomial logistic regression (often just called 'multinomial regression') is used to predict a nominal dependent variable given one or more independent variables. It is sometimes considered an extension of binomial logistic regression to allow for a dependent variable with more than two categories. As with other types of regression, multinomial logistic regression can have nominal and/or continuous independent variables and can have interactions between independent variables to predict the dependent variable.

2) Presentation of the Data and Research Question

The data were collected on 200 high school students and are scores on various tests, including a video game and a puzzle. The outcome measure in this analysis is the student's favourite flavor of ice cream - vanilla, chocolate or strawberry - from which we are going to see what relationships exists with video game scores (**video**), puzzle scores (**puzzle**) and gender (**female**).

3) Assumptions

- **Assumption 1:** Your **dependent variable** should be measured at the **nominal** level with more than or equal to **three** values. Examples of **nominal variables** include ethnicity (e.g., with three categories: Caucasian, African American and Hispanic), political party (e.g. Lib Dems, Labour, Conservatives).
- **Assumption 2:** You have **one or more independent variables** that are **continuous**, **ordinal** or **nominal** (including **dichotomous variables**). However, ordinal independent variables must be treated as being either continuous or categorical.
- **Assumption 3:** You should have **independence of observations** and the dependent variable should have **mutually exclusive and exhaustive categories** (i.e. no individual belonging to two different categories!).
- **Assumption 4:** There should be **no multicollinearity**. Multicollinearity occurs when you have two or more independent variables that are highly correlated with each other.

- **Assumption 5:** There needs to be a **linear relationship between any continuous independent variables and the logit transformation of the dependent variable**.
- **Assumption 6:** There should be **no outliers**, **high leverage values** or **highly influential points** for the scale/continuous variables.

4) Procedure on SPSS

We first select ***Analyze -> Regression -> Multinomial Logistic…***

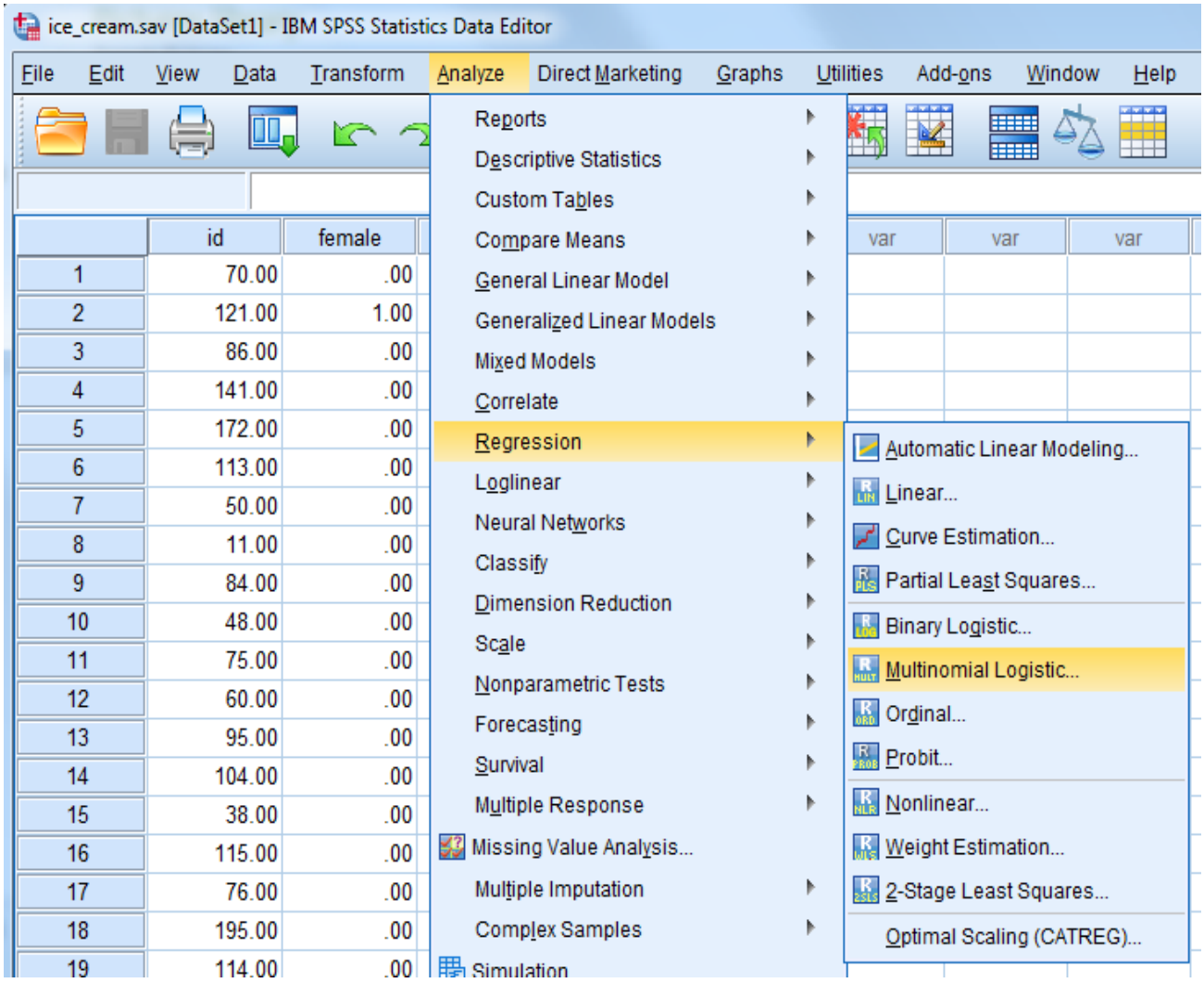


Figure 1. Selecting Multinomial Logistic Regression

We then enter the variable "ice_cream" as our dependent variable, that is, a categorical variable taking three values: "vanilla", "chocolate" and "strawberries" (see Figure 2). Transfer the categorical variable "gender" in the Factor box and then we transfer the continuous variables, i.e. the "score on video game" and the "score on puzzle", in the Covariate(s) box.

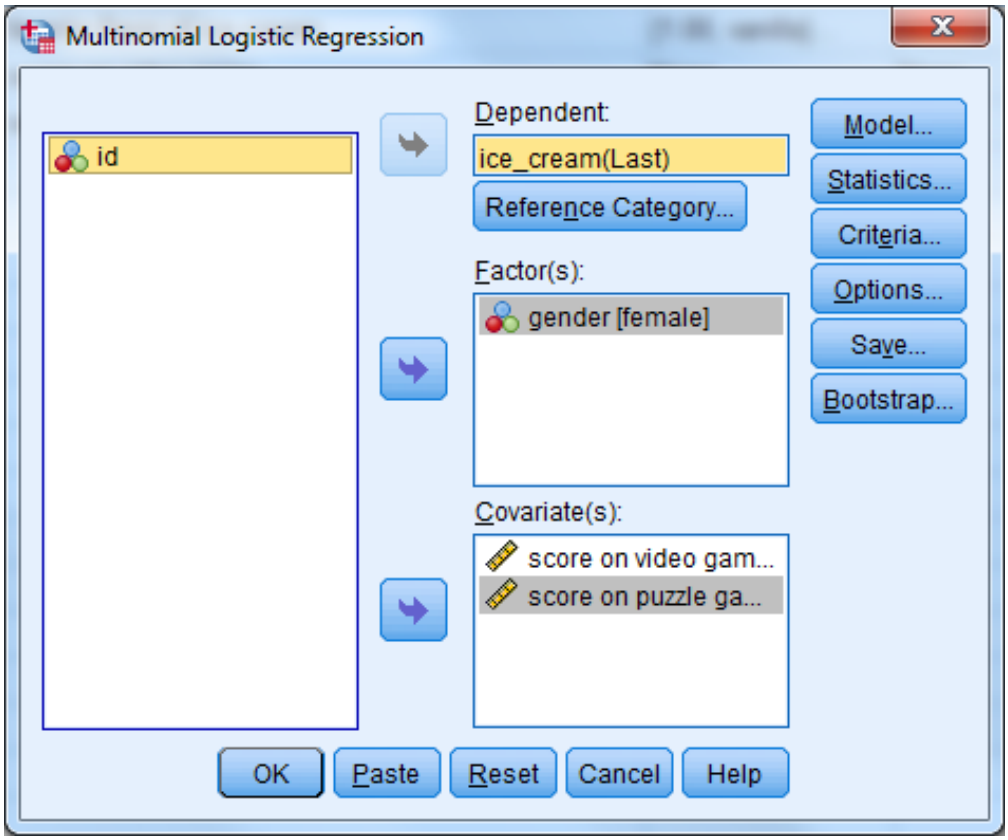
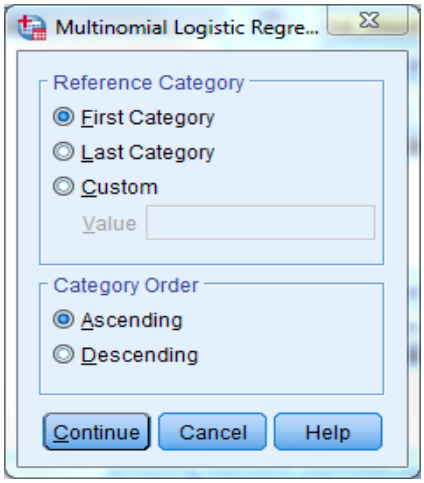


Figure 2. Setting the regression model.

In this regression model we need to specify the reference category of our dependent variable (see Figure 3). For this, click on "Reference Category…" and then select which category of "ice_cream" should be taken as the reference. If you do not specify it, the procedure will automatically choose the last category by default (i.e. "ice_cream = strawberries") as the reference category. In our case, we took "ice_cream = vanilla" as our reference category because we generally want the reference category to be the category with the highest number of people in it (see Figure 4).



"vanilla" is the first value of ice_cream, therefore we will select "First Category". We want to treat each category in the ascending order, i.e. chocolate and then strawberry, so we will choose "Ascending" in the Category Order.

Figure 3. Choosing the reference category in the dependent variable.

favorite flavor of ice cream

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	vanilla	95	47.5	47.5	47.5
	chocolate	47	23.5	23.5	71.0
	strawberry	58	29.0	29.0	100.0
	Total	200	100.0	100.0	

Figure 4. Distribution of the variable ice_cream.

We will be presented with the **Multinomial Logistic Regression: Statistics** dialogue box (Figure 5), as shown below. Tick all the options below:

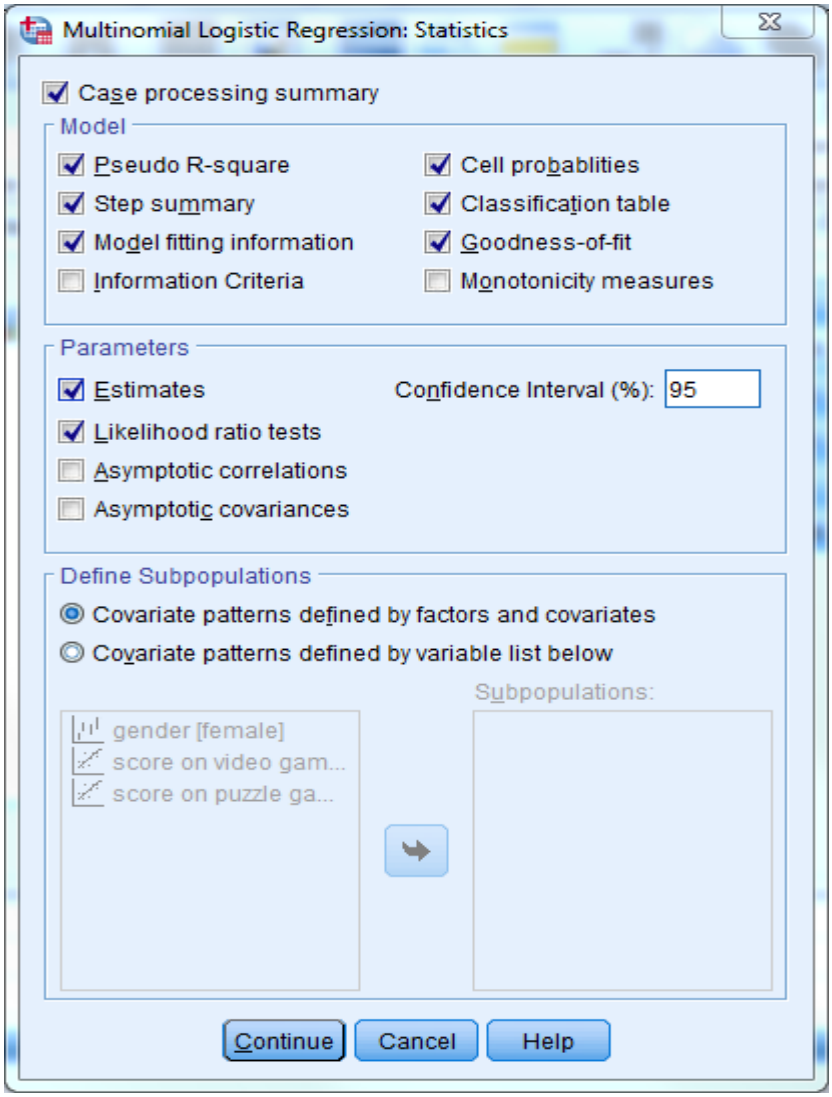


Figure 5. Statistics Dialog box.

5) **Results**

A first way to assess the goodness of fit is to consider whether the variables we added statistically significantly improve the model compared to the intercept alone (i.e., with no variables added). We can see from the "**Sig.**" column that $p = 0.000$ (meaning in fact that $p < 0.001$), which means that the full model statistically significantly predicts the dependent variable better than the intercept-only model.

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	365.736			
Final	332.641	33.095	6	.000

The **Goodness-of-Fit** table provides two measures that can be used to assess how well the model fits the data, as shown below. The first row, labelled "**Pearson**", presents the Pearson chi-square statistic. A statistically significant result (i.e., $p < 0.05$) indicates that the model does not fit the data well. You can see from the table above that the *p*-value is 0.240 (from the "**Sig.**" column) and is, therefore, not statistically significant. Based on this measure, the model fits the data well. The second statistic is the "**Deviance**" and in the same way, we consider that that model fits the data well if the test shows no significance (i.e. p-value > 0.05).

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	294.296	278	.240
Deviance	287.613	278	.333

The table below (**Likelihood Ratio Tests**) shows which of your independent variables are statistically significant. You can see that video (the "**video**" row) was not statistically significant because $p = 0.174$ (the "**Sig.**" column). On the other hand, the puzzle and female variable (the "**puzzle**" and the "**female**" row) was statistically significant because $p = 0.002$ for "**puzzle**" and p = 0.078 for "**female**". There is not usually any interest in the model intercept (i.e. the "**Intercept**" row).

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	332.641[a]	.000	0	.
video	336.139	3.498	2	.174
puzzle	345.602	12.962	2	.002
female	337.730	5.090	2	.078

The **Likelihood Ratio Tests** 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 there were three categories of the dependent variable, you can see that there are two sets of logistic regression coefficients (sometimes called two logits). The first set of coefficients is found in the "chocolate" row (representing the comparison of the Chocolate category to the reference category, Vanilla). The second set of coefficients is found in the "strawberry" row (this time representing the comparison of the strawberry category of favourite ice cream to the reference category, vanilla).

Parameter Estimates

favorite flavor of ice cream[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B): Lower Bound	Upper Bound
chocolate	Intercept	2.729	1.139	5.740	1	.017			
	video	-.024	.021	1.262	1	.261	.977	.937	1.018
	puzzle	-.039	.020	3.978	1	.046	.962	.926	.999
	[female=.00]	-.817	.391	4.362	1	.037	.442	.205	.951
	[female=1.00]	0[b]	.	.	0	.	.	.	.
strawberry	Intercept	-4.090	1.209	11.448	1	.001			
	video	.023	.021	1.206	1	.272	1.023	.982	1.066
	puzzle	.043	.020	4.675	1	.031	1.044	1.004	1.085
	[female=.00]	.033	.350	.009	1	.925	1.033	.520	2.052
	[female=1.00]	0[b]	.	.	0	.	.	.	.

a. The reference category is: vanilla.

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

As you can see, each dummy variable has a coefficient for the female variable. However, there is a statistical significance value for chocolate but not strawberry for the female variable. This would suggest that there is a statistical association between gender and preferring chocolate rather than vanilla, but that there is no association between gender and preferring strawberries rather than vanilla.

You can see that "video" for both sets of coefficients is not statistically significant (p = 0.261 and p = 0.272, respectively; the "Sig." column). On the other hand the "puzzle" for both sets of coefficients is statistically significant (p = 0.046 and p = 0.031, respectively; the "Sig." column). This was presented in the previous table (i.e., the Likelihood Ratio Tests table).

By looking at the value of the coefficients B, if a subject were to increase his **puzzle** score by one point, the multinomial log-odds of preferring chocolate to vanilla would be expected to decrease by 0.039 unit while holding all other variables in the model constant. On the other hand if a subject were to increase his **puzzle** score by one point, the multinomial log-odds for preferring strawberry to vanilla would be expected to increase by 0.043 unit while holding all other variables in the model constant.