Investigating a Model with Three Categories

In this last handout, we will continue with multinomial outcomes—in this case, investigating an outcome with three categories. It is important to note that methods designed to be used with ordinal variables cannot be used with nominal outcomes, since they do not have ordered categories (Agresti, 2007). In situations where there are more than two possible outcomes that are not ordered, or where the ordinal nature of the data is not clear (or key assumptions associated with ordinal data not supported), multinomial logistic regression is a useful procedure. The approach is similar to binary logistic regression, but is more general because the dependent variable is not restricted to two categories. Once again, because the outcome represents a probability between 0 and 1, linear regression would not be appropriate.

Similar to the binary case, the multinomial approach also makes use of the logit link function. The scores are transformed into latent continuous scores that describe the log odds of being in a particular category versus the reference group. This is an example of a multinomial distribution since the variance of the observed proportion depends only on the population proportion π_i. Because the variance is determined by the predicted value of π_i, it is not modeled as a separate term in the model. Similar to the binomial distribution is selected, the variance is set to a scale factor of 1.0, which suggests it does not need to be interpreted (Hox, 2002).

The probability of being in the *K* different categories of the dependent variable can be summarized as follows:
Prob(Y = 1) = π_{1i},
Prob(Y = 2) = π_{2i},
Prob(Y = 3) = $\pi_{3i} = 1 - (\pi_{1i} - \pi_{2i})$.

Note that there are *K*-1 probabilities required to specify the multinomial outcome (since the last can be calculated once the first two are known). The sampling model binomial, with a set of dummy variables constructed such that $Y_{ki} = 1$ if $\pi_i = k$ and $Y_{ki} = 0$ otherwise.

The link function is a multinomial logit link. For each category $k = 1,\ldots K\text{-}1$, the underlying predicted log odds can be defined as follows:

$$\eta_{ki} = \log\left(\frac{\pi_{ki}}{\pi_{Ki}}\right) = \log\left(\frac{\text{Prob}(P = k)}{\text{Prob}(P = K)}\right); \qquad (3)$$

that is, the predicted underlying outcome (η_{ki}) is the log odds of individual *i* being in category *k* relative to category *K* (i.e., defined as the reference category). Often, either the first or last category is selected as the reference category.

For three nominal categories, therefore, there will be two equations for η_{1i} and η_{2i} :

$$\eta_{1i} = \beta_{0(1)} + \sum_{q=1}^{Q_1} \beta_{q(1)} X_{qi}$$

$$\eta_{2i} = \beta_{0(2)} + \sum_{q=1}^{Q_1} \beta_{q(2)} X_{qi}$$

.

Discriminant Analysis

Let's examine the discriminant analysis results for classifying individuals into clerk (1), custodian (2), and manager (3) job categories. We have 474 individuals[1]. Note that if you are using this for data set for assignment, you can continue on by adding gender and minority for the second model after you work through this first model.

For this example, we will use three variables likely to be related: (education, beginning salary, and experience). We will calculate the prior probabilities from the data (category 1 = 0.776, category 2 = 0.057, category 3 = 0.177). In the following table, we can see that the two functions are significant, and when the first is removed the second is still significant. This suggests both functions are working well to classify individuals into job categories.

Table 1. Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	.274	607.924	6	.000
2	.765	125.629	2	.000

The standardized coefficients suggest that beginning salary dominates the classification on the first function (which primarily separates managers from the other two categories), which previous experience dominates in separating category 2 (custodians) from category 1 (clerks).

[1] Download the data set (*Logistic3CategoryData.zip*) from the class web page. Instructions for replicating the corresponding models (tables 1-5) are provided at the end of this handout.

Table 2. Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
Educational Level (years)	.273	-.317
Beginning Salary	.871	.312
Previous Experience (months)	-.224	.837

The classification rate is 87.3%, which suggests that the functions do quite well in classifying individuals based on their beginning salary, education, and experience.

Table 3. Classification Results[b,c]

			Predicted Group Membership			
		Employment Category	Clerical	Custodial	Manager	Total
Original	Count	Clerical	342	19	2	363
		Custodial	14	13	0	27
		Manager	25	0	59	84
	%	Clerical	94.2	5.2	.6	100.0
		Custodial	51.9	48.1	.0	100.0
		Manager	29.8	.0	70.2	100.0
Cross-validated[a]	Count	Clerical	342	19	2	363
		Custodial	14	13	0	27
		Manager	25	0	59	84
	%	Clerical	94.2	5.2	.6	100.0
		Custodial	51.9	48.1	.0	100.0
		Manager	29.8	.0	70.2	100.0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
b. 87.3% of original grouped cases correctly classified.
c. 87.3% of cross-validated grouped cases correctly classified.

Multinomial Logistic Regression

We can now estimate the same model using multinomial logistic regression. Table 4 displays the two sets of coefficients comparing custodians to the reference group (clerks) and managers to the reference group (clerks). Comparing custodians to clerks, we can see that custodians make significantly higher salaries than clerks (since the log odds is positive). Note that the odds ratio for beginning salary is not very revealing, since it represents the increased odds of being a

custodian versus clerk for a one-dollar increase. Custodians also have significantly higher months of experience (odds ratio = 1.012) but lower educational levels (odds ratio = 0.703).

Table 4. Parameter Estimates

Employment Category[a]		B	Std. Error	Wald	Df	Sig.	Exp(B)	95% Confidence Interval for Exp(B) Lower Bound	Upper Bound
Manager	Intercept	-18.450	3.868	22.754	1	.000			
	salbegin	.001	.000	39.141	1	.000	1.001	1.000	1.001
	Educ	.346	.259	1.776	1	.183	1.413	.850	2.350
	prevexp	-.013	.005	5.506	1	.019	.988	.977	.998
Custodial	Intercept	-3.983	1.559	6.528	1	.011			
	salbegin	.000	.000	6.944	1	.008	1.000	1.000	1.000
	Educ	-.352	.106	11.045	1	.001	.703	.571	.866
	prevexp	.012	.002	30.389	1	.000	1.012	1.008	1.016

a. The reference category is: Clerical.

Regarding the comparison between managers and clerks, we can see mangers have significantly higher beginning salaries and significantly lower experience levels, but years of education is not significant in predicting group membership. Note that the very low intercepts are the result of being the "mythical" individual who has no education (0), no experience (0), and no beginning salary (0). We could change the coding such that we centered the predictors on the sample averages (e.g., by using z-scores) or the lowest education in the sample (8 years).

Below we can see the classification table. We classified 91.8% correctly which was a bit better than the discriminant analysis (87.3%). The model did better at classifying clerical and managers, but not as well as the discriminant analysis in classifying custodians.

Table 5. Classification

	Predicted			
Observed	Manager	Custodial	Clerical	Percent Correct
Manager	74	0	10	88.1%
Custodial	0	9	18	33.3%
Clerical	5	6	352	97.0%
Overall Percentage	16.7%	3.2%	80.2%	91.8%

References

Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: John Wiley & Sons, Inc.

Hox, J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). NY: Routledge Academic.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). New York: Chapman and Hall.

Defining the Discriminant Analysis Model (Tables 1, 2, 3) with IBM SPSS Menu Commands

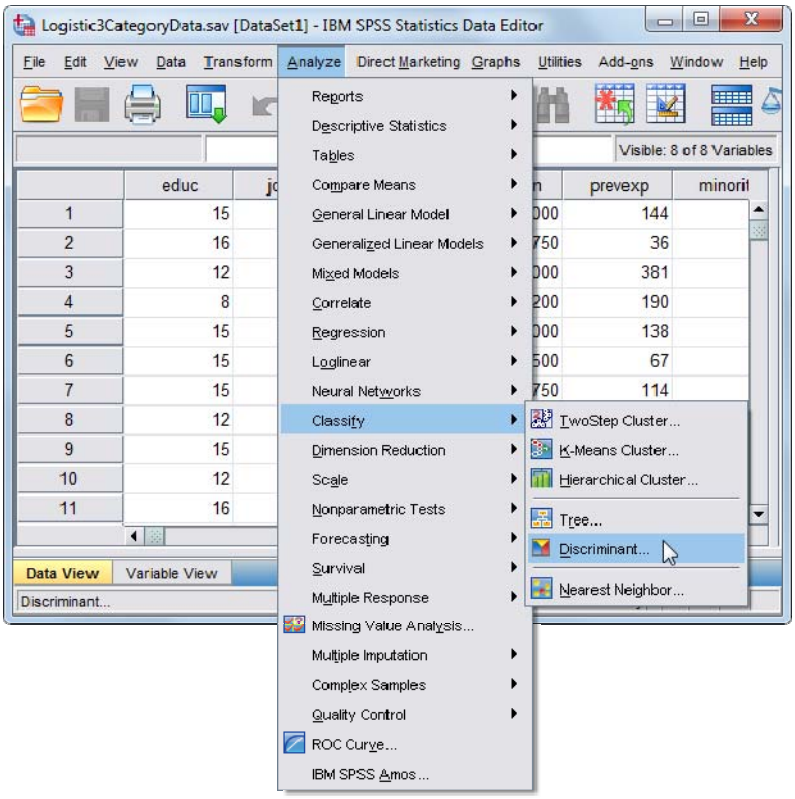
IBM SPSS syntax:

```
DISCRIMINANT
  /GROUPS=jobcat(0 3)
  /VARIABLES=educ salbegin prevexp
  /ANALYSIS ALL
  /PRIORS SIZE
  /STATISTICS=TABLE CROSSVALID
  /CLASSIFY=NONMISSING POOLED.
```

Launch the IBM SPSS application program and select the *Logistic3CategoryData.sav* data file.

1. Go to the toolbar, select ANALYZE, CLASSIFY, DISCRIMINANT.

This command opens the *Discriminant Analysis* main dialog box.

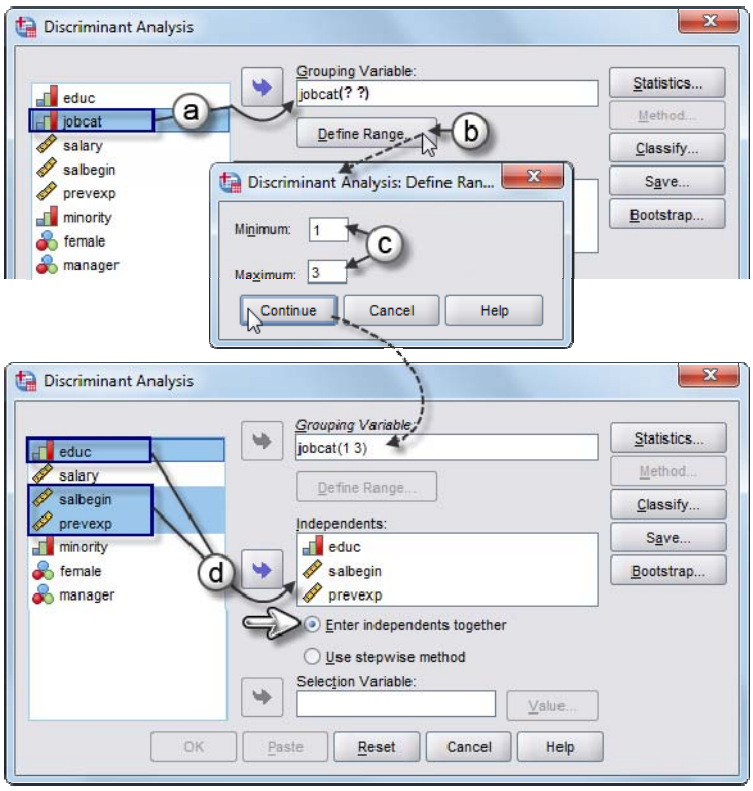


2. In the *Discriminant Analysis* main dialog box we will specific the grouping variable and independent variables in the model.

a. We will specify *jobcat* as the grouping variable. Click to select *jobcat* then click the right arrow button (or drag the variable) into the *Grouping Variable* box.

b. Now we need to define the range of the grouping variable (*jobcat*). Click DEFINE RANGE button to access the dialog box.

c. Remember that *jobcat* has 3 categories (coded 1,2,3). In the *Discriminant Analysis Define Range* dialog box, enter the minimum and maximum values of *jobcat* (1,3). Then click the CONTINUE button to close the dialog box.

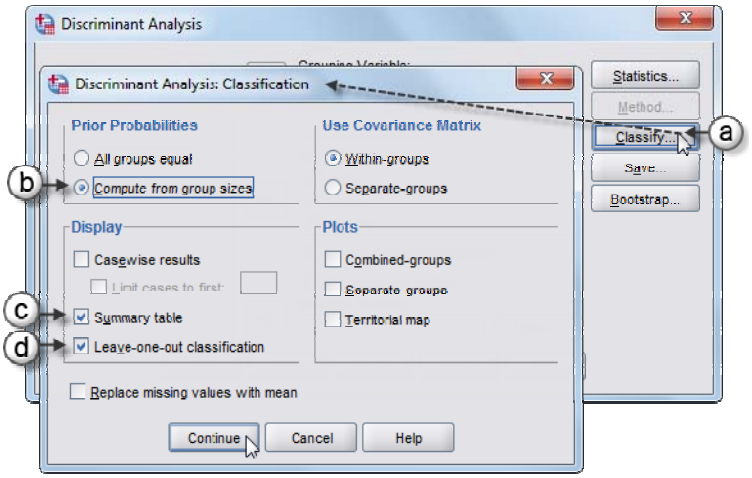


d. Now designate the predictors for the model. Click to select *educ, salbegin* and *prevexp* then click the right arrow button (or drag the variables) into the *Independents* box.

Note: The default is *Enter independents together* which we'll retain for this analysis. A stepwise analysis would require using the stepwise method instead.

Note: Although not necessary for this example if you would like to include the means in the output, click the STATISTICS button to access the dialog box and select *Means.*

3a. From the *Discriminant Analysis* main dialog box click the CLASSIFY button to access the dialog box.



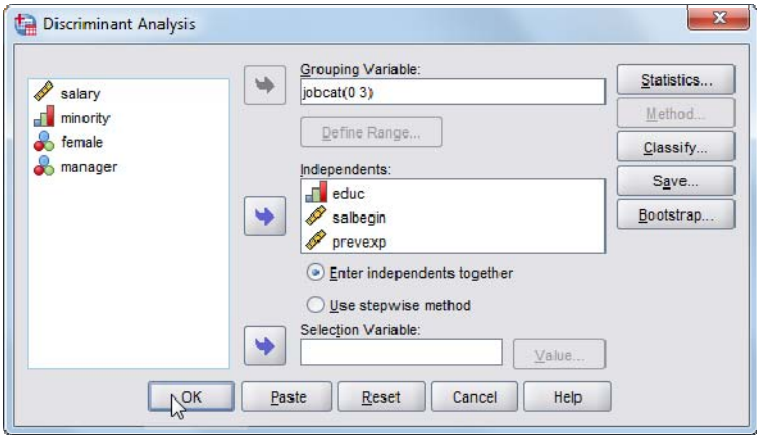
b. For this example change the default prior probabilities setting by selecting *Compute from group means*. This option takes the observed group size in the sample to determine the prior probabilities of group membership (IBM SPSS, 2011).

c. For the display we will select *Summary table* which will be included in the output.

d. We will also select *Leave-one-out classification.* This option denotes that each case in the analysis "is classified by the functions derived from all cases other than that case" (IBM SPSS, 2011).

Click the CONTINUE button to return to the *Discriminant Analysis* main dialog box.

5. From the *Discriminant Analysis* main dialog box click the OK button to generate the output results.



Defining the Multinomial Logistic Regression Analysis Model (Tables 4, 5) with IBM SPSS Menu Commands

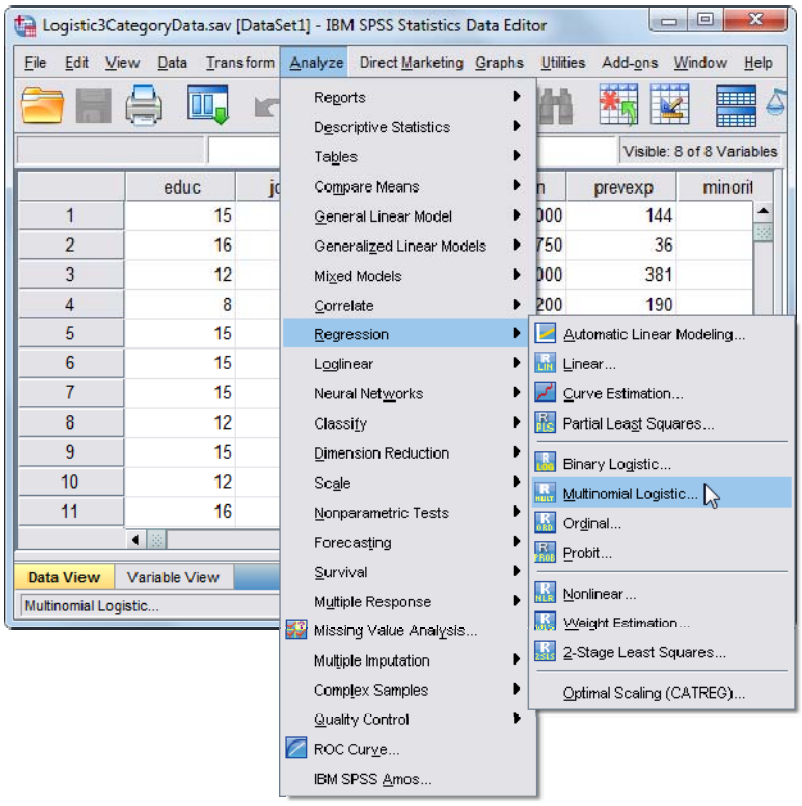
IBM SPSS syntax:

```
NOMREG jobcat (BASE=LAST ORDER=DESCENDING) WITH
salbegin educ prevexp
  /CRITERIA CIN(95) DELTA(0) MXITER(100) MXSTEP(5)
CHKSEP(20) LCONVERGE(0) PCONVERGE(0.000001)
    SINGULAR(0.00000001)
  /MODEL
  /STEPWISE=PIN(.05) POUT(0.1) MINEFFECT(0)
RULE(SINGLE) ENTRYMETHOD(LR)
REMOVALMETHOD(LR)
  /INTERCEPT=INCLUDE
  /PRINT=CLASSTABLE PARAMETER SUMMARY LRT CPS
STEP MFI.
```

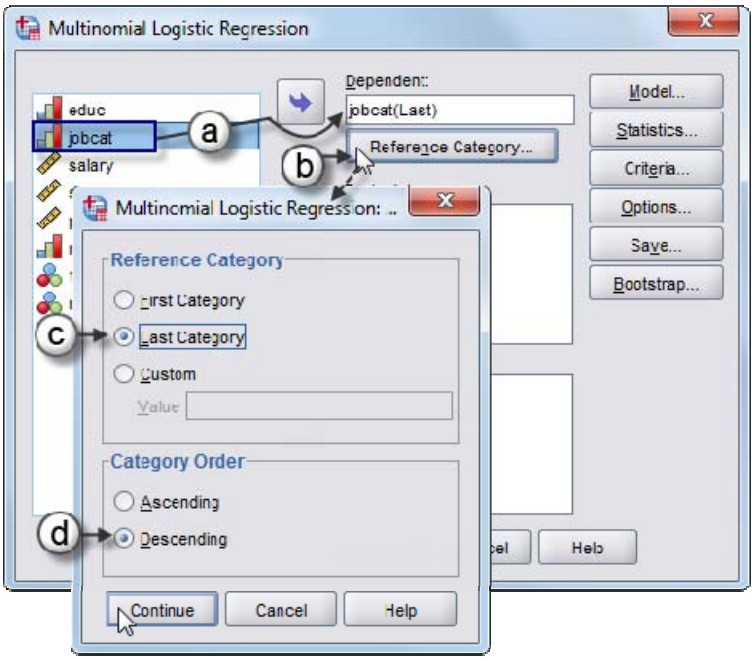
(Continue using the *Logistic3CategoryDat.sav* data file.)

1. Go to the toolbar, select REGRESSION, MULTINOMIAL LOGISTIC REGRESSION.

This command opens the *Multinomial Logistic Regression* main dialog box.



2. In the *MultinomialLogistic Regression* main dialog box we will specific the dependent variable and independent variables in the model.



a. We will specify *jobcat* as the dependent variable. Click to select *jobcat* then click the right arrow button (or drag the variable) into the *Dependent* box.

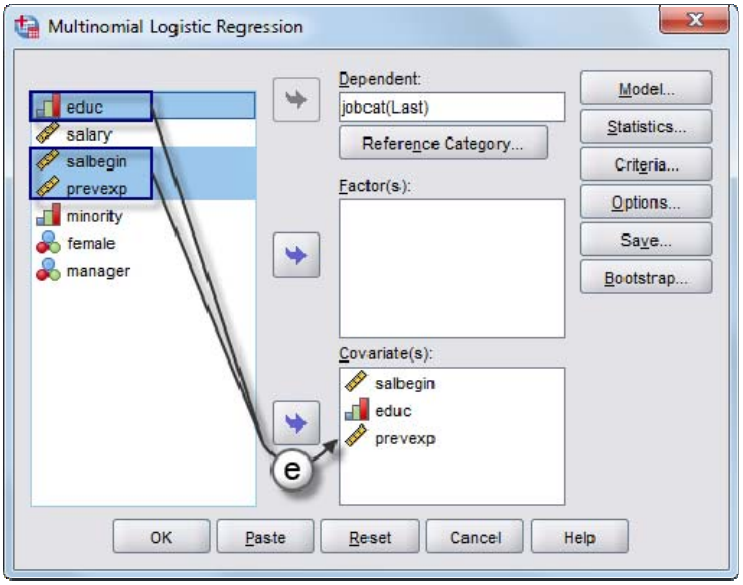
Note that "Last" appears after *jobcat.* This is because the default setting makes the last category of *jobcat* (clerical) the reference category.

b. Since *jobcat* is categorical we may designate reference categories and category order by clicking the REFERENCE CATEGORY button.

c. As noted in Step 2a, "Last Category" is the default setting and will be retained for this model.

d. We will change the category order by clicking: *Descending.* This option will use the highest value of *jobcat* to define the first category and the lowest value to define the last category.

e. Next we need to add the predictor variables to the model. Click to select *salbegin, educ,* and *prevexp* then click the right arrow button to add them in the order shown in the *Covariates* box.

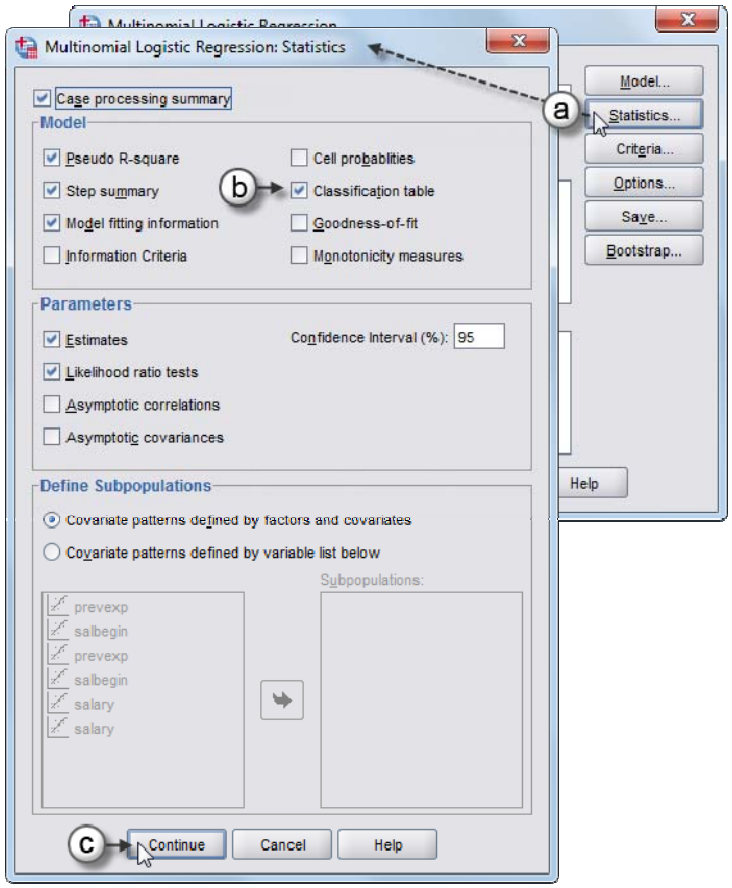


Note that *salbegin, educ,* and *prevexp* are continuous variables so were entered into the *Covariates* box. Categorical variables would be entered in the *Factor(s)* box.

3a. We will retain the default settings for the assorted modeling options but make one change for the display output. From the *Multinomial Logistic Regression* main dialog box, click the STATISTICS button to access the dialog box.

b. For the output display we will select *Summary table* to include it in the output. .

Click the CONTINUE button to return to the *Multinomial Logistic Regression* main dialog box.



4. From the *Multinomial Logistic Regression* main dialog box, click the OK button to generate the output results.