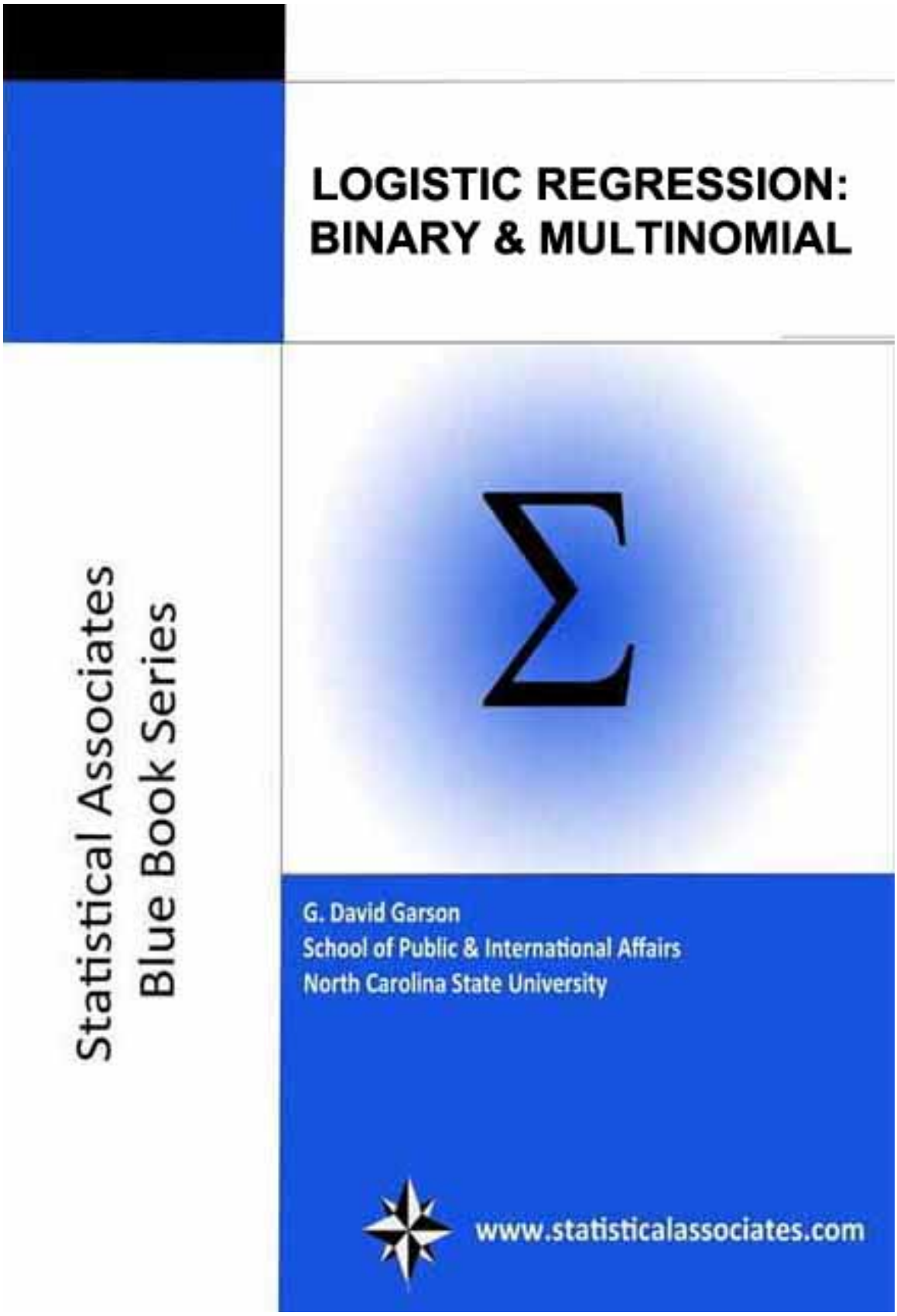


ISBN: 978-1-62638-024-0

Contact:

G. David Garson, President
Statistical Publishing Associates
274 Glenn Drive
Asheboro, NC 27205 USA

Email: sa.associates@gmail.com
Web: www.statisticalassociates.com

Table of Contents

Overview

Binary logistic regression is a form of regression which is used when the dependent variable is a true or forced dichotomy and the independent variables are of any type. Multinomial logistic regression exists to handle the case of dependent variables with more classes than two, though it is sometimes used for binary dependent variables as well since it generates somewhat different output, described below.

Logistic regression can be used to predict a categorical dependent variable on the basis of continuous and/or categorical independent variables; to determine the effect size of the independent variables on the dependent variable; to rank the relative importance of independent variables; to assess interaction effects; and to understand the impact of covariate control variables. The impact of predictor variables is usually explained in terms of odds ratios, which is the key effect size measure for logistic regression.

Logistic regression applies maximum likelihood estimation after transforming the dependent into a logit variable. A logit is the natural log of the odds of the dependent equaling a certain value or not (usually 1 in binary logistic models, or the highest value in multinomial models). Logistic regression estimates the odds of a certain event (value) occurring. This means that logistic regression calculates changes in the log odds of the dependent, not changes in the dependent itself as does OLS regression.

Logistic regression has many analogies to OLS regression: logit coefficients correspond to b coefficients in the logistic regression equation; the standardized logit coefficients correspond to beta weights; and a pseudo R^2 statistic is available to summarize the overall strength of the model. Unlike OLS regression, however, logistic regression does not assume linearity of relationship between the raw values of the independent variables and raw values of the dependent; does not require normally distributed variables; does not assume homoscedasticity; and in general has less stringent requirements.

Logistic regression does, however, require that observations be independent and that the independent variables be linearly related to the logit of the dependent. The predictive success of logistic regression can be assessed by looking at the classification table, showing correct and incorrect classifications of the

dichotomous, ordinal, or polytomous dependent variable. Goodness-of-fit tests such as the likelihood ratio test are available as indicators of model appropriateness, as is the Wald statistic to test the significance of individual independent variables.

Procedures related to logistic regression but not treated in the current volume include generalized linear models, ordinal regression, log-linear analysis, and logit regression, briefly described below.

Binary and multinomial logistic regression may be implemented in stand-alone statistical modules described in this volume or in statistical modules for generalized linear modeling (GZLM), available in most statistical packages. GZLM provides allows the researcher to create regression models with any distribution of the dependent (ex., binary, multinomial, ordinal) and any link function (ex., log for logilinear analysis, logit for binary or multinomial logistic analysis, cumulative logit for ordinal logistic analysis, and many others). Similarly, generalized linear mixed modeling (GLMM) is now available to handle multilevel logistic modeling. These topics are also treated separately in their own volumes in the Statistical Associates 'Blue Book' series.

When multiple classes of a multinomial dependent variable can be ranked by order, then ordinal logistic regression is preferred to multinomial logistic regression since ordinal regression has higher power for ordinal data. (Ordinal regression is discussed in the separate Statistical Associates' "Blue Book" volume, *Ordinal Regression.*) Also, logistic regression is not used when the dependent variable is continuous, nor when there is more than one dependent variable (compare logit regression, which allows multiple dependent variables).

Note that in its "Complex Samples" add-on module, SPSS supports "complex samples logistic regression" (CSLOGISTIC). This module is outside the scope of this volume but operates in a largely similar manner to support data drawn from complex samples and has a few capabilities not found in its ordinary LOGISTIC procedure (ex., nested terms). Likewise, all packages described in this volume have additional options and capabilities not covered in this volume, which is intended as an introductory rather than comprehensive graduate-level discussion of logistic regression.

Logit regression, treated in the Statistical Associates 'Blue Book' volume, *Loglinear Analysis*, is another option related to logistic regression. For problems with one dependent variable and where both are applicable, logit regression has numerically equivalent results to logistic regression, but with different output options. For the same class of problems, logistic regression has become more popular among social scientists. Loglinear analysis applies logistic methods to the analysis of categorical data, typically crosstabulations.

Data examples

The example datasets used in this volume are listed below in order of use, with versions for SPSS (.sav), SAS (.sas7bdat), and Stata (.dta).

The sections on binary and multinomial regression use survey data in a file called "GSS93subset". Variables are described below.

- Click here to download GSS93subset.sav for SPSS.
- Click here to download GSS93subset.sas7bdat for SAS.
- Click here to download GSS93subset.dta for Stata.

The section on ROC curves uses data in a file called "auto", dealing with characteristics of 1978 automobiles. It is supplied as a sample data file with Stata.

- Click here to download auto.sav for SPSS.
- Click here to download auto.sas7bdat for SAS.
- Click here to download auto.dta for Stata.

The section on conditional matched pairs logistic regression uses data in a file called "BreslowDaySubset", dealing with causes of endometrial cancer. See Breslow & Day (1980).

- Click here to download BreslowDaySubset.sav for SPSS.
- Click here to download BreslowDaySubset2.sav for SPSS. This contains differenced data required by the SPSS approach, as discussed in the SPSS conditional logistic regression section.
- Click here to download BreslowDaySubset.sas7bdat for SAS.
- Click here to download BreslowDaySubset.dta for Stata.

Key Terms and Concepts

Binary, binomial, and multinomial logistic regression

Though the terms "binary" and "binomial" are often used interchangeably, they are not. *Binary logistic regression*, discussed in this volume, deals with dependent variables which have two values, usually coded 0 and 1 (ex., for sex, 0 = male and 1 = female). These two values may represent a true dichotomy, as for gender, or may represent a forced dichotomy, such as high and low income. In contrast, *binomial logistic regression* is used where the dependent variable is not a binary variable per se, but rather is a count based on a binary variable.

To take a classic binomial example, subjects may be told to flip a coin 100 times, with each subject tallying the number of "heads", with the tally being the dependent variable in binomial logistic regression. More realistically, the dependent variable is apt to be a tally of successes (ex., spotting a signal amid visual noise in a psychology experiment). Binomial logistic regression is implemented in generalized linear modeling of count variables, as discussed in the separate Statistical Associates 'Blue Book' volume on “Generalized Linear Models”. For instance, to implement binomial regression in SPSS, under the menu selection Analyze > Generalized Linear Models > Generalized Linear Models, select “Binary logistic” under the Type of Model" tab, then under the Response tab, enter a binomial count variable (ex., "successes") as the dependent variable and also select the "Number of events occurring in a set of trials" radio button to enter the number of trials (ex.,, 100) or if number of trials varies by subject, the name of the number-of-trials variable.

Multinomial logistic regression extends binary logistic regression to cover categorical dependent variables with two or more levels. Multinomial logistic regression does not assume the categories are ordered (ordinal regression, another variant in the logistic procedures family, is used if they are, as discussed above). Though typically used where the dependent variable has three classes or more, researchers may use it even with binary dependent variables because output tables differ between multinomial and binary logistic regression procedures.

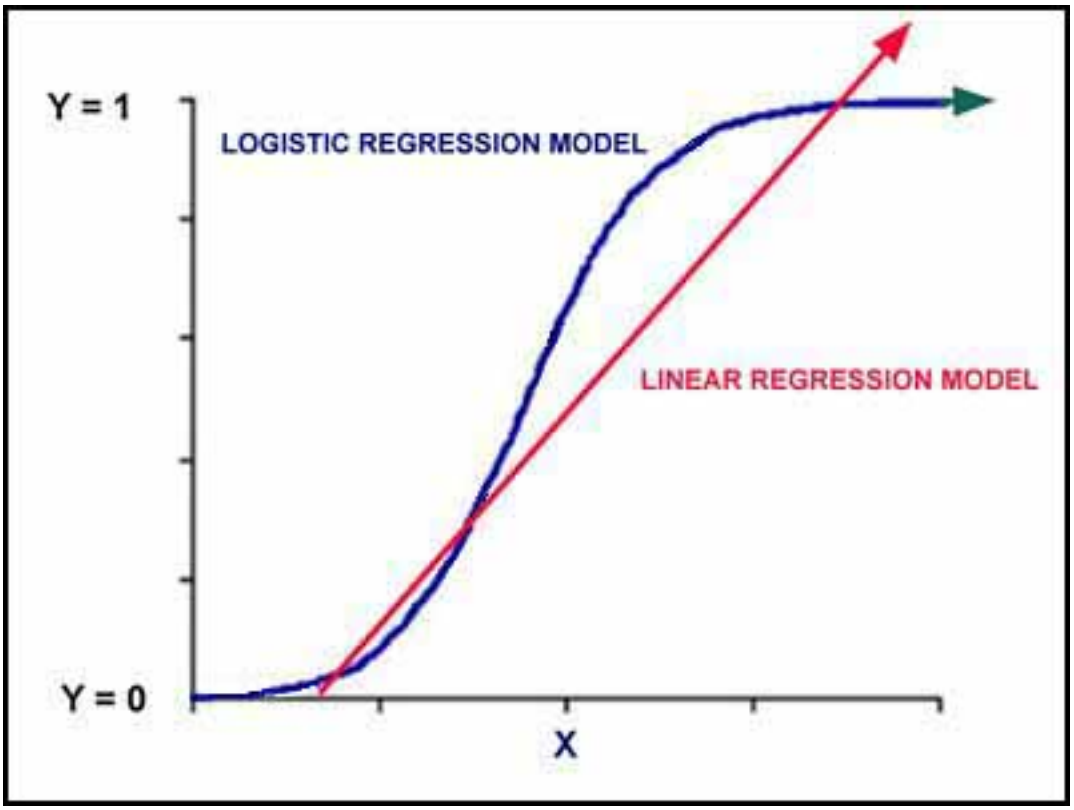
To implement binary logistic regression in SPSS, select Analyze > Regression > Binary Logistic. PROC LOGISTIC is used in SAS to implement binary logistic

regression, as described below. In STATA, binary logistic regression is implemented with the `logistic` or `logit` commands, as described below.

To implement multinomial logistic regression in SPSS, select Analyze > Regression > Multinomial Logistic. PROC CATMOD is used in SAS to implement multinomial logistic regression, as described below. In STATA, multinomial logistic regression is implemented with the `mlogit` command, as described below.

The logistic model

The logistic curve, illustrated below, is better for modeling binary dependent variables coded 0 or 1 because it comes closer to hugging the y=0 and y=1 points on the y axis, as illustrated below. Even more, the logistic function is bounded by 0 and 1, whereas the OLS regression function may predict values above 1 and below 0. Logistic analysis can be extended to multinomial dependents by modeling a series of binary comparisons: the lowest value of the dependent compared to a reference category (by default the highest category), the next-lowest value compared to the reference category, and so on, creating k - 1 binary model equations for the k values of the multinomial dependent variable. Ordinal regression also decomposes into a series of binary logistic comparisons.



The logistic equation

Logistic regression centers on the following terms:

- *Odds:* An odds is a ratio formed by the probability that an event occurs divided by the probability that the event does not occur. In binary logistic regression, the odds is usually the probability of getting a "1" divided by the probability of getting a "0". That is, in binary logistic regression, "1" is predicted and "0" is usually the reference category. In multinomial logistic regression, any lower value may be predicted and the highest-coded value is usually the reference category.
- *Odds ratio:* An odds ratio is the ratio of two odds, such as the ratio of the odds for men and the odds for women. Odds ratios are the main effect size measure for logistic regression, reflecting in this case what difference gender makes as a predictor of some dependent variable. An odds ratio of 1.0 (which is 1:1 odds) indicates the variable has no effect. The further from 1.0 in either direction, the greater the effect.
- *Log odds:* The log odds is the coefficient predicted by logistic regression and is called the "logit". It is the natural log of the odds of the dependent

variable equaling some value (ex., 1 rather than 0 in binary logistic regression). The log odds thus equals the natural log of the probability of the event occurring divided by the probability of the event not occurring:

ln(odds(event)) = ln(prob(event)/prob(nonevent))

- *Logit*: The "logit function" is the function used in logistic regression to transform the dependent variable prior to attempting to predict it. Specifically, the logit function in logistic regression is the log odds, explained above. The "logit" is the predicted value of the dependent variable. "Logit coefficients" are the b coefficients in the logistic equation used to arrive at the predicted value. Some texts label these logistic b coefficients as "logits," but this terminology is not recommended in this volume.
- *Parameter estimates*: These are the logistic (logit or b) regression coefficients for the independent variables and the constant in a logistic regression equation, much like the b coefficients in OLS regression. Synonyms for parameter estimates are unstandardized logistic regression coefficients, logit coefficients, log odds-ratios, and effect coefficients. Parameter estimates are on the right-hand side of the logistic regression equation and logits are on the left-hand side.

The logistic regression equation itself is:

$$z = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k$$

- Where z is the log odds of the dependent variable = ln(odds(event)). The "z" is the "logit", also called the log odds.
- The b_0 term is the constant or intercept term in the equation. It reflects the log odds (logit estimate) of the dependent variable when model predictors are evaluated at zero. As often 0 will be outside the range of predictor variables. Intercepts will be more interpretable if predictors are centered around their means prior to analysis because then the intercepts are the log odds of the dependent when predictors are at their mean values. In binary logistic regression, there is one intercept estimate. In multinomial logistic regression, there are (k - 1) intercepts, where k is the number of categories of the dependent variable and 1 is subtracted for the reference value.

- There are k independent (X) variables, some of which may be interaction terms.
- The "b" terms are the logistic regression coefficients, also called parameter estimates. Some textbooks call these “logits” but that usage is disparaged in this volume.
- Exp(b) = the odds ratio for an independent variable = the natural log base e raised to the power of b. The odds ratio of an independent variable is the factor by which the independent variable increases or (if negative) decreases the log odds of the dependent variable The term “odds ratio” usually refers to odds ratios for independent variables. See further discussion of interpreting b parameters below.

To convert the log odds (which is z, which is the logit) back into an odds ratio, the natural logarithmic base e is raised to the zth power: odds(event) = exp(z) = the odds ratio for the dependent variable. Exp(z) is thus the estimate of the odds(event). For binary logistic regression, it usually is the estimate for the odds that the dependent = 1. For multinomial logistic regression, it usually is the estimate for the odds that the dependent equals the a given value rather than the highest-coded value.

Put another way, logistic regression predicts the log odds of the dependent event. The "event" is a particular value of y, the dependent variable. By default the event is y = 1 for binary dependent variables coded 0,1, and the reference category is 0. For multinomial values event is y equals the value of interest and the reference category is usually the highest value of y. Note this means that for a binary dependent variable coded (0, 1), the reference category is 0 in binary logistic regression but 1 when run in multinomial logistic regression. That is, multinomial regression flips the reference value from lowest to highest.

Beware that what value is predicted and what value is the default reference category in logistic regression may vary by software package and by whether binary or multinomial regression is requested. This is discussed in the section which follows.

Logits and link functions

Overview

As discussed above, logits are the log odds of the event occurring (usually, that the dependent = 1 rather than 0). Parameter estimates (b coefficients) associated with explanatory variables are estimators of the change in the logit caused by a unit change in the independent. In SPSS output, the parameter estimates appear in the "B" column of the "Variables in the Equation" table. Logits do not appear but must be estimated using the logistic regression equation above, inserting appropriate values for the constant and X variable(s). The b coefficients vary between plus and minus infinity, with 0 indicating the given explanatory variable does not affect the logit (that is, makes no difference in the probability of the dependent value equaling the value of the event, usually 1); positive or negative b coefficients indicate the explanatory variable increases or decreases the logit of the dependent. Exp(b) is the odds ratio for the explanatory variable, discussed below.

Link functions

Parameter estimates on the right-hand side of the logistic regression formula are related to the logit values being estimated on the left-hand side by way of a link function.

(1) OLS regression uses an identity link function, meaning the predicted dependent variable is a direct function of the values of the independent variables.
(2) Binary logistic regression uses the logit link to create a logistic model whose distribution hugs the 0 and 1 values of the Y axis for a binary dependent and, moreover, does not extrapolate out of range (below 0 or above 1).
(3) Multinomial logistic regression also uses the logit link, setting up a series of binary comparisons between the given level of the dependent and the reference level (by default, the highest coded level).
(4) Ordinal logistic regression uses a cumulative logit link, setting up a series of binary comparisons between the given level or lower compared to higher levels. (Technically, the SPSS Analyze, Regression, Ordinal menu choice runs the PLUM procedure, which uses a logit link, but the logit link is used to fit a cumulative logit model. The SPSS Analyze, Generalized Linear Models, Generalized linear Models menu choice for ordinal regression assumes an ordered multinomial distribution with a cumulative logit link).

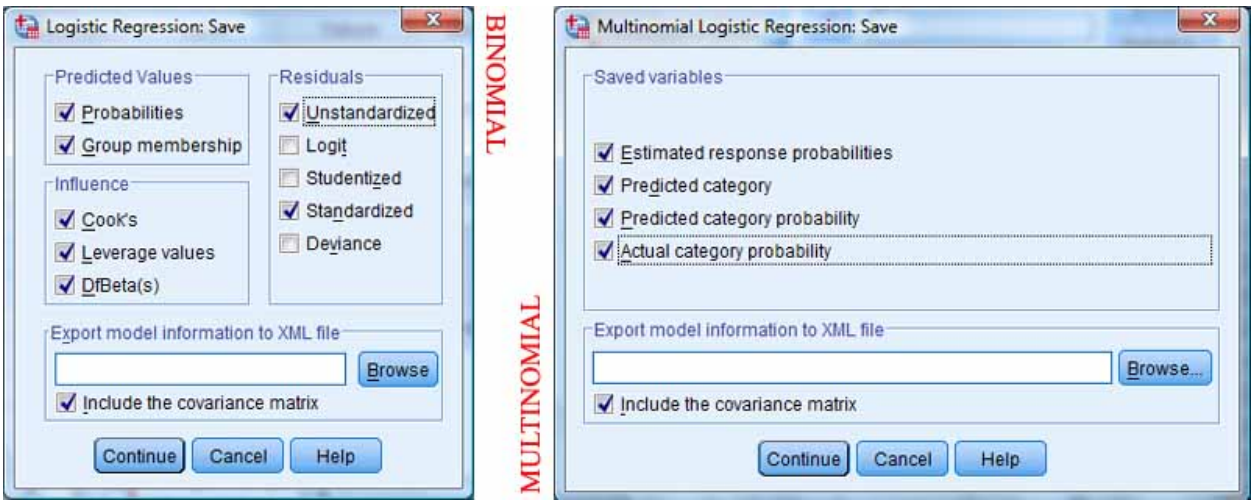
Logit coefficients, and why they are preferred over odds ratios in modeling

Note that for the case of decrease the odds ratio can vary only from 0 to .999, while for the case of increase it can vary from just over 1.0 to infinity. This asymmetry is a drawback to using the odds ratio as a measure of strength of relationship. Odds ratios are preferred for interpretation, but logit coefficients are preferred in the actual mathematics of logistic models. The odds ratio is a different way of presenting the same information as the unstandardized logit (effect) coefficient. Odds ratios were defined above and are discussed at greater length in sections below.

Saving predicted probabilities

Predicted probability values and other coefficients may be saved as additional columns added to the dataset

(1) SPSS: Use the "Save" button dialogs of binary and multinomial logistic regression dialogs in SPSS, illustrated below.



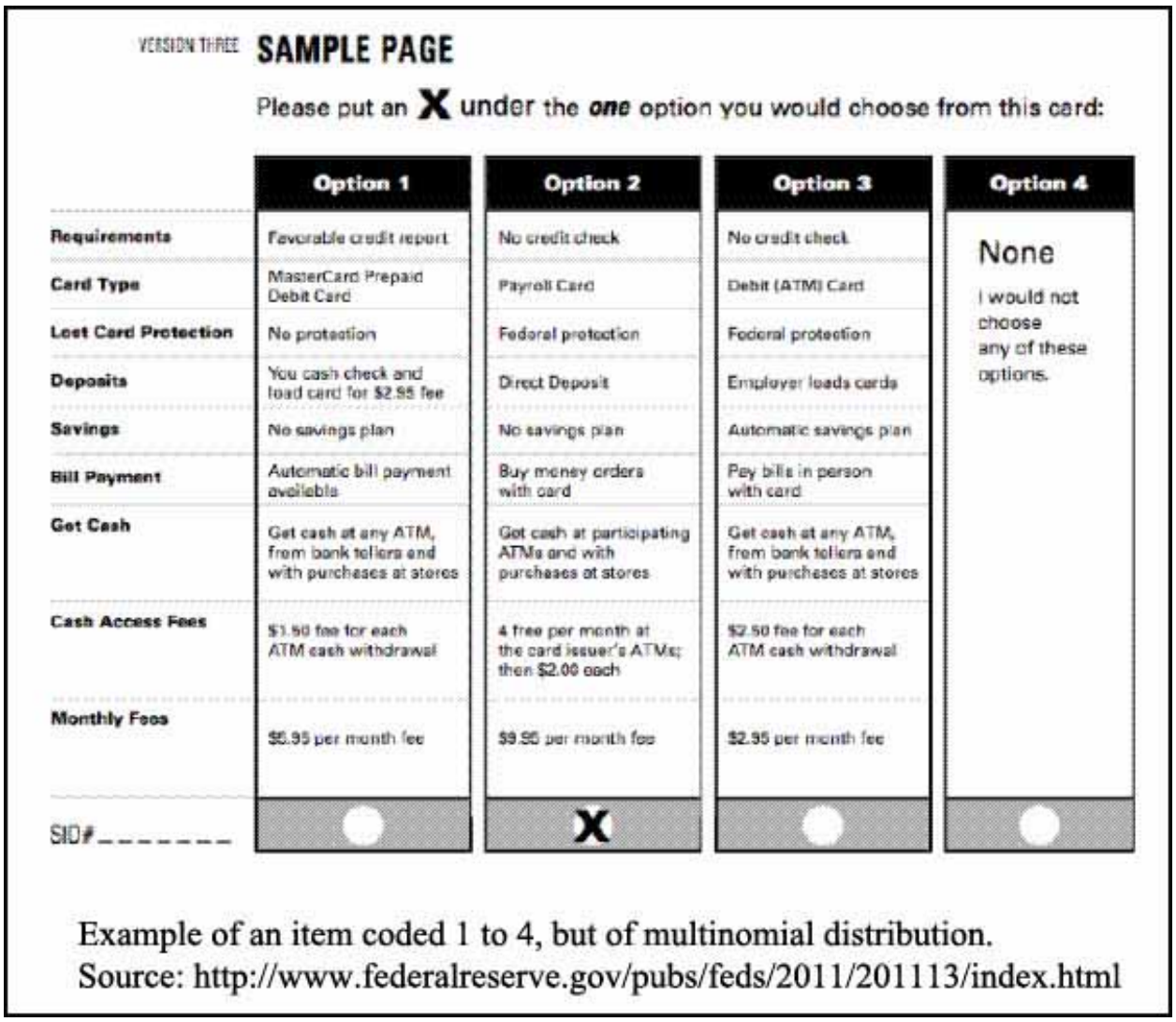
(2) SAS: Use the OUTPUT= statement to save predicted values or other coefficients. For instance, the statement below creates a new SAS dataset called "mydata" containing the variables predprob, lowlimit, and highlimit. These are respectively the predicted probability of the case and the lower and upper confidence limits on the predicted probability.

```
OUTPUT OUT=mydata PREDICTED=predprob LOWER=lowlimit UPPER=highlimit ;
```

(3) Stata: The post-estimation command `predict prob, pr` adds a variable called "prob" to the dataset, containing predicted probabilities. File > Save or File >Save As saves the dataset.

The dependent variable

Binary and multinomial logistic regression support only a single dependent variable. For binary logistic regression, this response variable can have only two categories. For multinomial logistic regression, there may be two or more categories, usually more, but the dependent variable is never a continuous variable.



VERSION THREE **SAMPLE PAGE**

Please put an **X** under the ***one*** option you would choose from this card:

	Option 1	Option 2	Option 3	Option 4
Requirements	Favorable credit report	No credit check	No credit check	None
Card Type	MasterCard Prepaid Debit Card	Payroll Card	Debit (ATM) Card	I would not choose any of these options.
Lost Card Protection	No protection	Federal protection	Federal protection	
Deposits	You cash check and load card for $2.95 fee	Direct Deposit	Employer loads cards	
Savings	No savings plan	No savings plan	Automatic savings plan	
Bill Payment	Automatic bill payment available	Buy money orders with card	Pay bills in person with card	
Get Cash	Get cash at any ATM, from bank tellers and with purchases at stores	Get cash at participating ATMs and with purchases at stores	Get cash at any ATM, from bank tellers and with purchases at stores	
Cash Access Fees	$1.50 fee for each ATM cash withdrawal	4 free per month at the card issuer's ATMs; then $2.00 each	$2.50 fee for each ATM cash withdrawal	
Monthly Fees	$5.95 per month fee	$9.95 per month fee	$2.95 per month fee	
SID#_____	○	X	○	○

Example of an item coded 1 to 4, but of multinomial distribution.
Source: http://www.federalreserve.gov/pubs/feds/2011/201113/index.html

It is important to be careful to specify the desired reference category of the dependent variable, which should be meaningful. For example, the highest category (4 = None) in the example above would be the default reference category in multinomial regression, but "None" could be chosen for a great variety of reasons and has no specific meaning as a reference. The researcher may wish to select another, more specific option as the reference category.

The dependent reference default in binary logistic regression

SPSS

In SPSS, by default binary logistic regression predicts the "1" value of the dependent, using the "0" level as the reference value. That is, the lowest value is the reference category of the dependent variable. In SPSS, the default reference level is always the lowest-coded value and cannot be changed. Of course, the researcher could flip the value order by recoding beforehand.

SAS

In SAS, binary logistic regression is implemented with PROC LOGISTIC, whose MODEL command sets the dependent variable reference category. By default, the lower-coded dependent variable category is the predicted (event) level and the higher-coded category is the reference level. This default corresponds to the "EVENT=FIRST" specification in the MODEL statement. However, this may be overridden as in the SAS code segment below, which makes the higher-coded (last) dependent value the default.

```
MODEL cappun (EVENT=LAST) = race sex degree
```

Assuming cappun is a binary variable coded 0, 1, an alternative equivalent SAS syntax is:

```
MODEL cappun (EVENT='1') = race sex degree
```

Note the default reference level in binary logistic regression in SAS is the opposite of that in SPSS, which will cause the signs of parameter estimates to be flipped unless EVENT=LAST is specified in the MODEL statement.

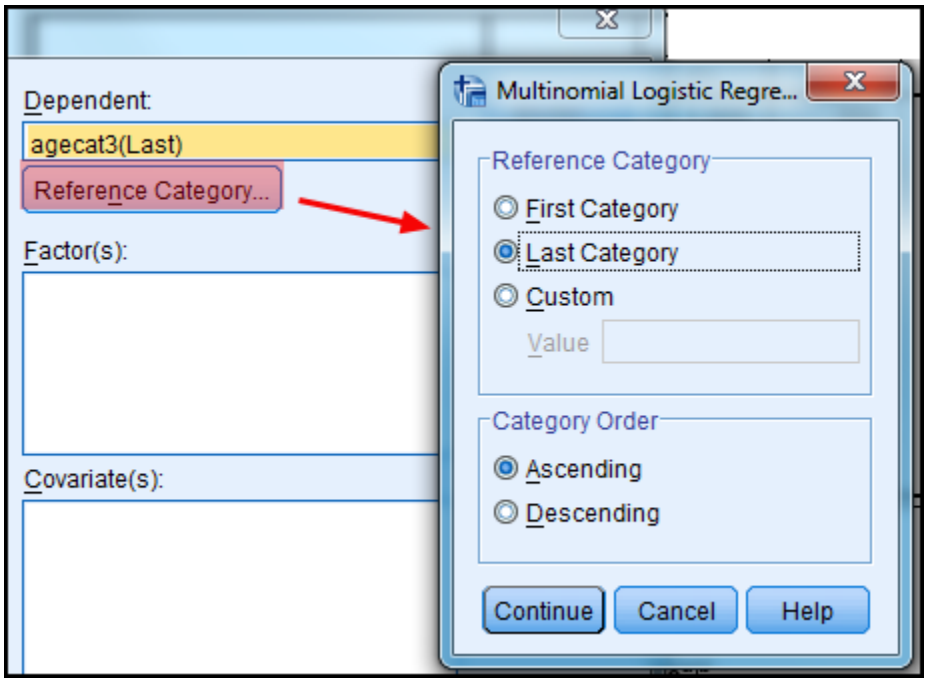
Stata

In binary logistic regression in Stata, the higher (1) level is the predicted level and the lower (0) level is the reference level by default. As in SPSS, Stata does not provide an option to change the reference level.

The dependent reference default in multinomial logistic regression

SPSS

By default, multinomial logistic regression in SPSS uses the highest-coded value of the dependent variable as the reference level. For instance, given the multinomial dependent variable "Degree of interest in joining" with levels 0=low interest, 1 = medium interest, and 2=high interest, 2 = high interest will be the reference category by default. For each independent variable, multinomial logistic output will show estimates for (1) the comparison of low interest with high interest, and (2) the comparison of medium interest with high interest. That is, the highest level is the reference level and all other levels are compared to it by default. However, in SPSS there is a "Reference Category" button in the multinomial regression dialog and it may be used to select a different reference category, as illustrated below.



Alternatively, in the In SPSS syntax window, set the reference category in multinomial regression simply by entering a base parameter after the dependent in the variable list, as in the code segment below:

```
NOMREG depvar (base=2) WITH indvar1 indvar2 indvar3 /
PRINT = PARAMETER SUMMARY.
```

When setting custom or base values, a numeral indicates the order number (ex., 2 = the second value will be the reference, assuming ascending order). For actual numeric values, dates, or strings, put quotation marks around the value.

The chart below summarizes the SPSS logistic regression defaults for dependent variables.

BINARY LOGISTIC REGRESSION: DEPENDENTS	OUTCOME:
Binary variable is entered as a dependent	Highest is predicted, lowest is reference
	The reference level cannot be changed in SPSS.
MULTINOMIAL LOGISTIC REGRESSION: DEPENDENTS	
Binary or multinomial variable entered as dependent	Highest is reference, all others compared to it by default.
	Click "Reference Category" button in SPSS to override the default.

SAS

Multinomial logistic regression is most easily implemented in SAS under PROC CATMOD in conjunction with a “RESPONSE LOGITS“ statement, though it may be implemented under PROC LOGISTIC as well.

If implemented in PROC LOGISTIC, the reference level in multinomial regression parallels that in binary logistic regression, discussed above. By default, the highest-coded level of the multinomial dependent variable is the reference level and all lower levels are predicted levels. This can be changed in the MODEL statement with any of the following specifications:

- EVENT=: Can be used for binary regression as described above but has no effect when there are more than two response levels.
- REF=: The term following the MODEL command word is the multinomial dependent variable. The REF= specification can set the reference level. For instance, MODEL y(REF='0') will set the '0' level of the multinomial dependent variable y as the reference level.
- DESCENDING: Flips the order, making the lowest-coded value the reference level.
- ORDER=: May be of the following FORMATTED, INTERNAL, DATA, or FREQ types. FORMATTED is the default, following the order specified by a prior PROC FORMAT procedure. If there is no such procedure, FORMATTED defaults to sort by internal order. INTERNAL is the unformatted value, used by SPSS and Stata. DATA is the order of appearance in the data set. FREQ is the order by descending frequency count.

If implemented in PROC CATMOD, the reference level in multinomial regression by default is the highest response level, as it is in SPSS multinomial regression. PROC CATMOD computes contrasts of the log of each response level with the log of the probability of the highest response category. PROC CATMOD does not support a DESCENDING option to flip the order. The work-around is to use PROC SORT to sort the dependent variable in descending order, then use the ORDER=DATA clause in the PROC CATMOD command, forcing the levels to follow the order in the data:

```
PROC SORT DATA = <insert dataset>;
BY DESCENDING <insert depvar>;
RUN;
PROC CATMOD ORDER = DATA  DATA = <insert dataset>;
```

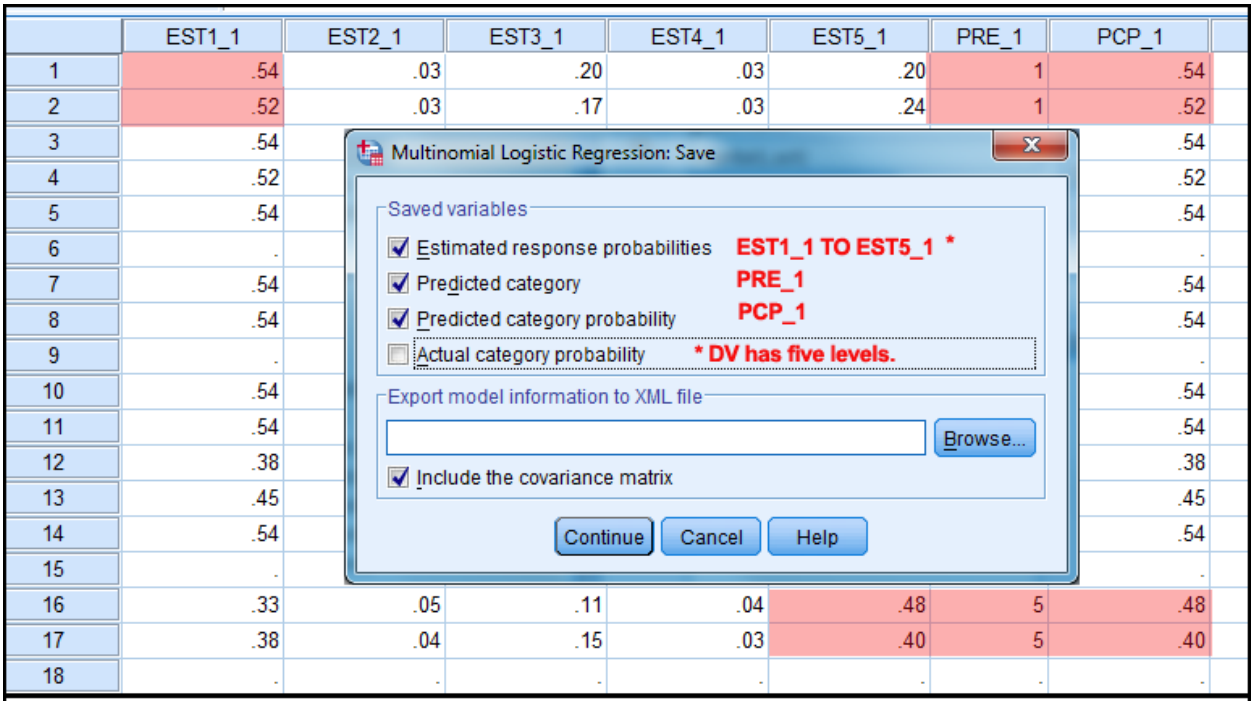
Stata

Multinomial logistic regression is implemented in Stata using the `mlogit` command. Unlike SPSS or SAS, the default reference category for multinomial regression in Stata is the category with the highest frequency. However, this can be reset by adding the `baseoutcome(#)` specification. The syntax format below, for instance, would set level 4 as the reference level for depvar, the multinomial dependent variable:


```
mlogit depvar indepvars, baseoutcome(4)
```

Dependent variable group membership

Particularly in multinomial logistic regression, it is common to label the levels of the dependent variable as "groups". The odds ratio for the dependent variable can be used to predict the probability of membership in any group. Any given observation may be assigned to the group with the highest probability. For binary logistic regression, this is the group with a highest estimated response probability, which will be over 50%. For multinomial regression, illustrated below using the SPSS multinomial regression "Save" dialog, the assigned group is the one with the highest estimated response probability, even if below 50%. Predicted category values (PRE_1 in the figure below) can be used as variables in future analysis. Saving response values is discussed further below.



	EST1_1	EST2_1	EST3_1	EST4_1	EST5_1	PRE_1	PCP_1
1	.54	.03	.20	.03	.20	1	.54
2	.52	.03	.17	.03	.24	1	.52
3	.54						.54
4	.52						.52
5	.54						.54
6	.						.
7	.54						.54
8	.54						.54
9	.						.
10	.54						.54
11	.54						.54
12	.38						.38
13	.45						.45
14	.54						.54
15	.						.
16	.33	.05	.11	.04	.48	5	.48
17	.38	.04	.15	.03	.40	5	.40
18	.	.	.	.	.	.	.

The highest estimated response probability in any row is the predicted category probability, determining the predicted category.

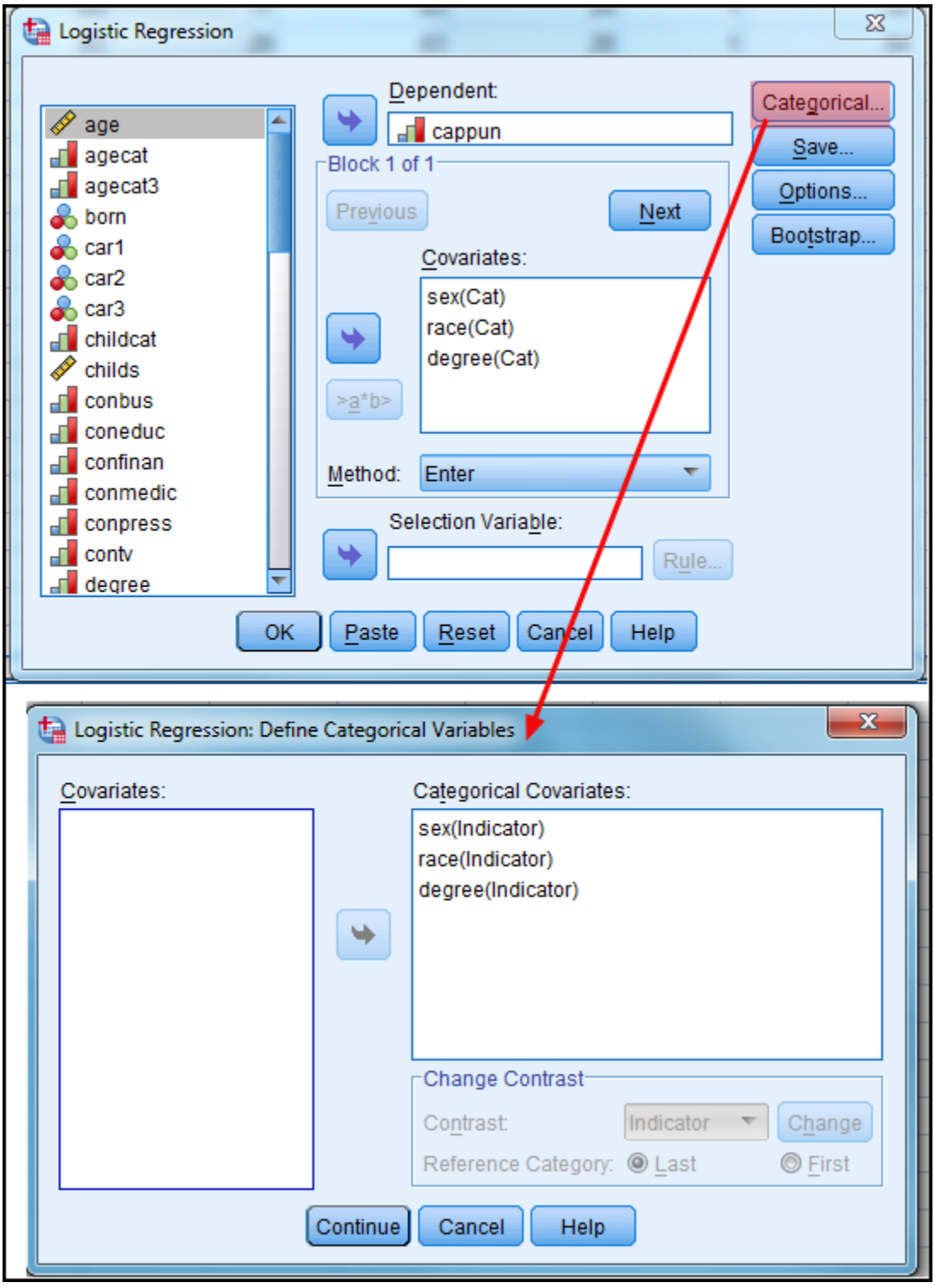
Factors: Declaring

Overview

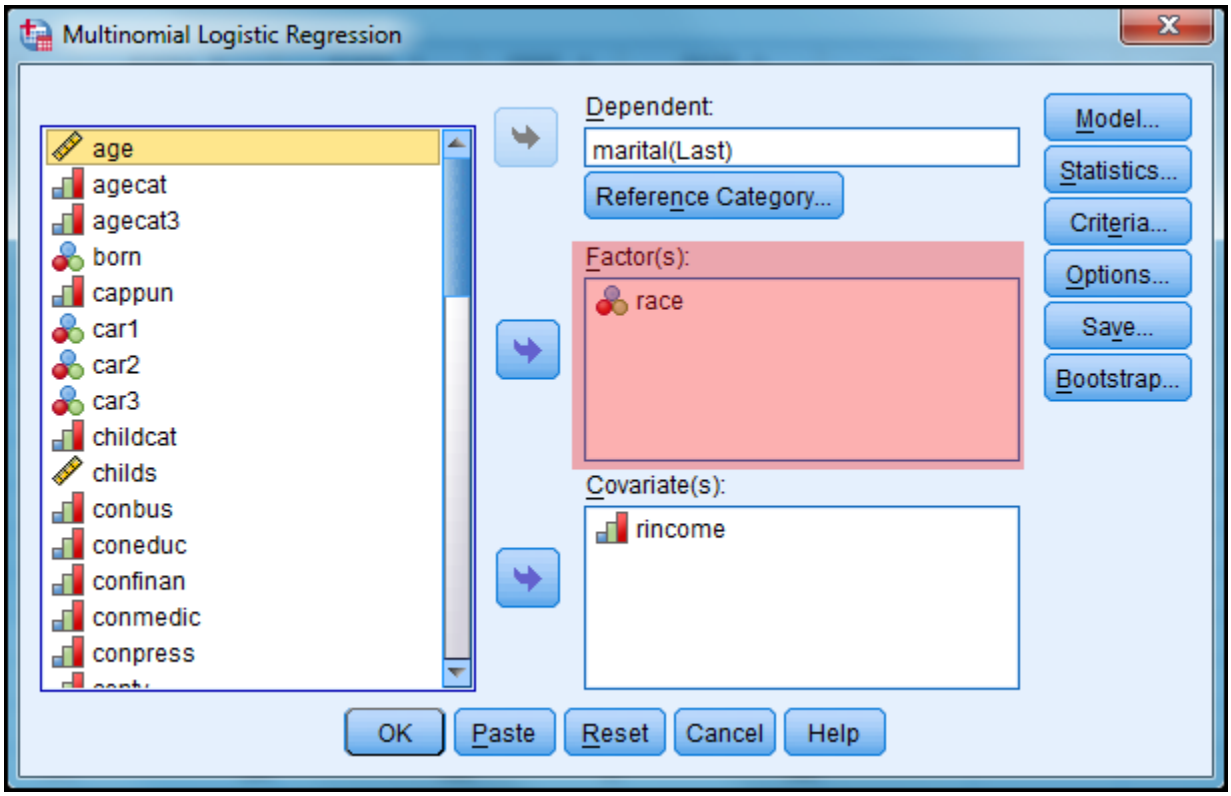
Sometimes called "design variables," factors are nominal or ordinal categorical predictor variables. In logistic regression, the levels of a factor are entered, in effect, as dummy variables in the logistic model, though the researcher does not do this manually. The researcher must, however, tell the statistics package which independent variables are to be treated as factors and how this is done varies from statistical package to statistical package.

SPSS

In SPSS binary logistic regression, categorical independent variables must be declared by clicking on the "Categorical" button in the main “Logistic Regression” dialog screen, shown below. In SPSS parlance for binary logistic regression, all predictor variables are “covariates” but factors are “categorical covariates” which have been declared so using the “Categorical” button.



In SPSS multinomial regression, there are separate places in the main dialog screen for factors (categorical variables) and for covariates (continuous variables), shown below.



SAS

Factors are declared using the CLASS statement in PROC LOGISTIC for either binary or multinomial logistic regression. In the example SAS code below, three variables (sex, race, degree) are all declared to be factors (categorical variables).

```
PROC LOGISTIC DATA=<insert datafile>;
CLASS sex race degree / <insert options>;
```

In PROC CATMOD, in contrast, all predictor variables are treated as factors, called "classification variables" in the context of multinomial logistic regression. If there are continuous covariates, PROC CATMOD will model them as categorical classification variables as well, making the computation process inefficient.

PROC LOGISTIC is preferred over PROC CATMOD for multinomial models with continuous covariates. SAS documentation warns that "Computational difficulties might occur if you have a continuous variable with a large number of unique values" and PROC CATMOD's default estimation by the "weighted least squares method is inappropriate since there are too many zero frequencies …PROC CATMOD is not designed optimally for continuous variables; therefore, it might be

less efficient and unable to allocate sufficient memory to handle this problem, as compared with a procedure designed specifically to handle continuous data" such as PROC LOGISTIC. If PROC CATMOD is nonetheless used, the researcher should declare the continuous variable in a DIRECT statement and request maximum likelihood (ML) estimation be used, as in the SAS code below.

```
PROC CATMOD DATA=<insert dataset name>
DIRECT educ;
/* Use DIRECT to list covariates & any dummy variables */
RESPONSE logits;
/* Specifies a multinomial logistic model */
MODEL income4 = educ sex race / ML=NR;
/* Models income as predicted by educ as a covariate */
/* and by sex and race as factors, then requests estimation*/
/* be done by maximum likelihood using the usual */
/* Newton-Raphson algorithm*/
```

Stata

Stata's `logistic` command for binary logistic regression takes the following form:

```
logistic depvar indepvars [, options]
```

The list of independent variables (indepvars above) may include factors as well as covariates. Variables are declared categorical by use of the "i." prefix, signifying that such variables must be represented by indicator variables. For instance, "educ" will be treated as a continuous covariate, but "i.educ" will be treated as a categorical variable (factor). The term "c.educ" declares "educ" to be a continuous variable, equivalent to "educ".

In Stata's `mlogit` command for multinomial logistic regression, factors are declared in the same manner. The general form of the command is:

```
mlogit depvar indepvars [, options]
```

The "indepvars" list may contain factors using the "i." prefix.

Factors: Reference levels

Overview

Reference levels make all the difference in how logistic regression results are interpreted. The dependent variable has a reference level (discussed above) and categorical predictor variables (factors) also have reference levels, discussed in this section.

Reference levels for predictor variables parallels the familiar practice in OLS regression of leaving out one level of a categorical variable when creating a set of dummy variables, where the left-out category is the reference level. A parameter estimate (b coefficient) will be output for all categories except the reference category.

Statistical packages offer different types of contrasts (coding) for factors and have different defaults. Some statistical packages allow the order to be flipped so "first" and "last" are reversed.

Indictor coding, also called reference coding, is the most common type of contrast. Other types are discussed in the FAQ section below. If the presence of the case in the given category is contrasted with absence of membership in the category, indicator coding is being used. The researcher sets first or last as the reference category ("Never married" = 5 is the reference category below). This is the most common type.

Indicator coding, last as reference

Categorical Variables Codings

			Parameter coding			
		Frequency	(1)	(2)	(3)	(4)
Marital status	Married	1259	1.000	.000	.000	.000
	Widowed	242	.000	1.000	.000	.000
	Divorced	412	.000	.000	1.000	.000
	Separated	79	.000	.000	.000	1.000
	Never married	607	.000	.000	.000	.000

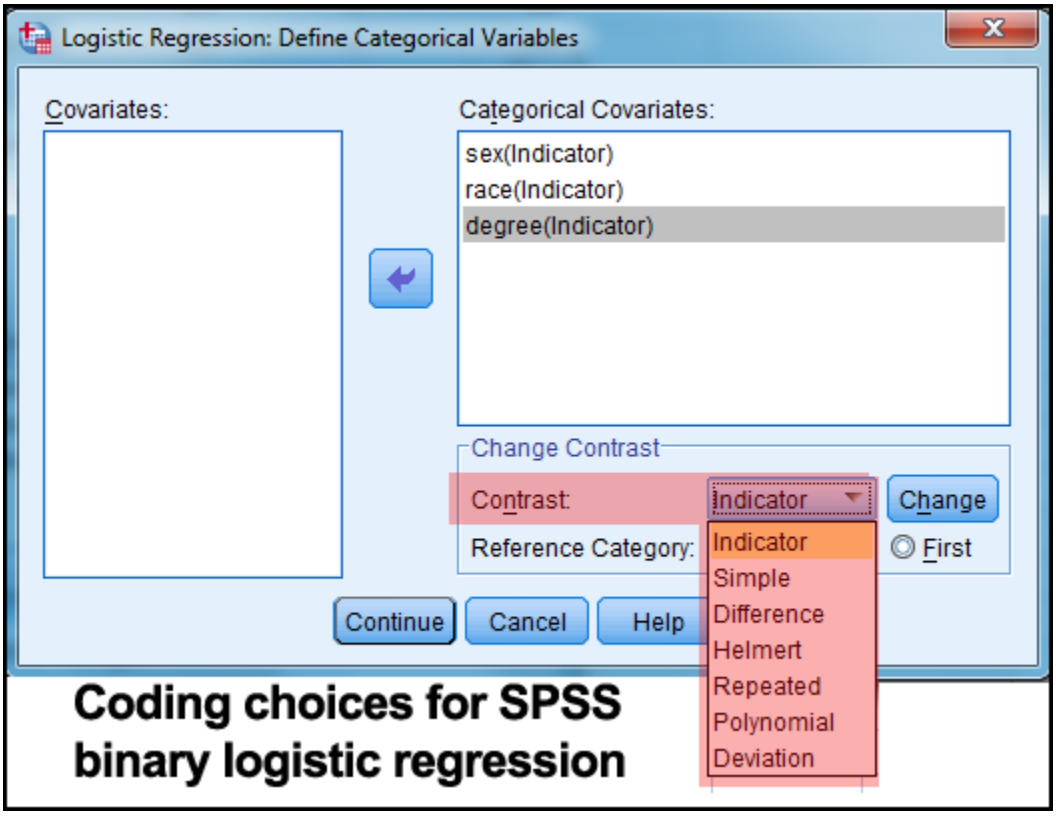
SPSS

Setting the reference category (contrast) for a factor (categorical predictor) in SPSS varies by type of analysis. Specifically, dichotomous and categorical variables may be entered as factors but are handled differently by default in binary vs. multinomial regression.

Binary logistic regression

In binary logistic regression, after declaring a variable to be categorical, the user is given the opportunity to declare the first or last value as the reference category.

Indicator coding is the default, but this may be changed by the researcher. In SPSS binary regression, different coding contrasts are selected when variables are declared to be categorical as described above. Warning: the researcher must click the "Change" button after selecting the contrast from the drop-down menu: do not just click the "Continue" button.



Multinomial logistic regression

In SPSS multinomial logistic regression, the last (highest-coded) category is the reference category. That is, indicator contrasts/coding is used and alternative parameterizations are not available. Factor reference categories cannot be changed. If the researcher prefers another category to be the factor reference category in multinomial logistic regression in SPSS, the researcher must either (1) recode the variables beforehand to make the desired reference category the last one, or (2) must create dummy variables manually prior to running the analysis.

The table below summarizes SPSS treatment of factors:

REFERENCE CATEGORIES		BINOMIAL	MULTINOMIAL
DEPENDENT	FACTOR	LOWEST IS REFERENCE	HIGHEST IS REFERENCE (but any may be selected)
	COVARIATE	NOT APPLICABLE	NOT APPLICABLE
INDEPENDENT	FACTOR	HIGHEST IS REFERENCE (but lowest may be selected)	HIGHEST IS REFERENCE
	COVARIATE	LOWEST IS REFERENCE	LOWEST IS REFERENCE

SAS

PROC LOGISTIC

SAS's PROC LOGISTIC offers an array of contrast options. Options are set by a PARAM= clause in the CLASS statement. Effect coding, also called deviation coding and discussed below, is the default if not overridden by a PARAM= specification. Note effect (deviation) coding is different from the SPSS default, which is reference (indicator) coding. See further discussion of coding/contrast alternatives below in the FAQ section.

PARAM=	Coding
EFFECT	Effect coding (deviation coding), which is the default in SAS
GLM	Less-than-full-rank reference cell coding (this *keyword* can be used only in a global option)
ORDINAL THERMOMETER	Cumulative parameterization for an ordinal CLASS variable
POLYNOMIAL POLY	Polynomial coding
REFERENCE REF	Reference cell coding (indicator coding)
ORTHEFFECT	Orthogonalizes PARAM=EFFECT coding
ORTHORDINAL ORTHOTHERM	Orthogonalizes PARAM=ORDINAL coding
ORTHPOLY	Orthogonalizes PARAM=POLYNOMIAL coding
ORTHREF	Orthogonalizes PARAM=REFERENCE coding

In PROC LOGISTIC, the CLASS statement declares categorical variables. And its ORDER=option sets the sorting order for factor levels and thus determines want the "last" category is. The default is ORDER=INTERNAL, which orders levels by their originally coded values and corresponds to the SPSS default also. The alternatives are:

- ORDER=FREQ: Levels are ordered by frequency. This is the default multinomial regression in Stata
- ORDER=DATA: Levels are ordered by their appearance in the dataset.
- ORDER=FORMATTED: Levels are ordered as specified by a prior PROC FORMAT procedure in SAS. If there has been no such prior procedure, SAS defaults to ORDER=INTERNAL

The researcher may determine if the first or last ordered value is the reference level in PROC LOGISTIC by adding a REF=FIRST or REF=LAST specification to the CLASS statement. REF=LAST is the default.

PROC CATMOD

PROC CATMOD, often used for multinomial logistic regression, defaults to deviation (effect) coding with the last (highest-coded) value as the reference

category. Deviation coding is discussed below in the FAQ section. Note this differs from the SPSS default, which is indicator coding. In the SAS MODEL statement for PROC CATMOD, using the option PARAM=REFERENCE rather than accepting the default, which is PARAM=EFFECT, the researcher can substitute reference coding, also called indicator coding and also discussed below.

In the PROC CATMOD command statement itself (not in a CLASS statement as in PROC LOGISTIC), the ORDER=option sets the sorting order for factor levels and thus determines what the "last" category is. The default is ORDER=INTERNAL, which orders levels by their originally coded values and corresponds to the SPSS default also. The alternatives are those described above for PROC LOGISTIC.

The PARAM= and ORDER= specifications appear in the SAS multinomial regression example below.

Stata

Binary logistic regression

In the `logistic` command the lowest-coded level of a factor is the reference level by default (the opposite of SPSS). If the variable "marital" is marital status, coded 1=Married, 2=Widowed, 3=Divorced, 4=Separated, and 5=Never Married, by default "Married" will be the reference level and there will be no row for it in output of parameter estimates or odds ratios. However, this can be changed using prefixes to "marital":

`logistic depvar i.marital` : The "I" prefix declares marital to be a categorical variable, with the lowest level by default the reference level.

`logistic depvar ib5.marital` : The "ib5" prefix overrides the default and makes the highest level, which is 5, the reference level. The "i" may be omitted.

Indicator coding is used in the Stata default, as in the SPSS default.

Multinomial logistic regression

In the `mlogit` command, the lowest factor level is the reference level by default. Other levels may be specified instead in the same manner as for binary logistic regression. Indicator coding is used by default.

Covariates

Overview

Covariates are continuous independent variables in most contexts.

SPSS

In the SPSS binary logistic regression dialog all independent variables are entered as covariates, then the researcher clicks the "Categorical" button to declare any of those entered as "categorical covariates". Others remain "covariates" and are assumed to be continuous. In SPSS multinomial regression, factors and covariates are entered in separate entry boxes, avoiding this complexity.

In output, for covariates, a single parameter estimate (coefficient) is output and is interpreted similar to OLS regression, where increasing the independent variable from 0 to 1 corresponds to a b increase in the dependent, except that in logistic regression the increase is in the log odds of the dependent variable, not the raw value of the dependent variable. The table below summarizes the options as found in SPSS.

In SPSS multinomial logistic regression, a variable such as the binary variable "sex", may be entered as a factor or as a covariate. As a factor, the higher value (sex = 2) is the reference level but as a covariate the lower value (sex = 1) is the reference. Consequently, the signs of the parameter estimates are reversed when comparing the two, shown in the figure below. Entered as a factor sex has a positive estimate of .462 and entered as a covariate it has a negative estimate of -.462. Also, the intercepts change, but as these are rarely interpreted in logistic regression, this makes little difference.

Multinomial Regression in SPSS
Entering a predictor as a factor vs. a covariate

Variables

Sex = 1 = Male
Sex = 2 = Female

Happiness
1= Very Happy
2 = Pretty Happy
3 = Not Very Happy

Factor coding: B is positive .512; the odds ratio for the "1 = Very Happy" level of the dependent, Happiness, is 1.668. The odds of being very happy rather than not very happy increase by a factor of 1.668 by being male rather than female. The lowest value of sex (Male = 1) is predicted and the highest (Female = 2) is the reference.

Covariate coding: B is negative .512; the odds ratio for "Very Happy" is .600. The odds of being very happy rather than not very happy decrease by a factor of .600 by being female rather than male. The highest value of sex (Female = 2) is predicted and the lowest (Male = 1) is the reference.

Parameter Estimates: Sex entered as a factor.

General Happiness[a]		B	Std. Error				Exp(B)	95% Confidence Interval for Exp (B) Lower Bound	Upper Bound
Very Happy	Intercept	.846	.113						
	[sex=1]	.512	.191				1.668	1.147	2.425
	[sex=2]	0[b]	.	.	.	.	.	.	.
Pretty Happy	Intercept	1.492	.105	203.570	1	.000			
	[sex=1]	.462	.180	6.568	1	.010	1.587	1.115	2.259
	[sex=2]	0[b]	.	.	0	.	.	.	.

Intercept = 1.492
Sex(1) = +0.462
SEX AS FACTOR

a. The reference category is: Not Too Happy.
b. This parameter is set to zero because it is redundant.

Parameter Estimates: Sex entered as a covariate.

General Happiness[a]		B	Std. Error				Exp(B)	95% Confidence Interval for Exp (B) Lower Bound	Upper Bound
Very Happy	Intercept	1.869	.328						
	sex	-.512	.191				.600	.412	.872
Pretty Happy	Intercept	2.416	.312	60.103		.000			
	sex	-.462	.180	6.568	1	.010	.630	.443	.897

Intercept - 2.416
Sex = -0.462
SEX AS COVARIATE

a. The reference category is: Not Too Happy.

SAS

In PROC LOGISTIC in SAS, categorical variables are declared in the CLASS statement, as described previously. All other variables after the equal sign in the MODEL statement are assumed to be continuous covariates.

```
MODEL cappun (EVENT=LAST) = race sex degree
```

In PROC CATMOD in SAS, all variables in the MODEL statement are assumed to be categorical unless listed in a separate DIRECT statement. Those after a DIRECT statement are assumed to be continuous covariates or interaction terms.

```
DIRECT income;
```

Stata

In Stata, under both the `logistic` and the `mlogit` commands for binary and multinomial regression respectively, all variables listed after the dependent variable (which is assumed to be binary or multinomial, of course) are predictor variables and are assumed to be covariates unless prefixed as categorical (with an "I". "b", or "ib" prefix, as discussed above).

Interaction Terms

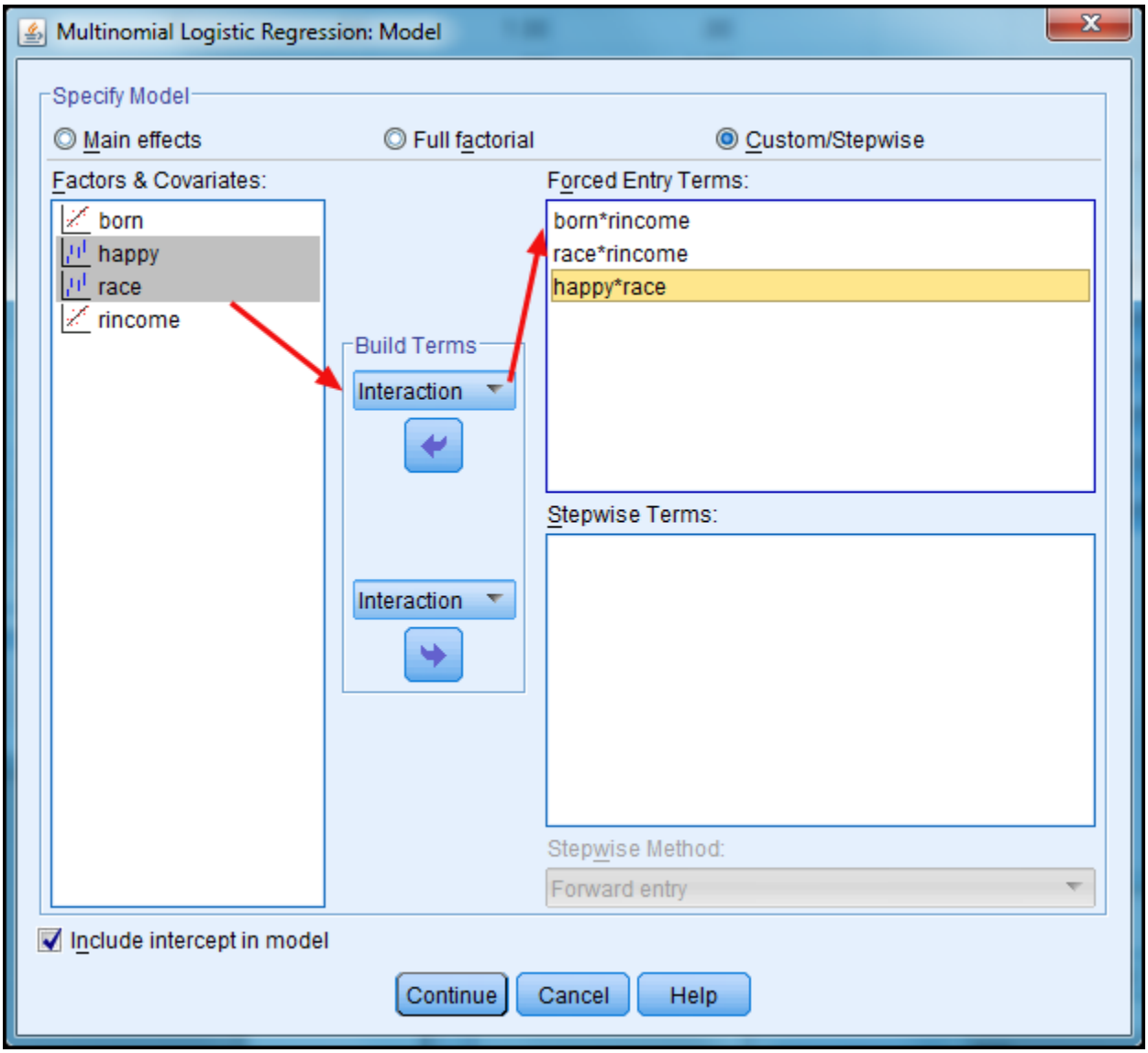
Overview

Interaction terms represent possible effects on the dependent variable over and above the separate effects of each variable individually. Interactions are represented by the crossproducts of predictor variables (ex., income*race). Warning: when interaction terms involving covariates are added to a model, centering of the covariates is strongly recommended to reduce multicollinearity and aid interpretation.

SPSS

For binary logistic regression, interaction terms such as age*income must be created manually beforehand. The binary logistic regression dialog does not support interaction terms but binary dependent variables may be entered in its multinomial logistic regression module which does directly support inclusion of interaction terms under its "Model" button.

Construction of interaction terms in SPSS multinomial logistic regression is illustrated in the figure below. In SPSS multinomial logistic regression, interaction terms are entered under the "Model" button, where interactions of factors and covariates are constructed automatically.



Interactions may be part of nested terms, which are supported by SPSS complex samples logistic regression but not regular SPSS logistic regression. See discussion of nested terms in SAS, below.

SAS

Interaction terms in PROC LOGISTIC are entered as crossproducts in the MODEL statement, as illustrated below. The crossproduct/interaction term does not appear in the CLASS statement, even if one of its components is a factor, though that component must be in the CLASS statement.

```
model RecoveryTime= Treatment Sex Age Treatment*Sex;
```

Interaction terms in PROC CATMOD are entered in the MODEL statement in the same manner. A full factorial model is shown below. The crossproduct/interaction term does not appear in the DIRECT statement, even if one of its components is a covariate, thought that component should be in the DIRECT statement.

```
model y=a b c a*b a*c b*c a*b*c;
```

Interactions may also be parts of a nested effect in SAS. Recall a nested term occurs when one variable is grouped by another. For instance, candy_brand(store_id) means that candy brands are grouped within store id's. blocks. In tabular terms, a list of candy brands might be columns and rows might be store id's. If nested, each candy brand would have an entry (ex., "1") in only one row. In such a table, an values of an interaction (ex., candy_brand*main_ingredient) could be the columns representing a term grouped within store_id.

A nested term is entered as a main effect or interaction term followed by parentheses containing a main effect of interaction term. For instance:

```
a(b*c)
a*b(c)
a*b(c*d)
```

Note that no effect may be nested within a continuous variable (a covariate).

Stata

In Stata, interactions of entered using the "##" operator rather than the usual, as in the example below in which the dichotomous variable hadcrime is predicted from owngun, marital (marital status), and the interaction of the factors marital and owngun:

```
. logistic hadcrime i.marital i.owngun i.marital##i.owngun
```

Factor by covariate and covariate by covariate interactions are specified in the same manner.

Stata's `nlogit` command supports nested logit models. As noted in Stata online documentation (enter `help nlogit` in Stata), "These models relax the assumption of independently distributed errors and the independence of irrelevant alternatives inherent in conditional and multinomial logit models by

clustering similar alternatives into nests." The `nlogit` procedure is outside the scope of this volume.

Estimation

Overview

The estimation method determines the algorithms by which a logistic procedure calculates parameter estimates and related coefficients. Often, as noted below, the researcher will have no choice but to use maximum likelihood estimation. Where there is a choice, very often the method of estimation will result in differences in estimates small enough that substantive research conclusions will not be affected.

Maximum likelihood estimation (ML)

ML is the method used to calculate the logistic (logit) coefficients in most default implementations of logistic regression and it may well be the only available method. ML estimation must be used in binary logistic regression if any of the predictor variables are continuous, though ML can also handle categorical predictors. ML methods seek to maximize the log likelihood (LL), which reflects how likely it is (that is, reflects the odds) that the observed values of the dependent variable may be predicted from the observed values of the independent variables.

There are a variety of reasons why ML estimation is preferred. In statistical terms, ML estimates are "consistent", which means the likelihood they are close to the true value increases asymptotically (as sample size grows). This implies that for large samples, ML estimates are unbiased. ML estimates are also statistically "efficient," meaning they have as low or lower standard errors as any alternative estimation method. Finally, ML estimates for large samples have an approximately normal sampling distribution (they are "asymptotically normal"), meaning that computation of significance and confidence levels based on the assumption of normality may be applied.

The price of all these desirable statistical attributes is that ML estimation requires larger samples than its alternatives. Nonetheless, it is common research practice to use ML estimation even with small and medium samples. While ML estimates

with such samples have diminished assurance of being efficient and unbiased, they are widely considered acceptable for hypothesis testing.

- For SPSS, ML using the usual Newton-Raphson algorithm is the default an only estimation method for both binary and multinomial logistic regression.
- For SAS, ML is the default and only estimation method for binary logistic regression in PROC LOGISTIC. While available for multinomial logistic regression in PROC CATMOD, it is not the default for SAS (see WLS discussion below).
- For Stata, ML using the usual Newton-Raphson algorithm is the default an only estimation method for both binary (the `logistic` command) and multinomial (the `mlogit` command) logistic regression. However, Stata offers ordinary and robust ML: by adding the robust term as an option in the `logistic` or `mlogit` commands, robust standard errors are computed which affects (usually slightly) the parameter confidence limits but not the parameter estimates themselves.

```
. logistic depvar indepvar, robust coef
```

Weighted least squares estimation (WLS)

WLS estimation weights data to adjust for heteroscedasticity, presumably the reason SAS makes this type of estimation its default. Stata addresses heteroscedasticity by providing a `robust` option in its `logistic` and `mlogit` commands. SPSS offers a robust option only if logistic regression is implemented in its generalized linear models (GZLM) module, which is discussed in the separate Statistical Associates "Blue Book" volume on "Generalized Linear Models."

WLS estimation may be used when all the predictor variables are categorical (Allison, 2012: 20). SAS's PROC CATMOD, for multinomial logistic regression, is intended for this situation and uses WLS estimation as its default, though this may be made explicit by putting the term WLS (or GLS, for generalized least squares, to which WLS is equivalent in this context) in the options section of the MODEL statement for PROC CATMOD. Alternatively, ML estimation may be specified by putting the term ML in the options section. ML is equivalent to ML=NR, which specifies the usual Newton-Raphson algorithm. ML=IPF specifies the less-used iterative proportional fitting algorithm. ML is preferred in PROC CATMOD for log-linear models and the analysis of generalized logits, otherwise the default WLS estimation is used.

Although PROC CATMOD can incorporate continuous predictor variables using the DIRECT statement, there needs to be several cases per covariate pattern or error messages may well occur when using the default WLS estimation and convergence may not be attained (Menard, 2010: 221, note 4).

WLS is not offered by SPSS or Stata for either binary or multinomial logistic regression.

Ordinary least squares estimation (OLS)

OLS methods seek to minimize the sum of squared distances of the data points to the regression line. OLS estimation can be seen as a subtype of ML for the special case of a linear model characterized by normally distributed disturbances around the regression line, where the computed parameter estimates maximize the likelihood of obtaining the least sum of squared disturbances. When error is not normally distributed or when the dependent variable is not normally distributed, ML estimates are preferred because they are unbiased beyond the special case handled by OLS.

In SPSS, SAS, and Stata, OLS estimation is not available for binary or multinomial logistic regression.

A basic binary logistic regression model in SPSS

Example

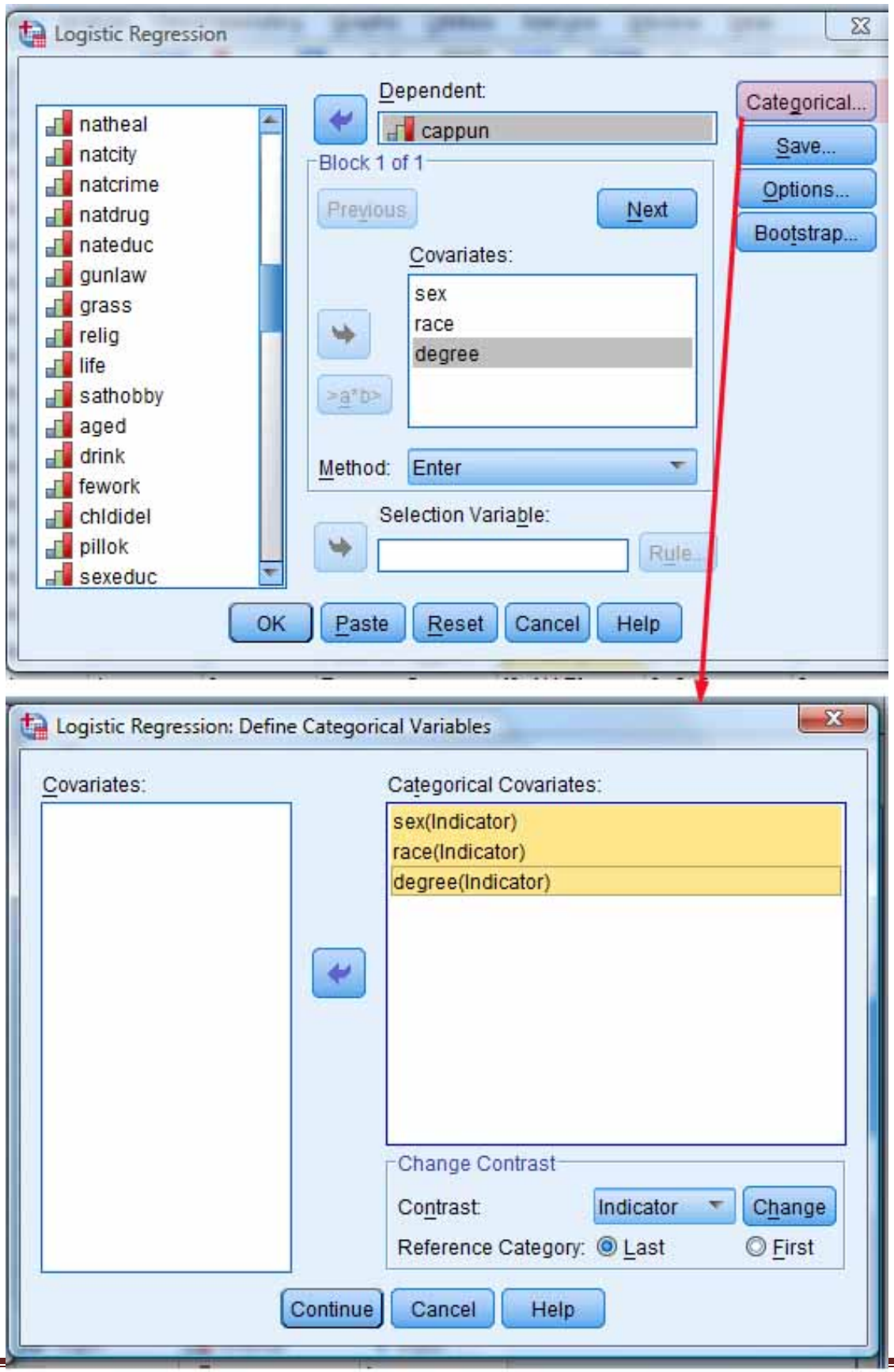
This example uses the file "GSS93subset.sav" (for download, see above). In this simple model, attitude toward capital punishment (cappun, coded 1= favor death penalty, 2= oppose) is the dependent. As a binary variable, cappun=2=Oppose is modeled and cappun=1=Favors is the reference. The dependent is predicted from sex (sex, coded 1=male, 2=female), race (race, coded 1=white, 2=black, 3=other), and highest degree earned by the respondent (degree, coded 0=< high school, 1= high school, 2= junior college, 3=bachelor, 4=graduate).

SPSS input

In SPSS, binary logistic regression is found in the menu system under Analyze > Regression > Binary Logistic. This leads to the main SPSS binary logistic regression

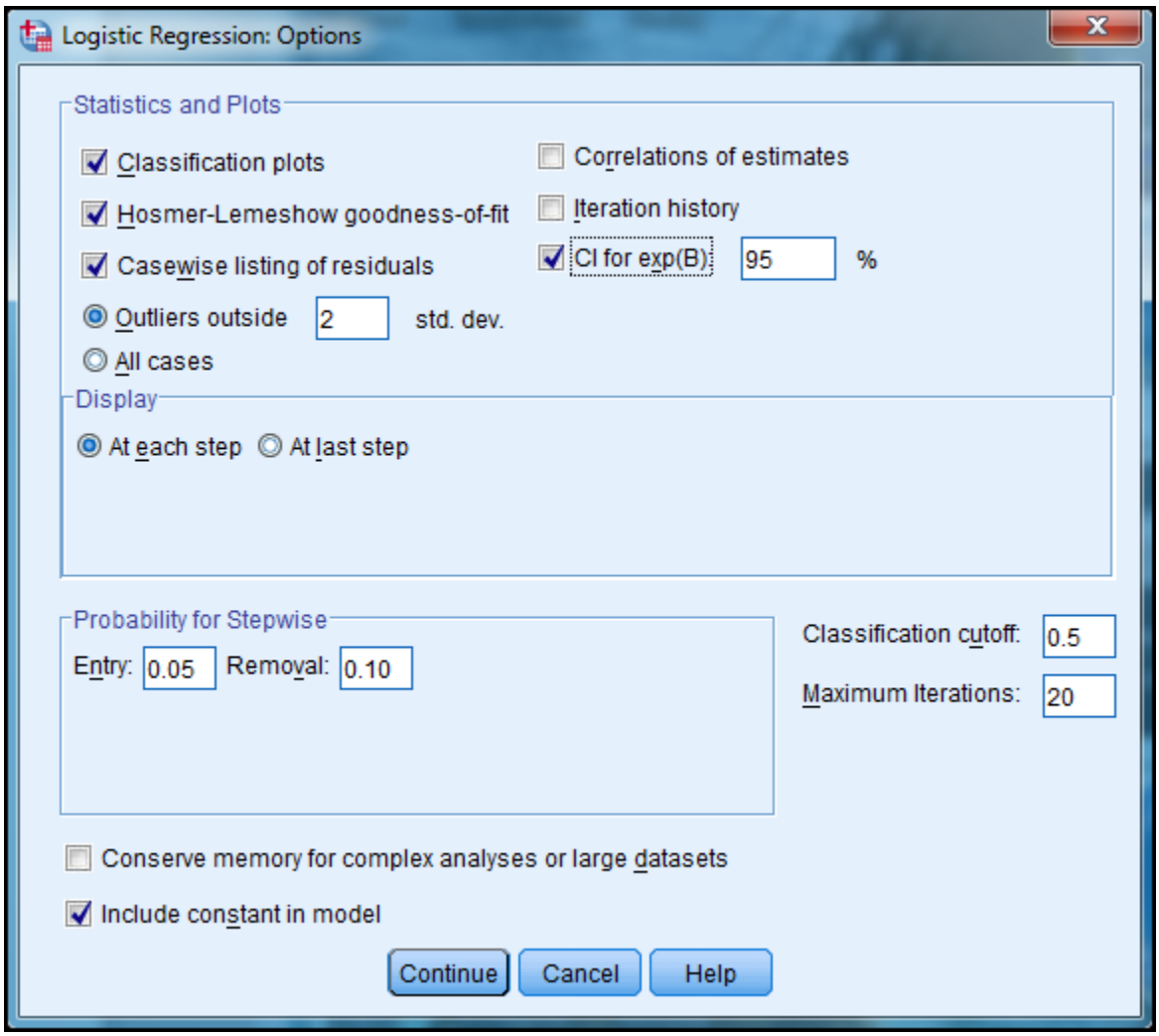
dialog shown below, displayed along with the dialog for the “Categorical” button. Select the dependent variable and the covariates (both continuous and categorical), but then use the “Categorical” button to declare which are categorical variables. Use the Options button to select desired output, then click the Continue and OK buttons to proceed.

Model selections are illustrated in the figure below. The "Categorical" button is pressed and all predictor variables are entered as categorical. Note at the bottom of the figure, that by default all categorical variables use indicator type contrasts in which the highest (last) value is the reference category.



Output selection

Model output is selected under the "Options" button, with common selections shown in the figure below.



Default output. Even if no options are selected under the Options button, SPSS generates the outputs discussed in the output section below.

SPSS output

Parameter estimates and odds ratios

The "Variables in the Equation" table gives parameter estimates and their significance levels for each level of each categorical predictor and for each

continuous covariate, if any. It also gives odds ratios. Note that the subscripts are numbered in ascending order starting with 1, and reference levels are the highest-coded values and are omitted.

Sex, coded 1=male, 2=female, for male in comparison to the reference level, 2=female, is significant at less than the better than the.001 level. Note that unlike race, there is no overall significance level for sex since, being binary, the overall level for sex is identical to the significance level for sex(1) = male. The negative sign indicates males are significantly less likely to oppose the death penalty.

Race displays significance levels for race overall, significant at less than the .001 level; the race(1) significance is that for 1=white in comparison with 3=other race, and is significant at the .004 level; and the race(2) significance is that for 2=black in comparison with 3=other, and is not significant. That race is significant overall means that at least one of its levels is significant.

Degree is also significant overall. Degree(1) displays the significance for the lowest-coded level of degree, which was 0=< high school compared to 4=graduate, and it is significant at the .001 level. Other levels of degree are also significant, except degree(4), which corresponds to degree=3=bachelor compared to 4=graduate, is not significant.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	sex(1)	-.612	.140	19.065	1	.000	.542	.412	.714
	race			42.943	2	.000			
	race(1)	-.784	.270	8.426	1	.004	.457	.269	.775
	race(2)	.383	.313	1.503	1	.220	1.467	.795	2.708
	degree			37.846	4	.000			
	degree(1)	-.858	.269	10.190	1	.001	.424	.250	.718
	degree(2)	-1.009	.238	18.059	1	.000	.364	.229	.581
	degree(3)	-.992	.359	7.642	1	.006	.371	.184	.749
	degree(4)	-.087	.260	.112	1	.738	.917	.551	1.525
	Constant	.323	.332	.948	1	.330	1.381		

a. Variable(s) entered on step 1: sex, race, degree.

Odds ratios are also shown in the "Variables in the equation" table, in the Exp(b) column. Odds ratios are variable (not model) effect size measures in logistic regression, with values above 1.0 reflecting positive effects and those below 1.0

reflecting negative effects, as discussed below. Effects which are non-significant will have odds ratios closer to 1.0.

The odds ratio is the factor by which the odds of being in the predicted level of the binary dependent variable are multiplies when the independent variable increases one unit. Thus, as sex increases one unit, going from 1=male to 2=female, the odds of opposing the death penalty are reduced by almost half (that is, they are multiplied by .54).

Confidence limits on odds ratios are not default output but appear in the "Variables in the equation" table if "CI for exp(b)" is selected under the "Options" button. Confidence limits by default are for the 95% confidence level, but the level may be altered by the researcher.

Omnibus tests of model coefficients

The "Omnibus Tests" table below shows that the researcher's model is significantly better than the intercept-only (block 0) model. See further discussion of the omnibus test below.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	89.011	7	.000
	Block	89.011	7	.000
	Model	89.011	7	.000

Model summary

Nagelkerke's pseudo R-square, also called Cragg & Uhler's R-square, is an effect size measure for the model as a whole. It is not interpreted as "percent of variance explained" but rather in terms of weak, moderate, or strong effect size. Nagelkerke's pseudo-R-square norms the Cox and Snell coefficient to vary between 0 and 1. Most would term the effect size weak for these data. Weak effect size suggests the need to re-specify the model with more and/or better predictors. Adding political party id and attitude toward country western music, for example, would increase Nagelkerke's R-squared to .149 when predicting

attitude toward capital punishment (not shown). See further discussion of pseudo R-squared measures below.

The Cox and Snell pseudo R-square is another approximation to R-squared in regression but is less reported as it does not vary from 0 to 1.

The -2 log likelihood in the "Model Summary" may be used later in likelihood ratio tests of differences between nested models. It is also known as -2LL, model chi-square, deviance, or the likelihood ratio.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1391.247[a]	.062	.095

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Classification table

This table is an alternative method of assessing model effect size. Here it can be seen that the binary logistic model has 77.5% accuracy. However, this percentage should be compared to the percent correct simply by always guessing the most numerous category, which would then give a percentage correct of (1058+14)/(1058+14+297+16) = 77.4%. That is, the binary model improves upon guessing knowing the marginals only by .1%. This measure of effect size does not take into account the size of the error: close but wrong predictions count the same as far-off wrong predictions. Effect size estimated through the classification table tends to be lower that effect size estimated by Nagelkerke's R-squared. For the distribution in the example data, it is much lower. See further elaboration through SPSS ROC curve analysis below, as well as general discussion of classification tables below.

Classification Table[a]

Observed			Predicted		
			Favor or Oppose Death Penalty for Murder		Percentage Correct
			Favor	Oppose	
Step 1	Favor or Oppose Death Penalty for Murder	Favor	1058	14	98.7
		Oppose	297	16	5.1
	Overall Percentage				77.5

a. The cut value is .500

Measures of association for the classification table

While binary logistic regression does not offer measures of association for the classification table (contrast multinomial regression, which offers measures of monotone association), they are easily calculated by entering the table frequencies above into SPSS, with a weighting column; then selecting Data, Weight Cases, to invoke the weighting variable; then selecting Analyze, Descriptives, Crosstabs and selecting desired measures of association, as illustrated below, based on the classification table above. These measures of association are discussed in a separate "blue book" volume on the topic, but in general the higher the association coefficient, the better the model predicts the binary dependent. Here, for instance, Somers' d symmetric at .068 indicates weak association and hence weak model performance.

Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	.006	.016	.365	.715
		ROW Dependent	.006	.017	.365	.715
		COLUMN Dependent	.000	.000	.[c]	.[c]
	Goodman and Kruskal tau	ROW Dependent	.012	.007		.000[d]
		COLUMN Dependent	.012	.007		.000[d]
	Uncertainty Coefficient	Symmetric	.015	.009	1.727	.000[e]
		ROW Dependent	.009	.005	1.727	.000[e]
		COLUMN Dependent	.047	.026	1.727	.000[e]
Ordinal by Ordinal	Somers' d	Symmetric	.068	.021	2.932	.003
		ROW Dependent	.314	.092	2.932	.003
		COLUMN Dependent	.038	.013	2.932	.003
Nominal by Interval	Eta	ROW Dependent	.109			
		COLUMN Dependent	.109			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Cannot be computed because the asymptotic standard error equals zero.

d. Based on chi-square approximation

e. Likelihood ratio chi-square probability.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	.109			.000
	Cramer's V	.109			.000
	Contingency Coefficient	.109			.000
Ordinal by Ordinal	Kendall's tau-b	.109	.033	2.932	.003
	Kendall's tau-c	.027	.009	2.932	.003
	Gamma	.606	.118	2.932	.003
Measure of Agreement	Kappa	.056	.019	4.076	.000
N of Valid Cases		1387			

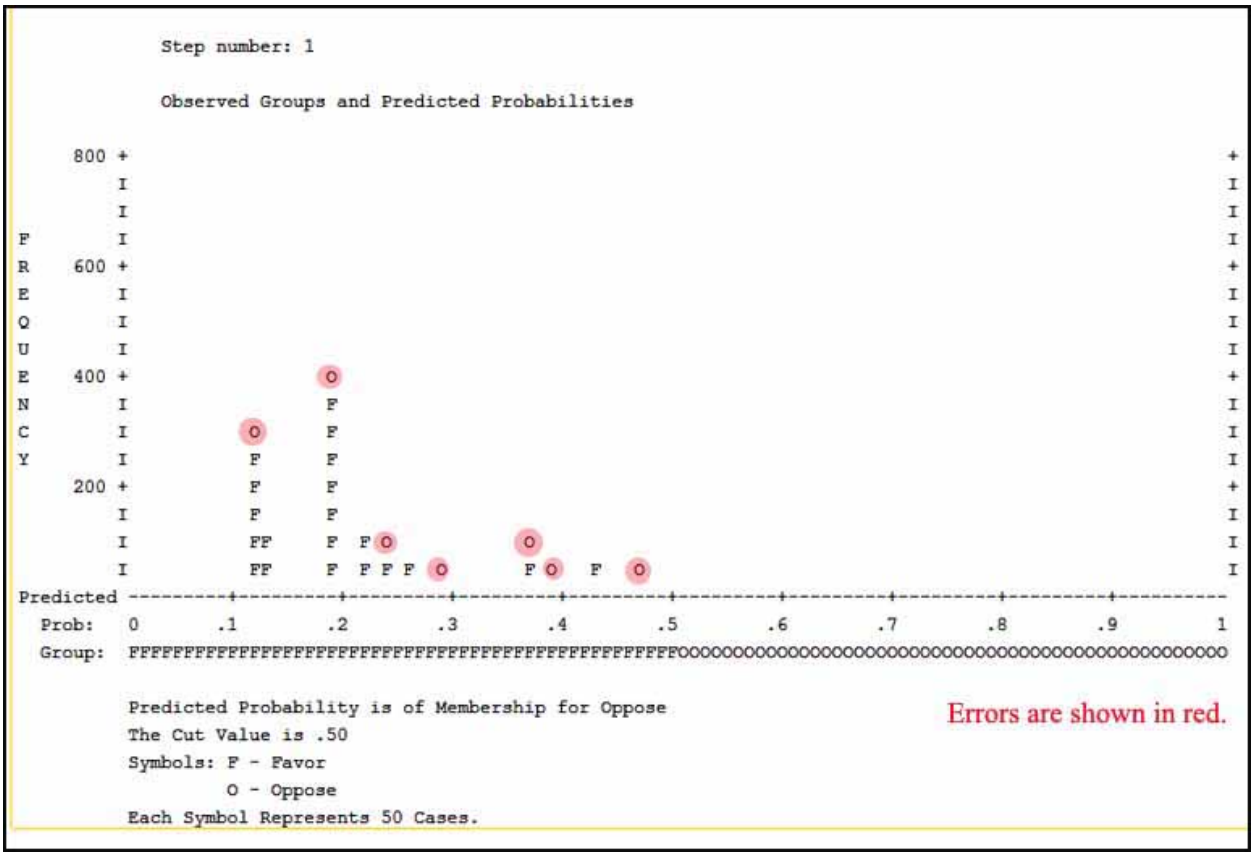
a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Classification plot

The classification plot is optional output. It is also called the classplot, the histogram of predicted probabilities, or the plot of observed groups and predicted probabilities. It is output when the researcher chooses "Classification plots" under the “Options” button of the SPSS logistic regression dialog. Classification plots are an alternative way of assessing correct and incorrect predictions under logistic regression. The X axis is the predicted probability from 0.0 to 1.0 of the dependent variable (cappun in this example) being classified a certain value (here, "F" for favoring the death penalty or "O" for opposing).

In the figure on the following page, the line of F's and O's under the X axis indicates that a prediction of 0 to .5 corresponded to the case being classified as "F"=favoring, and .5 to 1 corresponded to being classified as "O" opposing the death penalty. The Y axis is frequency: the number of cases classified. As noted below the table, each symbol represents 50 respondents. In the figure below, O's on the left are prediction errors. F's on the right would also be prediction errors, but for these data only 14 respondents who were "F" were predicted to be "O" - not enough for a single symbol where each symbol is 50 respondents.



In a classification plot, the researcher looks for two things: (1) a U-shaped rather than normal distribution is desirable. A U-shaped distribution indicates the predictions are well-differentiated. A normal distribution indicates many predictions are close to the cut point, which is not as good a model fit.; and (2) there should be few errors. The plot will also tell such things as how well the model classifies difficult cases (ones near p = .5). Above, there are many errors in predicting "F" = favoring the death penalty, and the model poorly predicts "O" = opposing the death penalty.

Hosmer-Lemeshow test of goodness of fit

The Hosmer-Lemeshow test is also optional output in SPSS. The two tables below are output after checking "Hosmer-Lemeshow goodness of fit" under the Options button in SPSS. Also called the chi-square test. this test is not available in multinomial logistic regression. Though not the default in SPSS, many statisticians consider this the recommended test for overall fit of a binary logistic regression model. It is considered more robust than the traditional chi-square test, particularly if continuous covariates are in the model or sample size is small. A finding of non-significance, as in the illustration below, corresponds to the researcher concluding the model adequately fits the data. This test is preferred over the omnibus test, classification tables, and classification plots when assessing model fit, with non-significance indicated good fit.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.060	6	.417

Contingency Table for Hosmer and Lemeshow Test

		Favor or Oppose Death Penalty for Murder = Favor		Favor or Oppose Death Penalty for Murder = Oppose		
		Observed	Expected	Observed	Expected	Total
Step 1	1	225	223.188	26	27.812	251
	2	112	106.088	9	14.912	121
	3	299	299.238	69	68.762	368
	4	103	109.726	35	28.274	138
	5	84	86.199	29	26.801	113
	6	92	95.877	42	38.123	134
	7	90	87.363	49	51.637	139
	8	67	64.321	54	56.679	121

How the Hosmer-Lemeshow test works

Hosmer and Lemeshow's goodness of fit test divides subjects into deciles based on predicted probabilities as illustrated above, then computes a chi-square from observed and expected frequencies. Then a probability (p) value is computed

from the chi-square distribution to test the fit of the logistic model. If the H-L goodness-of-fit test significance is higher than .05 (that is, is non-significant), as we want for well-fitting models, we fail to reject the null hypothesis that there is no difference between observed and model-predicted values, implying that the model's estimates fit the data at an acceptable level.

That is, well-fitting models show non-significance on the H-L goodness-of-fit test, indicating model prediction is not significantly different from observed values. This does not mean that the model necessarily explains much of the variance in the dependent, only that however much or little it does explain is significant. As the sample size gets large, the H-L statistic can find smaller and smaller differences between observed and model-predicted values to be significant. The H-L statistic assumes sampling adequacy, with a rule of thumb being enough cases so that no group has an expected value < 1 and 95% of cells (typically, 10 decile groups times 2 outcome categories = 20 cells) have an expected frequency > 5. Collapsing groups may not solve a sampling adequacy problem since when the number of groups is small, the H-L test will be biased toward non-significance (will overestimate model fit). This test is not to be confused with a similarly named, obsolete goodness of fit index discussed below.

Casewise listing of residuals for outliers > 2 standard deviations

The optional “Casewise List” table lists cases for which the model is not working well. Below, all outliers are respondents who opposed the death penalty but were predicted to favor it. Again, the model works poorly for this group. Note that SPSS multinomial logistic regression does not output this outlier-related list, so the researcher must run a binary logistic regression for each dummy variable representing a level of the multinomial dependent.

Casewise List[b]

Case	Selected Status[a]	Observed: Favor or Oppose Death Penalty for Murder	Predicted	Predicted Group	Temporary Variable: Resid	Temporary Variable: ZResid
1	S	O**	.111	F	.889	2.833
86	S	O**	.111	F	.889	2.833
232	S	O**	.113	F	.887	2.808
239	S	O**	.111	F	.889	2.833
330	S	O**	.111	F	.889	2.833
345	S	O**	.111	F	.889	2.833
358	S	O**	.111	F	.889	2.833
361	S	O**	.127	F		6
		Output excised				
970	S			F		2.833
1052	S	O**	.111	F	.889	2.833
1058	S	O**	.111	F	.889	2.833
1075	S	O**	.111	F	.889	2.833
1209	S	O**	.111	F	.889	2.833
1239	S	O**	.111	F	.889	2.833
1263	S	O**	.111	F	.889	2.833
1340	S	O**	.111	F	.889	2.833
1350	S	O**	.127	F	.873	2.626
1473	S	O**	.127	F	.873	2.626
1483	S	O**	.111	F	.889	2.833

a. S = Selected, U = Unselected cases, and ** = Misclassified cases.

b. Cases with studentized residuals greater than 2.000 are listed.

A basic binary logistic regression model in SAS

Example

This example uses the file "GSS93subset.sas7bdat" (for download, see above). In this simple model, attitude toward capital punishment (cappun, coded 1= favor death penalty, 2= oppose) is the dependent. As a binary variable, cappun=2=Oppose is modeled and cappun=1=Favors is the reference. The dependent is predicted from sex (sex, coded 1=male, 2=female), race (race, coded

1=white, 2=black, 3=other), and highest degree earned by the respondent (degree, coded 0=< high school, 1= high school, 2= junior college, 3=bachelor, 4=graduate).

SAS input

The same example is used as for the SPSS binary model above, in which capital punishment (1=favor, 2=oppose_ is predicted from sex, race, and highest educational degree. Essentially the same binary logistic model as run above for SPSS is run for SAS using the SAS commands below. Statements within the slash/asterisks are comments not implemented by PROC LOGISTIC.

```
PROC IMPORT OUT= WORK.binarylogistic
            DATAFILE= "C:\\Data\GSS93 subset.sav"
            DBMS=SPSS REPLACE;
RUN;
TITLE "PROC LOGISTIC BINARY LOGISTIC REGRESSON EXAMPLE" JUSTIFY=CENTER;
/* Optional title on each page */
PROC LOGISTIC DATA=binarylogistic;
/* Use the work data file from PROC IMPORT */
CLASS sex race degree / PARAM=REF REF=LAST ORDER=INTERNAL;
/* CLASS declares variables as categorical */
/* PARAM=REF requests dummy rather than default effects coding */
/* REF = LAST makes highest predictor category the reference as in SPSS */
/* ORDER = INTERNAL keeps categories ordered as coded as in SPSS*/
MODEL cappun (EVENT=LAST) = race sex degree
/ SELECTION=STEPWISE SLSTAY=.10 INFLUENCE CTABLE PPROB=.5 EXPB RSQUARE
LACKFIT;
/* Logistic model predicting cappun = 2, the higher level */
/* EVENT=LAST makes cappun=1 the reference & cappun=2 predicted as in SPSS */
/* SELECTION=STEPWISE uses SPSS default model-building */
/* SLSTAY=.10 uses SPSS default for keeping variable in model */
/* INFLUENCE requests residual diagnostics */
/* CTABLE requests a classification table based on response probabilities */
/* PPROB sets the cutting point for CTABLE */
/* RSQUARE requests generalized R-square */
/* LACKFIT requests the Hosmer-Lemeshow test */
RUN;
```

Reconciling SAS and SPSS output

SAS defaults for binary logistic regression differ from those in SPSS and must be overridden to obtain comparable output. The following settings are required to reconcile SAS and SPSS output.

- Variables declared categorical in SPSS must be listed in the CLASS statement in SAS.
- PARAM=REF is needed in the CLASS statement to override the SAS default of effect coding, instead using what SAS calls reference cell coding, equivalent to what SPSS calls indicator coding. Though not illustrated here, the researcher should compare the Class Level Information table in SAS with the Categorical Variables Coding table in SPSS to assure that coding is the same. Alternative coding schemes are discussed in the FAQ section below.
- REF=LAST is the default in both SPSS and SAS for reference coding, but is inserted above for clarity: for the categorical predictors, the highest-coded category is the reference level.
- ORDER=INTERNAL is needed in the CLASS statement to assure categories are ordered as coded, not by external formatted values, frequency, or order of appearance in the data. The SPSS default is as coded. The SAS default is ORDER-FORMATTED, as by a PROC FORMAT statement. If there is none, SAS reverts to ORDER-INTERNAL.
- In the MODEL statement, there must be an EVENT=LAST statement to override the SAS default, which is EVENT=FIRST. EVENT=LAST specifies that the higher code of the binary dependent is predicted and the lower code is the reference. Note options pertaining to the dependent variable are in the MODEL statement and options pertaining to the factors/independent variables are in the CLASS statement.
- In the MODEL statement, SELECTION=STEPWISE is needed to override the SAS default of SELECTION=NONE. Failure to do this may make parameter estimates diverge considerably. Going along with this, SLSTAY must be set to .10 to override the SAS default of .05.
- As PROC LOGISTIC lacks an output option for the classification table, the whole section of syntax above starting with OUTPUT OUT= is needed to create the table manually.

SAS output

Parameter estimates

PROC LOGISTIC outputs the two tables below, which correspond to the "Variables in the Equation" table in SPSS, discussed above. The "Type 3 Analysis of Effects"

table gives Wald significance levels for each factor and covariate overall, while the "Analysis of Maximum Likelihood Estimates" table gives significance levels for each level of each factor and for the intercept. Coefficients and probability levels are almost identical to those in SPSS, with minor rounding and algorithmic differences at the third decimal place. Interpretation is the same as described above for the parallel SPSS output.

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
race	2	42.9432	<.0001
sex	1	19.0644	<.0001
degree	4	37.8454	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		1	0.3229	0.3318	0.9476	0.3303	1.381
race	white	1	-0.7841	0.2701	8.4258	0.0037	0.457
race	black	1	0.3833	0.3126	1.5029	0.2202	1.467
sex	Male	1	-0.6119	0.1402	19.0644	<.0001	0.542
degree	Less than HS	1	-0.8581	0.2688	10.1899	0.0014	0.424
degree	High school	1	-1.0094	0.2375	18.0586	<.0001	0.364
degree	Junior college	1	-0.9920	0.3588	7.6423	0.0057	0.371
degree	Bachelor	1	-0.0867	0.2595	0.1115	0.7384	0.917

Odds ratio estimates

PROC LOGISTIC also output odds ratios identical to those shown in the Exp(b) column of the "Variables in the Equation" table in SPSS, discussed and illustrated above .

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
race white vs other	0.457	0.269	0.775
race black vs other	1.467	0.795	2.708
sex Male vs Female	0.542	0.412	0.714
degree Less than HS vs Graduate	0.424	0.250	0.718
degree High school vs Graduate	0.364	0.229	0.581
degree Junior college vs Graduate	0.371	0.184	0.749
degree Bachelor vs Graduate	0.917	0.551	1.525

Global null hypothesis tests

Taken from the final stepwise step, the SAS tables below correspond to the "Omnibus Tests of Model Coefficients" table in SPSS, though default SAS output provides additional tests. Global/omnibus tests test the significance of the overall model. A finding of not significant would mean none of the predictors explain the dependent variable to a significant degree. This corresponds to the global F test in regression. The "Likelihood Ratio" test below is the same as the omnibus test in the "Model" row in SPSS and is the most common of the global tests of the model.

SAS also reports global tests by the score and Wald statistics, which almost always yield similar findings. By all three tests, the model is significant for the example data. If the stepwise option is taken in PROC LOGISTIC, SAS also gives the residual chi-square test (not shown) for steps prior to the last step. Residual chi-square is non-significant for a good final model. Specifically, non-significance in the residual chi-square test means that coefficients of variables not in the model do not differ significantly from 0.0. That is, nonsignificance upholds the validity of stopping the stepwise procedure at the given step.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	89.0105	7	<.0001
Score	91.9912	7	<.0001
Wald	84.6029	7	<.0001

Model fit statistics

The SAS tables below correspond to the "Model Summary" table in SPSS. The colored-in cell for the -2 Log L row and the Intercept and Covariates column corresponds to -2Log Likelihood in SPSS and is used in likelihood ratio tests when comparing models. R-square in SAS corresponds to Cox & Snell R-square in SPSS. Max re-scaled R Square in SAS corresponds to Nagelkerke R Square in SPSS. Both are approximations to R-squared in regression but are "pseudo R-squared" because a variance-explained interpretation does not apply. Here the effect size is weak.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1482.257	1407.247
SC	1487.491	1449.114
-2 Log L	1480.257	1391.247

R-Square	0.0622	Max-rescaled R-Square	0.0948

Cox & Snell Nagelkerke

The classification table

PROC LOGISTIC in SAS supports classification tables in a different way, with somewhat different results, than SPSS. A notable difference is that classifications are "bias corrected," meaning they are made leaving out the current case being classified, causing cell count to differ slightly from SPSS above. Also, more information is provided and in a different format. One thing that does not change is the baseline comparison, which is the percent correct by chance, defined as the percent correct if the most frequent category is always predicted. For the example data this is the 1,074 people out of 1,385 who favor the death penalty, or 77.4%. A good model should have a higher percent correct than 77.4%.

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	14	1058	14	299	77.4	4.5	98.7	50.0	22.0

The SAS classification table above corresponds to the SPSS-format classification table below (created with Excel).

		PREDICTED		
		Non-event	Event	Row total
		cappun = 1 = favor	cappun = 2 =oppose	
OBSERVED	Non-event	1058	14	1072
	Event	299	14	313
	Column total	1357	28	1385

The SAS classification table is requested by adding the CTABLE PPROB=.5 options to the MODEL statement, generating the table below. CTABLE requests a classification table and PPROB sets the classification cutting point. The percent correct is 77.4%, no better than chance, reflecting a weak model. SAS prints associated statistics:

- *Correct*: Correct predictions (of cappun=2) as a percent of sample size = 100*((1058+14)/1385) = 77.4

- *Sensitivity*: Correct predictions of events (of cappun=2) as percent of total predicted events. Thus 100*(14/313)= 4.5. Sensitivity in SAS is the same as in SPSS.
- *Specificity:* Correct predictions of non-events (of cappun=1) as percent of total predicted non-events . Thus 100*(1058/(1072)= 98.7. Specificity in SPSS ROC curve analysis is 1.0 minus what SAS reports as specificity: 1.0 - .987 = .013 .
- *False POS:* Incorrect predictions of events divided by total predictions of events, times 100: 100* (14/28) = 50
- *False NEG:* Incorrect predictions of non-events divided by total predictions of non-events, times 100: 100* (299/1357) = 22

It may be that the model classifies better if a cutting point other than .50 is selected. If the model is re-rerun, omitting the PPROB=.5 option, a much longer classification table is produced. The longer table shows results for a wide range of probability levels for the cutting point. A better cutting point than .50 would be one which increased the percent correct and which has more equal values for sensitivity and specificity.

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.100	313	0	1072	0	22.6	100.0	0.0	77.4	.
0.120	284	251	821	29	38.6	90.7	23.4	74.3	10.4
0.140	278	337	735	35	44.4	88.8	31.4	72.6	9.4
0.160	278	337	735	35	44.4	88.8	31.4	72.6	9.4
0.180	272	337	735	41	44.0	86.9	31.4	73.0	10.8
0.200	202	669	403	111	62.9	64.5	62.4	66.6	14.2
0.220	171	744	328	142	66.1	54.6	69.4	65.7	16.0
...rows omitted...									
0.460	16	1042	30	297	76.4	5.1	97.2	65.2	22.2
0.480	16	1058	14	297	77.5	5.1	98.7	46.7	21.9
0.500	14	1058	14	299	77.4	4.5	98.7	50.0	22.0
0.520	12	1059	13	301	77.3	3.8	98.8	52.0	22.1
...rows omitted...									
0.680	0	1071	1	313	77.3	0.0	99.9	100.0	22.6
0.700	0	1072	0	313	77.4	0.0	100.0	.	22.6

In the expanded classification table above, the row highlighted in red is the previously-discussed table for a cutting point of .5. The row highlighted in green shows that if the cutting point is .2, then sensitivity and specificity are relatively equal (the model is equal in predict cappun=0 and predicting cappun=1), but the percent correct drops to 62.9%. The row highlighted in blue shows that if the cutting point is set at .48, there are marginal benefits: the percent correct goes up to 77.% percent and the gap between sensitivity and specificity is reduced by a small amount. To capture this marginal improvement in a single table, the model may be re-run with CTABLE PPROB=.48 in the MODEL statement.

The performance of the table is rated higher by the pseudo R^2 measures discussed in the previous section and by measures of association discussed in the following section than by the classification table because the pseudo R^2 measures give credit for predictions which are near-misses, whereas the classification table gives no credit.

The association of predicted probabilities and observed responses table

Although PROC LOGISTIC does not directly support output of the classification table, it does output the output of a table which is similar in function though not in coefficients, illustrated below. The higher the coefficients for the measures of association below, the better the logistic model is performing. Warning: These coefficients are not based on the frequencies in the classification table, which would yield quite different measures of association illustrated above for SPSS output.

Rather, the SAS table below is based on predicted response values (probabilities) paired with observed values, not on classification categories. Thus, for instance, where gamma based on the classification table based on predictions (as in SPSS) would be .61, below it is reported as .39, with the former reflecting predictions and the latter reflecting response probabilities from the classification table produced by the CTABLE option in the MODEL statement and partially illustrated below the association of predicted probabilities and observed responses table. (SAS could also generate classification-based measures of association using PROC CROSSTAB.)

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	61.9	Somers' D	0.344
Percent Discordant	27.5	Gamma	0.385
Percent Tied	10.6	Tau-a	0.121
Pairs	335536	c	0.672

Hosmer and Lemeshow test of goodness of fit

PROC LOGISTIC outputs the two tables below, which correspond to the "Hosmer and Lemeshow test" tables in SPSS, discussed and interpreted above. Where SPSS output gives the partition for both levels of the dependent variable and SAS only for the predicted level (here, cappun = 2 = Oppose death penalty), the chi-square

value and the probability level is identical, except for rounding. The finding of non-significance below indicates acceptable fit.

Partition for the Hosmer and Lemeshow Test

Group	Total	cappun = Oppose		cappun = Favor	
		Observed	Expected	Observed	Expected
1	251	26	27.81	225	223.19
2	121	9	14.91	112	106.09
3	368	69	68.76	299	299.24
4	138	35	28.27	103	109.73
5	113	29	26.80	84	86.20
6	134	42	38.12	92	95.88
7	139	49	51.64	90	87.36
8	121	54	56.68	67	64.32

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
6.0598	6	0.4165

Regression diagnostics table

When the INFLUENCE option is in the MODEL statement of PROC LOGISTIC, SAS outputs far more diagnostics than in the "Casewise List" table illustrated above for SPSS. Though too small to be readable in the illustration below due to its being a very wide table with a large number of columns, the columns in the regression diagnostics table are these:

1) Case number
2) Next 8 columns: observed values for each factor level
3) Pearson residual
4) Deviance residual
5) Hat matrix diagonal
6) Next 8 columns: DfBeta influence statistic for each factor level
7) Confidence interval displacement C
8) Confidence interval displacement CBar
9) Delta deviance
10) Delta chi-square


PARTIAL TABLE OUTPUT

Though not illustrated here, these case statistics can be saved using the OUTPUT= statement, and the researcher can plot residuals, DfBeta statistics, or delta deviance against case number to identify highly influence observations in the dataset. Also not illustrated here, SAS outputs a variety of diagnostic plots for residuals and influence statistics, enabling the researcher to identify graphically the distribution of influential observations.

SAS has a large number of other PROC LOGISTIC options not illustrated in this brief example.

A basic binary logistic regression model in STATA

Overview and example

The same example is used as for the SPSS and SAS binary models above, in which capital punishment (cappun: 1=favor, 2=oppose; recoded to 0, 1 below) is predicted from sex, race, and highest educational degree. Note that the "oppose" value is predicted and "favor" is the reference value, following the SPSS example.

This example uses the file "GSS93subset.dta" (for download, see above). The dependent variable, cappun, is predicted from sex (sex, coded 1=male, 2=female), race (race, coded 1=white, 2=black, 3=other), and highest degree earned by the respondent (degree, coded 0=< high school, 1= high school, 2= junior college, 3=bachelor, 4=graduate).

In STATA, binary logistic regression is implemented with the `logistic` or `logit` commands. The two commands are equivalent, but `logistic` reports odds ratios by default (to get coefficients, add "`coef`" as an option) whereas `logit` reports parameter coefficients by default (to get odds ratios, add "`or`" as an option). Odds ratios, as noted below, are equal to the natural logarithm base e raised to the power of the parameter coefficient.

Data setup

First, of course, the example data file is opened in the usual way:

```
. use "C:\Data\GSS93subset.dta", clear
```

The logistic command for binary logistic regression requires that the dependent variable be coded from 0, where 0 indicates non-occurrence of the event of interest. If one has coded (1,2) rather than (0,1), an "outcome does not vary" error message will appear. For the present example, the dependent variable was "cappun", coded (1, 2). Where 1 was favoring the death penalty and 2 was opposing it. Given this coding, the command below was issued:

```
. replace cappun = cappun - 1
```

Unfortunately, the **`replace`** command does not shift the value labels. To do this, we issue the following commands. The first creates a label type called "cappunlabel". The second applies this label type to the now-recoded variable "cappun".

```
. label define cappunlabel 0 "Favor" 1 "Oppose"
. label values cappun cappunlabel
```

Below we follow the SPSS example, where opposing the death penalty is the predicted value of the dependent (1, after recoding).

Stata input

The command below generates the same parameter coefficients as in the SPSS example above.

```
. logistic cappun ib2.sex ib3.race ib4.degree, coef
```

Comments:

- `logistic` is used with the `coef` option to generate parameter coefficients rather than the default odds ratios. Omit `coef` to get odds ratios.
- cappun Is the binary dependent variable. By default, the higher value (oppose death penalty) is the predicted level, the same as in SPSS.
- For ib2.sex, the "I" prefix enters sex as a factor rather than covariate, to parallel the SPSS example. Stata defaults to using the lowest level (female) as the reference level and SPSS uses the highest (male). The "b2" prefix tells Stata that the second level (male) should be the reference level, similar to SPSS.
- The predictor variables race and degree are also declared to be categorical with the "I" prefix and their default reference categories are flipped to make their highest levels the reference,

After logistic or logit is run successfully, a number of postestimation commands are available in Stata, some of which are listed below (type `help logistic` from the Stata command line to see them all).

- `estat classification`: generate a classification table

To obtain some model fit statistics not found in Stata but in a user-supported extension to Stata called "fitstat", install fitstat by issuing the command:

```
net install fitstat.pkg
```

Then issue the `fitstat` command as one would any postestimation command. For additional information, type `help fitstat` after installing.

Stata output

Parameter estimates

The parameter estimates, which are the same as for SPSS and SAS, show sex, race and degree are all significant predictors of opposing the death penalty. However,

being black compared to being "other" was not significant, nor was having a bachelor's degree compared to having a graduate degree. Being male makes one significantly less likely to oppose the death penalty.

```
. logistic cappun ib2.sex ib3.race ib4.degree, coef

Logistic regression                             Number of obs   =       1385
                                                LR chi2(7)      =      89.01
                                                Prob > chi2     =     0.0000
Log likelihood = -695.62328                     Pseudo R2       =     0.0601

------------------------------------------------------------------------------
      cappun |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |
        Male |  -.6119507   .1401528    -4.37   0.000    -.8866452   -.3372562
             |
        race |
       white |   -.784128   .2701361    -2.90   0.004    -1.313585    -.254671
       black |   .3832899   .3126486     1.23   0.220    -.2294901      .99607
             |
      degree |
  Less than HS |  -.8581098   .2688167    -3.19   0.001    -1.384981   -.3312387
   High school |  -1.009401   .2375313    -4.25   0.000    -1.474954   -.5438486
Junior college |  -.9920108   .3588425    -2.76   0.006    -1.695329   -.2886925
      Bachelor |  -.0866659   .2595031    -0.33   0.738    -.5952826    .4219509
             |
         _cons |   .3229485   .3317581     0.97   0.330    -.3272855    .9731825
------------------------------------------------------------------------------
```

Odds ratios

Odds ratios are variable (not model) effect size measures in logistic regression, with values above 1.0 reflecting positive effects and those below 1.0 reflecting negative effects, as discussed below. Effects which are non-significant will have odds ratios closer to 1.0. In Stata, odds ratios are obtained simply by omitting the "`coef`" option from the `logistic` command, as shown in the figure below.

```
. logistic cappun ib2.sex ib3.race ib4.degree

Logistic regression                               Number of obs   =       1385
                                                  LR chi2(7)      =      89.01
                                                  Prob > chi2     =     0.0000
Log likelihood = -695.62328                       Pseudo R2       =     0.0601
```

cappun	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex						
Male	.542292	.0760038	-4.37	0.000	.4120357	.713726
race						
white	.4565176	.1233219	-2.90	0.004	.2688545	.7751715
black	1.467103	.4586878	1.23	0.220	.7949388	2.70762
degree						
Less than HS	.4239627	.1139683	-3.19	0.001	.2503286	.7180337
High school	.3644371	.0865652	-4.25	0.000	.2287892	.5805098
Junior college	.3708303	.1330697	-2.76	0.006	.1835388	.7492426
Bachelor	.9169834	.2379601	-0.33	0.738	.5514067	1.524934
_cons	1.381194	.4582224	0.97	0.330	.7208779	2.646353

The odds ratio is the factor by which the odds of being in the predicted level of the binary dependent variable are multiplies when the independent variable increases one unit. Thus, as sex increases one unit, going from 1=male to 2=female, the odds of opposing the death penalty are reduced by almost half (that is, they are multiplied by .54).

Likelihood ratio test of the model

What SPSS calls the omnibus test an SAS calls the global test of the model is labeled in Stata as "LR chi2" and is a likelihood ratio test comparing likelihood ratio chi-square for the researcher's model with the same coefficient for the null model. Below it is significant at better than the .001 level, indicating that the researcher's model is significantly different from the null model.


```
. logistic cappun ib2.sex ib3.race ib4.degree, coef

Logistic regression                     Number of obs   =       1385
                                        LR chi2(7)      =      89.01
                                        Prob > chi2     =     0.0000
Log likelihood = -695.62328             Pseudo R2       =     0.0601
```

Model fit statistics

While there is no R-squared in logistic regression interpretable as "percent of variance explained by the model," various pseudo-R-squared coefficients have been devised. These are interpreted in terms of "weak", "moderate", or "strong" rather than percent of variance explained. The "Pseudo R2" reported by Stata near the top of its default logistic regression output and shown in the figure above is Cox and Snell pseudo R-square, which here reflects a "weak" level of explanation of attitudes on capital punishment. As the Cox and Snell coefficient does not vary from 0 to 1, a normed version is usually used instead, called Nagelkerke's pseudo R-square (a.k.a., Cragg & Uhler's R-square).

Nagelkerke's coefficient is not supported by Stata itself but can be printed using the user-supported postestimation command `fitstat`.

To install: net install `fitstat.pkg`

To run: enter `fitstat` at the Stata command prompt. Output for the example is shown below. Nagelkerke's pseudo R-square is labeled by its other name, Cragg & Uhler's R-square, in `fitstat` output. For the example data, the same Nagelkerke/Cragg & Uhler coefficient of 0.095 is reported as in SPSS and SAS. The "Maximum Likelihood R2" of 0.062 reported by Fitstat in Stata is the same as "Cox & Snell R-Square" reported by SPSS above. Both reflect a relatively weak relationship.

```
. net install fitstat.pkg
checking fitstat consistency and verifying not already installed...
. fitstat

Measures of Fit for logistic of cappun

Log-Lik Intercept Only:     -740.129     Log-Lik Full Model:         -695.623
D(1374):                    1391.247     LR(7):                       89.011
                                         Prob > LR:                    0.000
McFadden's R2:                 0.060     McFadden's Adj R2:            0.045
Maximum Likelihood R2:         0.062     Cragg & Uhler's R2:           0.095
McKelvey and Zavoina's R2:     0.099     Efron's R2:                   0.064
Variance of y*:                3.651     Variance of error:            3.290
Count R2:                      0.775     Adj Count R2:                 0.006
AIC:                           1.020     AIC*n:                     1413.247
BIC:                       -8547.521     BIC':                        -38.376
```

Note that for binary logistic regression fitstat also reports "D" (deviance) as 1391.247, which is the same as "-2LL" in SAS and as "-2 log likelihood" in the SPSS "Iteration history" table for the last iteration.

The AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) values in the Stata Fitstat output above can be used to compare models, with a lower coefficient representing better fit. SPSS does not output AIC or BIC for binary logistic regression, only for multinomial logistic regression. SAS outputs AIC (see above) but computes it with a slightly different algorithm, yielding for this example an "AIC with intercept and covariates" of 1407.247, compared to Stata's "AIC*n" of 1413.247. SAS generates BIC in its PROC GENMOD procedure, not illustrated in this volume.

The classification table

The `estat classification` command, run after the logistic command has executed successfully, generates the classification table. The percent correctly classified is 77.5%, the same as for SPSS and SAS above. This hit rate, sometimes called the confusion rate, is a measure of overall model effectiveness. However, as a model effect size measure, it bears two caveats:

1. The hit rate must be compared with the rate of hits by chance. The most common definition of "by chance" is the number of hits one would get simply by guessing the most numerous category, calculated for the table

below as 1072/1385 = 77.4% By this standard, the hit rate of 77.5% is only trivially greater. Stata prints out a number of coefficients based on the classification table but the hit rate by chance is not one of them.

2. The classification table rewards correct classifications but gives no credit at all for close but wrong classifications. The classification table underestimates effect size if the researcher considers "being close" is desirable as well as "being correct".

```
. estat classification

Logistic model for cappun

                 ---------- True ----------
Classified |         D            ~D |      Total
-----------+--------------------------+-----------
     +     |        16            14 |         30
     -     |       297          1058 |       1355
-----------+--------------------------+-----------
   Total   |       313          1072 |       1385

Classified + if predicted Pr(D) >= .5
True D defined as cappun != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)    5.11%
Specificity                     Pr( -|~D)   98.69%
Positive predictive value       Pr( D| +)   53.33%
Negative predictive value       Pr(~D| -)   78.08%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    1.31%
False - rate for true D         Pr( -| D)   94.89%
False + rate for classified +   Pr(~D| +)   46.67%
False - rate for classified -   Pr( D| -)   21.92%
--------------------------------------------------
Correctly classified                        77.55%
--------------------------------------------------
```

Classification plot

Classification plots are not directly supported by Stata. A third-party "Classplot" extension is available but as this author could not get it to work even using its own help file examples, it is not discussed here.

Measures of association

In Stata, the immediate form of the `tabulate` command, `tabi`, may be used to generate measures of association:

```
. tabi 16 14 \ 297 1058, all exact
```

Stata outputs such symmetric association measures as Cramer's V, gamma, and Kendall's tau b. The values are identical to those for SPSS above and indicate a generally weak relationship.

```
. tabi 16 14 \ 297 1058, all exact

           |        col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |        16         14 |        30
         2 |       297      1,058 |     1,355
-----------+----------------------+----------
     Total |       313      1,072 |     1,385

          Pearson chi2(1) =  16.5589   Pr = 0.000
 likelihood-ratio chi2(1) =  13.6708   Pr = 0.000
                Cramér's V =   0.1093
                     gamma =   0.6056  ASE = 0.118
           Kendall's tau-b =   0.1093  ASE = 0.033
            Fisher's exact =                 0.000
   1-sided Fisher's exact =                 0.000

. tabi  1058 14 \ 297 16 , all exact

           |        col
       row |         1          2 |     Total
-----------+----------------------+----------
         1 |     1,058         14 |     1,072
         2 |       297         16 |       313
-----------+----------------------+----------
     Total |     1,355         30 |     1,385

          Pearson chi2(1) =  16.5589   Pr = 0.000
 likelihood-ratio chi2(1) =  13.6708   Pr = 0.000
                Cramér's V =   0.1093
                     gamma =   0.6056  ASE = 0.118
           Kendall's tau-b =   0.1093  ASE = 0.033
            Fisher's exact =                 0.000
   1-sided Fisher's exact =                 0.000
```

To obtain the same table format as in SPSS binary logistic regression output above, the second `tabi` command is issued. As can be seen, the measures of association are identical either way.

Hosmer-Lemeshow test

The Hosmer-Lemeshow test is an alternative to the omnibus or global chi-square test of the model. It is often preferred when there are few covariates in the model. A finding of non-significance, as in example output below, indicates the

researcher should not reject the model as poor fit to the data. See further discussion in the SPSS section above.

```
. estat gof, group (10) table

Logistic model for cappun, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)
(There are only 8 distinct quantiles because of ties)
```

Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
1	0.1108	26	27.8	225	223.2	251
2	0.1125	3	3.3	26	25.7	29
5	0.1869	75	80.4	385	379.6	460
6	0.2109	35	28.3	103	109.7	138
7	0.2387	27	24.9	78	80.1	105
8	0.3348	43	39.4	97	100.6	140
9	0.3867	50	52.3	91	88.7	141
10	0.6696	54	56.7	67	64.3	121

```
      number of observations =      1385
            number of groups =         8
Hosmer-Lemeshow chi2(6)      =      3.71
                  Prob > chi2 =      0.7152
```

SPSS, SAS, and Stata all partition the data differently to form the groups required by the Hosmer-Lemeshow procedure. Each yields 8 groups with 6 degrees of freedom. Each reports a non-significant Hosmer-Lemeshow chi-square for these data and thus each leads to the same substantive conclusion (fail to reject the model). However, the different algorithms for partitioning the data do mean that the probability levels differ. SPSS and SAS are quite similar, whereas the Stata version of the test diverges from the other two. See output for SPSS above and for SAS above.

Residuals and regression diagnostics

The `predict` post-estimation command can be used in Stata to generate a variety of statistics related to residuals and regression diagnostics. Options for the `predict` command are shown below.

```
Syntax for predict

        predict [type] newvar [if] [in] [, statistic nooffset rules asif]

    statistic             Description
    -------------------------------------------------------------------------------
    Main
      pr                  probability of a positive outcome; the default
      xb                  linear prediction
      stdp                standard error of the prediction
    * dbeta               Pregibon (1981) Delta-Beta influence statistic
    * deviance            deviance residual
    * dx2                 Hosmer, Lemeshow,& Sturdivant Delta chi-sq influence coef.
    * ddeviance           Hosmer, Lemeshow,& Sturdivant Delta-D influence statistic
    * hat                 Pregibon (1981) leverage
    * number              sequential number of the covariate pattern
    * residuals           Pearson residuals; adjusted for # sharing covariate pattern
    * rstandard           standardized Pearson residuals; adjusted for# sharing
                            covariate pattern
      score               first derivative of the log likelihood with respect to xb
    -------------------------------------------------------------------------------
```

The following examples illustrate some common options. All assume a prior successful logistic regression run since `predict` is a post-estimation command.

1. Create a new variable called "predicted" containing predicted values. The research may set any unused name for this variable. . The new variable is added to the end of the researcher's working dataset. These predicted values are the same as those for SPSS in its "Casewise List" table above, though this table only lists cases with large studentized residuals.

```
.predict predicted
```

2. Create a new variable called "res1" containing Pearson residuals. Studentized residuals, which reflect the change in model deviance if the case is excluded, are not available.

```
. predict res1, residuals
```

3. Create a new variable called "res2" containing standardized Pearson residuals.

```
. predict res2, rstandard
```

4. Create a new variable called "db" containing delta-beta influence statistics, used to spot highly influential cases. Delta-beta is analogous to Cook's D in OLS regression.

```
. predict db, dbeta
```

Note Stata has a third-party program for calculating a type of dfbeta influence statistic, which is specific to each predictor variable. To install, type `. net install ldfbeta`. To run, type `ldfbeta`.

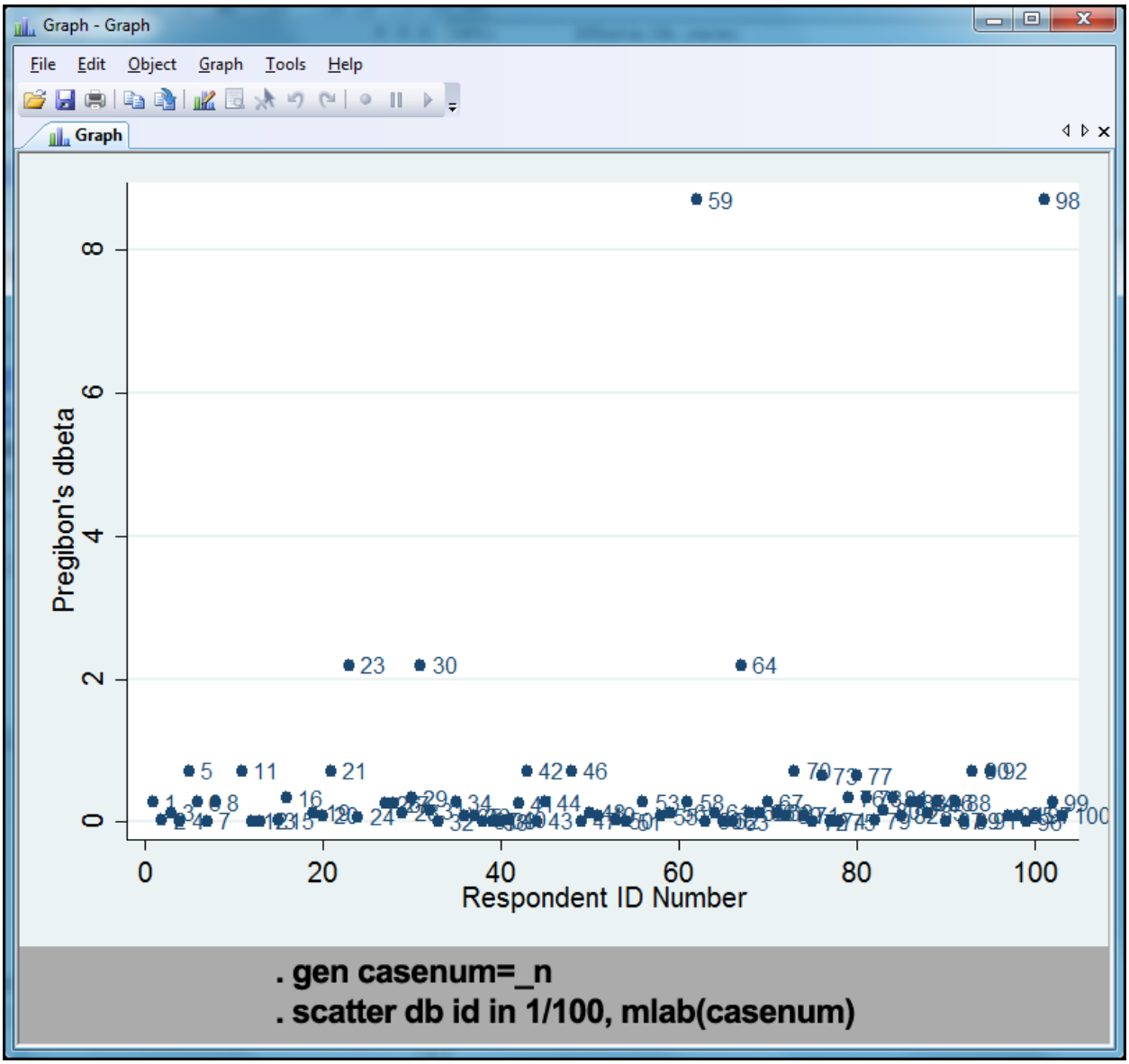
Graphing residuals

Any predicted variable may be plotted. To take just one illustration, the figure below plots the delta-beta influence statistic on the y axis against case number on the X axis. For readability reasons, only the first hundred cases are plotted. It can be seen that cases 59 and 98 are highly influential by the delta-beta criterion. While it is not usually recommended to remove highly influential cases from the analysis as that biases estimates, it may well be that the set of highly influential cases calls for additional analysis in its own right.

To obtain this graph in Stata, the following steps were taken:

1. Run the logistic model as discussed above: **`.logistic cappun ib2.sex ib3.race ib4.degree, coef`**
2. Create a new variable called "db" to contain the delta-beta influence statistic: **`. predict db, dbeta`**
3. Create a new variable called "casenum" containing case numbers for purposes of labeling points in the scatterplot: `. gen casenum=_n`
4. Generate the scatterplot: `. scatter db id in 1/100,mlab(casenum)`

In the `scatter` command, `db` is the delta-beta influence statistic; `id` is the id number; the `in 1/100` term limits the graph to cases 1-100; and `mlab(casenum)` labels the cases by the casenum variable.



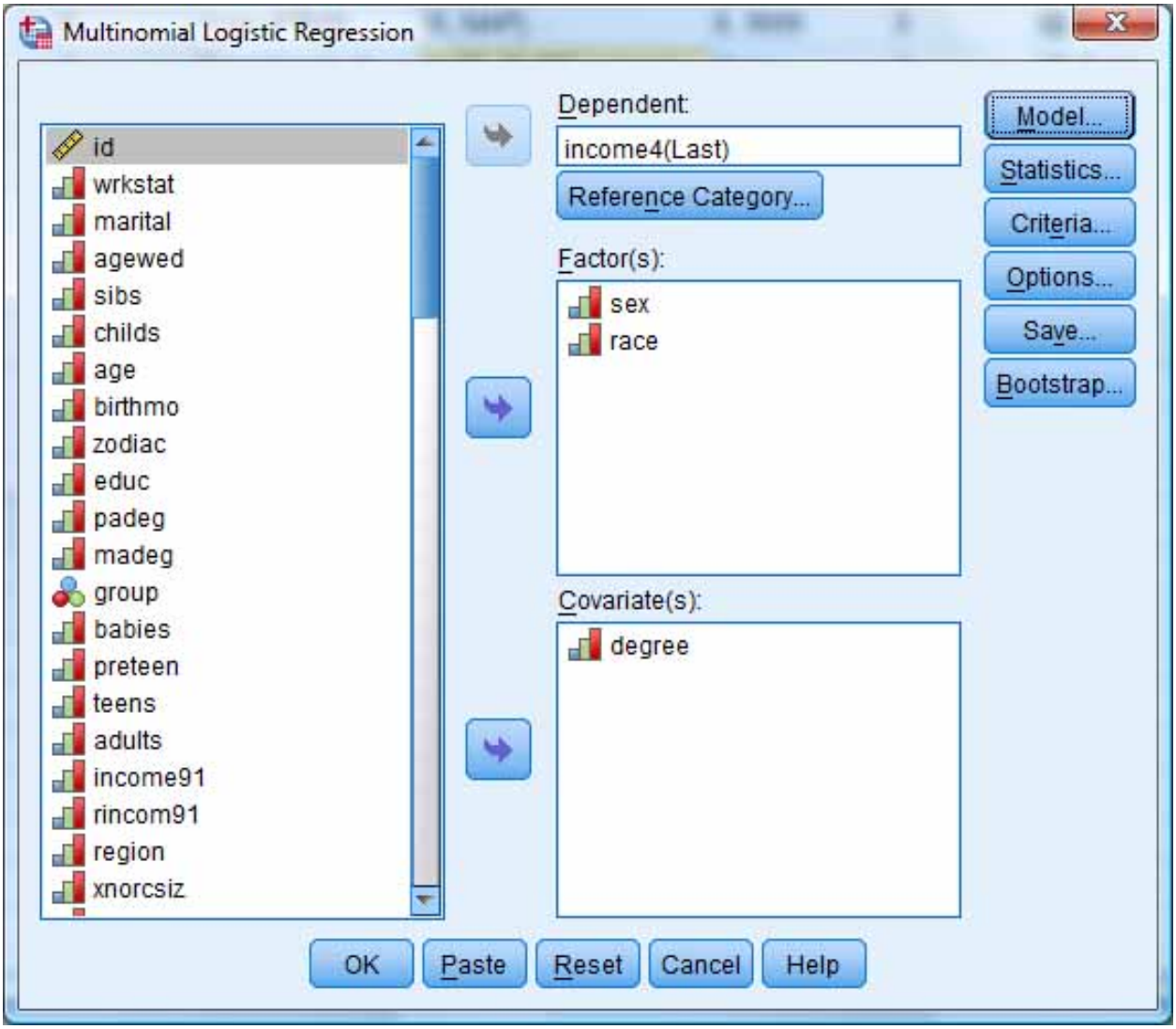
A basic multinomial logistic regression model in SPSS

Example

This example parallels the binary example, using the same data file, but a multinomial variable is used as the dependent variable: income4, which is coded 1 = $24,999 or less, 2 = $25,000 - $39,999, 3 = $40,000 to $59,000, and 4 = $60,000 or more. Income category is predicted from highest educational degree, sex, race, and the interaction of sex by race. While income level is used here for pedagogical

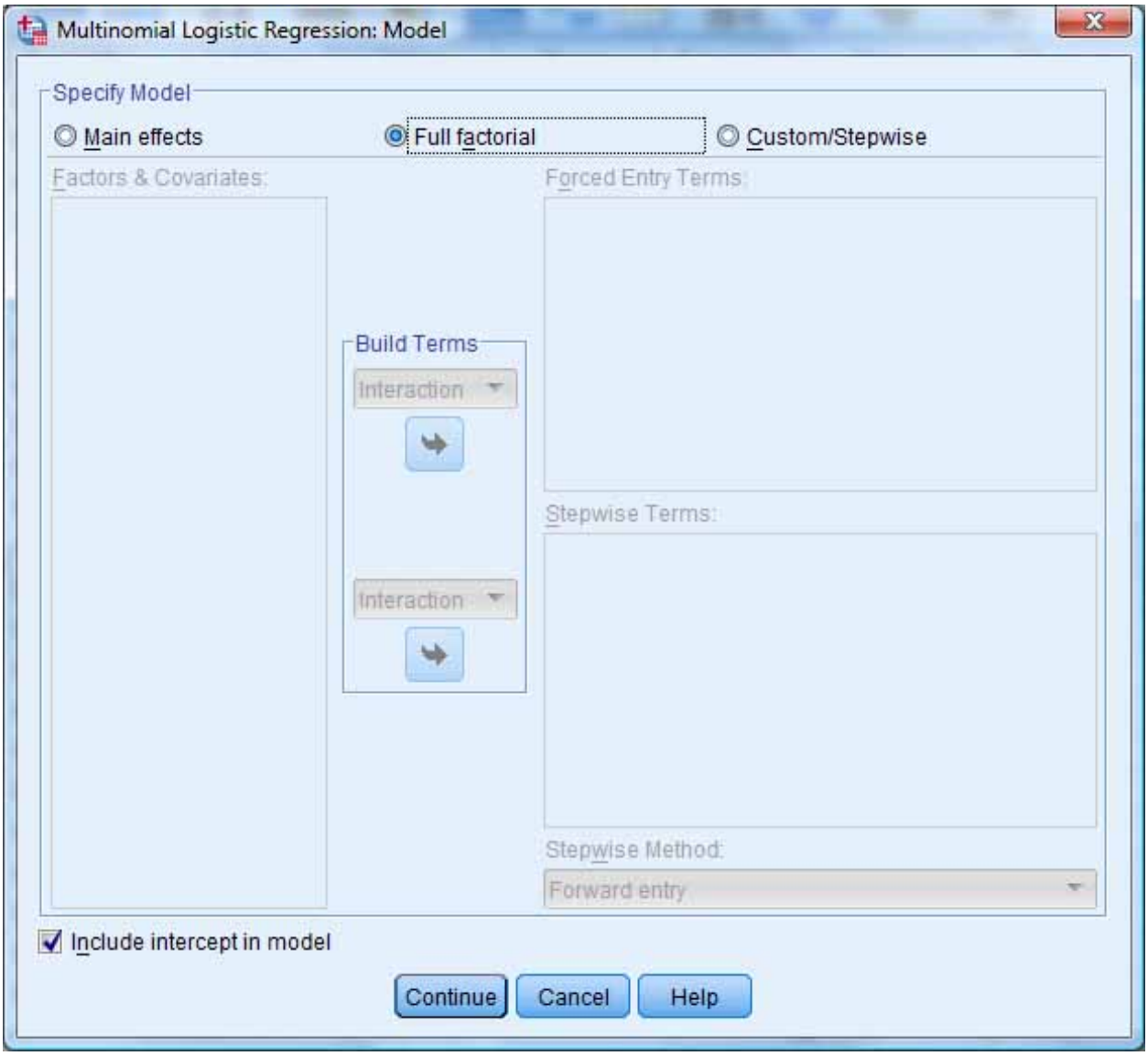
reasons, note that its categories are ordinal and ordinal regression would have more power than multinomial regression (less chance of type II errors, which are false negative findings).

As illustrated below, the SPSS multinomial regression dialog has separate areas for specification of factors and covariates. Here, sex and race are again specified as factors but this time degree is specified as a covariate. By default, the highest value of the dependent variable (income4 = 4 = $60,000 or more) is the reference value, but the researcher may alter this by clicking the "Reference Category" button. In SPSS, select Analyze > Regression > Multinomial Logistic.



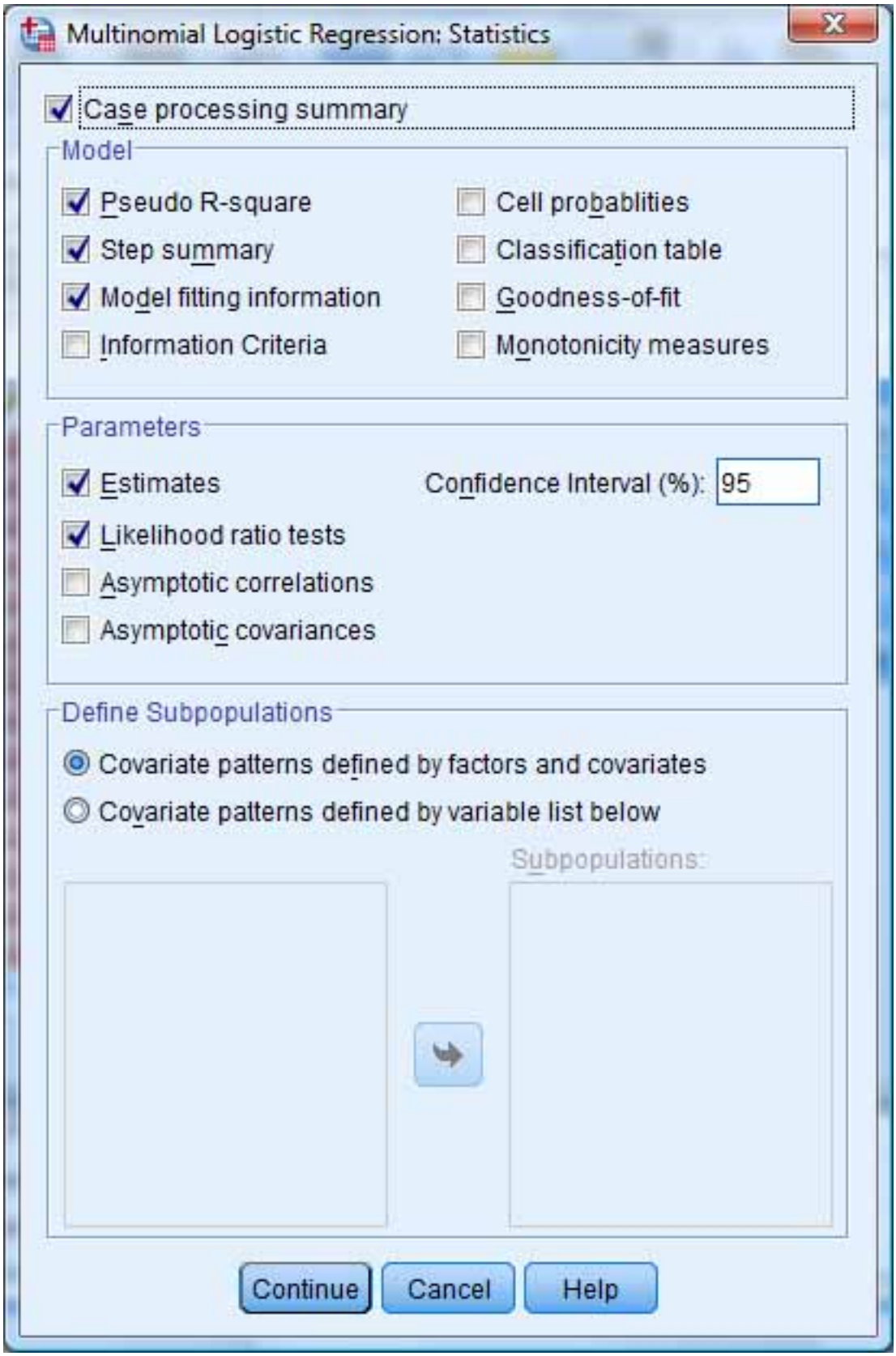
Model

The Model button defaults to modeling the main effects only (here, of sex, race, and degree). For instructional purposes, however, a full factorial model is selected. A full factorial model will model all main effects and all interaction terms except interactions between covariates and factors (not degree*sex, for instance). Here the only interaction will be sex*race. The researcher may also select "Custom/Stepwise" to request a custom model. Note that the “Include intercept in model” checkbox is left checked (the default).



SPSS statistical output

Clicking the "Statistics" button on the main "Multinomial Logistic Regression" dialog shown above leads to the statistics dialog shown below. Default selections for statistical output are shown in the figure below, each of which is then discussed in turn. In this example, however, all statistical outputs in the "Model" panel were selected. Not just default output.



Step summary

A step summary table is output if stepwise logistic regression is specified in the Model dialog, not done for this example. Such a table would show model improvement at each step.

Model fitting information table

In the "Final" row, "-2 Log Likelihood" column, in the figure below, one finds what is variously called the likelihood ratio, deviance, -2LL, model chi-square, or chi-square goodness of fit. A well-fitting model is significant, as it is below, allowing the researcher to conclude that fit is significantly better than the intercept-only model. As goodness of fit improves in successive models, -2LL will decrease. The likelihood ratio is also used as the basis for likelihood ratio test of the difference in fit between nested models, discussed below.

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	581.979	597.910	575.979			
Final	353.463	464.985	311.463	264.516	18	.000

The likelihood ratio test subtracts -2LL for the final model from -2LL for the intercept-only model to get the chi-square value shown highlighted (575.979 – 311.463 = 264.517 (with full precision, the value is 264.516 as reported above). At 18 degrees of freedom, 264.516 is higher than the critical chi-square value at better than the .001 level. The researcher's model is significantly better than the null model. Readers may compare the same calculation in Stata below.

AIC and BIC are not part of default output but appear in the "Model Fitting Information" table above if "Information Criteria" is checked under the Statistics button. Where the likelihood ratio is used to compare successive nested models, AIC and BIC can also be used to compare non-nested models, where lower AIC or BIC is better fit. AIC and BIC are discussed below.

Goodness of fit tests

These are alternative tests of the overall model, comparing the researcher's model to the saturated model, which, by definition, fits the data perfectly. Nonsignificance corresponds to upholding overall fit for the researcher's model, meaning the researcher's model is not significantly different from the saturated model. While the goodness of fit test comes in two flavors, deviance and Pearson chi-square tests, deviance goodness of fit is more commonly reported.

For the example data, goodness of fit is not upheld since deviance chi-square is still significant. For large samples, such as the example file, small differences may be significant and thus misleading. As noted below in further discussion of this test, when this test conflicts with the likelihood ratio test above, the likelihood ratio test is preferred.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	80.709	69	.158
Deviance	93.651	69	.026

Likelihood ratio tests

A table such as that on the following page is output if under the Statistics button, the researcher selects "Likelihood ratio tests" in the "Parameters" area. These likelihood ratio tests test the researcher's full model compared to a reduced model in which the row term is left out.

In the column for -2LL for the intercept, the -2LL coefficient (311.463) is identical to and has the same meaning as in the "Final" row of the "Model Fitting Information" table above. However, for the likelihood ratio tests here, omitting the intercept does not change degrees of freedom between the full and reduced models, so df for the test is 0 and a p value cannot be computed.

Below that, the likelihood ratios for the predictor variables are shown. These likelihood ratio tests compare the researcher's full model with a reduced model leaving out a given effect. By this test, degree is significant in explaining income level, but the interaction of sex*race is not. As explained in table footnote a,

significance for sex and race cannot be differentiated from the overall model effect reflected in the intercept row, which was significant in the "Model Fitting Information" table above. As seen in the "Parameter Estimates" table next, this goes along with the fact that all levels of sex and race were non-significant.

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Intercept	311.463[a]	.000	0	.
degree	530.558	219.095	3	.000
sex	311.463[a]	.000	0	.
race	311.463[a]	.000	0	.
sex * race	321.601	10.137	6	.119

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.

Parameter estimates

For this example, three sets of parameter estimates are output since the dependent variable, income4 had four levels, with the fourth ($60,000 and up) being the reference level. This output is analogous to that seen in the "Variables in the Equation" table for binary regression above and interpretation is similar. The "Sig." column is the significance test for each covariate and each level of each factor or interaction term.

The "Exp(b)" column is the odds ratio, which is the main effect size measure for logistic regression, which the two right-most columns are the confidence limits on

the odds ratio. Coefficients are not computed for reference levels. The coefficients which are computed are interpreted in comparison with reference levels. For example, a unit change in degree (ex., going from < HS to HS level), multiplies the odds of being the in the low income group ($0 - $24,999) rather than the high-income reference group by a factor of .404 (that is, it reduces the odds), controlling for other variables in the model. When the odds ratio = 1.0, the predictor variable has no effect. The lower below 1.0, the more the negative effect on the odds. The higher above 1.0, the more the positive effect. Interpreting parameter estimates and odds ratios is discussed further below.

Parameter Estimates

Total Family Income[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
24,999 or less	Intercept	1.639	.395	17.227	1	.000			
	degree	-.906	.071	160.924	1	.000	.404	.351	.465
	[sex=1]	.321	.685	.219	1	.640	1.379	.360	5.283
	[sex=2]	0[b]	.	.	0	.	.	.	.
	[race=1]	-.080	.391	.042	1	.837	.923	.429	1.985
	[race=2]	.556	.480	1.342	1	.247	1.743	.681	4.463
	[race=3]	0[b]	.	.	0	.	.	.	.
	[sex=1] * [race=1]	-.286	.704	.165	1	.685	.751	.189	2.985
	[sex=1] * [race=2]	-.787	.820	.920	1	.337	.455	.091	2.272
	[sex=1] * [race=3]	0[b]	.	.	0	.	.	.	.
	[sex=2] * [race=1]	0[b]	.	.	0	.	.	.	.
	[sex=2] * [race=2]	0[b]	.	.	0	.	.	.	.
	[sex=2] * [race=3]	0[b]	.	.	0	.	.	.	.
25,000 to 39,999	Intercept	.108	.429	.063	1	.802			
	degree	-.266	.065	16.843	1	.000	.767	.675	.870
	[sex=1]	-.211	.842	.063	1	.802	.810	.156	4.216
	[sex=2]	0[b]	.	.	0	.	.	.	.
	[race=1]	-.112	.424	.069	1	.792	.894	.390	2.053
	[race=2]	.541	.527	1.056	1	.304	1.719	.612	4.828
	[race=3]	0[b]	.	.	0	.	.	.	.
	[sex=1] * [race=1]	.670	.860	.607	1	.436	1.954	.362	10.537
	[sex=1] * [race=2]	-.403	.996	.164	1	.686	.668	.095	4.706
	[sex=1] * [race=3]	0[b]	.	.	0	.	.	.	.
	[sex=2] * [race=1]	0[b]	.	.	0	.	.	.	.
	[sex=2] * [race=2]	0[b]	.	.	0	.	.	.	.
	[sex=2] * [race=3]	0[b]	.	.	0	.	.	.	.
40,000 to 59,999	Intercept	-.336	.460	.533	1	.465			
	degree	-.158	.068	5.393	1	.020	.854	.747	.976
	[sex=1]	-.324	.945	.118	1	.732	.723	.113	4.611
	[sex=2]	0[b]	.	.	0	.	.	.	.
	[race=1]	-.067	.453	.022	1	.883	.935	.385	2.273
	[race=2]	-.622	.675	.850	1	.357	.537	.143	2.015
	[race=3]	0[b]	.	.	0	.	.	.	.
	[sex=1] * [race=1]	.767	.963	.635	1	.426	2.154	.326	14.234
	[sex=1] * [race=2]	1.361	1.146	1.409	1	.235	3.898	.412	36.849
	[sex=1] * [race=3]	0[b]	.	.	0	.	.	.	.
	[sex=2] * [race=1]	0[b]	.	.	0	.	.	.	.
	[sex=2] * [race=2]	0[b]	.	.	0	.	.	.	.
	[sex=2] * [race=3]	0[b]	.	.	0	.	.	.	.

a. The reference category is: 60,000 or more.

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

Pseudo R-square

There is no true counterpart in logistic regression to R-square in OLS regression. The output below parallels that in binary logistic regression, discussed above. Here Nagelkerke's R-square is .174, reflecting a weak (some would say moderate)

effect size for sex, race, and degree as predictors of income level. See further discussion of pseudo R-squared measures below.

Pseudo R-Square

Cox and Snell	.162
Nagelkerke	.174
McFadden	.067

Classification table

Selecting "Classification table" under Statistics generates a classification table similar to that illustrated and discussed above for binary logistic regression. See also further discussion of classification tables below.

Observed and expected frequencies

Selecting this under Statistics generates a table showing observed and predicted cell frequencies for every cell in factor space, including reference levels of the multinomial dependent and of the factors. Covariates, like degree in this example, do not appear in the table though they are predictors of cell count. Cells with the largest residuals in the "Pearson Residual" column are cells for which the model is working least well, which may be informative when diagnosing model shortcomings. Pearson residuals are simply standardized residuals, which are raw residuals divided by an estimate of their standard deviation. Pearson residuals have a mean of 0 and a standard deviation of 1.

Asymptotic correlation matrix

This table (not shown) displays for each level of income4 except the reference level, the correlation of every covariate and every factor level (except reference levels) with every other. For instance, for the lowest level of income4, being white (race = 1) rather than other race (the reference level, race = 3) correlates with degree at the .063 level. There is also a similar “Asymptotic covariance matrix” option (recall covariances are unstandardized correlations).

A basic multinomial logistic regression model in SAS

Example

The same example is used as for the SPSS multinomial model above. The example involves predicting the variable "income4" (four levels of income) from degree, sex, race, and the interaction of sex and race. Sex and race are entered as categorical (factor) variables and degree is entered as a continuous (covariate) variable.

Note that for multinomial logistic regression, PROD CATMOD is used. If one were to run the model below in PROC LOGISTIC in the manner illustrated for binary logistic regression, one would obtain results for ordinal (logistic) regression, which are different. However, PROC LOGISTIC could also implement multinomial logistic regression by adding the LINK=GLOGIT option in the MODEL statement, overriding the default cumulative logit link (not illustrated here).

SAS syntax

The following syntax generated the multinomial logistic regression tables in this section.

```
PROC IMPORT OUT= WORK.multinomiallogistic
      DATAFILE= "C:\\Data\GSS93subset.sav"
      DBMS=SPSS REPLACE;
RUN;
TITLE "PROC CATMOD MULTINOMIAL LOGISTIC REGRESSION EXAMPLE"
JUSTIFY=CENTER;
/* Optional title on each page */
PROC CATMOD DATA=multinomiallogistic ORDER=INTERNAL;
/* Use the work data file from PROC IMPORT */
/* ORDER = INTERNAL keeps categories ordered as coded as in SPSS, but
this is a SAS default*/
DIRECT degree;
/* Above, list covariates and dummy variables, if any */
RESPONSE logits;
/* Above, this specifies a multinomial logistic model */
MODEL income4 = degree sex race sex*race
/ ML=NR CLPARM PARAM=REFERENCE NOPROFILE;
/* Above, a logistic model predicting income4=4, the highest level */
/* ML=NR asks for default maximum likelihood estimates using the
Newton-Raphson algorithm */
/* By default highest-coded category of income4 will be the reference,
as in SPSS */
/* PARAM=REFERENCE requests reference parameterization rather than SAS
default PARAM=EFFECT */
```



```
/* CLPARM requests Wald confidence limits on parameter estimates */
/* NOPROFILE suppresses the profile table */
RUN;
```

SAS statistical output

Overview

SAS output for multinomial regression generally parallels that for SPSS above. To minimize redundancy, the reader is referred to that section for fuller explanation. Note that SAS terminology differs from SPSS terminology. In particular, "likelihood ratio test" has different meanings. There are two major likelihood ratio tests:

1. *Comparing the researcher's model with the null model.* In SPSS language, this is the test of "model fit".
2. *Comparing the researcher's model with the saturated model.* In SPSS language, this is the test of "goodness of fit". In SAS language, it is a "likelihood ratio test" in the "Maximum Likelihood Analysis of Variance" table, for the "Likelihood Ratio" test row.

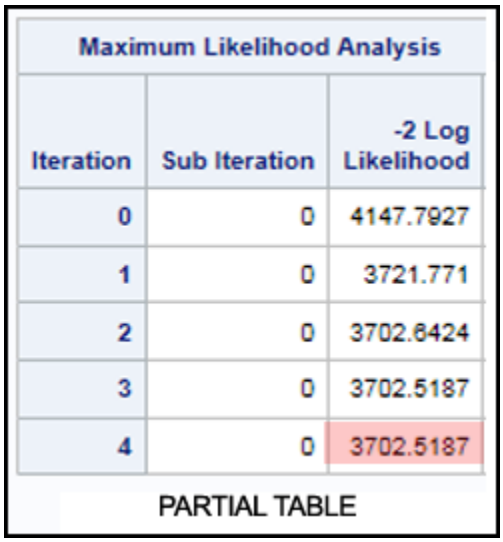
Model fit

In SPSS above, model fit was defined by a likelihood ratio test comparing the researcher's full model with the null (intercept-only) model. This is not part of SAS default output, which instead prints the goodness of fit test discussed in the section which follows.

An approximation of the model fit likelihood ratio test may be undertaken by adding the "ITPRINT" option to PROC CATMOD's MODEL statement. This will print the -2 log likelihood (-2LL) value for each iteration, including the last.

```
MODEL income4 = degree sex race sex*race
/ ML=NR ITPRINT CLPARM PARAM=REFERENCE NOPROFILE;
```

For the full model, the final -2LL is 3702.5187, shown in partial output below. This value is consistent with output from SPSS and Stata.



Maximum Likelihood Analysis

Iteration	Sub Iteration	-2 Log Likelihood
0	0	4147.7927
1	0	3721.771
2	0	3702.6424
3	0	3702.5187
4	0	3702.5187

PARTIAL TABLE

To get -2LL for the null model, the null model is run with this code:

```
PROC CATMOD DATA=multinomiallogistic ORDER=INTERNAL;
RESPONSE logits;
MODEL income4 =
/ ML=NR ITPRINT PARAM=REFERENCE NOPROFILE;
RUN;
```

The -2LL value for the null model obtained in this manner is 3977.092, different from the 3967.034 value for SPSS and Stata. The difference gives a chi-square value of 274.573 as opposed to the 264.516 value in SPSS and Stata. The p value is the same for all three packages, however: 0.000, indicating the full model is significantly different from the null model by the likelihood ratio test.

Goodness of fit tests

The bottom "Likelihood Ratio" row in the table below is the same chi-square value (93.65) and same significance level (.026) as in the row for deviance chi-square in the SPSS "Goodness of Fit" table above. This is an overall fit test for the model, with non-significance upholding the model (not the case here), meaning the researcher's model is not significantly different from the saturated model. Since it is still significant by this test, the example model fails this goodness of fit test. However, for large samples, as here, even small differences may be found to be significant, limiting the utility of this test, which can differ (and does for these data) from the likelihood ratio test, which is preferred and was output and illustrated for SPSS multinomial regression above.

The other rows give Wald chi-square tests for each effect in the model. Here, only the effect of degree is significant. For these tests, significance is desirable. The Wald chi-square coefficients in the SAS Anova table below differ from the corresponding likelihood ratio coefficients in the "Likelihood Ratio Tests" table in SPSS, above, but the findings are the same: of the predictor variables, only degree is significant.

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
Intercept	3	26.63	<.0001
degree	3	166.96	<.0001
sex	3	0.81	0.8470
race	6	9.77	0.1349
sex*race	6	9.53	0.1461
Likelihood Ratio	69	93.65	0.0258

PROC CATMOD does not support information theory measures of goodness of fit such as AIC or BIC, discussed in the SPSS and Stata sections.

Parameter estimates

The “Analysis of Maximum Likelihood Estimates” table, illustrated below, contains the same estimates and same significance levels as in the "Parameter Estimates" table in SPSS, albeit in different format and order, as discussed in greater detail above. This table also shows that degree is a significant predictor of the odds of being in any lower income level compared to being in the highest level. Sex and race are not significant for any level, controlling for other variables in the model. PROC CATMOD does not have an option to exponentiate the parameter estimates into odds ratios, as in the corresponding SPSS table. Of course, this can be done with ease manually by applying the Exp(b) function in a spreadsheet or SAS.

Analysis of Maximum Likelihood Estimates								
Parameter		Function Number	Estimate	Standard Error	Chi-Square	Pr > ChiSq	95% Confidence Limits	
Intercept		1	1.6388	0.3949	17.23	<.0001	0.8649	2.4127
		2	0.1076	0.4293	0.06	0.8021	-0.7338	0.9489
		3	-0.3358	0.4597	0.53	0.4652	-1.2369	0.5653
degree		1	-0.9059	0.0714	160.93	<.0001	-1.0459	-0.7660
		2	-0.2657	0.0647	16.84	<.0001	-0.3926	-0.1388
		3	-0.1580	0.0680	5.39	0.0202	-0.2913	-0.0246
sex	Male	1	0.3210	0.6854	0.22	0.6395	-1.0223	1.6644
	Male	2	-0.2110	0.8417	0.06	0.8021	-1.8607	1.4388
	Male	3	-0.3242	0.9453	0.12	0.7316	-2.1770	1.5285
race	white	1	-0.0803	0.3907	0.04	0.8372	-0.8460	0.6855
	white	2	-0.1117	0.4239	0.07	0.7922	-0.9425	0.7192
	white	3	-0.0669	0.4530	0.02	0.8826	-0.9546	0.8209
	black	1	0.5556	0.4797	1.34	0.2468	-0.3846	1.4959
	black	2	0.5415	0.5270	1.06	0.3042	-0.4914	1.5744
	black	3	-0.6222	0.6749	0.85	0.3565	-1.9450	0.7005
sex*race	Male white	1	-0.2859	0.7038	0.17	0.6846	-1.6654	1.0936
	Male white	2	0.6700	0.8596	0.61	0.4358	-1.0149	2.3549
	Male white	3	0.7674	0.9634	0.63	0.4257	-1.1208	2.6556
	Male black	1	-0.7869	0.8203	0.92	0.3374	-2.3946	0.8208
	Male black	2	-0.4033	0.9960	0.16	0.6855	-2.3554	1.5488
	Male black	3	1.3605	1.1460	1.41	0.2351	-0.8856	3.6066

PROC CATMOD also does not output Nagelkerke and other pseudo-R-squared measures as in SPSS above, nor AIC, BIC, and -2LL with respect to the intercept-only model (SPSS "Model Fitting Information" table above), nor with respect to reduced models dropping one effect (SPSS "Likelihood Ratio Tests" table above).

Pseudo R-Square

PROC CATMOD does not support pseudo- R^2 measures. However, PROC LOGISTIC, which does support such measures, could also implement multinomial logistic

regression by adding the LINK=GLOGIT option in the MODEL statement to run a generalized logits model, overriding the default cumulative logit link (not illustrated here).

Classification table

PROC CATMOD does not support the CTABLE option found in PROC LOGISTIC nor classification tables. However, PROC LOGISTIC, which does support CTABLE, could also implement multinomial logistic regression by adding the LINK=GLOGIT option in the MODEL statement to run a generalized logits model, overriding the default cumulative logit link (not illustrated here).

Observed and predicted functions and residuals

Adding the PREDICT option to the MODEL statement in PROC CATMOD generates a table, only the starting portion of which is shown here. PRED= PROB is a synonym for PREDICT. For each combination of sex, race, and degree, the observed and predicted function values are shown for each of the three prediction equations created for the four-level dependent variable, income4.

Maximum Likelihood Predicted Values for Response Functions

sex	race	degree	Function Number	Observed		Predicted		Residual
				Function	Standard Error	Function	Standard Error	
Male	white	Less than HS	1	1.92991	0.356707	1.593685	0.159257	0.336224
			2	0.747214	0.404651	0.454922	0.173962	0.292292
			3	-0.25131	0.503953	0.040564	0.187312	-0.29188
Male	white	High school	1	0.638489	0.180312	0.687751	0.127948	-0.04926
			2	0.412532	0.188045	0.189204	0.136536	0.223328
			3	0.138836	0.199487	-0.11739	0.14644	0.256231
Male	white	Junior college	1	0	0.471405	-0.21818	0.132583	0.218183

Partial Table

The same general information is also output in probability form:

Maximum Likelihood Predicted Values for Probabilities

sex	race	degree	income4	Observed Probability	Observed Standard Error	Predicted Probability	Predicted Standard Error	Residual
Male	white	Less than HS	24,999 or less	0.6392	0.0488	0.5764	0.0296	0.0628
			25,000 to 39,999	0.1959	0.0403	0.1846	0.0211	0.0113
			40,000 to 59,999	0.0722	0.0263	0.122	0.0168	-0.05
			60,000 or more	0.0928	0.0295	0.1171	0.015	-0.024
Male	white	High school	24,999 or less	0.341	0.0293	0.3911	0.0233	-0.05

Partial Table

If the option PRED=FREQ is substituted for PREDICTED (or PRED=PROB) in the MODEL statement, the same information is displayed in frequency terms:

Maximum Likelihood Predicted Values for Frequencies

sex	race	degree	income4	Observed Frequency	Observed Standard Error	Predicted Frequency	Predicted Standard Error	Residual
Male	white	Less than HS	24,999 or less	62	4.729813	55.90854	2.87137	6.091465
			25,000 to 39,999	19	3.908753	17.90274	2.04399	1.097264
			40,000 to 59,999	7	2.548499	11.82949	1.627527	-4.82949
			60,000 or more	9	2.857437	11.35924	1.45182	-2.35924
Male	white	High school	24,999 or less	89	7.658416	102.0671	6.073262	-13.0671

Partial Table

Regardless of the format of the observed/predicted/residual table, the category combinations with the lowest residuals are those for which the model is working best and combinations with the highest residuals are those poorly handled by the model.

Correlation matrix of estimates

This table (not shown) is displayed if the CORRB option is in the MODEL statement of PROC CATMOD. The "Correlation Matrix of the Maximum Likelihood Estimates" table displays for each level of income4 except the reference level, the correlation of every covariate and every factor level (except reference levels) with every other. For instance, for the lowest level of income4, being white (race = 1) rather than other race (the reference level, race = 3) correlates with degree at the .063 level. There is also a similar covariance matrix invoked by including the COVB

option in the MODEL statement (recall covariances are unstandardized correlations).

A basic multinomial logistic regression model in STATA

Example

The same example is used as for the SPSS multinomial model above, and for SAS. The example involves predicting the variable "income4" (four levels of income) from degree, sex, race, and the interaction of sex and race. Sex and race are entered as categorical (factor) variables and degree is entered as a continuous (covariate) variable.

The example data file is opened in the usual way:

```
. use "C:\Data\GSS93subset.dta", clear
```

Stata data setup

The mlogit command for multinomial logistic regression requires that the dependent variable be coded from 0, where 0 indicates non-occurrence of the event of interest. If the researcher has coded from 1 rather than from 0, an "outcome does not vary" error message will appear. For the present example, the dependent variable, income4, was coded from 1. To remedy this, the following Stata command is issued:

The logistic command for binary logistic regression requires that the dependent variable be coded from 0, where 0 indicates non-occurrence of the event of interest. If one has coded (1,2) rather than (0,1), an "outcome does not vary" error message will appear. For the present example, the dependent variable was "cappun", coded (1, 2). Where 1 was favoring the death penalty and 2 was opposing it. Given this coding, the command below was issued:

```
. replace income4 = income4 - 1
```

As mentioned previously, the **replace** command does not shift the value labels. To do this, we issue the following commands. The first creates a label type called "incomelabel". The second applies this label type to the now-recoded variable "income4".

```
. label define incomelabel 0 "24,999 or less" 1 "25,000
to 39,999" 2 “40,000 to 59,999” 3 “60,000 or more”

. label values income4 incomelabel
```

Stata syntax

For multinomial logistic regression in Stata, the `mlogit` command is used. Multinomial logit regression is identical to multinomial logistic regression in Stata. The operative command for this example is given below. This command is entered in a single command line.

```
mlogit income4 degree ib2.sex ib3.race ib2.sex#ib3.race,
baseoutcome(4) baselevels
```

where…

- `mlogit`: Invokes multinomial logistic regression
- `income4`: The first-listed variable is the dependent variable
- `degree`: A continuous predictor (a covariate)
- `ib2.sex ib3.race`: The “i” prefix tells Stata that sex and race are categorical variables (factors), to be split into constituent indicator variables. The “b#” prefixes declare level 2 (female) to be the reference category for sex and level 3 (“other race”) to be the reference category for race. This makes Stata conform to the SPSS and SAS practice of having the highest-coded value be the reference value, whereas by default Stata uses the lowest-coded value.
- `ib2.sex#ib3.race`: This is how the sex by race interaction term is added to the model.
- `, baseoutcome(4) baselevels`: The comma tells Stata that options for the mlogit command follow. The `baseoutcome(4)` option sets the highest level of the dependent variable (income4) as the reference (base) level. This makes Stata conform to SPSS and SAS practice, whereas by default Stata uses the level with the highest frequency on the logic that it is the most informative reference. The `baselevels` option causes the reference levels to be included in the parameter estimates table, overriding the default, which is to omit them. While estimates are not computed for reference levels, having them in the table reminds the researcher of what has been requested.

Stata statistical output

Overview

Stata output for multinomial regression generally parallels that for SPSS above and for SAS above. To minimize redundancy, the reader is referred to these sections for fuller explanation.

Base Stata is limited in the goodness of fit coefficients it reports. For instance, among the pseudo-R-square measures, it reports only McFadden's R-square (labeled "Pseudo R2" near the top of default mlogit output). For this reason, use of the "fitstat" extension command is recommended (see previous discussion above).

Model fit

Model fit refers to a likelihood ratio test comparing the researcher's full model with the intercept-only (null model). In the fitstat output below, the log likelihood (LL) for the full model is minus 1851.259 and LL for the intercept-only model is minus 19.83.517. The difference is: -1983.517 – (-1851.259) = -132.258. LL is multiplied by -2 to make it conform to a chi-square distribution. Thus -2LL = -2* -132.58 = 264.516, which is the likelihood ratio with 18 degrees of freedom reported as LR(18) in fitstat output below. That its probability is 0.000 means that the researcher's full model is significantly different from the null model. LR(18) here in Stata is the same as likelihood ratio chi-square in the SPSS "Model Fitting Table" above.

Likelihood ratio chi-square is reported near the top of default mlogit output, illustrated below, and also appears in fitstat output above. This is identical to the likelihood ratio test for the final model in the SPSS "Model Fitting Information" table discussed previously above and in the .

```
. mlogit income4 degree ib2.sex ib3.race ib2.sex#ib3.race ,baseoutcome(4) baselevels

Multinomial logistic regression                 Number of obs   =       1496
                                                LR chi2(18)     =     264.52
                                                Prob > chi2     =     0.0000
Log likelihood = -1851.2594                     Pseudo R2       =     0.0667
```

The same model fitting value (264.516) is reported by fitstat:

```
. fitstat

Measures of Fit for mlogit of income4

Log-Lik Intercept Only:     -1983.517     Log-Lik Full Model:        -1851.259
D(1444):                     3702.519     LR(18):                      264.516
                                          Prob > LR:                     0.000
McFadden's R2:                 0.067     McFadden's Adj R2:             0.040
Maximum Likelihood R2:*        0.162     Cragg & Uhler's R2:**          0.174
Count R2:                      0.406     Adj Count R2:                  0.026
AIC:                           2.544     AIC*n:                      3806.519
BIC:                       -6853.916     BIC':                       -132.926
```

* Cox and Snell R-square
** Nagelkerke R-square

The `estat ic` postestimation command also generates the same model LL (-1851.259), shown below.

```
. estat ic

Akaike's information criterion and Bayesian information criterion

------------------------------------------------------------------------------
       Model |     Obs  ll(null)  ll(model)     df         AIC        BIC
-------------+----------------------------------------------------------------
           . |    1496 -1983.517 -1851.259     21    3744.519     3856.04
------------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

AIC and BIC

AIC is the Akaike information criterion. BIC is the Bayesian information criterion. Both are measures of model error, where lower is less error. The coefficients have no intrinsic meaning, but when comparing models, lower is better fit.

AIC equals -2LL+ 2df. For these data, AIC = -2* -1851.259 + 2*21 = 3744.519.

BIC equals -2LL + ln(N)*df. For these data, BIC = -2*-1851.259 + ln(1496)*21 = 3856.04.

Comparing AIC and BIC in SPSS

In the SPSS "Model Fitting Criteria" table above, rather different values for AIC and BIC are reported (SAS PROC CATMOD does not output AIC or BIC). In SPSS output, AIC is 353.463 and BIC is 464.985.

The Stata and SPSS values appear quite different but are related. The difference has to do with how the log likelihood (LL) is computed, not with the formulas for AIC and BIC. The log likelihood test of the full model versus the null model involves differencing -2LL for the two models, then using the difference chi-square to find the p value of the difference. If significant, the researcher's model differs significantly from the intercept-only model. SPSS and SAS arrive at the same -2LL difference value based on different ways of calculating LL, and therefore the likelihood ratio test is the same for both.

- In Stata: 264.516 = -2((LL_{null}) – (LL_{model})) = -2*((-1983.517)-(-1851.259))
- In SPSS: 264.516 = -(2LL_{null} – 2LL_{model}) = (575.979 – 311.463)

Even though the likelihood ratio test is the same, the LL values used in the AIC and BIC formulas differ, so therefore the AIC and BIC values differ even though SPSS uses the same formulas as given above for AIC and BIC in Stata. As the AIC and BIC values are not used in their own right but only when comparing models, with lower being better, the same conclusions will be reached with either the Stata or SPSS values.

Pseudo R-square

Pseudo R-square is an overall effect size measure for a multinomial logistic regression model. In fitstat output above, the same three pseudo R-square measures (Cox and Snell, Nagelkerke, and McFadden pseudo R-square) are reported, at the same levels, as in SPSS output. Of the three, Nagelkerke's pseudo R-square is most often reported, labeled as "Cragg & Uhler's R2" by fitstat. At .174, it reflects a weak or weak-to-moderate relationship of the predictor variables to the dependent variable. Note it cannot be interpreted as "percent of variance explained" but rather is reported in weak-moderate-strong terms.

Goodness of fit test

The model goodness of fit test is a test of the difference between the researcher's full model and the saturated model. The saturated model by definition yields estimates which perfectly fit the data. Ideally, the researcher's model would not be significantly different. The likelihood ratio test comparing -2LL for the

researcher's model and the saturated model should be non-significant. However, for large datasets, small differences may well be significant. When the goodness of fit test conflicts with the model fit test described above, the latter is preferred. Perhaps for this reason, Stata reports the model fit test (comparing the null model) rather than the goodness of fit test (comparing the saturated model).

Likelihood ratio tests

It is also possible to compare the researcher's full model with a reduced model dropping one of the effects. A table showing the significance of dropping each effect may be used to understand which are significant and which are not. In SPSS, this is the "Likelihood Ratio Tests" table above. An alternative method of assessing individual effects in SAS is its "Maximum Likelihood Analysis of Variance" table above. However, judgments about individual effects may also be made using the significance levels associated with model parameters, discussed in the section which follows. Stata users generally take this path.

Parameter estimates

In multinomial regression there are four sets of parameter estimates, including four intercepts (the "_cons" rows for the constant), one for each level of the dependent variable (income4) . The parameter estimates (coefficients) are the same as previously reported for SPSS and SAS. For each level of income, they show only degree to be a significant predictor. Race, sex, and the sex*race interaction are not significant.

income4	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
24_999_or_less						
degree	-.9059341	.0714144	-12.69	0.000	-1.045904	-.7659645
sex						
Male	.3210323	.6854152	0.47	0.640	-1.022357	1.664421
Female	0	(base)				
race						
white	-.0802624	.3907154	-0.21	0.837	-.8460506	.6855258
black	.5556336	.4796848	1.16	0.247	-.3845313	1.495798
other	0	(base)				
sex#race						
Male#white	-.2859292	.703858	-0.41	0.685	-1.665465	1.093607
Male#black	-.7868648	.8202631	-0.96	0.337	-2.394551	.8208212
_cons	1.638845	.3948557	4.15	0.000	.8649418	2.412748
PARTIAL OUTPUT						
40_000_to_59_999						
degree	-.1579585	.0680205	-2.32	0.020	-.2912763	-.0246407
sex						
Male	-.3242184	.9453058	-0.34	0.732	-2.176984	1.528547
Female	0	(base)				
race						
white	-.0668662	.4529532	-0.15	0.883	-.9546382	.8209058
black	-.6222411	.6750674	-0.92	0.357	-1.945349	.7008668
other	0	(base)				
sex#race						
Male#white	.7674216	.9634078	0.80	0.426	-1.120823	2.655666
Male#black	1.36052	1.146102	1.19	0.235	-.8857976	3.606838
_cons	-.3357733	.4597457	-0.73	0.465	-1.236858	.5653118
60_000_or_more	(base outcome)					

Odds ratios/ relative risk ratios

In the context of multinomial logistic regression, Stata calls odds ratios "relative risk ratios" (RRR). They are computed manually by applying the exp(b) function to any given parameter estimate as b. Stata makes it simple to compute relative risk

coefficients simply by adding the option "rrr" to the mlogit command line. As illustrated below, the first column then becomes relative risk ratios rather logistic regression coefficients. These are the same as odds ratios reported as "exp(B)" in the SPSS parameter estimates table earlier above.

income4	RRR	Std. Err.	z	P>|z|	[95% Conf.	Interval]
24_999_or_less						
degree	.4041642	.0288631	-12.69	0.000	.3513741	.4648853
sex						
Male	1.37855	.9448791	0.47	0.640	.3597461	5.282615
Female	1	(base)				
race						
white	.9228741	.3605812	-0.21	0.837	.4291063	1.984815
black	1.743045	.8361121	1.16	0.247	.6807697	4.462899
other	1	(base)				
sex#race						
Male#white	.7513158	.5288197	-0.41	0.685	.1891026	2.985022
Male#black	.4552699	.3734411	-0.96	0.337	.0912136	2.272365
_cons	5.149217	2.033198	4.15	0.000	2.374868	11.1646

PARTIAL OUTPUT

Given degree as a predictor variable, a unit change in degree (ex., going from < HS to HS level), multiplies the odds of being the in the low income group ($0 - $24,999) rather than the high-income reference group by a factor of .404 (that is, it reduces the odds), controlling for other variables in the model. When RRR = 1.0, the predictor variable has no effect. The lower below 1.0, the more the negative effect on the odds. The higher above 1.0, the more the positive effect.

Classification table

Unlike its counterparts in SPSS and SAS, the mlogit command in Stata does not support classification tables. Creation of such tables requires manually entering a series of commands to create the row and column variables for the table, then issuing the tab command. These steps are not discussed in this volume but may be found in numerous places on the Internet by simple search for "mlogit classification table".

Observed and expected frequencies

The `mlogit` command supports the postestimation `predict` command discussed above with reference to binary logistic regression. The reader is referred to this section for discussion of observed and expected values and of residuals. Residual analysis, of course, is central to understanding where the model works most and least well.

Asymptotic correlation matrix

This table (not shown) is displayed if the postestimation command `estat vce, correlation` is issued. The resulting table, labeled "Correlation matrix of coefficients of mlogit model", displays for each level of income4 except the reference level, the correlation of every covariate and every factor level (except reference levels) with every other. For instance, for the lowest level of income4, being white (race = 1) rather than other race (the reference level, race = 3) correlates with degree at the .063 level. A similar covariance matrix may be created by issuing the postestimation command `estat vce, covariance` (or just `estat vce` as it is the default).

ROC curve analysis

Overview

ROC curve analysis stands for "receiver operating characteristic" curve analysis, a term derived from radar-related signal detection theory in World War II. The curve in question is the function plotted when "sensitivity" is plotted on the Y axis against "specificity" on the X axis. In binary logistic regression, one of the dependent variables is predicted (it is the "event") and the other is the reference level (it is the "non-event"). Usually the even it coded 1 and the non-event is coded 0.

Sensitivity refers to the rate of classification of events, equal to the number of correctly classified events divided by the total number of events. Specificity is the rate of classification of non-events, equal to the number of correctly classified non-events divided by the total number of non-events. The sensitivity at any specificity may be calculated from the classification assignment probabilities in

the classification table. Sensitivity and specificity were also defined and illustrated above, in the section on classification tables in SAS binary logistic regression.

There are two major uses of ROC curves in logistic regression..

Comparing models

ROC curves may be plotted for alternative models as an aid in selecting the desirable model, as illustrated in SPSS, SAS, and Stata sections below. Note, however, that ROC curve analysis is based on classification information, where classifications are based on the computed case probability in relation to the researcher-set classification cutting point. This means two models will be ranked the same if they classify the same, even if one model has many more near misses than the other. The ROC curve, like the classification table, gives no credit for how close the computed probability is to the cutting point, only whether the resulting classification is correct or not. By comparison, pseudo R^2 measures do give credit for distance of the computed probability to the cutting point. ROC curve analysis is appropriate when the research purpose is prediction. When the research purpose is causal modeling, ROC curve analysis should be used in conjunction with pseudo R^2 measures and analysis of residuals.

Optimal classification cutting points

The default classification cutting point of .5 for binary outcomes does not necessarily result in optimal classifications. When the research purpose is prediction, the optimal cutting point may be seen as the one which returns the highest percent of correct classifications. When the research purpose is modeling, however, the optimal classification cutting point may be seen as the one which has relatively equal sensitivity and specificity (see definition above), meaning the model predicts “1” outcomes and “0” outcomes equally well, even if this does not result in the highest overall percent correct. At other times a compromise between the two criteria may be warranted. Note also that selecting cutting points in this manner is a data-driven solution, specific to the given data sample at hand, and may not generalize to other samples. See Meyers, Gamst, & Guarino for further illustration (2013: 545-556, 567-570).

Example

The examples below use the data file auto.sav, one of the SPSS example files. The file contains attribute information on 74 makes of cars sold in 1978. See above for access.

Two models are compared, both seeking to predict the binary dependent variable, foreign (labeled origin), coded 0 = Domestic, 1 = Foreign:

Model 1 predicts region of origin from miles per gallon (mpg), automobile weight (weight), and automobile length (length), on the hypotheses that US domestic cars in 1978 were long, heavy gas-guzzlers.

Model 2 predicts region of origin using two additional predictors, turning radius (turn) and gear ratio (gear_ratio), on the hypothesis that foreign cars were not just smaller and more efficient, but also exhibited sportier handling.

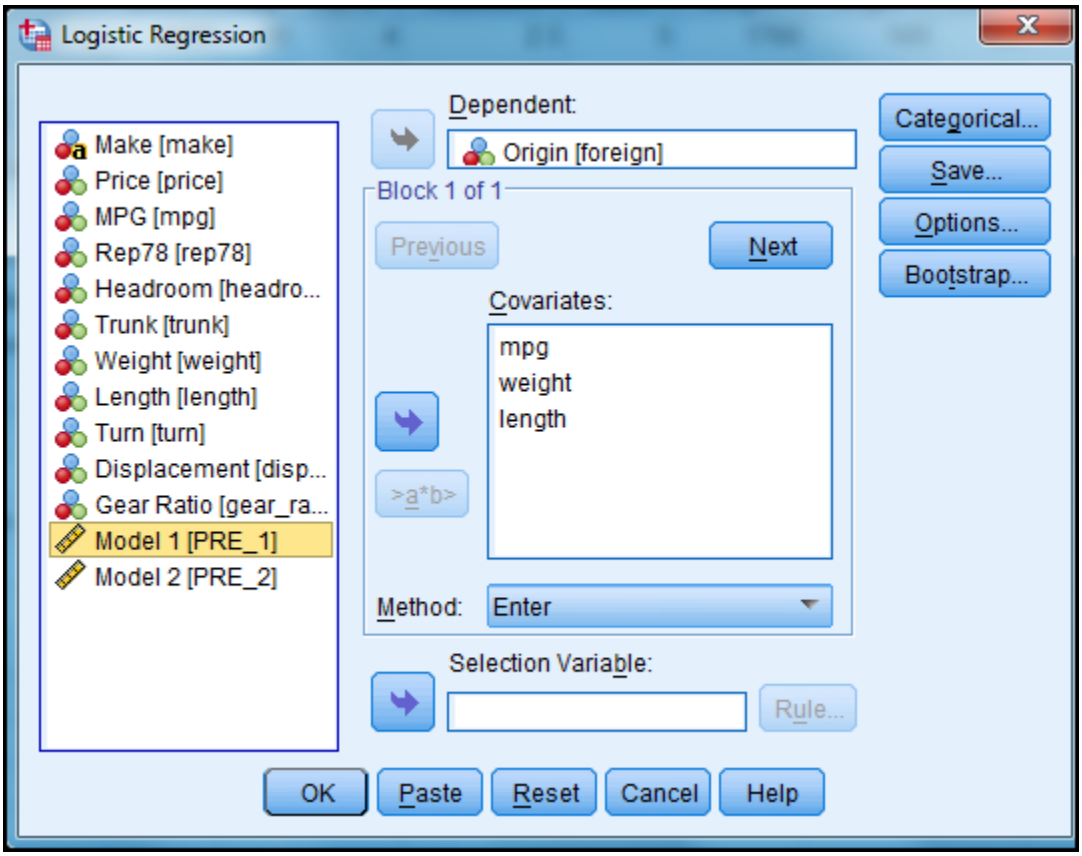
All predictor variables were continuous in data level.

SPSS

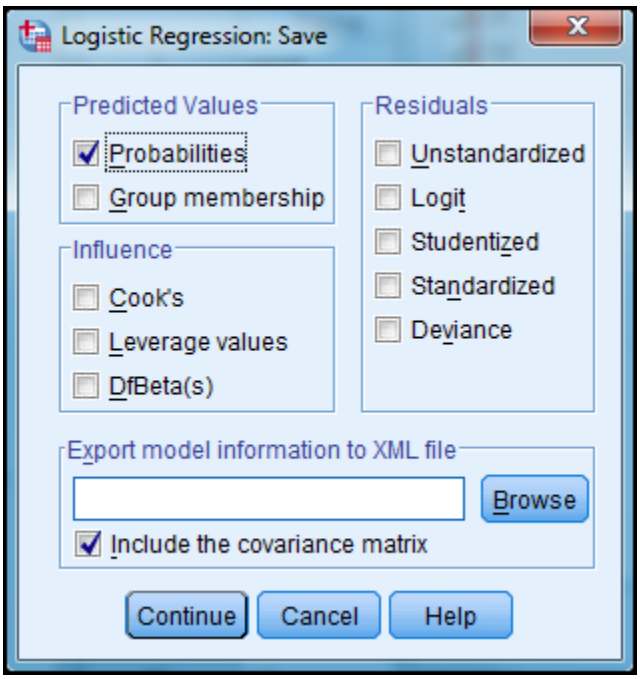
Comparing models

Load auto.sav into SPSS and follow these steps:

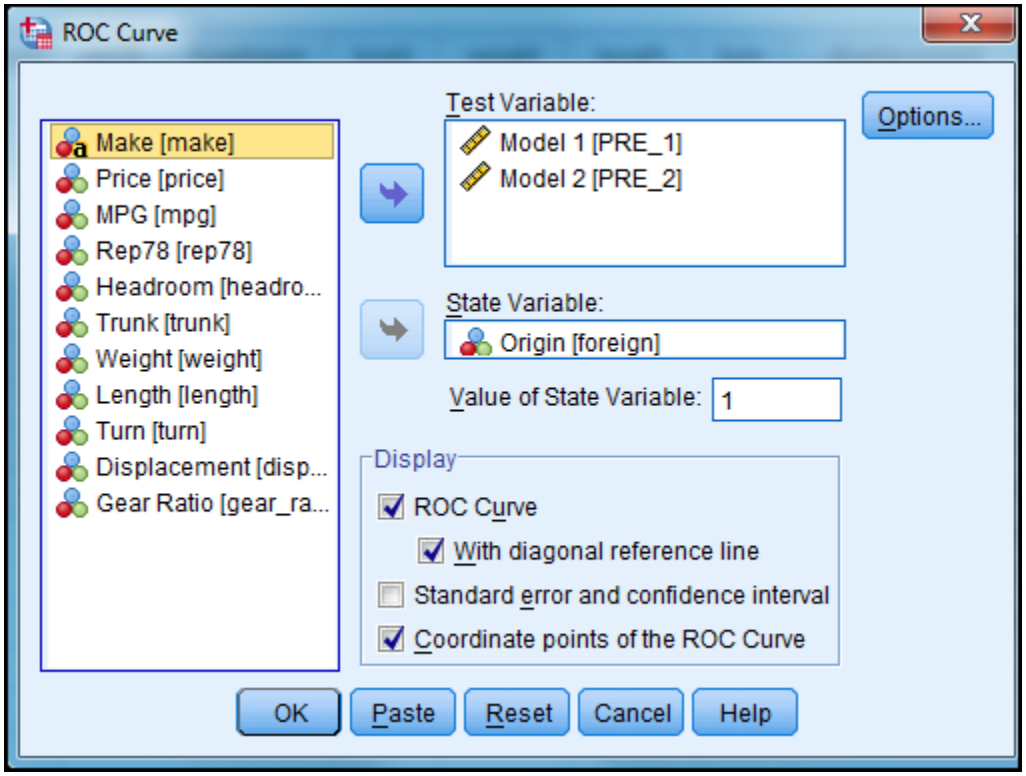
1. Select Analyze > Regression > Binary Logistic.
2. In the ensuing "Logistic Regression" dialog, enter origin as "Dependent" and enter mpg, weight, and length as "Covariates", as in the figure below. Note categorical predictors are also permitted if declared categorical by clicking the "Categorical" button.



3. Click the Save button and check "Probabilities" as in the figure below, then click Continue.

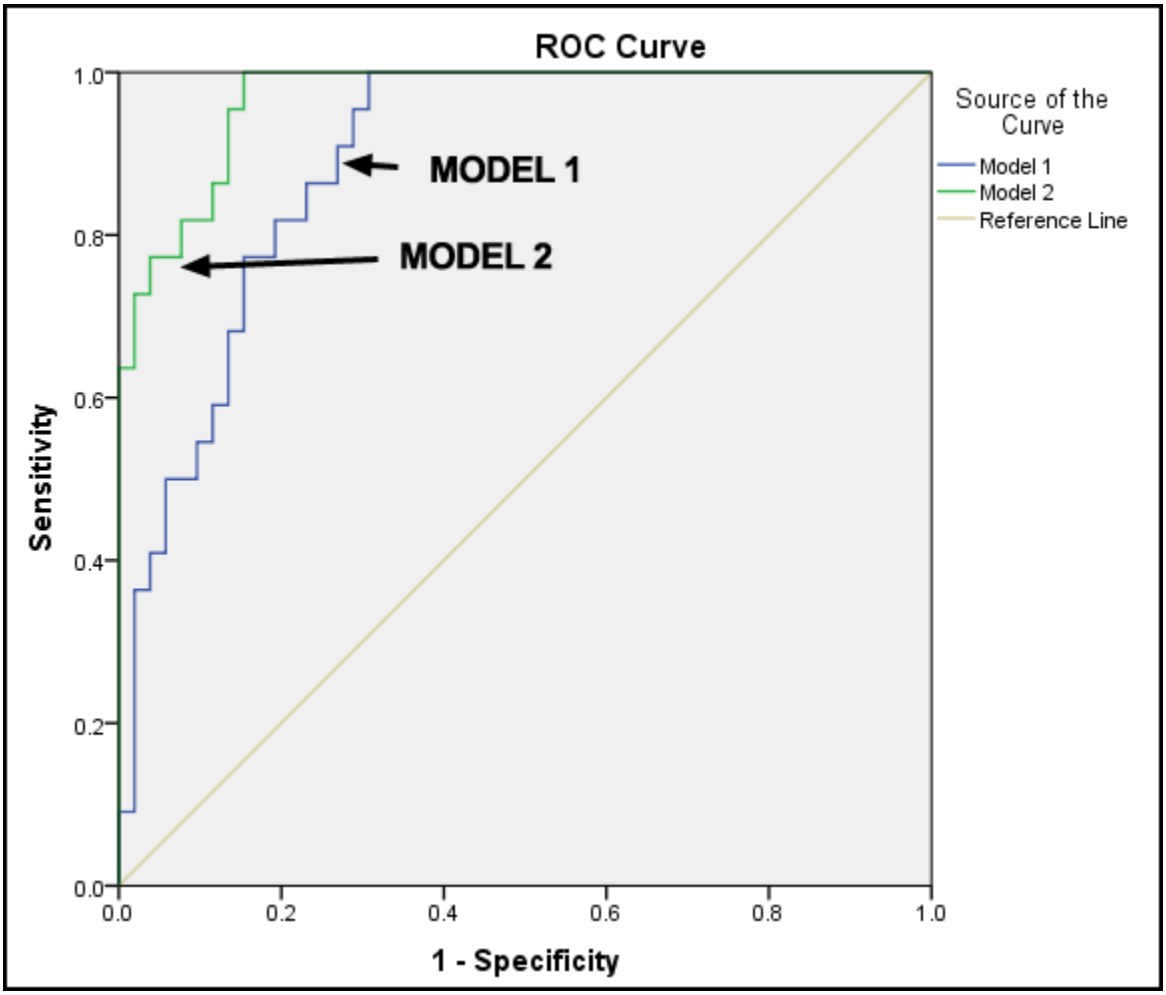


4. Click OK to run Model 1. Note that probabilities for model 1 are saved as a new variable, "PRE_1". In the SPSS data editor, this may be given the label "Model 1".
5. Repeat steps 1 – 4, for Model 2, entering turn and gear_ratio as additional covariates. The new probabilities variable will be "PRE_2". This may be given the label "Model 2".
6. Select Analyze > ROC Curve and enter both PRE_1 and PRE_2 in the "Test Variable" box and enter gender as the "State Variable". Set the value of the state variable to 1 (foreign is coded 0 = domestic, 1= foreign), and here we predict foreign with domestic as the reference category).



7. Check "With diagonal reference line" and click OK to run the ROC Curve procedure.

The resulting ROC curve appears as shown below.



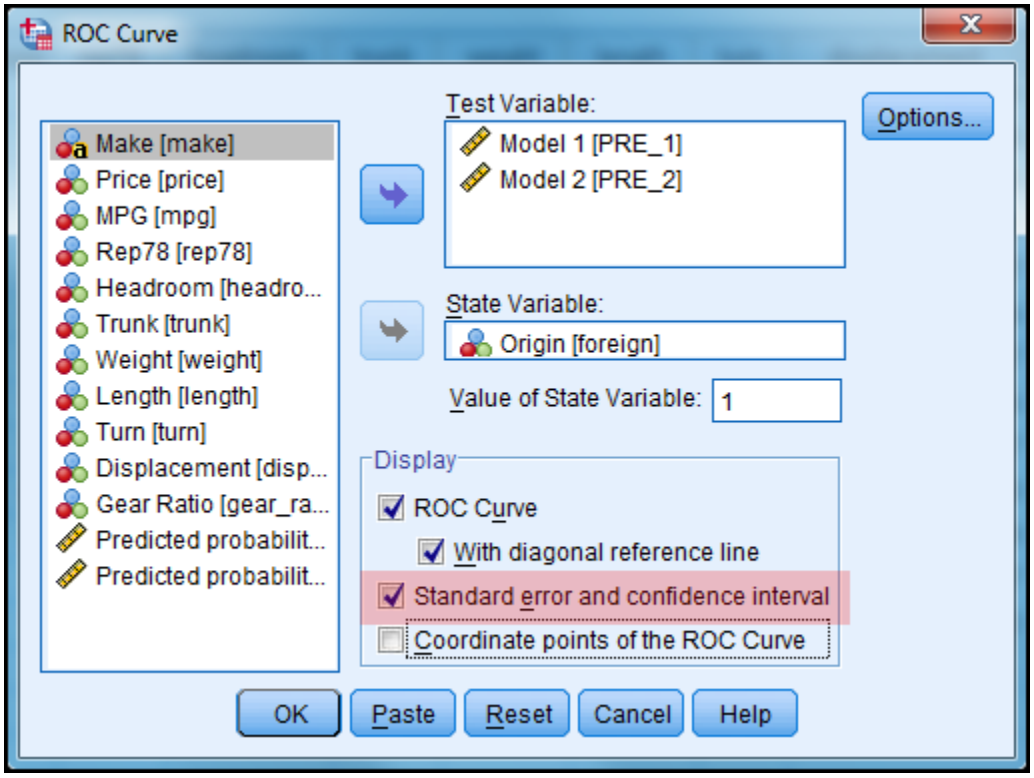
This figure illustrates the extent to which Origin = 1 = foreign may be predicted from either of two models. In the plot above, the more the model's curve is further from the 45-degree base line (that is, the greater the area under the ROC curve) the better the model's ability to classify. As would be expected, Model 2 is shown to have better classification ability.

The degree to which Model 2 is better than Model 1 is indicated by it being associated with greater area under the ROC curve:

Area Under the Curve

Test Result Variable(s)	Area
Model 1	.893
Model 2	.969

Is the area under the normal curve significant? This question is answered by checking the "Stqandard error and confidence interval" checkbox in the "ROC Curve" dialog, as shown below.



The resulting table is depicted below, showing the areas under the ROC curves for both Model 1 and Model 2 are significant.

Area Under the Curve

Test Result Variable(s)	Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
				Lower Bound	Upper Bound
Model 1	.893	.036	.000	.823	.963
Model 2	.969	.016	.000	.936	1.000

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

SPSS does not support a test of the significance of the difference in areas under the ROC curve for two models from the same sample, as does Stata (see below). The procedure for doing this manually was described by Hanley & McNeil (1983).

Note that in the "Method" area of the "Logistic Regression" main dialog, it is possible to select stepwise logistic regression rather than the default "Enter"

method. If this is done, Model 2 drops weight and length as not useful, leaving Model 2 with only mpg, turn, and gear_ratio as predictor variables.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	gear_ratio	5.837	1.306	19.976	1	.000	342.704
	Constant	-19.306	4.211	21.021	1	.000	.000
Step 2[b]	turn	-.382	.157	5.916	1	.015	.683
	gear_ratio	4.781	1.421	11.317	1	.001	119.249
	Constant	-1.695	7.068	.058	1	.810	.184
Step 3[c]	mpg	-.238	.107	4.975	1	.026	.788
	turn	-.611	.222	7.550	1	.006	.543
	gear_ratio	6.616	2.037	10.548	1	.001	746.704
	Constant	6.356	8.116	.613	1	.434	575.948

a. Variable(s) entered on step 1: gear_ratio.

b. Variable(s) entered on step 2: turn.

c. Variable(s) entered on step 3: mpg.

Optimal classification cutting points

The foregoing ROC analysis assumed a default classification cutting point of .5. Probabilities lower than the cutting point were classified as 0 (domestic) and probabilities above the cutting point were classified as 1 (foreign) in origin. To assess whether .5 is the optimal cutting point, it is necessary in SPSS to check output for "Coordinate points of the ROC Curve", as shown previously above in the ROC main dialog screen.

The resulting "Coordinates of the Curve" table shows sensitivity and specificity (see above) for both Model 1 and Model 2. For discussion purposes, here we consider just Model 1, the classification table for which was as below:

Classification Table[a]

Observed			Predicted		
			Origin		Percentage Correct
			Domestic	Foreign	
Step 1	Origin	Domestic	46	6	88.5
		Foreign	9	13	59.1
	Overall Percentage				79.7

a. The cut value is .500

The resulting "Coordinates of the Curve" table is quite large. Only partial output is seen below. At a cutting point of .5, sensitivity is shown to be .591 and specificity is shown to be around .115. The calculations are these:

- *Sensitivity* = 13/22 = .591. Of the 22 events (foreign cars), 13 or 59.1% were correctly predicted
- *Specificity as usually defined* = 46/52 = .885. Of the 52 non-events (domestic cars), 46 of 88.5% were correctly predicted.
- *Specificity as defined by SPSS* -= 1 = .885 = .115. Of the 52 non-events (domestic cars), 6 or 11.5% were incorrectly predicted.

Coordinates of the Curve

Test Result Variable(s)	Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
Model 1	.0000000	1.000	1.000
	.0009465	1.000	.981
	.0012666	1.000	.962
	.0023634	1.000	.942
	.0035807	1.000	.923
	[illegible]	1.00	[illegible]
	.4795564	.591	.135
	.5201919	.591	.115
	.5757511	.545	.115
	.6151616	.545	.096
	.6397050	.500	.096
	.6444411	.500	.077
	.6631544	.500	.058
	.6789510	.455	.058
	.6955052	.409	.058
	.7124068	.409	.038
	.7170409	.364	.038
	.7494012	.364	.019
	.7899094	.318	.019
	.8245211	.273	.019
	.8496367	.227	.019
	.8574897	.182	.019
	.8697896	.136	.019
	.8759510	.091	.019
	.8840605	.091	.000
	.8950403	.045	.000
	1.0000000	.000	.000

As cutting points go down toward 0, sensitivity goes up until 100% of event are correctly classified. At the same time, specificity as usually defined, not as SPSS defines it, goes down. SPSS, however, flips specificity and reports 1 minus specificity as usually defined so that as cutting points go down toward 0, sensitivity (percent of correct classification of events) goes up toward 1.0 and SPSS-defined specificity (percent of incorrect classification on non-events) also goes up toward 1.0.

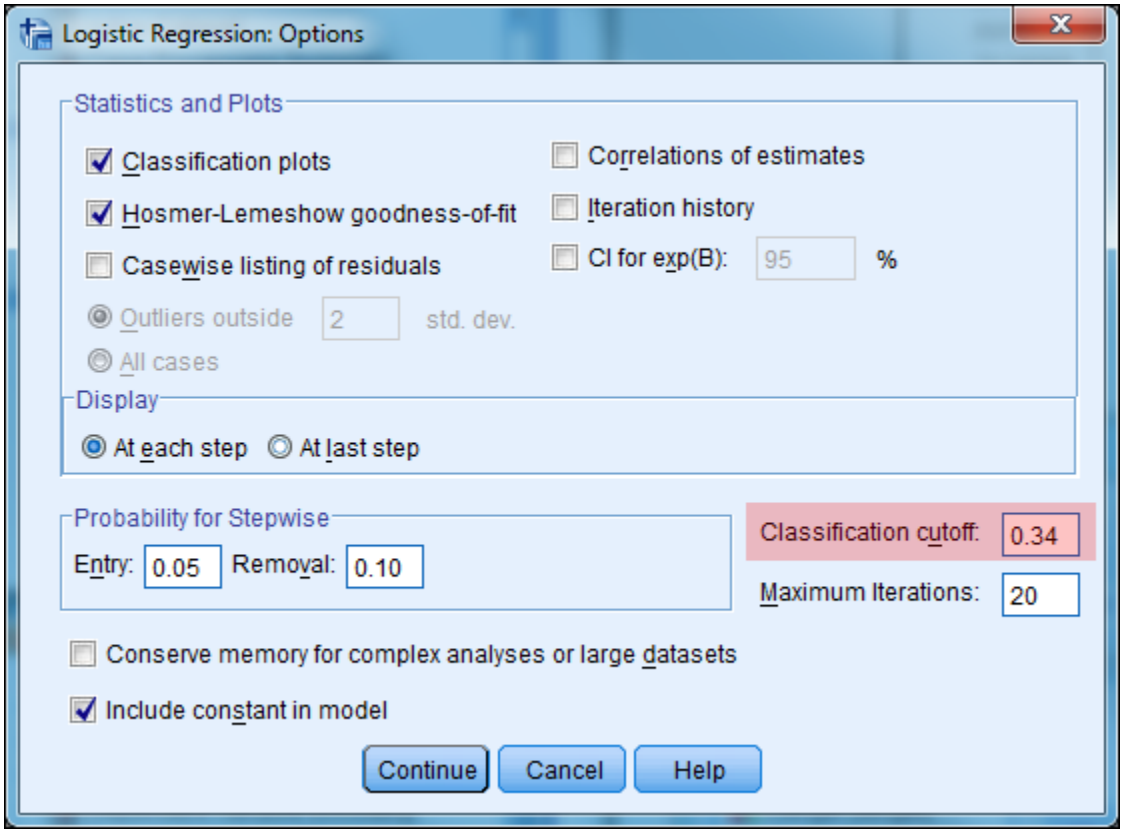
By either SPSS-defined specificity or by the usual definition of specificity, the researcher may use the table above to see the consequences of shifting the cutting point from the default .5. However, if the usual definition of specificity is employed (number of non-events correctly classified as a percent of all non-

events), then starting at the top, specificity will be ordered low to high and sensitivity will be ordered high to low.

.3138686	.818	.212
.3438202	.818	.192
.3664890	.773	.192
.3974794	.773	.173

Sometimes researchers define as optimum not the model with the highest percent correct but rather the model which operates equally well in classifying both events an non-events. In the table above, excerpted from the "Coordinates of the Curve" table, a cutting point of around .34 will have a sensitivity of around .82 (82% of events – foreign cars in this example – classified correctly, and around .81 (1 - .192 in the table above) specificity as usually defined (81% of non-events – domestic cars in this example – also classified correctly.

The binary logistic model for Model 1 can be re-run using .34 as the cutting point, under the "Options" button dialog shown below.



The resulting classification table is shown below. Compared to the original classification table above, not only has the overall percentage correct increased from 79.7% to 81.1%, but the model performs approximately equally well in classifying domestic and foreign cars.

Classification Table[a]

Observed			Predicted		
			Origin		Percentage Correct
			Domestic	Foreign	
Step 1	Origin	Domestic	42	10	80.8
		Foreign	4	18	81.8
	Overall Percentage				81.1

a. The cut value is .340

Side note: Where do the cutoff values in the first column of decimals in the "Coordinates of the Curve" table above come from? For model 1, test values are the PRE_1 probabilities. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

SAS

Overview

The ROC curve analysis example is the same as that in SPSS and SAS above, auto.sas7bdat, which is a SAS sample data file the availability of which is treated above. This data file contains variables pertaining to 74 car models sold in 1978. The binary dependent variable is "foreign", which is origin of the automobile, coded 0 = domestic, 1 = foreign.

Two models are compared:

Model 1 predicts "foreign" from miles per gallon (mpg), automobile weight (weight), and automobile length (length), on the hypotheses that US domestic cars in 1978 were long, heavy gas-guzzlers.

Model 2 predicts "foreign" using two additional predictors, turning radius (turn) and gear ratio (gear_ratio), on the hypothesis that foreign cars were not just smaller and more efficient, but also exhibited sportier handling.

All predictor variables were continuous in data level.

The commented SAS code for this example is shown below. It is essentially the same as for the SAS binary logistic regression example discussed previously above, except, of course, using the "auto" example data.

```
LIBNAME in "C:\Data";
TITLE "PROC LOGISTIC BINARY LOGISTIC OF AUTO DATA" JUSTIFY=CENTER;
 /* Optional title on each page */
PROC LOGISTIC DATA=in.auto;
 /* reads in auto.sas7b.dat */
MODEL foreign (EVENT=LAST) = mpg length weight
 / SELECTION=NONE SLSTAY=.10 CTABLE PPROB=.5;
 /* Logistic model predicting foreign = 1 (foreign origin), the higher level
*/
 /* EVENT=LAST makes foreign=0 the reference & foreign=1 predicted as in SPSS
*/
 /* SELECTION=NONE uses SPSS default "Enter" method, also the SAS default */
 /* SLSTAY=.10 uses SPSS default for keeping variable in model */
 /* CTABLE requests a classification table based on response probabilities */
 /* PPROB sets the cutting point for CTABLE */
RUN;
```

In this output, sensitivity is the number of correct classifications of the predicted dependent level (the "event", by default the higher level, hence foreign=1=foreign origin) as a percent of total events. Specificity is the number of correct classifications of the reference dependent level (the "non-event, here foreign=0=domestic origin). See expanded explanation in the SPSS and SAS sections above.

Model 1 output for the PROC LOGISTIC syntax above is shown below.

Classification Table

Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.500	13	46	6	9	79.7	59.1	88.5	31.6	16.4

The number of correct and incorrect classifications of events (foreign=1) and non-events (foreign-0) is the same as for the SPSS ROC example above.

Similar output for Model 2 may be created by adding the additional predictors in the MODEL statement and re-running the SAS syntax.

It is also possible to run a stepwise logistic regression model:

```
MODEL foreign (EVENT=LAST) = mpg length weight turn gear_ratio
   / SELECTION=STEPWISE SLSTAY=.10 CTABLE PPROB=.5;
```

The STEPWISE option in the MODEL statement drops variables which do not improve the model. In Model 2, only gear_ratio, turn, and mpg are found to be predictive and length and weight are dropped as redundant with these three in predictive utility.

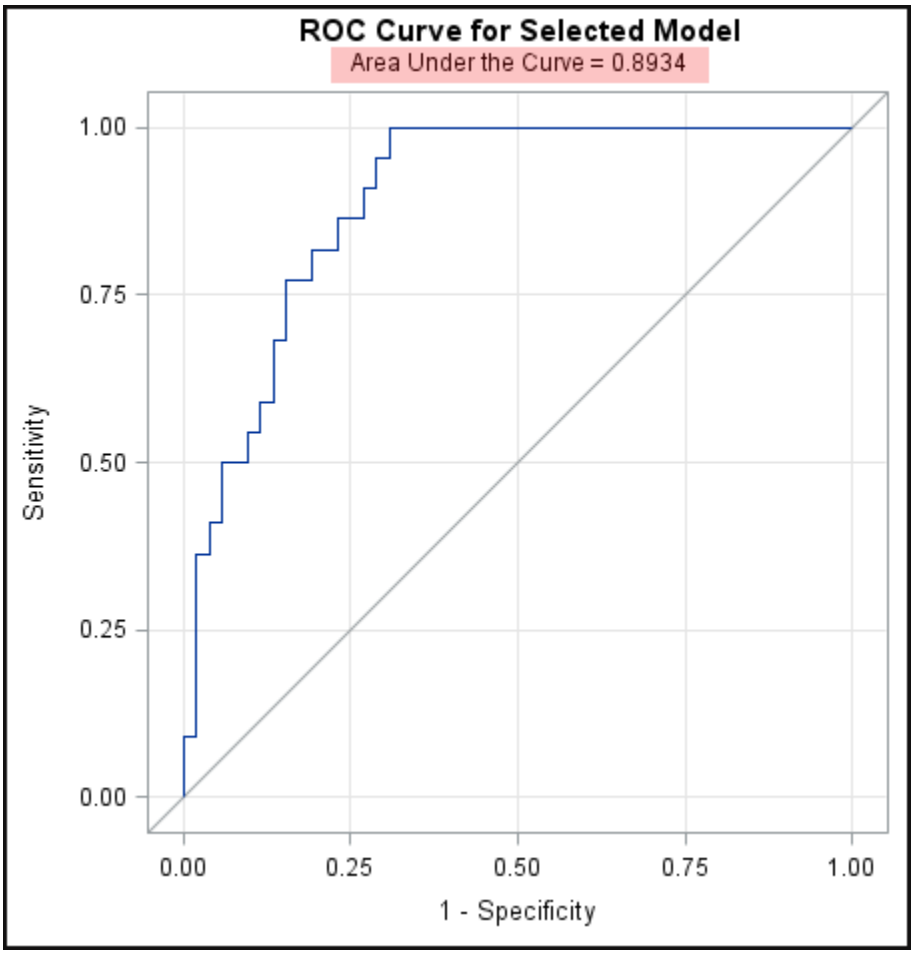
Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq	Variable Label
1	gear_ratio		1	1	36.9601		<.0001	Gear Ratio
2	turn		1	2	7.6588		0.0056	Turn
3	mpg		1	3	5.7933		0.0161	MPG

Comparing Models

When comparing models using ROC curve analysis, the model with greater area under the ROC curve is the better-classifying model. ROC curve plots are output when PLOTS=ROC is added to the PROC LOGISTIC statement:

```
PROC LOGISTIC DATA=in.auto PLOTS=ROC;
 /* reads in auto.sas7b.dat and requests ROC plots */
```



Above, for Model 1, the area under the ROC curve is .8934. After adding turn and gear_ratio as additional predictors to create model 2, the area under the curve (not shown) is .9685, indicating Model 2 is the better-classifying model. If stepwise selection is requested, a similar plot (not shown) is produced but with separate progressively higher curves for each variable added (each step), thereby showing graphically how much adding each variable contributes to greater area under the ROC curve.

Using ROC statements it is possible to determine if the area under the ROC curve is significant for either Model 1 or Model 2, and if the difference in areas between models is significant. The following syntax illustrates the ROC option:

```
PROC LOGISTIC DATA=in.auto PLOTS=ROC;
MODEL foreign (EVENT=LAST) = mpg length weight turn gear_ratio
 / SELECTION=NONE SLSTAY=.10 CTABLE NOFIT;
 /* Model 2 is modeled since any variable in the ROC statements must be */
 /* in the MODEL statement also */
 /* The NOFIT option suppresses some of the usual output. */
 ROC 'Model 1' mpg length weight;
```

```
 ROC 'Model 2' mpg length weight turn gear_ratio;
 /* ROC curves are requested for both models */
 ROCCONTRAST reference ('Model 1') / ESTIMATE e;
 /* Model 1 is made the reference category. */
 /* ESTIMATE asks for estimates of difference of areas. */
RUN;
```

The output form the ROC statement includes the "ROC Association Statistics" table below. If 0 is not within the 95% confidence limits, the row model's area is significant at the .05 level, as both are here.

ROC Association Statistics

ROC Model	Mann-Whitney Area	Mann-Whitney Standard Error	Mann-Whitney 95% Wald Confidence Limits		Somers' D (Gini)	Gamma	Tau-a
Model 1	0.8934	0.0356	0.8235	0.9632	0.7867	0.7867	0.3332
Model 2	0.9685	0.0161	0.9370	1.0000	0.9371	0.9371	0.3969

The ROC output also includes the "ROC Contrast" tables below, which test the significance of the difference in areas under the ROC curve of Model 2 compared to Model 1 (the reference model). For this example, The difference is significant at the .0178 level, which is the same result as in Stata below.

ROC Contrast Test Results

Contrast	DF	Chi-Square	Pr > ChiSq
Reference = Model 1	1	5.6149	0.0178

ROC Contrast Estimation and Testing Results by Row

Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
Model 2 - Model 1	0.0752	0.0317	0.0130	0.1374	5.6149	0.0178

Optimal classification cutting points

For classification table and ROC analysis, binary logistic regression defaults to a classification probability cutting point of 0.5. This default cutting point may not be optimal by either of the two most common definitions: (1) the cutting point which maximizes percent classified correctly overall; or (2) the cutting point at which the model classifies both dependent variables equally well (non-events as well as events).

As in the SPSS section on optimal classification cutting points above, in this section for SAS we discuss only Model 1.

As a default of PROC LOGISTIC if the PPROB= option is omitted from the MODEL statement, SAS prints a table showing specificity and sensitivity for a wide range of classification cut points, discussed above for a previous example. This may be used for purposes of selecting the optimal cutting point.

Classification Table

	Correct		Incorrect		Percentages				
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	22	0	52	0	29.7	100.0	0.0	70.3	.
0.020	22	10	42	0	43.2	100.0	19.2	65.6	0.0
0.040	22	20	32						0.0
0.060									
0.320	18	41	11	4	79.7	81.8		37.9	8.9
0.340	17	41	11	5	78.4	77.3	78.8	39.3	10.9
0.360	16	41	11	6	77.0	72.7	78.8	40.7	12.8
0.480	13	44	8	9	77.0	59.1	84.6	38.1	17.0
0.500	13	45	7	9	78.4	59.1	86.5	35.0	16.7
0.520	12	45	7	10	77.0	54.5	86.5	36.8	18.2
0.900	0	51	1	22	68.9	0.0	98.1	100.0	30.1
0.920	0	51	1	22	68.9	0.0	98.1	100.0	30.1
0.940	0	52	0	22	70.3	0.0	100.0	.	29.7

Partial output

At the default classification cutoff probability of 0.5, the percent correct for Model 1 is 78.4% (pink row), with a sensitivity of 50.1 and a specificity of 86.5. This is not the highest percent correct in the table. If the cut probability is .320, for instance, overall percent correct goes up slightly, to 79.7% (green cell). Another criterion for optimal cutting point, however, is the point at which the model has approximately equal sensitivity and specificity (predicts events and non-events approximately equally well). For Model 1, a cutting point of .34 (blue row) achieves this goal, with overall percent correct of 78.4%, sensitivity of 77.3%, and specificity of 78.8%. The model may be re-run, adding PPROB=.34 as an option in the MODEL statement:

```
MODEL foreign (EVENT=LAST) = mpg length weight
 / SELECTION=NONE SLSTAY=.10 CTABLE PPROB=.34;
```

Stata

Overview

The ROC curve analysis example is the same as that in SPSS and SAS above, auto.dta, which is a Stata sample data file which can be invoked with the command `sysuse auto.dta`. This data file contains variables pertaining to 74 car models sold in 1978. The binary dependent variable is "foreign", which is origin of the automobile, coded 0 = domestic, 1 = foreign.

Two models are compared:

Model 1 predicts "foreign" from miles per gallon (mpg), automobile weight (weight), and automobile length (length), on the hypotheses that US domestic cars in 1978 were long, heavy gas-guzzlers.

Model 2 predicts "foreign" using two additional predictors, turning radius (turn) and gear ratio (gear_ratio), on the hypothesis that foreign cars were not just smaller and more efficient, but also exhibited sportier handling.

All predictor variables were continuous in data level.

The classification table for Model 1 is created as default output of the post-estimation command `estat classification`, following the binary logistic regression command `logistic foreign mpg weight length`, illustrated in the Stata output below for Model 1.

In this output, sensitivity is the number of correct classifications of the predicted dependent level (the "event", by default the higher level, hence foreign=1=foreign origin) as a percent of total events. Specificity is the number of correct classifications of the reference dependent level (the "non-event, here foreign=0=domestic origin). See expanded explanation in the SPSS and SAS sections above.

```
. logistic foreign mpg weight length

Logistic regression                                     Number of obs   =         74
                                                        LR chi2(3)      =      35.72
                                                        Prob > chi2     =     0.0000
Log likelihood = -27.175156                             Pseudo R2       =     0.3966

------------------------------------------------------------------------------
     foreign | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |   .8448593   .0777007    -1.83   0.067     .7055055    1.011739
      weight |      .9961   .0019217    -2.03   0.043     .9923406    .9998736
      length |   1.000034   .0589616     0.00   1.000     .8908986    1.122538
       _cons |   895179.3    6896050     1.78   0.075     .2481103    3.23e+12
------------------------------------------------------------------------------

. estat classification

Logistic model for foreign

              -------- True --------
Classified |         D            ~D |      Total
-----------+--------------------------+-----------
         + |        13             6 |         19
         - |         9            46 |         55
-----------+--------------------------+-----------
     Total |        22            52 |         74

Classified + if predicted Pr(D) >= .5
True D defined as foreign != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   59.09%
Specificity                     Pr( -|~D)   88.46%
Positive predictive value       Pr( D| +)   68.42%
Negative predictive value       Pr(~D| -)   83.64%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   11.54%
False - rate for true D         Pr( -| D)   40.91%
False + rate for classified +   Pr(~D| +)   31.58%
False - rate for classified -   Pr( D| -)   16.36%
-------------------------------------------------
Correctly classified                        79.73%
```

A POST-ESTIMATION COMMAND

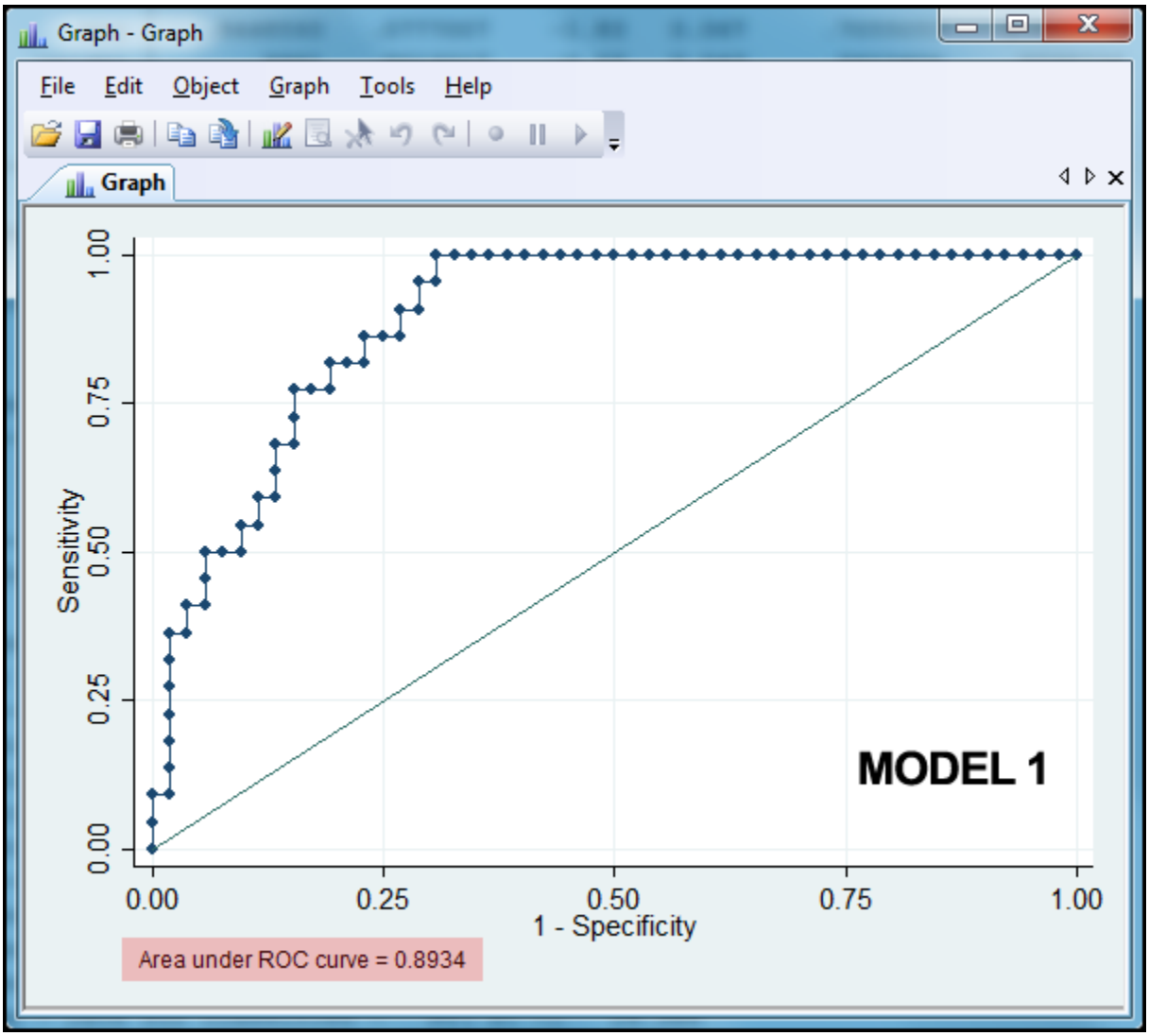
CUTTING POINT IS .5

FOREIGN = 1 IS BEING PREDICTED

SENSITIVITY

SPECIFICITY AS USUALLY DEFINED (LIKE SAS, OPPOSITE OF SPSS)

The ROC curve itself is created simply by issuing the post-estimation command `lroc`, generating the plot below. The area under the ROC curve is high (.8934), in the direction of 1.0, indicating a good model. This value is identical to that in SPSS and SAS.



Comparing Models

A central use of ROC curve analysis is to compare two models. The model with the greater area under the ROC curve is the model which classifies the dependent variable (here, "foreign", meaning domestic or foreign). Model 2 adds the predictors turning radius (turn) and gear ratio (gear_ratio): `logistic foreign mpg weight length turn gear_ratio.`

Issuing the `estat classification` post-estimation command as before generates the table below, in which it can be seen that the percent correct has increased from 79.73% to 87.84% as a result of adding the predictors turn and gear_ratio. That is, Model 2 classifies better than does Model 1.

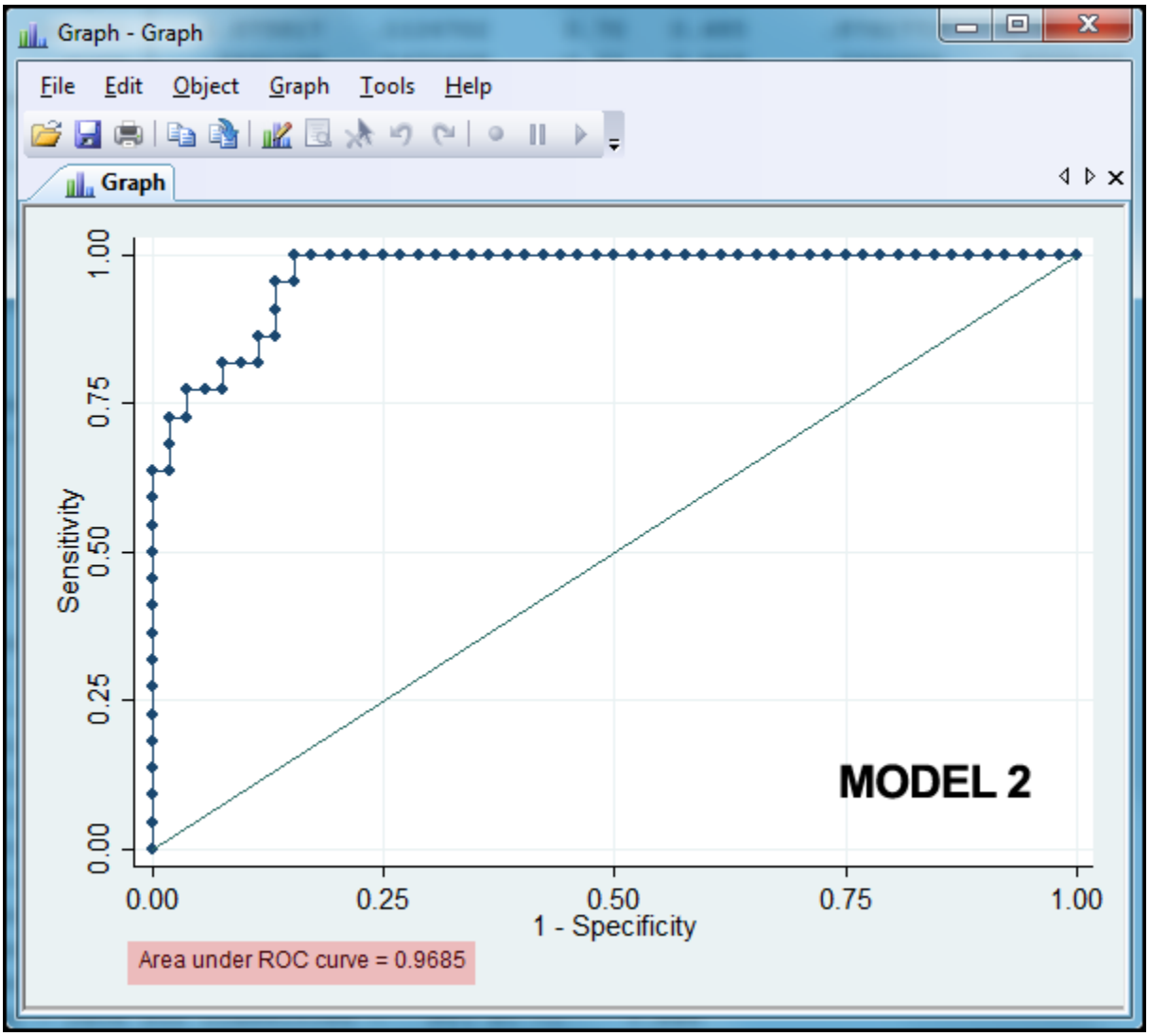
```
. estat classification

Logistic model for foreign

                 ---------- True ----------
Classified |         D              ~D      |      Total
-----------+--------------------------------+-----------
     +     |        18              5       |         23
     -     |         4              47      |         51
-----------+--------------------------------+-----------
   Total   |        22              52      |         74

Classified + if predicted Pr(D) >= .5
True D defined as foreign != 0
--------------------------------------------------------------
Sensitivity                     Pr( +| D)   81.82%
Specificity                     Pr( -|~D)   90.38%
Positive predictive value       Pr( D| +)   78.26%
Negative predictive value       Pr(~D| -)   92.16%
--------------------------------------------------------------
False + rate for true ~D        Pr( +|~D)    9.62%
False - rate for true D         Pr( -| D)   18.18%
False + rate for classified +   Pr(~D| +)   21.74%
False - rate for classified -   Pr( D| -)    7.84%
--------------------------------------------------------------
Correctly classified                        87.84%
--------------------------------------------------------------
```

That Model 2 is the better-classifying model is also seen in the ROC curve, generated by the post-estimation command, `lroc`. It can be seen in the plot below that the area under the ROC curve has increased from .8934 to .9689, also indicating Model 2 is the better model. These are the same values as reported by SPSS and SAS.



Although the ROC plots above show that Model 2 is better than Model 1 in terms of classification, how do we know that the values for the two areas under the ROC curve are far enough apart to be sure that Model 2 is significantly better than Model 1? Cleves (2002) has outlined how Stata's `roccomp` procedure will compute the significance of the difference. This procedure will also create a single plot with both ROC curves. The steps for the example data are:

1. `logistic foreign mpg weight length` // Run Model 1
2. `lroc, nograph` // Create ROC output for Model 1 but skip the plot
3. `predict xb1, xb` // Here "xb" is an automatically created variable containing the linear predictions for the most recent model (here, Model 1). Here the predict command adds a variable called xb1 to the dataset, where xb1 contains the values of xb, which are the Model 1 predicted

values for the dependent variable "foreign", which is origin of the automobile (domestic or foreign).

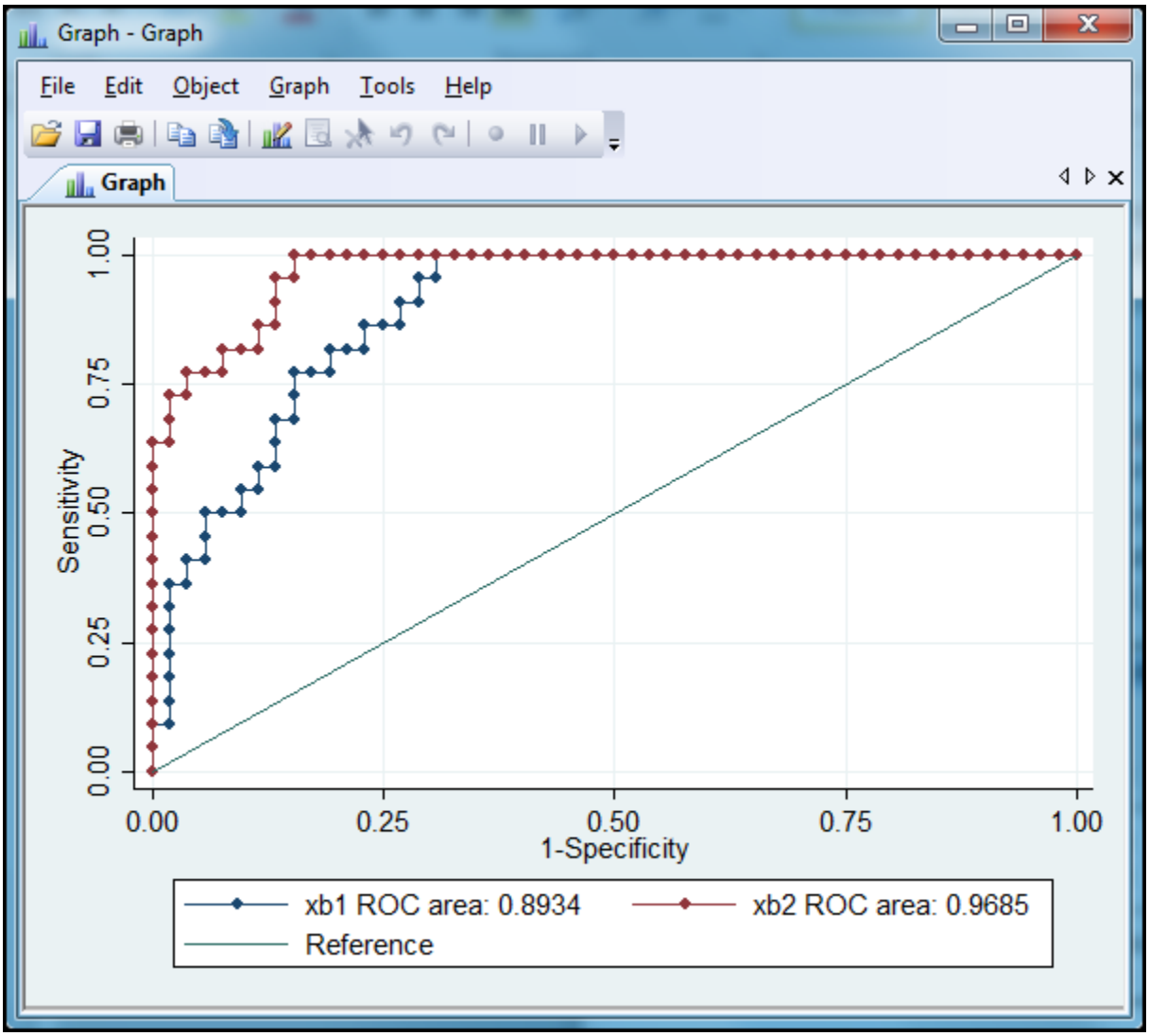
4. `logistic foreign mpg weight length turn gear_ratio` // Run Model 2
5. `lroc, nograph` // Create ROC output for Model 2 but skip the plot
6. `predict xb2, xb` // Repeat step 3 for Model 2, storing the Model 2 predictions in a new variable called xb2.
7. `roccomp foreign xb1 xb2,graph summary` / Use the roccomp command to plot and test the difference in areas under the ROC curves of Model 1 and Model 2. Output is as below:

```
. roccomp foreign xb1 xb2,graph summary

                                  ROC                    -Asymptotic Normal--
                    Obs       Area     Std. Err.      [95% Conf. Interval]
---------------------------------------------------------------------------
xb1                  74     0.8934       0.0356        0.82350     0.96321
xb2                  74     0.9685       0.0161        0.93698     1.00000
---------------------------------------------------------------------------
Ho: area(xb1) = area(xb2)
    chi2(1) =     5.61       Prob>chi2 =   0.0178
```

Based on the `roccomp` output, we can say that Model 2 is a better classifier than Model 1 and that the difference is significant at the 0.0178 level. This is the same result as for SAS output above. We can also say the areas under the curve for either model are significant at the .05 level since 0 is not within the 95% confidence intervals for either model. Note that while the two models are nested in this example, that is not a requirement of the `roccomp` procedure.

The `roccomp` plot is shown below. The higher red line is Model 2 and the lower blue line is Model 1.



Optimal classification cutting points

For classification table and ROC analysis, binary logistic regression defaults to a classification probability cutting point of 0.5. This default cutting point may not be optimal by either of the two most common definitions: (1) the cutting point which maximizes percent classified correctly overall; or (2) the cutting point at which the model classifies both dependent variables equally well (non-events as well as events). The logistic procedure in Stata does not provide equivalents to the SPSS "Coordinates of the Curve" or the SAS "Classification Table" output used to select optimal cutting points. Moreover, the Stata logistic procedure does not provide for changing the default classification cutting point. See above for how this is done in SPSS or above for how it is done in SAS.

Conditional logistic regression for matched pairs

Overview

Matched case-control pairs, where for every subject (case) there is a matched pair (control), may be analyzed in SPSS using multinomial logistic regression, in SAS using PROC LOGISTIC, and in Stata using the `clogit` (conditional logit) command. Most output tables previously discussed (for example, parameter estimates, likelihood ratio tests, pseudo-R^2, AIC and BIC) apply, but note that classification tables and ROC analyses are invalid for conditional logistic estimates.

Example

The examples in this section use a subset of cases from Breslow & Day (1980), in the file "BreslowDaySubset" with .sav, .sas7bdat, and .dta extensions for SPSS, SAS, and Stata respectively. See above for access. The data concern a study of endometrial cancer as possibly caused by gall bladder disease and hypertension. There were 63 matched pairs in the subset. The variables are these:

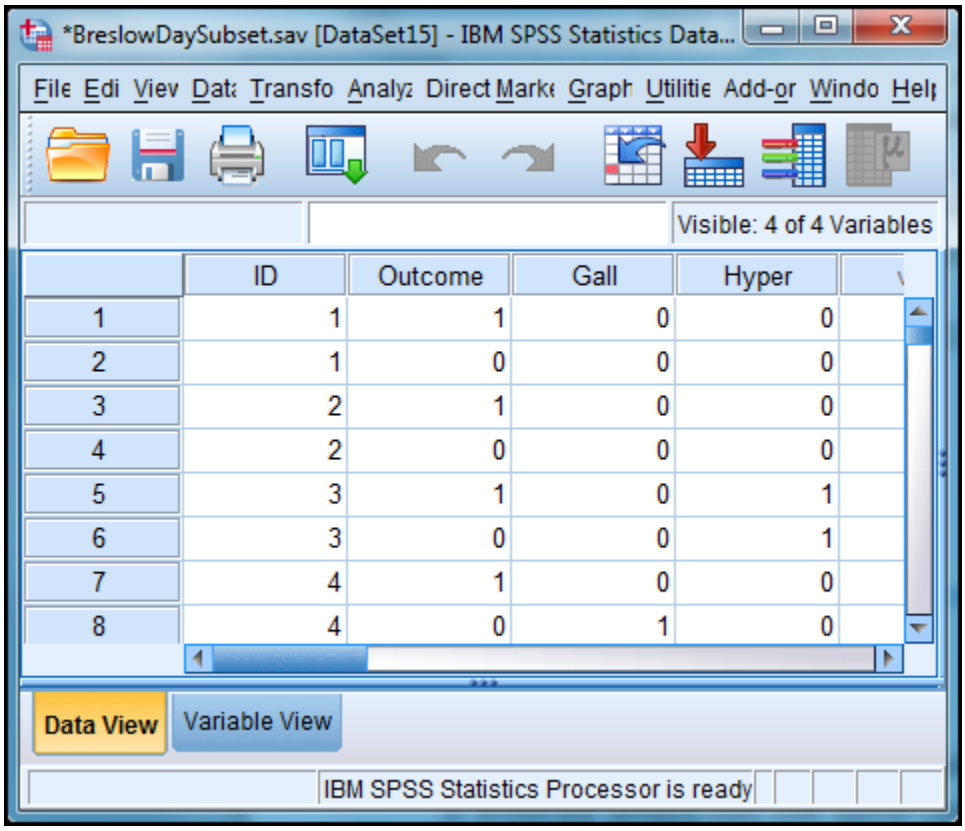
ID	Each case and matched control pair has the same id. For instance, there are two data rows for ID=3, one for the subject and one for the subject's matched pair.
Outcome	Coded 0 = no cancer, 1 = cancer
Gall	Coded 0 = no gall bladder disease, 1 = gall bladder disease
Hyper	Coded 0 = no hypertension, 1 = hypertension

SAS support's online "Example 51.11" uses these data with documentation here.

Logistic regression results assume a sample of sufficient size: estimates are asymptotic by default, if exact procedures are not invoked. For pedagogical purposes, however, a subset of the entire Breslow-Day data was used. See discussion of sample size in the "Assumptions" section below.

Data setup

The example data have 63 matched pairs for a total of 126 data rows. Initial raw data are shown below using the SPSS data editor. Each subject and matched pair share the same ID, is separate data rows.



Conditional logistic regression in SPSS

Overview

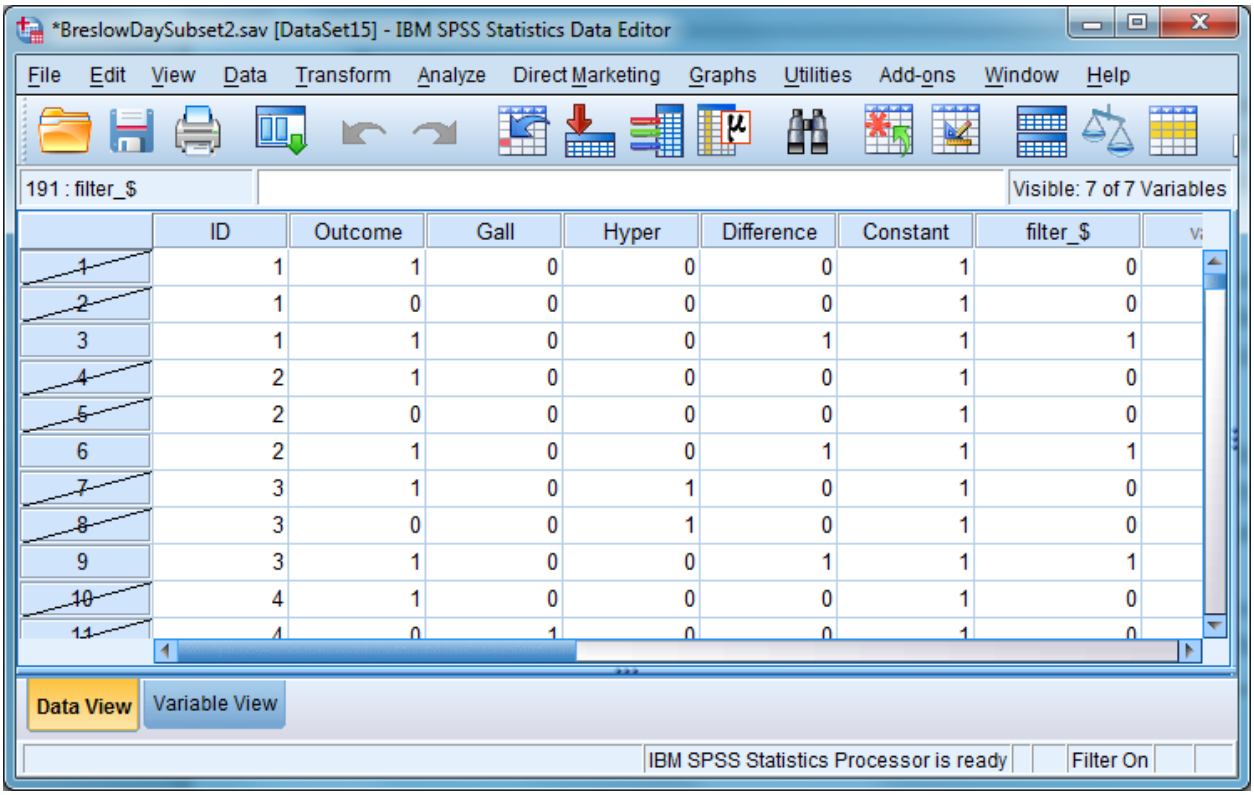
Conditional logistic regression for matched pairs data may be accomplished in SPSS in its multinomial logistic regression procedure, but pair data must be differenced as described below; a constant must be added; predictor variables must be entered as covariates; and the intercept must be omitted.

Because SPSS does not support a stratification variable in multinomial logistic regression, a different but statistically equivalent approach must be taken, described by Breslow (1982) as conditional analysis using transformed data. This approach requires the following steps:

1. For each ID value, a third data row is added. To identify the added rows, a new variable (here called Difference) is added, coded 1 if the row is a difference row and otherwise coded 0. Thus there are three data rows per ID.

2. For each predictor variable (here, for Gall and Hyper), the value of the control member of the matched pair is subtracted from the subject's corresponding value and the difference is entered in the third (Difference) row for that ID.
3. A new variable is added whose value cannot vary. Here the variable "Constant" is added, set to 1 for all data rows.
4. In SPSS, select Data > Select Cases > If, and make the "If" expression "Difference=1". This causes SPSS to act as if the Difference=1 rows are the only rows in the dataset when multinomial regression is run. SPSS adds the variable "filter_$" to flag the selected cases.

A portion of the transformed data appears as below. This is the BrewslowDayImputed2.sav data file available above.

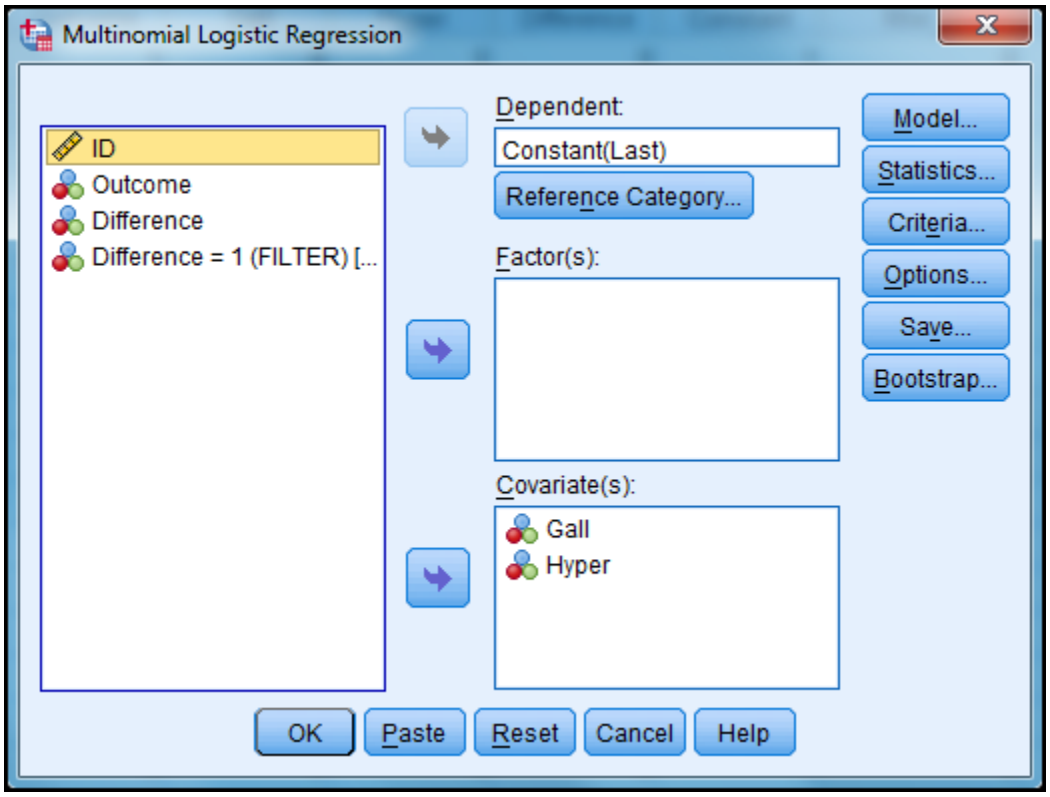


	ID	Outcome	Gall	Hyper	Difference	Constant	filter_$
1	1	1	0	0	0	1	0
2	1	0	0	0	0	1	0
3	1	1	0	0	1	1	1
4	2	1	0	0	0	1	0
5	2	0	0	0	0	1	0
6	2	1	0	0	1	1	1
7	3	1	0	1	0	1	0
8	3	0	0	1	0	1	0
9	3	1	0	0	1	1	1
10	4	1	0	0	0	1	0
11	4	0	1	0	0	1	0

SPSS input

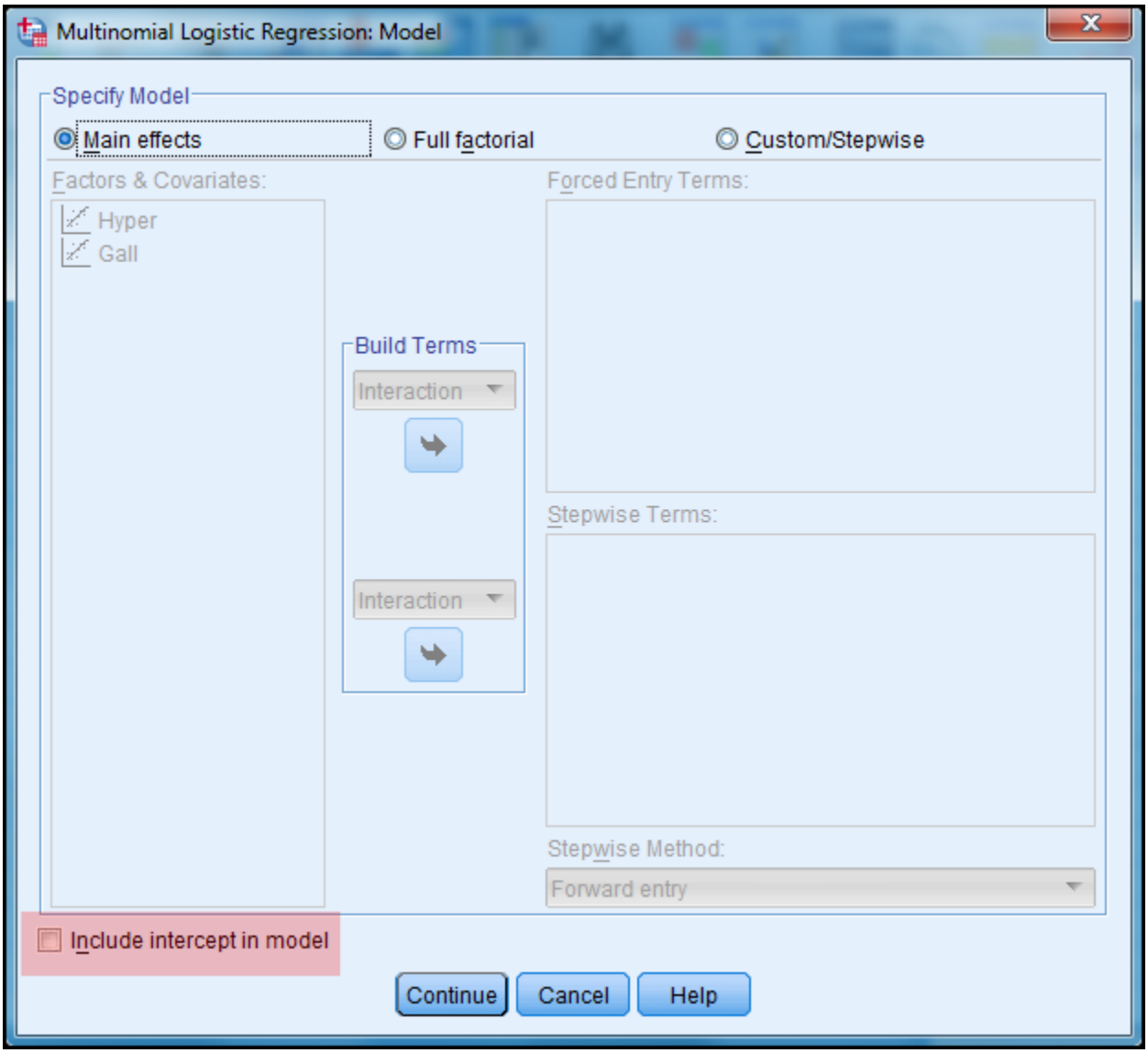
After selecting Analyze > Regression > Multinomial logistic, a dependent variable is entered which has the same constant value for all cases. This is what tells SPSS

that a conditional logistic model is being requested. In this example the dependent is the variable "Constant". The real dependent, of course, is whatever was used to differentiate Subject and Match pairs (case/control pairs), in this case Outcome = 1 = has cancer.

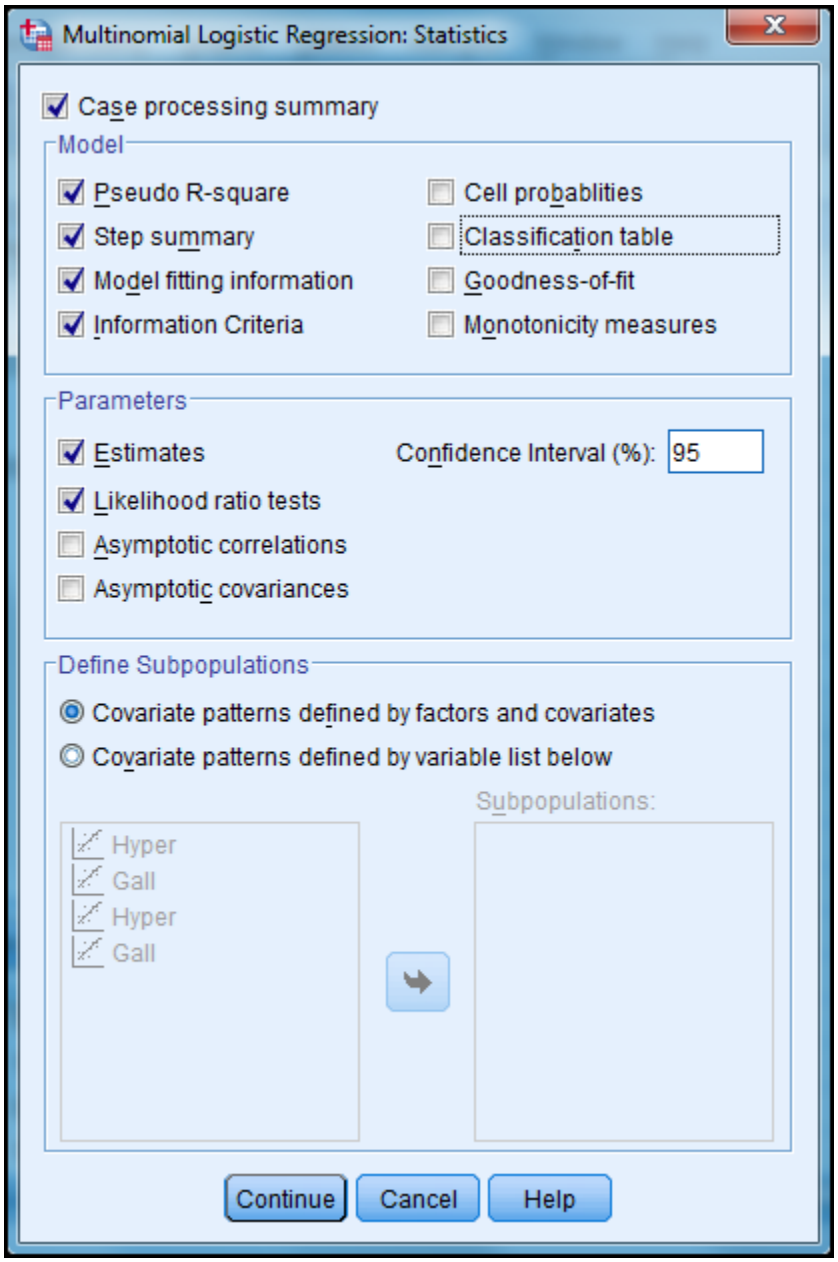


All predictors are entered as covariates. If there were factors (categorical variables), they would have had to have been transformed into 0,1 dummy variables. The variables used for matching (ex., age, gender) are not entered into the model and indeed, may not even be in the data set. If these matching variables are in the dataset, they may be used as part of interaction effects but may not be used as main effects since by definition their "Diff" value is 0 for every subject.

Under the Model button, the researcher must uncheck the "Include intercept" checkbox found in the dialog window for the "Model" button, illustrated below.



Output options are selected under the "Statistics" button dialog, shown below.



SPSS output

Although the usual multinomial regression output will be generated, the classification table is invalid for a conditional logistic model. The tables illustrated below, however, are valid (Model Fitting Information, Pseudo R-Square, Likelihood Ratio Tests, and Parameter Estimates tables) are valid for conditional logistic models.

The likelihood ratio test of the significance of the final model overall is .103, indicating the model is not significant, even at an exploratory (.10) level. We fail reject the null hypothesis that gall bladder disease, controlling for hypertension, is related to endometrial cancer. This may be in part due to small sample size used in this data subset. This is suggested by the fact that model effect size such as Nagelkerke's R^2 (.903), while weak, does not closely approach 0. Even for a small sample, that model effect size is weak suggests the need for better model specification (more and/or better predictor variables).

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Null	87.337			
Final	82.788	4.549	2	.103

Pseudo R-Square

Cox and Snell	.070
Nagelkerke	.093
McFadden	.052

The next two tables help assess the individual predictor variables, which were Gall and Hyper. If there is a conflict, the likelihood ratio test in the upper portion of the figure is preferred over the Wald test in the lower portion. For both tests, hypertension is not a significant control variable. Also in both tests, gall bladder disease is a significant cause of endometrial cancer at an exploratory level (.05<p<=.10) but not at a confirmatory level (p <=.05). This again suggests that with a larger sample size, gall bladder disease would be found to be significant at a confirmatory level.

Likelihood Ratio Tests

Effect	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	df	Sig.
Gall	86.527	3.739	1	.053
Hyper	83.654	.866	1	.352

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.

Parameter Estimates

Constant		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B): Lower Bound	Upper Bound
1	Gall	.970	.531	3.343	1	.067	2.639	.933	7.468
	Hyper	.348	.377	.853	1	.356	1.416	.677	2.965

The "Exp(B)" column in the SPSS "Parameter Estimates" table contains the odds ratios. We may say that the odds of being in the cancer group (Outcome=1) compared to being in the non-cancer group (Outcome=0) are increased (multiplied) by a factor of 2.639 when the subject is in the gall bladder disease group (Gall=1) compared to the group without this disease (Gall=0), controlling for hypertension.

Conditional logistic regression in SAS

Overview

Conditional logistic regression may be implemented using the same PROC LOGISTIC procedure as for other logistic models described earlier. In data setup, it is not necessary to go through the differencing process described above for the same model in SPSS. Discussion redundant with that in the SPSS section above is omitted here in the SAS section on conditional logistic regression.

The example model predicts endometrial cancer (Outcome =1) from gall bladder disease (Gall), controlling for hypertension (Hyper). See discussion above.

SAS input

The commented SAS syntax is below, largely repeating SAS syntax for other models above. There is no CLASS statement because in this model, both predictor variables are treated as covariates. The STRATA statement makes the model conditional on ID, which clusters matched pairs together.

```
LIBNAME in "C:\Data";
TITLE "PROC LOGISTIC Conditional Logistic Regression for Matched Pairs"
JUSTIFY=CENTER;
 /* Optional title on each page */
PROC LOGISTIC DATA=in.BreslowDaySubset;
 /* reads in BreslowDaySubset.sas7b.dat */
STRATA ID;
 /* STRATA makes this a conditional logistic regression model */
 /* Recall each ID has a subject and a control match row */
MODEL outcome (EVENT=LAST)= gall hyper
 / SELECTION=NONE;
 /* Cancer outcome predicted by gall bladder disease and hypertension */
  /* EVENT=LAST makes outcome=0 the reference & outcome=1 predicted as in
SPSS */
 /* SELECTION=NONE uses SPSS default "Enter" method, also the SAS default */
 /* CTABLE is not allowed with STRATA and is omitted */
RUN;
```

SAS output

SAS output is identical to that in SPSS above. Although the usual multinomial regression output will be generated, note that the classification table is invalid for a conditional logistic model.

The likelihood ratio test of the significance of the final model overall is .1029, the same as in SPSS, indicating the model is not significant, even at an exploratory (.10) level. We fail reject the null hypothesis that gall bladder disease, controlling for hypertension, is related to endometrial cancer. This may be in part due to small sample size used in this data subset.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
AIC	87.337	86.788
SC	87.337	92.460
-2 Log L	87.337	82.788

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.5487	2	0.1029
Score	4.3620	2	0.1129
Wald	4.0060	2	0.1349

The next two tables help assess the individual predictor variables, which were Gall and Hyper. The "Analysis of Conditional Maximum Likelihood Estimates" gives the same estimates of the logistic regression coefficients for Gall and Hyper as did the "Parameter Estimates" table in SPSS. The same table gives Wald tests of each effect, showing hypertension is not a significant control variable. Gall bladder disease (controlling for hypertension) is a significant cause of endometrial cancer at an exploratory level (.05<p<=.10) but not at a confirmatory level (p <=.05). This suggests that with a larger sample size, gall bladder disease would be found to be significant at a confirmatory level.

Analysis of Conditional Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Gall	1	0.9704	0.5307	3.3432	0.0675
Hyper	1	0.3481	0.3770	0.8526	0.3558

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Gall	2.639	0.933	7.468
Hyper	1.416	0.677	2.965

The "Odds Ratio Estimates" table gives the same odds ratios for Gall and Hyper as did the "Exp(B)" column in the SPSS "Parameter Estimates" table. We may say that the odds of being in the cancer group (Outcome=1) compared to being in the

non-cancer group (Outcome=0) are increased (multiplied) by a factor of 2.639 when the subject is in the gall bladder disease group (Gall=1) compared to the group without this disease (Gall=0), controlling for hypertension.

Conditional logistic regression in Stata

Overview

Conditional logistic regression in Stata is implemented with the conditional logistic command, `clogit`. As Stata supports stratification for clustered variables like ID, which clusters matched pairs, it is not necessary to use the differencing procedure described above for SPSS. Rather, the Stata approach is similar to that for SAS. Stata yields the same statistical results as SPSS and Stata.

The example model predicts endometrial cancer (Outcome =1) from gall bladder disease (Gall), controlling for hypertension (Hyper). See discussion above.

Stata input

To implement the matched-pairs conditional logistic regression model for the example data, the following command is used:

```
. clogit Outcome Gall Hyper, strata(ID)
```

The command above yields logistic regression coefficients as estimates. To estimate odds ratios instead, add the "or" option:

```
. clogit Outcome Gall Hyper, strata(ID) or
```

The `strata` option makes the model conditional on ID, which clusters matched pairs together. A synonym for `strata` is `group`.

Stata output

The default `clogit` output for the example data appears below. The following observations may be made on the basis of this table:

1. The likelihood ratio test of the significance of the model as a whole is 0.1029, the same as in the SPSS "Model Fitting Information" table the SAS "Testing Global Null Hypothesis: BETA=0" table above. The model is not significant, even at the exploratory .10 level. We fail to reject the null

hypothesis that gall bladder diseases is unrelated to cancer outcome, controlling for hypertension.

2. "Pseudo R2" is the same as McFadden's coefficient in the SPSS "Pseudo R-Square" table above. At .0521, it indicates a weak model and suggests the need for better specification by adding more and/or different predictors to the model.
3. The z tests of the significance of the logistic coefficients show neither Gall nor Hyper to be significant at the usual confirmatory .05 level. However, at 0.067, Gall is significant at the exploratory level. With a larger sample, it may become significant at the confirmatory level.

```
. clogit Outcome Gall Hyper, strata(ID)

Conditional (fixed-effects) logistic regression   Number of obs    =        126
                                                  LR chi2(2)       =       4.55
                                                  Prob > chi2      =     0.1029
Log likelihood = -41.393921                       Pseudo R2        =     0.0521

------------------------------------------------------------------------------
     Outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Gall |   .9704079   .5307309     1.83   0.067    -.0698055    2.010621
       Hyper |   .3480657   .3769635     0.92   0.356    -.3907691      1.0869
------------------------------------------------------------------------------
```

By adding "or" as an option, odds ratios are output in lieu of logistic coefficients:

```
. clogit Outcome Gall Hyper, strata(ID) or

Conditional (fixed-effects) logistic regression   Number of obs    =        126
                                                  LR chi2(2)       =       4.55
                                                  Prob > chi2      =     0.1029
Log likelihood = -41.393921                       Pseudo R2        =     0.0521

------------------------------------------------------------------------------
     Outcome | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Gall |   2.639021    1.40061     1.83   0.067     .9325752    7.467956
       Hyper |   1.416325   .5339029     0.92   0.356     .6765363     2.96507
------------------------------------------------------------------------------
```

The odds ratios are the same as those in the "Parameter Estimates" table in SPSS or the "Odds Ratio Estimates" table in SAS. We may say that the odds of being in the cancer group (Outcome=1) compared to being in the non-cancer group

(Outcome=0) are increased (multiplied) by a factor of 2.639 when the subject is in the gall bladder disease group (Gall=1) compared to the group without this disease (Gall=0), controlling for hypertension.

More about parameter estimates and odds ratios

For binary logistic regression

Parameter estimates (b) and odds ratios (Exp(b)) may be output in binary logistic regression for dichotomous, categorical, or continuous predictor variables. Their interpretation is similar in each case, though often researchers will not interpret odds ratios in terms of statements illustrated below, but instead will simply use odds ratios as effect size measures and comment on their relative sizes when comparing the effects of independent variables, or will comment on the change in an odds ratio between a model and some nested alternative model.

Example 1

In the example below, the binary variable capital punishment (1=favor, 2=oppose) is the dependent. In binary logistic regression, by default the higher category of the dependent (2= oppose) is predicted and the lower category (1 = favoring) is the comparison reference by default (the researcher could override this). It does not matter if the dependent is coded 0,1 or 1, 2. Capital punishment attitude is predicted from sex (a dichotomy), marital status (a categorical variable with 5 levels, with the reference category = 5 = never married)), and age (a continuous variable). The example is illustrated using SPSS.

- *For dichotomies*. Odds ratios will differ depending on whether a dichotomy is entered as a dichotomous covariate or as a categorical variable. In this example, sex is a dichotomy, with 1=male and 2=female. In binary regression, if a dichotomy is declared categorical, as it is in the upper portion of the figure below, then the prediction is for the lower category and the higher category (sex = 2 = female) is the reference. For categorical coding of sex, we can say that the odds of opposing capital punishment compared to favoring it are decreased by a factor of .571 by being male rather than female, based on the odds ratio found in the Exp(b) column for sex(1), where sex = 1 = male. Or we could say that the odds a male opposes capital punishment are .571 the odds a female respondent opposes it. If, on

the other hand, sex is entered as a dichotomous covariate rather than as a factor, as in the lower portion of the figure below, then the reference category is the lower category (usually the 0 category, but here with (1,2) coding, the lower category is 1 = male). For covariate coding of sex, the odds ratio is 1.751. We can say that the odds of opposing the death penalty compared to favoring it are increased by a factor of 1.751 by being female rather than male, controlling for other variables in the model. Or we could say that the odds a female opposes the death penalty are 1.75 the odds a male opposes it, controlling for other variables in the model.

Binary Logistic Regression

Dependent: 1 = favor capital punishment (the reference)
2 = oppose capital punishment (the predicted)

Sex entered as a categorical variable :
1 = Male
2 = Female (the default reference)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	sex(1)	-.560	.138	16.401	1	.000	.571
	marital			9.986	4	.041	
	marital(1)	-.533	.181	8.658	1	.003	.587
	marital(2)	-.333	.312	1.141	1	.286	.717
	marital(3)	-.448	.231	3.765	1	.052	.639
	marital(4)	-.103	.406	.065	1	.799	.902
	age	.001	.005	.016	1	.900	1.001
	Constant	-.662	.209	10.033	1	.002	.516

a. Variable(s) entered on step 1: sex, marital, age.

Above: The odds of opposing capital punishment compared to favoring it are reduced by a factor of .571 by being male rather than female.

Sex entered as a dichotomous covariate:
1 = Male (the reference)
2 = Female

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	sex	.560	.138	16.401	1	.000	1.751
	marital			9.986	4	.041	
	marital(1)	-.533	.181	8.658	1	.003	.587
	marital(2)	-.333	.312	1.141	1	.286	.717
	marital(3)	-.448	.231	3.765	1	.052	.639
	marital(4)	-.103	.406	.065	1	.799	.902
	age	.001	.005	.016	1	.900	1.001
	Constant	-1.783	.305	34.221	1	.000	.168

a. Variable(s) entered on step 1: sex, marital, age.

Above: The odds of opposing capital punishment compared to favoring it are increased by a factor of 1.751 by being female rather than male.

File: GSS93 subset.sav

- *For categorical variables*: Categorical variables must be interpreted in terms of the left-out reference category, as in OLS regression. In the example above, the categorical variable "marital" has five levels: 1=married, 2=widowed, 3=divorced, 4=separated, 5=never married. "Never married", being the highest-coded level, is therefore the default reference category. If b is positive for a level of a categorical variable, then when the respondent is in that category there is a positive effect on the dependent variable compared to being in the reference category. Also, the odds ratio, which is the Exp(b) column, will be greater than 1.0. Here, however, all categories have a negative b coefficient. For these data, being in any of the first four marital status categories compared to being never married (marital=5) decreases the odds of opposing capital punishment. (Recall binary logistic regression by default predicts the higher category of the dependent, which is cappun = 2 = opposing capital punishment.) For the first category of marital (1=married) in the example above, the odds ratio is .587. We would therefore say that the odds of opposing capital punishment compared to favoring it are decreased by a factor of .587 when the respondent is married compared to being never married, controlling for other variables in the model. Similar statements might be made about the other levels of marital, all making comparison to the reference category, never married, except for the fact that the other three contrasts are not significant (marital 2, 3, or 4 versus 5).

- *For continuous covariates*: For continuous covariates, when the parameter estimate, b, is transformed into an odds ratio, it may be simply expressed similar to the manner in other forms of regression. Thus in the example above, the covariate age has an odds ratio of 1.001. Since this is very close to 1.0 (no effect for odds ratios), it is not surprising that this is very non-significant. Were it significant, we would say that the odds of opposing capital punishment compared to favoring it increase by a factor of 1.001 for each year age increases, controlling for other variables in the model.

Example 2

In the figure on the next page, gunown is the dependent variable. Since its highest-coded value is 2 = do not own a gun in the home, not owning is being

predicted. Being white (race = 1) compared to other race (race = 3 = other race, the reference category) reduces the odds of not owning a gun in the home to a significant degree. "Reduces" corresponds to a negative parameter estimate and an odds ratio below 1.0. Being African-American (race=2) rather than other race has no significant effect.

All three independent variables in the model (sex, race, degree) are significant, but not all categories of each independent variable are significant. No category of any of the independent variables modeled significantly increases the odds of not having a gun in the home compared to the respective reference categories when other variables in the model are controlled - that is, there are no significant odds ratios above 1.0. (The chart output is from Excel).

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	sex(1)	-.614	.102	36.251	1	.000	.541
	race			70.698	2	.000	
	race(1)	-.950	.118	65.007	1	.000	.387
	race(2)	.109	.146	.564	1	.453	1.116
	degree			33.922	4	.000	
	degree(1)	-.479	.238	4.057	1	.044	.619
	degree(2)	-.736	.208	12.496	1	.000	.479
	degree(3)	-.913	.262	12.098	1	.001	.401
	degree(4)	-.066	.233	.081	1	.776	.936
	Constant	2.173	.233	86.886	1	.000	8.783

a. Variable(s) entered on step 1: sex, race, degree.

PREDICTING GUNOWN

BINARY LOGISTIC REGRESSION

so the highest coded category of gun ownership is the predicted category, and the lowest is the reference category.

What crosstabs would show:
More males than females own a gun in the home.
More whites than African-Americans own a gun in the home.
Gun ownership is highest among those with a junior college degree.

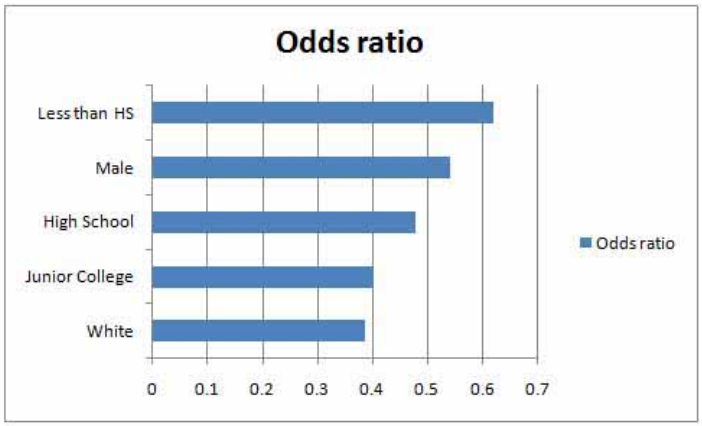
Coding:
Gunown: 1 = Have gun in home, 2 = Do not. The dependent.
Sex: 1 = male, 2 = female. Indicator coding.
Race: 1 = white, 2 = African-American, 3 = other. Deviation coding.
Degree: 0 = less than HS, 1 = HS, 2 = junior college, 3 = Bachelor, 4 = Grad. Indicator coding.

What is being predicted: Odds of not having a gun (Gunown = 2) compared to owning (= 1)

In English:

*** The odds of not owning a gun are decreased by a factor of .541 if the subject is male compared being female (is in the first category of sex compared to the reference category, which is the highest category), controlling for other variables in the model.**
*** The odds of not owning a gun are decreased by a factor of .387 if the subject is white (the first category of race) compared to the grand mean for all races (deviation coding is used), controlling for all other variables in the model.**
*** The odds of not owning a gun a increased non-significantly if the subject is African-American.**
*** The odds of not owning a gun are decreased the most (factor of .401) if the subject has a junior college degree (the third category of Degree) compared to a graduate degree.**

Graphic presentation:



The lower the odds ratio below 1.0, the greater the effect of that variable in reducing the odds of not owning a gun in the home compared to owning, controlling for all other variables in the model. Only significant variable effects are shown.

The comparison category for Male is Female. The comparison category for White is the mean of all races. The comparison category for degree categories shown in the chart is graduate degree holders.

For multinomial logistic regression

Example 1

In SPSS multinomial logistic regression, the dependent can be multinomial (the usual) but also can be binary. In SPSS, select Analyze, Regression, Multinomial Logistic. In SPSS syntax, this is the NOMREG procedure. Example 1 illustrates important differences between using a binary dependent in binary logistic regression and using the same binary dependent in the SPSS multinomial logistic regression procedure. SPSS will yield substantively identical results for the same model and data, but the logistic coefficients may differ because different reference categories may be used in binary compared to multinomial models. Significance levels will not be affected. In the example below, the dependent variable is the binary variable "cappun", coded 1 = favors capital punishment and 2 = opposes. The predictor variable is "degree", coded ordinally from 0 = less than high school to 4 = graduate degree. Crosstabs would show that those with bachelor's and graduate degrees are more likely to oppose capital punishment. Both binary and multinomial logistic regression output is shown in the figure below.

With "degree" entered as a covariate....

- In binary regression, the b coefficient for degree is positive .221 and the odds ratio is 1.248. A unit increase in degree increases by a factor of 1.248 the odds of being opposed to capital punishment rather than in favor. For the dependent, what is predicted is the highest value (2 = opposes). For the independent, there is no reference category as it is treated as a covariate.

- In multinomial regression, the b coefficient for degree is negative (-.221) and the odds ratio is .801. A unit increase in degree decreases by a factor of .801 the odds of being in favor of capital punishment rather than opposed. For the dependent, what is predicted is the lowest value (1 = favors). For the independent, there is no reference category as it is treated as a covariate.

With "degree" entered as a factor....

- In binary regression, there is no b coefficient for degree as a whole. Rather, there are b coefficients for each level of degree except the highest. Thus note that in binary output for a factor (unlike multinomial output), SPSS will count levels from 1 to 5 even if the actual codes on file are from 0 to 4. Degree= 1 in the table is actually coded 0, for instance, and the highest coded value (4 = graduate) is the omitted reference category 5. Having less than a high school degree (being in the degree(1) level) rather than having a graduate degree reduces by a factor of .557 the odds of being opposed to capital punishment rather than in favor. For the dependent, what is predicted is the highest value, cappun = 2 = opposes capital punishment. For the independent, degree, the highest value (4 = graduate) is the reference category.

- In multinomial regression, there are b coefficients and odds ratios for degree(0) through degree(4), corresponding to the original codes 1 through 5. Having less than a high school degree (degree(0)) rather than having a graduate degree (4) increases by a factor of 1.796 the odds of favoring capital punishment rather than opposing. For the dependent, what is predicted is the lowest value, cappun = 1 = favors capital punishment. For the independent, degree, the highest value is still the reference category.

Comparison of Binomial and Multinomial Defaults

Dependent variable	Independent variable
cappun	degree
1 = favors capital punishment	0 = less than high school
2 = opposes	1 = high school
	2 = junior college
	3 = bachelor
	4 = graduate

Binomial logistic regression with degree entered as a covariate (the default)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	degree	.221	.053	17.603	1	.000	1.248
	Constant	-1.564	.106	219.124	1	.000	.209

a. Variable(s) entered on step 1: degree.

Binomial logistic regression with degree entered as a factor, with the default indicator coding and default highest category as reference

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	degree			25.940	4	.000	
	degree(1)	-.586	.258	5.157	1	.023	.557
	degree(2)	-.775	.228	11.600	1	.001	.461
	degree(3)	-.739	.347	4.540	1	.033	.478
	degree(4)	-.028	.253	.012	1	.912	.973
	Constant	-.693	.207	11.211	1	.001	.500

a. Variable(s) entered on step 1: degree.

Multinomial logistic regression with degree entered as a covariate

Parameter Estimates

Favor or Oppose Death Penalty for Murder[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
Favor	Intercept	1.564	.106	219.124	1	.000			
	degree	-.221	.053	17.603	1	.000	.801	.723	.889

a. The reference category is: Oppose.

Multinomial logistic regression with degree entered as a factor

Parameter Estimates

Favor or Oppose Death Penalty for Murder[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp (B)	
								Lower Bound	Upper Bound
Favor	Intercept	.693	.207	11.211	1	.001			
	[degree=0]	.586	.258	5.157	1	.023	1.796	1.083	2.978
	[degree=1]	.775	.228	11.600	1	.001	2.172	1.390	3.393
	[degree=2]	.739	.347	4.540	1	.033	2.094	1.061	4.132
	[degree=3]	.028	.253	.012	1	.912	1.028	.627	1.687
	[degree=4]	0[b]	.	.	0	.	.	.	.

Example 2

For the multinomial logistic output below, the multinomial variable work status (1=full time, 2=part time, 3 = other, 4 = unemployed) is predicted from sex (a dichotomy, with 1=male, 2=female), marital status (a categorical variable with 5 levels, with 1 = married and the reference category = 5 = never married)), and age (a continuous variable).

- *For dichotomies*. Sex (1=male, 2=female) may be entered as a factor or as a covariate. If entered as a factor, the higher value is predicted and the lower category is the reference. If entered as a covariate, the lower value is predicted and the higher category is the reference. Here, sex is entered as a covariate, so male is predicted and female is the reference. The major difference from the binary situation is that sex now has three b coefficients with three odds ratios (exp(b)), one for each level of the multinomial dependent (Work) except its reference category (4=unemployed). In the figure below, for Work = 2 = part-time, the odds ratio for sex is 4.552, where sex is entered as a covariate dichotomy. We can therefore say that the odds of being part-time rather than unemployed is increased by a factor of about 4.6 by being male rather than female, controlling for other variables in the model. We cannot make a corresponding statement about full time work as that odds ratio is non-significant.

- *For categorical predictors (factors)*. Recall that the categorical variable "marital" has five levels: 1=married, 2=widowed, 3=divorced, 4=separated, 5=never married. The highest category, "Never married," is therefore the default reference category. For instance, for full-time work, the odds ratio for 1=married is 4.7. Therefore we can say that the odds of being full-time rather than unemployed is increased by a factor of 4.7 by being married rather than never married, controlling for other variables in the model.

- *For continuous covariates*. For covariates, b is the amount, controlling for any other predictor variables in the model, that the log odds of the dependent variable changes when the continuous independent variable changes one unit. Most researchers report covariates in terms of odds ratios. Age is an example of a continuous covariate in the example below. Age is significant only for work status = other, where its odds ratio is 1.075.

The closer the odds ratio is to 1.0, the closer the predictor comes to being independent of the dependent variable. so we can say age is not a strong predictor. Age is not significantly related to the odds of being in full- or part-time work compared to being unemployed. The only significant statement would be that a 1 year increase in age increases by 7.5% the odds of being in the "Other" work status category rather than "Unemployed", controlling for other variables in the model. (The "Other" category includes retired, housework, school, temporarily not working, and other).

Parameter Estimates

Work status[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
Full time	Intercept	1.648	.571	8.339	1	.004			
	sex	.475	.286	2.753	1	.097	1.608	.917	2.818
	age	.007	.012	.318	1	.573	1.007	.983	1.032
	[marital=1]	1.554	.381	16.640	1	.000	4.728	2.241	9.974
	[marital=2]	-.114	.706	.026	1	.872	.893	.224	3.560
	[marital=3]	.726	.418	3.022	1	.082	2.067	.912	4.685
	[marital=4]	17.806	.278	4.117E3	1	.000	5.407E7	3.138E7	9.314E7
	[marital=5]	0[b]	.	.	0	.	.	.	.
Part time	Intercept	-1.561	.629	6.154	1	.013			
	sex	1.516	.309	24.109	1	.000	4.552	2.486	8.335
	age	.015	.013	1.389	1	.239	1.016	.990	1.042
	[marital=1]	1.010	.403	6.276	1	.012	2.746	1.246	6.051
	[marital=2]	-.035	.747	.002	1	.963	.966	.224	4.173
	[marital=3]	-.286	.462	.382	1	.536	.751	.304	1.859
	[marital=4]	17.124	.421	1.658E3	1	.000	2.734E7	1.199E7	6.234E7
	[marital=5]	0[b]	.	.	0	.	.	.	.
Other	Intercept	-3.060	.605	25.608	1	.000			
	sex	1.429	.296	23.388	1	.000	4.175	2.339	7.451
	age	.072	.013	32.786	1	.000	1.075	1.049	1.102
	[marital=1]	.988	.393	6.318	1	.012	2.685	1.243	5.798
	[marital=2]	.158	.708	.050	1	.823	1.171	.292	4.695
	[marital=3]	-.394	.437	.812	1	.368	.674	.286	1.589
	[marital=4]	17.362	.000	.	1	.	3.471E7	3.471E7	3.471E7
	[marital=5]	0[b]	.	.	0	.	.	.	.

a. The reference category is: Unemployed.

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

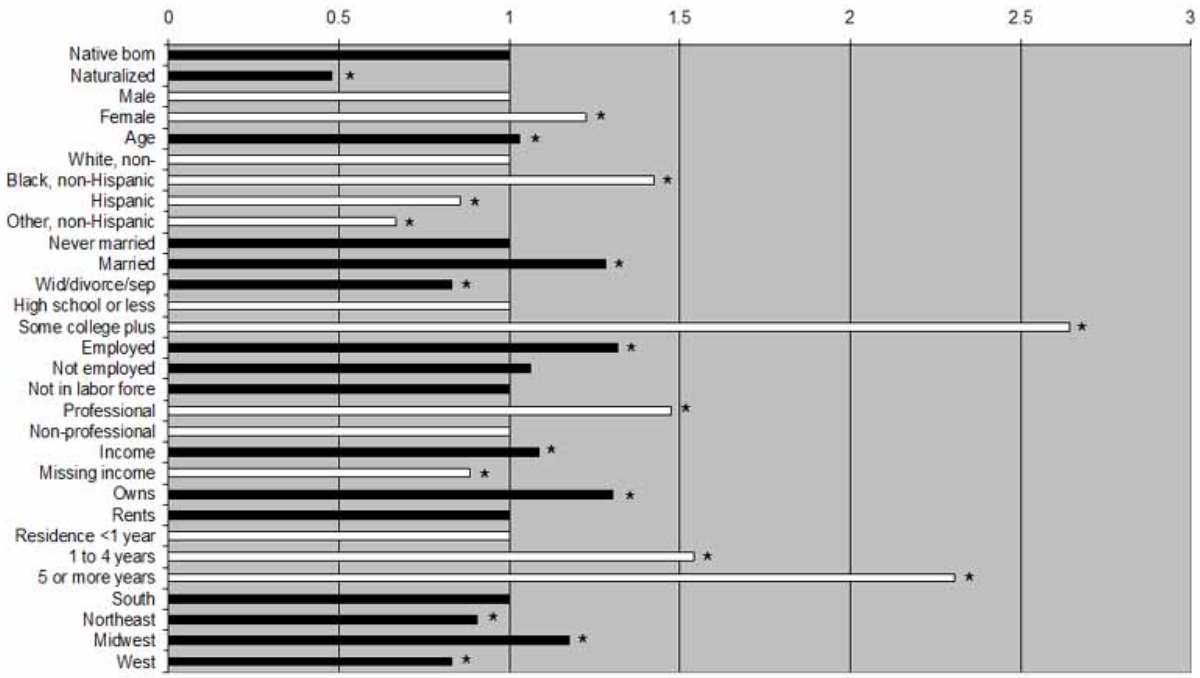
Coefficient significance and correlation significance may differ

Note that a logistic coefficient may be found to be significant when the corresponding correlation is found to be not significant, and vice versa. To make certain global statements about the significance of an independent variable, both the correlation and the parameter estimate (b) should be significant. Among the reasons why correlations and logistic coefficients may differ in significance are these:

(1) Logistic coefficients are partial coefficients, controlling for other variables in the model, whereas correlation coefficients are uncontrolled;
(2) Logistic coefficients reflect linear and nonlinear relationships, whereas correlation reflects only linear relationships; and
(3) A significant parameter estimate (b) means there is a relation of the independent variable to the dependent variable for selected control groups, but not necessarily overall.

Reporting odds ratios

Reporting odds ratios is illustrated below in a U. S. Census Bureau analysis of voter registration. In this figure the lengths of horizontal bars are proportionate to the size of the odds ratios for the corresponding values of such predictors as citizenship, gender, age, race, and other variables. For factors like citizenship, there is one odds ratio per level. For covariates like age, there is a single odds ratio. See the warning below the table, however!



*Retrieved 4/3/2011 from http://www.census.gov/hhes/www/socdemo/voting/publications/other/CrisseyFile_slides.ppt as an illustration of use of odds ratios to report effect size in logistic regression.

Warning. Recall that an odds ratio of 1.0 corresponds to no effect. An effect strongly increases the odds to the extent it is above 1.0. An effect strongly decreases the odds to the extent it is below 1.0. A chart like the one above may be misleading, making viewers think that the longer the line, the greater the effect. It would be much better to have a bidirectional graph, with 1.0 as the vertical center line and odds ratios greater than 1.0 shown going to the right of the center line. and odds ratios below 1.0 expressed as 1 minus the odds ratio and going to the left of the center line. The graph would have a label such as "Effect sizes based on odds ratios" but should not be labeled "Odds Ratios" due to the subtraction operation for values below 1.0. A graph footnote should explain the difference between the right- and left-hand sides of the graph.

Odds ratios: Summary

The odds ratio, labeled Exp(b) in SPSS output, is the natural log base, e, to the exponent, b, where b is the logistic regression parameter estimate. When b=0, Exp(b)=1, so therefore an odds ratio of 1 corresponds to an explanatory variable which does not affect the dependent variable. For continuous variables, the odds ratio represents the factor by which the odds(event) change for a one-unit change in the variable. An odds ratio greater than 1.0 means the independent variable increases the odds of the event being predicted. If the odds ratio is less than 1.0, then the independent variable decreases the odds(event). To convert between odds ratios and parameter estimates, the following formulas are used:

$$\text{odds ratio} = \exp(b)$$
$$b = \ln(\text{odds ratio})$$

Effect size

The odds ratio is a measure of effect size. The ratio of odds ratios greater than 1.0 is the ratio of relative importance of the independent variables which increase the odds associated with the dependent variable. Similarly, the ratio of odds ratios less than 1.0 is the same for negative effects. Note standardized logit coefficients may also be used, as discussed below, but then one is discussing relative importance of the independent variables in terms of effect on the dependent variable's log odds, which is less intuitive.

Confidence interval on the odds ratio

SPSS labels the odds ratio "Exp(B)" and prints "Low" and "High" confidence levels for it. If the low-high range contains the value 1.0, then being in that variable value category makes no difference on the odds of the dependent, compared to being in the reference (usually highest) value for that variable. That is, when the 95% confidence interval around the odds ratio includes the value of 1.0, indicating that a change in value of the independent variable is not associated in change in the odds of the dependent variable assuming a given value, then that variable is not considered a useful predictor in the logistic model.

Warning: very high or very low odds ratios

Extremely high odds ratios may occur when a factor cell associated with the odds ratio (a cell which would appear in a crosstabulation of the dependent variable with a factor) is zero. The same situation can be associated with extremely low odds ratios. What has happened is that the algorithm estimating the logistic coefficient (and hence also exp(b), the odds ratio) is unstable, failing to converge while attempting to move iteratively toward positive infinity (or negative infinity). When extremely high odds ratios occur, the researcher can confirm this cause by looking for extremely large standard errors and/or running the appropriate crosstabulation and looking for zero cells. Large odds ratio estimates may also arise from sparse data (ex., cells < 5); from small sample size in relation to a large number of covariates used as controls (see Greenland, Schwartzbaum, & Finkle, 2000); and from employing a predictor variable which has low variance. The presence of sparse or zero cells may require collapsing levels of the variable or omitting the variable altogether.

Comparing the change in odds for different values of X

The odds ratio is the factor by which odds(event) changes for a 1 unit change in X. But what if years_education was the X variable and one wanted to know the change factor for X=12 years vs. X=16 years? Here, the X difference is 4 units. The change factor is not Exp(b)*4. Rather, odds(event) changes by a factor of $Exp(b)^4$. That is, odds(event) changes by a factor of Exp(b) raised to the power of the number of units change in X.

Comparing the change in odds when interaction terms are in the model

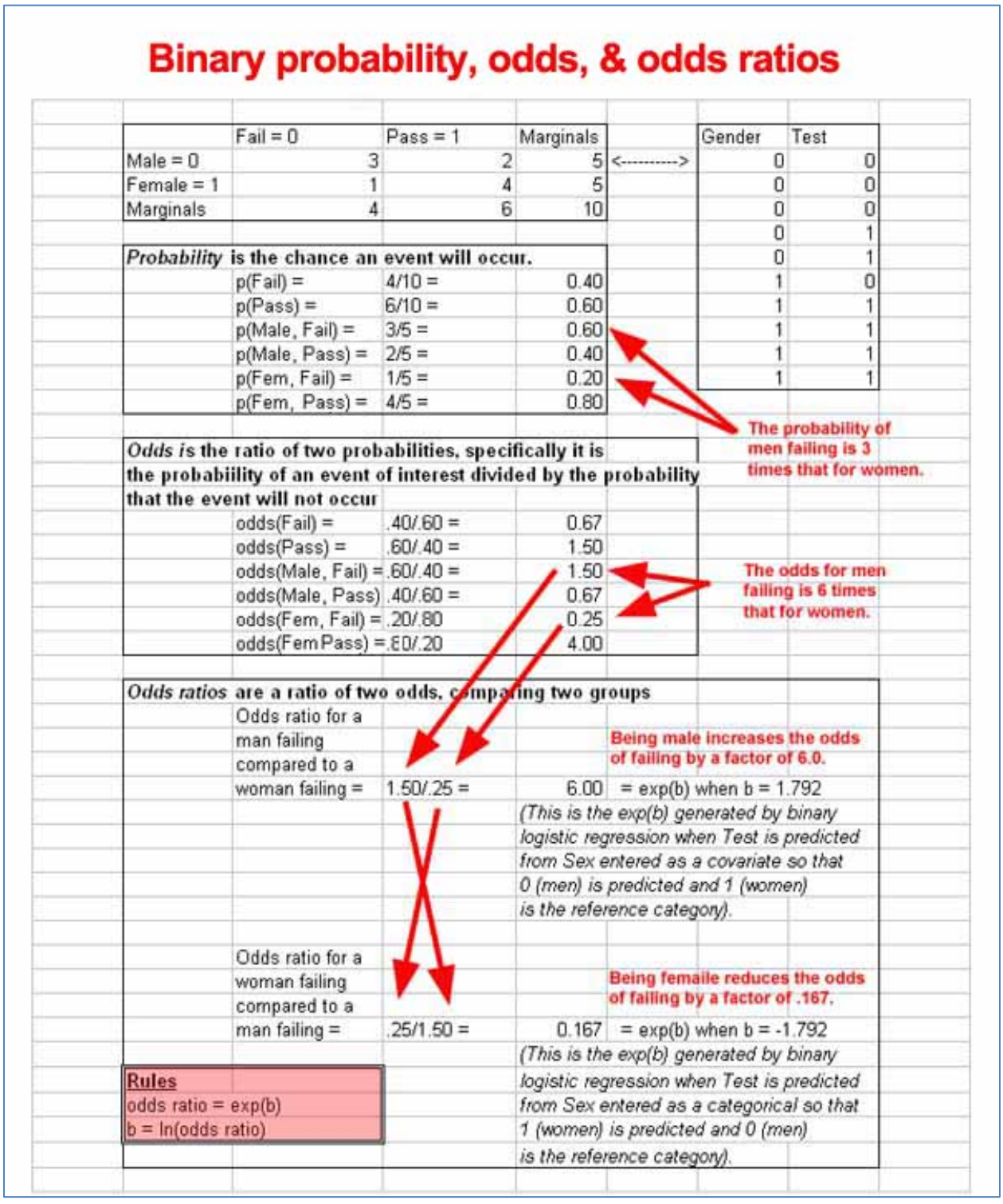
In general, the odds ratio represents the factor by which odds(event) is multiplied for a unit increase in the X variable. However, the effect of the X variable is not properly gauged in this manner if X is also involved in interaction effects which are also in the model. Before exponentiating, the b coefficient must be adjusted to include the interaction b terms. Let X_1 be years_education and let X_2 be a dichotomous variable called "school_type" coded 0=private school, 1=public school, and let the interaction term $X_1 * X_2$ also be in the model. Let the b coefficients be .864 for X_1, .280 for X_2, and .010 for X_1*X_2. The adjusted b, which we shall label b*, for years education = .864 + .010*school_type. For private schools, b* =.864. For public schools b* = .874. Exp(b*) is then the estimate of the

odds ratio, which will be different for different values of the variable (here, school_type) with which the X variable (here, years_education) is interacting.

Probabilities, logits, and odds ratios

Probabilities

As illustrated in the spreadsheet example below, a probability is the chance that an event will occur. An odds is the ratio of two probabilities. An odds ratio is the ratio of two odds. The figure below illustrates the transition from probabilities to odds to odds ratios for the example of sex as a predictor of passing or failing a test (the binary dependent variable).



In a second spreadsheet example, the transition from probabilities to odds to odds ratios is illustrated below for an example in which low, medium, and high

levels of training are used to predict low, medium, and high scores on a performance test (the multinomial dependent variable).

Multinomial probability, odds, & odds ratios

Score * Training Crosstabulation

Count

		Training			
		Low	Medium	High	Total
Score	Low	4	2	1	7
	Medium	2	2	2	6
	high	1	2	6	9
Total		7	6	9	22

By default, high score and high training are the reference categories.

Formula

A	p(Low training, low score) =	4/7 =	0.57
B	p(high training, low score)=	1/9 =	0.11
C	p(low training, high score) =	1/7 =	0.14
D	p(high training, high score) =	6/9 =	0.67

Probabilities given low or high training

E odds of getting a low score compared to getting a high score if in the low training group
= odds(p(low training, low score))/ p(low training, high score)
= odds(A/C) = .57/.14 = 4.00

F odds of getting a low score compared to getting a high score if in the high training group
= odds(p(high training, low score))/ p(high training, high score)
= odds(B/D) = .11/.67 = 0.17

Odds

G odds ratio for training = low
= E/F = 24.00

This is the factor by which being low training compared to high training increases the odds of scoring low compared to scoring high. It is also the Exp(B) computed by SPSS multinomial regression for Training=Low when Score=Low, for b = 3.178.

ln(24.00) = 3.178
exp(3.178) = 24.00

Odds ratio

Parameter Estimates (SPSS Multinomial Regression, Default Reference Groups)

Score[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% CI for Exp(B) Lower Bound	95% CI for Exp(B) Upper Bound
Low	Intercept	-1.792	1.080	2.752	1	.097			
	[Training= Low]	3.178	1.555	4.179	1	.041	24.000	1.140	505.194
	[Training= Medium]	1.792	1.472	1.482	1	.224	6.000	.335	107.420
	[Training= High]	0[b]	.	.	0	.	.	.	.
Medium	Intercept	-1.099	.816	1.810	1	.178			
	[Training= Low]	1.792	1.472	1.482	1	.224	6.000	.335	107.420
	[Training= Medium]	1.099	1.291	.724	1	.395	3.000	.239	37.672
	[Training= High]	0[b]	.	.	0	.	.	.	.

Probability interpretations. While logistic coefficients are usually interpreted via odds, not probabilities, it is possible to use probabilities. In the logistic regression formula discussed above, Exp(z) is odds(event). Therefore the quantity (1 - Exp(z)) is odds(nonevent). Therefore P(event) = Exp(z)/(1 + Exp(z)). Recall z = the constant plus the sum of crossproducts of the b coefficients times the values of their respective X (independent) variables. For dichotomous independents assuming the values (0,1), the crossproduct term is null when X = 0 and is b when X=1. For continuous independents, different probabilities will be computed depending on the value of X. That is, P(event) varies depending on the covariates.

Relative risk ratios (RRR)

The Stata mlogit multinomial logit regression procedure outputs the relative risk ratio, RRR, sometimes also labeled relative risk reduction or risk response ratio. RRR is sometimes interpreted as equivalent to an odds ratio, but it is not mathematically the same. Nonetheless, similar substantive conclusions are apt to flow from interpreting RRR as a variant of the odds ratio. RRR is the probability that selecting a given level of the predictor increases or decreases the dependent (in binary models, increases or decreases the probability that the dependent=1) relative to selecting the baseline level.

More about significance tests

Overview

Although described in examples above, this section summarizes the main significance tests associated with logistic regression. To minimize redundancy, discussion in the SAS and Stata sections which would repeat that in the SPSS section is omitted.

Significance of the model

SPSS

Omnibus tests of model coefficients

For binary logistic regression, SPSS outputs the "Omnibus Tests of Model Coefficients" table, Illustrated above. This likelihood ratio chi-square test tests if the model with the predictors is significantly different from the model with only

the intercept. The omnibus test may be interpreted as a test of the capability of predictors in the model jointly to predict the response (dependent) variable. A finding of significance corresponds to a conclusion that there is adequate fit of the data to the model, meaning that at least one of the predictors is significantly related to the response variable.

Hosmer-Lemeshow test

Also for binary logistic regression the Hosmer and Lemeshow test of goodness of fit test was illustrated and discussed above. This test is often preferred over the omnibus test as an overall test of the model. A well-fitting model is non-significant by this test.

Likelihood ratio test of the model

For multinomial regression, the likelihood ratio test appears in SPSS in the "Model Fitting Information" table, discussed and illustrated above. It is the preferred test of the overall model. A well-fitting model is significant by this test

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1.209E3	1.214E3	1.207E3			
Final	1.051E3	1.100E3	1.033E3	174.279	8	.000

The likelihood ratio test, also called the log-likelihood test, is based on the -2LL (deviance) statistic. The likelihood ratio test is a test of the significance of the difference between the likelihood ratio (-2LL) for the researcher's model (Final) minus the likelihood ratio for a reduced model (such as the “Intercept Only” model). The likelihood ratio test is generally preferred over its alternative, the Wald test, discussed below. There are three main forms of the likelihood ratio test::

1. *Testing the overall model*. Two models are referenced in the SPSS "Model Fitting Information" table: (1) the "Intercept Only" model, also called the null model, in which the intercept reflects the net effect of all variables not

in the model plus error; and (2) the "Final" model, also called the fitted model, which is the researcher's model with predictor variables. The logistic equation is the linear combination of predictor variables which maximizes the log likelihood that the dependent variable equals the predicted value/class/group. By default, the predicted value is the highest-coded level. The difference in the --2LL measures can be used to test how much the final model improves over the null model.

2. *The researcher's model vs. the null model.* The likelihood ratio test shown above in the "Final" row in the "Model Fitting Information" in SPSS tests the researcher's fitted model with the null (intercept-only) model. Deviance (-2LL) for the fitted model will be less than for the null model, and if it is enough less, it will be found to be significant, which means significantly different from the null model. That is, a finding of significance (p<=.05 is the usual cutoff) leads to rejection of the null hypothesis that all of the predictor effects are zero. When this likelihood test is significant, at least one of the predictors is significantly related to the dependent variable. Put another way, when this test is significant, the researcher rejects the null hypothesis that knowing the independents makes no difference in predicting the dependent in logistic regression.

3. *Testing the difference between any two nested models.* In the same manner, the likelihood ratio test can test the significance of the difference between any two models, provided one is nested under the other (that is, that the parameters in the "reduced model" are a subset of the parameters in the "full model"). The difference in the -2LL statistics can be tested using a chi-square table, with degrees of freedom equal to the difference in degrees of freedom (df) for the two models. For a given model, df = the number of terms in the model minus 1 (for the constant).

Warning: If the log-likelihood test statistic shows a small p value (<=.05) for a model with a large effect size, ignore contrary findings based on the Wald statistic discussed above as it is biased toward Type II errors in such instances - instead assume good model fit overall.

Goodness of fit test

For multinomial logistic regression, Pearson and deviance goodness of fit tests appear in the optional "Goodness of Fit" table in multinomial logistic regression in SPSS, illustrated below. This not default output but must be checked in the Statistics button dialog. These two tests each test overall model fit. Adequate fit corresponds to a finding of nonsignificance for these tests, as in the illustration below.

Both are chi-square methods, but the Pearson statistic is based on traditional chi-square and the deviance statistic is based on likelihood ratio chi-square. The deviance test is preferred over the Pearson (Menard, 2002: 47). Either test is preferred over classification tables when assessing model fit but are less preferred than the likelihood ratio test of the model, discussed above. Both tests usually yield the same substantive results. The goodness of fit deviance test compares the researcher's model with the saturated model, whereas the likelihood ratio test compares it with the intercept-only model. When the two conflict, the likelihood ratio test is preferred, though conflict suggests a weak model.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	584.307	606	.730
Deviance	639.453	606	.168

Goodness of Fit Index (obsolete)

Also known as *Hosmer and Lemeshow's goodness of fit index* or *C-hat,* this is an alternative to model chi-square for assessing the significance of a logistic regression model. This is different from Hosmer and Lemeshow's (chi-square) goodness of fit test discussed above. Menard (p. 21) notes it may be better when the number of combinations of values of the independents is approximately equal to the number of cases under analysis. This measure was included in SPSS output as "Goodness of Fit" prior to Release 10. However, it was removed from the reformatted output for SPSS Release 10 because, as noted by David Nichols, senior statistician for SPSS, it "is done on individual cases and does not follow a

known distribution under the null hypothesis that the data were generated by the fitted model, so it's not of any real use" (SPSSX-L listserv message, 3 Dec. 1999).

SAS

The "Testing Global Null Hypothesis: BETA=0" table shown earlier above is what SPSS calls the "omnibus test", based on likelihood ratio chi-square. A finding of significance demonstrates that the researcher's model is significantly different from the intercept-only model.

The "Hosmer and Lemeshow Goodness of Fit Test" table shown above is sometimes preferred as a test of the model. A finding of non-significance corresponds to finding the model has acceptable fit.

The "Maximum Likelihood Analysis of Variance" table shown above for multinomial logistic regression, in the "Likelihood Ratio" row, shows the likelihood ratio test of the model, where a finding of significance corresponds to acceptable fit.

Stata

In its default output shown above, Stata output the "Prob > chi2" statistic, which corresponds to the omnibus test in SPSS or the global test in SAS. A finding of significance corresponds to the model having acceptable fit.

Using the `estat gof` post-estimation command, it is also possible to obtain the Hosmer-Lemeshow goodness of fit test for the model, as illustrated above. A finding of non-significance corresponds to acceptable model fit.

Significance of parameter effects

SPSS

Wald tests

The SPSS "Variables in the Equation" table in binary logistic regression, illustrated and discussed above, contains significance tests for each predictor variable. Wald tests also appear in the "Parameter Estimates" table in multinomial regression, also illustrated and discussed above. The Wald statistic tests the null hypothesis that a particular parameter estimate is zero. The researcher may well want to

drop independents from the model when their effect is not significant by the Wald statistic. When Wald tests conflict with likelihood ratio tests (described below), the latter are preferred.

Warning: Computationally, the Wald statistic is the squared ratio of the unstandardized logistic coefficient to its standard error. Menard (2002: 39) warns that for large logit coefficients, standard error is inflated, lowering the Wald statistic and leading to Type II errors (false negatives: thinking the effect is not significant when it is). That is, there is a flaw in the Wald statistic such that very large effects may lead to large standard errors and small Wald chi-square values. For models with large logit coefficients or when dummy variables are involved, it is better to test the difference using the likelihood ratio test of the difference of models with and without the parameter. Also note that the Wald statistic is sensitive to violations of the large-sample assumption of logistic regression. Put another way, the likelihood ratio test is considered more reliable for small samples (Agresti, 1996). For these reasons, the likelihood ratio test of individual model parameters is generally preferred.

Stepwise or block tests

When predictors are entered stepwise or in blocks, the Hosmer-Lemeshow and omnibus tests may also be interpreted as tests of the change in model significance when adding a variable (stepwise) or block of variables (block) to the model. This is illustrated below for a model predicting having a gun in the home from marital status, race, and attitude toward capital punishment. In stepwise forward logistic regression, Marital is added first, then Race, then Cappun; Age is in the covariate list but is not added as the stepwise procedure ends at Step 3.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	2	1.000
2	.927	4	.921
3	4.059	6	.669

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	82.314	4	.000
	Block	82.314	4	.000
	Model	82.314	4	.000
Step 2	Step	61.009	2	.000
	Block	143.323	6	.000
	Model	143.323	6	.000
Step 3	Step	27.873	1	.000
	Block	171.196	7	.000
	Model	171.196	7	.000

Testing individual model parameters

By extension, if the reduced model is a given model dropping one parameter, then the likelihood ratio test of difference in models is a test of the significance of the dropped parameter, as discussed above. That is, the likelihood ratio test tests if the logistic regression coefficient for the dropped variable can be treated as 0, thereby justifying dropping the variable from the model. A non-significant likelihood ratio test indicates no difference between the full and the reduced models, hence justifying dropping the given variable so as to have a more parsimonious model that works just as well. This test is preferred over the Wald test for testing individual model parameters. This is what is done in the likelihood ratio tests in the table illustrated below. For this example, age could be dropped as non-significant, but dropping "marital" would result in the greatest loss of model fit.

Likelihood Ratio Tests

Effect	Model Fitting Criteria: AIC of Reduced Model	Model Fitting Criteria: BIC of Reduced Model	Model Fitting Criteria: -2 Log Likelihood of Reduced Model	Likelihood Ratio Tests: Chi-Square	Likelihood Ratio Tests: df	Likelihood Ratio Tests: Sig.
Intercept	1.051E3	1.100E3	1.033E3[a]	.000	0	.
age	1.052E3	1.095E3	1.036E3	3.083	1	.079
race	1.089E3	1.127E3	1.075E3	42.181	2	.000
marital	1.101E3	1.128E3	1.091E3	57.978	4	.000
cappun	1.077E3	1.120E3	1.061E3	28.211	1	.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.

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

Testing for interaction effects. A common use of the likelihood ratio test is to test the difference between a full model and a reduced model dropping an interaction effect. If model chi-square (which is -2LL for the full model minus -2LL for the reduced model) is significant, then the interaction effect is contributing significantly to the full model and should be retained.

Block chi-square is a synonym for the likelihood ratio test, referring to the change in -2LL due to entering a block of variables. The logistic regression dialog in SPSS allows the researcher to enter independent variables in blocks. Blocks may contain one or more variables.

Sequential logistic regression is analysis of nested models using block chi-square, where the researcher is testing the control effects of a set of covariates. The logistic regression model is run against the dependent for the full model with independents and covariates, then is run again with the block of independents dropped. If chi-square difference is not significant, then the researcher concludes that the independent variables are controlled by the covariates (that is, they have no effect once the effect of the covariates is taken into account). Alternatively, the nested model may be just the independents, with the covariates dropped. In that case a finding of non-significance implies that the covariates have no control effect.

Assessing dummy variables with sequential logistic regression. Similarly, running a full model and then a model with all the variables in a dummy set dropped (ex., East, West, North for the variable Region) allows assessment of dummy variables, using chi-square difference. Note that even though SPSS computes log-likelihood ratio tests of individual parameters for each level of a dummy variable, it is preferred to assess the dummy variables as a set using block chi-square rather than using tests of individual dummy parameters. Because all dummy variables associated with the categorical variable are entered as a block this is sometimes called the "block chi-square" test and its value is considered more reliable than the Wald test, which can be misleading for large effects in finite samples.

SAS

The "Analysis of Maximum Likelihood Estimates" table shown above for binary logistic regression, shows the significance of each continuous predictor variable and of each level of each categorical predictor variable. A finding of significance corresponds to an effect different from 0.

The "Maximum Likelihood Analysis of Variance" table shown above for multinomial logistic regression, in the rows for individual predictor effects, shows the likelihood ratio test of each effect, where a finding of significance corresponds to an effect different from 0.

The "Analysis of Maximum Likelihood Estimates" table, shown above for multinomial logistic regression, shows the significance of each continuous predictor variable and of each level of each categorical predictor variable. A finding of significance corresponds to an effect different from 0.

Stata

In its default output shown above for binary logistic regression, Stata shows the significance of each continuous predictor variable and of each level of each categorical predictor variable. A finding of significance corresponds to an effect different from 0.

For multinomial logistic regression, a similar table appears, illustrated above. Again, a finding of significance corresponds to an effect different from 0.

More about effect size measures

Overview

Researchers should report effect size as well as significance for logistic models. Below, effect size measures are reviewed, using SPSS output for illustration.

Effect size for the model

Pseudo R-squared

There is no widely-accepted direct analog to OLS regression's R^2. This is because an R^2 measure seeks to make a statement about the "percent of variance explained," but the variance of a dichotomous or categorical dependent variable depends on the frequency distribution of that variable. For a dichotomous dependent variable, for instance, variance is at a maximum for a 50-50 split and the more lopsided the split, the lower the variance. This means that R-squared measures for logistic regressions with differing marginal distributions on their respective dependent variables cannot be compared directly, and comparison of logistic R-squared measures with R^2 from OLS regression is problematic at best.

Nonetheless, a number of logistic R-squared measures have been proposed, all of which may be reported as approximations to OLS R^2, not as "percent of variance explained" but rather just characterized as weak, moderate, or strong. Be aware that many researchers consider these R-square substitutes to be of only marginal interest and that the classification rate, discussed below, is a preferable measure of effect size. Note that R^2-like measures below are not goodness-of-fit tests but rather attempt to measure strength of association. Unfortunately, the pseudo-R^2 measures reflect and confound effect strength with goodness of fit. For small samples, for instance, an R^2-like measure might be high when goodness of fit was unacceptable by the likelihood ratio test. SPSS supports three R^2-like measures: Cox and Snell's, Nagelkerke's, and McFadden's, as illustrated below. Output is identical for binary and multinomial logistic regression and in SPSS appears in the "Pseudo R Square" table.

Pseudo R-Square

Cox and Snell	.097
Nagelkerke	.133
McFadden	.078

- *Cox and Snell's R^2* is an attempt to imitate the interpretation of multiple R-Square based on the log likelihood of the final model vs. log likelihood for the baseline model, but its maximum can be (and usually is) less than 1.0, making it difficult to interpret. It is part of SPSS output in the "Model Summary" table. In Stata's fitstat extension output, it is "Maximum Likelihood R2".
- *Nagelkerke's R^2* is a modification of the Cox and Snell coefficient norming it to vary from 0 to 1. That is, Nagelkerke's R^2 divides Cox and Snell's R^2 by its maximum in order to achieve a measure that ranges from 0 to 1. Therefore Nagelkerke's R^2 will normally be higher than the Cox and Snell measure but will tend to run lower than the corresponding OLS R^2. Nagelkerke's R^2 is part of SPSS output in the "Model Summary" table and is the most-reported of the pseudo R^2 estimates. Stata reports this coefficient as Cragg & Uhler's R2 in output of its fitstat extension. See Nagelkerke (1991).
- *McFadden's R^2* is a less common pseudo-R^2 variant, based on log-likelihood kernels for the full versus the intercept-only models. It is the default pseudo R^2 coefficient in Stata. See McFadden (1974).
- *Pseudo-R^2* is a Aldrich and Nelson's coefficient which serves as an analog to the squared contingency coefficient, with an interpretation like R-square. Its maximum is less than 1. It may be used in either binary or multinomial logistic regression.
- *Hagle and Mitchell's Pseudo-R^2* is an adjustment to Aldrich and Nelson's Pseudo R-Square and generally gives higher values which compensate for the tendency of the latter to underestimate model strength.
- *R^2* is OLS R-square, which can be used in binary logistic regression (see Menard, p. 23) but not in multinomial logistic regression, although to do so violates its normal distribution assumption and therefore this measure is not recommended. To obtain R-square, save the predicted values from logistic regression and run a bivariate regression on the observed dependent values. Note that logistic regression can yield deceptively high

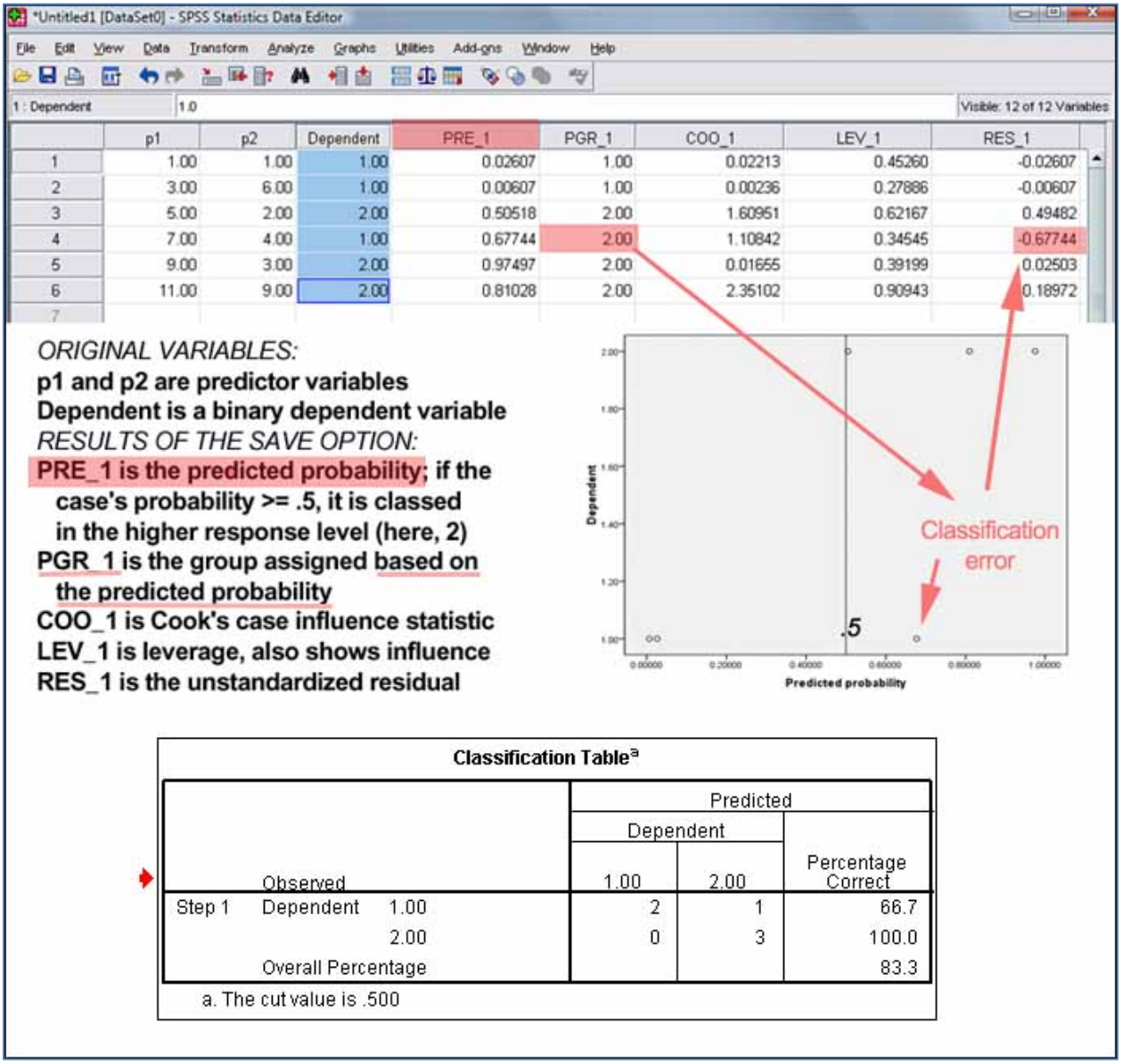
R^2 values when you have many variables relative to the number of cases, keeping in mind that the number of variables includes k-1 dummy variables for every categorical independent variable having k categories.

Classification tables

Classification tables are the 2 x 2 tables in the binary logistic regression or the 2 x n tables for multinomial logistic regression which tally correct and incorrect estimates. The columns are the predicted values of the dependent while the rows are the observed values of the dependent. In a perfect model, the overall percent correct will be 100%. Classification tables were illustrated and discussed above.

Classification tables vs. pseudo R-squared. Although classification hit rates (percent correct) as overall effect size measures are preferred over pseudo-R^2 measures, they to have some severe limitations for this purpose. Classification tables should not be used exclusively as goodness-of-fit measures because they ignore actual predicted probabilities and instead use dichotomized predictions based on a cutoff (ex., .5). For instance, in binary logistic regression, predicting a 0-or-1 dependent, the classification table does not reveal how close to 1.0 the correct predictions were nor how close to 0.0 the errors were. A model in which the predictions, correct or not, were mostly close to the .50 cutoff does not have as good a fit as a model where the predicted scores cluster either near 1.0 or 0.0. Also, because the hit rate can vary markedly by sample for the same logistic model, use of the classification table to compare across samples is not recommended.

Logistic classification. Using the logistic regression formula (discussed above with reference to the b parameters), logistic regression computes the predicted probabilities of being in the group of interest (for the illustration below, Dependent=2, since the binary dependent is coded 1 and 2). Any case whose predicted probability is .5 or greater is classified in the higher (2) class. For binary dependents, any residual greater than absolute(.5) represents a classification error.



What hit rate is "good"? The "hit rate" is simply the number of correct classifications divided by sample size, and is listed as "percentage correct" for the model. Researchers disagree about when the hit rate is "high enough" to consider the model "good". There is agreement, however, that the hit rate should be higher than the chance hit rate -- although there are two ways of defining "chance hit rate". Many researchers want the observed hit rate to be some percentage, say 25%, better than the chance rate to consider a model "good". That is, the researcher (1) picks one of the two definitions of "chance hit rate" and computes

it for the given data; (2) multiplies the chance hit rate by the improvement factor (ex., *1.25), giving the criterion hit rate; and (3) then if the observed hit rate is as high or higher than the criterion hit rate, the model is considered "good".

STEPWISE OUTPUT

Classification Table[a]

Observed			Predicted		
			Have gun in home		
			Yes	No	Percentage Correct
Step 1	Have gun in home	Yes	0	653	.0
		No	0	1213	100.0
	Overall Percentage				65.0
Step 2	Have gun in home	Yes	185	468	28.3
		No	206	1007	83.0
	Overall Percentage				63.9

a. The cut value is .500

Chance rate as proportional reduction in error. The figure above is actually two classification tables, with the upper section being for the baseline or intercept-only model, and the bottom section being for the researcher's final model. Such a double classification table is generated by SPSS when stepwise logistic regression is requested. The comparison is very useful. In the example above, in the bottom final model portion, the final model classifies 63.9% correct, which may seem moderately good. However, looking at the upper intercept-only portion, we see that the null model actually did better, classifying 65% correct. It did so simply by using the most numerous category (gunown = no) to classify all cases, thereby getting 1,213 correct. This is the proportional reduction in error (PRE) definition of "chance hit rate." PRE is actually negative for these data: (63.9 - 65)/65 = -1.1/65 = -.02. There is actually a 2% increase in error (rounded off) when employing the model compared to chance, by this definition of chance. The model percent correct of 63.9% is thus not as good as it might first appear.

- ✓ Note that while no particular split of the dependent variable is assumed, the split makes a difference. In the example above, where 65% of respondents say there is not a gun in the home, guessing "No" leads to

being right by chance 65% of the time. If the dependent is split 99:1 instead of 65:35, then one could guess the value of the dependent correctly 99% of the time just by always selecting the more common value. The more lopsided the split, the harder it is for the researcher's logistic model to do better than chance in a PRE definition of chance. This does not mean the model's predictor variables are non-significant, just that they do not move the estimates enough to make a difference compared to pure guessing.

Chance rate as proportional by chance. A second, more lenient definition of "chance" is the proportional by chance (PC) method. In this method, chance is defined as the sum of squared prior probabilities. That is, one squares the percentages of cases in each level of the dependent and sums. A third way of putting it is the PC is the sum of squared dependent marginals when the marginals are expressed as percentages. (The "Case Summary Table" of SPSS output provides the dependent marginals). For the table above, there are 653 gunown="yes" respondents, or 34.99%, and 1,213 (65.01%) gunown="no" respondents. The PC chance rate is thus $.3499^2 + .6501^2 = .1225 + .4226 = .5450$ (rounding error taken into account). For more on PC see Hair et al, 1987: 89-90; Hand, Mannila, & Smyth, 2001.

Improvement criteria Thus the proportional by chance (PC) definition of chance gives a chance hit rate of 54.5%. The proportional reduction in error (PRE) definition of chance gave a chance hit rate of 65.0%. The observed hit rate, 63.9 is in-between. Is that "good enough"? A common rule of thumb (improvement criterion) is that a model should do 25% better than chance. By this rule, the PC criterion level would be .545*1.25 = .681. The PRE criterion level would be .650*1.25 = .813. By either improvement criterion, the observed hit rate)63.9%) does not indicate a "good model" for these data. However, it is up to the researcher to define what is "good". The most liberal criterion would be any observed hit rate above the PC baseline (.545), which would make the model above a "good" one. The most conservative criterion would be any observed hit rate above the improvement criterion PRE rate (.813). "Good" is in the eye of the beholder and how "chance" is defined. Nearly all researchers would agree, however, that an observed hit rate below the PC baseline is a poor model.

Terms associated with classification tables:

- *Hit rate:* Number of correct predictions divided by sample size. The hit rate for the model should be compared to the hit rate for the classification table for the constant-only model (Block 0 in SPSS output). The Block 0 rate will be the percentage in the most numerous category (that is, the null model predicts the most numerous category for all cases).

- *Sensitivity:* Percent of correct predictions in the reference category of the dependent (ex., 1 for binary logistic regression).

- *Specificity:* Percent of correct predictions in the given category of the dependent (ex., 0 for binary logistic regression).

- *False positive rate:* In binary logistic regression, the number of errors where the dependent is predicted to be 1, but is in fact 0, as a percent of total cases which are observed 0's. In multinomial logistic regression, the number of errors where the predicted value of the dependent is higher than the observed value, as a percent of all cases on or above the diagonal.

- *False negative rate:* In binary logistic regression, the number of errors where the dependent is predicted to be 0, but is in fact 1, as a percent of total cases which are observed 1's. In multinomial logistic regression, the number of errors where the predicted value of the dependent is lower than the observed value, as a percent of all cases on or below the diagonal.

- *Measures of association related to the classification table*. Like any crosstabulation, the classification table may be the basis for computing various measures of association. Most must be computed outside logistic regression, as discussed in the separate "blue book" volume on measures of association.

- *Measures of monotone association*. In SPSS multinomial but not binary logistic regression, "Monotonicity measures" is an option in the Statistics button dialog. Recall, however, that one may enter binary dependents in the multinomial procedure in SPSS to obtain the table shown below and, indeed, SPSS will compute the table below only if the dependent is binary.

Measures include Somers' D, Goodman and Kruskal's gamma, Kendall's tau-a, and the Concordance Index, as illustrated below. For most measures, 1.0 indicates maximum effect and 0 indicates independence (no effect of the independents on the dependent). However, see more detailed discussion of each measures of strength of association in the separate "blue book" volume on measures of association.

Measures of Monotone Association

Pairs	Concordant	N	460453
		Percentage	68.2%
	Discordant	N	210414
		Percentage	31.2%
	Tied	N	3989
		Percentage	.6%
	Total	N	674856
		Percentage	100.0%
Measures	Somers' D		.371
	Goodman and Kruskal's Gamma		.373
	Kendall's Tau-a		.171
	Concordance Index C		.685

- *Lambda-p* is a PRE (proportional reduction in error) measure, which is the ratio of (errors without the model - errors with the model) to errors without the model. If lambda-p is .80, then using the logistic regression model will reduce our errors in classifying the dependent by 80% compared to classifying the dependent by always guessing a case is to be classed the same as the most frequent category of the dichotomous dependent. Lambda-p is an adjustment to classic lambda to assure that the coefficient will be positive when the model helps and negative when, as is possible, the model actually leads to worse predictions than simple guessing based on the most frequent class. Lambda-p varies from 1 to (1 - N), where N is the number of cases. Lambda-p = (f - e)/f, where f is the smallest row frequency (smallest row marginal in the classification table) and e is the number of errors (the 1,0 and 0,1 cells in the classification table).

- *Tau-p*. When the classification table has equal marginal distributions, tau-p varies from -1 to +1, but otherwise may be less than 1. Negative values mean the logistic model does worse than expected by chance. Tau-p can be lower than lambda-p because it penalizes proportional reduction in error for non-random distribution of errors (that is, it wants an equal number of errors in each of the error quadrants in the table.)

- *Phi-p* is a third alternative discussed by Menard (pp. 29-30) but is not part of SPSS output. Phi-p varies from -1 to +1 for tables with equal marginal distributions.

- *Binomial d* is a significance test for any of these measures of association, though in each case the number of "errors" is defined differently (see Menard, pp. 30-31).

- *Separation:* Note that when the independents completely predict the dependent, the error quadrants in the classification table will contain 0's, which is called *complete separation*. When this is nearly the case, as when the error quadrants have only one case, this is called *quasicomplete separation*. When separation occurs, one will get very large logit coefficients with very high standard errors. While separation may indicate powerful and valid prediction, often it is a sign of a problem with the independents, such as definitional overlap between the indicators for the independent and dependent variables.

The c statistic

The c statistic is a measure of the discriminative power of the logistic equation. It varies from .5 (the model's predictions are no better than chance) to 1.0 (the model always assigns higher probabilities to correct cases than to incorrect cases for any pair involving dependent=0 and dependent=1). Thus c is the percent of all possible pairs of cases in which the model assigns a higher probability to a correct case than to an incorrect case. The c statistic is not part of SPSS logistic output but may be calculated using the COMPUTE facility, as described in the SPSS manual's chapter on logistic regression. Alternatively, save the predicted probabilities and then get the area under the ROC curve. In SPSS, select Analyze, Regression, Binary (or Multinomial); select the dependent and covariates; click Save; check to save

predicted values (pre_1); Continue; OK. Then select Graphs, ROC Curve; set pre_1 as the test variable; select standard error and confidence interval; OK. In the output, c is labeled as "Area." It will vary from .5 to 1.0.

Information theory measures of model fit

These are variants on deviance which have no intrinsic meaning but when comparing models, lower is better fit and the size of the difference in information theory coefficients is a measure of effect size for the better model as compared to the less well fitting model. Unlike likelihood ratio comparisons, they may be used for non-nested as well as nested model comparisons.

- *The Akaike Information Criterion, AIC,* is a common information theory statistic used when comparing alternative models. Lower is better model fit. In SPSS multinomial logistic regression, AIC is output in the Step Summary, Model Fitting Information, and Likelihood Ratio Tests tables, as illustrated below. AIC is not output by SPSS binomial logistic regression. It is, however, also output by SAS's PROC LOGISTIC. AIC is discussed further in the separate "blue book" volume on structural equation modeling.

- *The Bayesian Information Criterion, BIC,* is a common information theory statistic used when comparing alternative models. Lower is better model fit. In SPSS multinomial logistic regression, BIC is output in the Step Summary, Model Fitting Information, and Likelihood Ratio Tests tables, as illustrated below. BIC is not output by SPSS binomial logistic regression. It is, however, also output by SAS's PROC LOGISTIC. BIC is discussed further in the separate "blue book" volume on structural equation modeling.

Step Summary

Model	Action	Effect(s)	Model Fitting Criteria			Effect Selection Tests		
			AIC	BIC	-2 Log Likelihood	Chi-Square[a]	df	Sig.
0	Entered	Intercept	1.468E3	1.478E3	1.464E3	.		
1	Entered	life	1.326E3	1.356E3	1.314E3	149.817	4	.000
2	Entered	age	1.417E3	2.128E3	1.125E3	189.567	140	.003
3	Entered	educ	1.434E3	2.321E3	1.070E3	55.059	36	.022

Stepwise Method: Forward Entry

a. The chi-square for entry is based on the likelihood ratio test.

Model Fitting Information

Model	Model Fitting Criteria			Likelihood Ratio Tests		
	AIC	BIC	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1.468E3	1.478E3	1.464E3			
Final	1.434E3	2.321E3	1.070E3	394.442	180	.000

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	1.434E3	2.321E3	1.070E3[a]	.000	0	.
age	1.353E3	1.557E3	1.269E3	198.929	140	.001
educ	1.417E3	2.128E3	1.125E3	55.059	36	.022
life	1.578E3	2.445E3	1.222E3[a]	151.706	4	.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.

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

- *The Schwartz Information Criterion, SIC* is a modified version of AIC and is part of SAS's PROC LOGISTIC output. Compared to AIC, SIC penalizes overparameterization more (rewards model parsimony). Lower is better model fit. It is common to use both AIC and SIC when assessing alternative logistic models.

- *Other information theory measures.* Other information theory measures are discussed further in the separate "blue book" volume on structural equation modeling.

Effect size for parameters

Odds ratios

*D*iscussed above, odds ratios are the primary effect size measure for logistic regression when comparing parameters (independent variables).

Standardized vs. unstandardized logistic coefficients in model comparisons

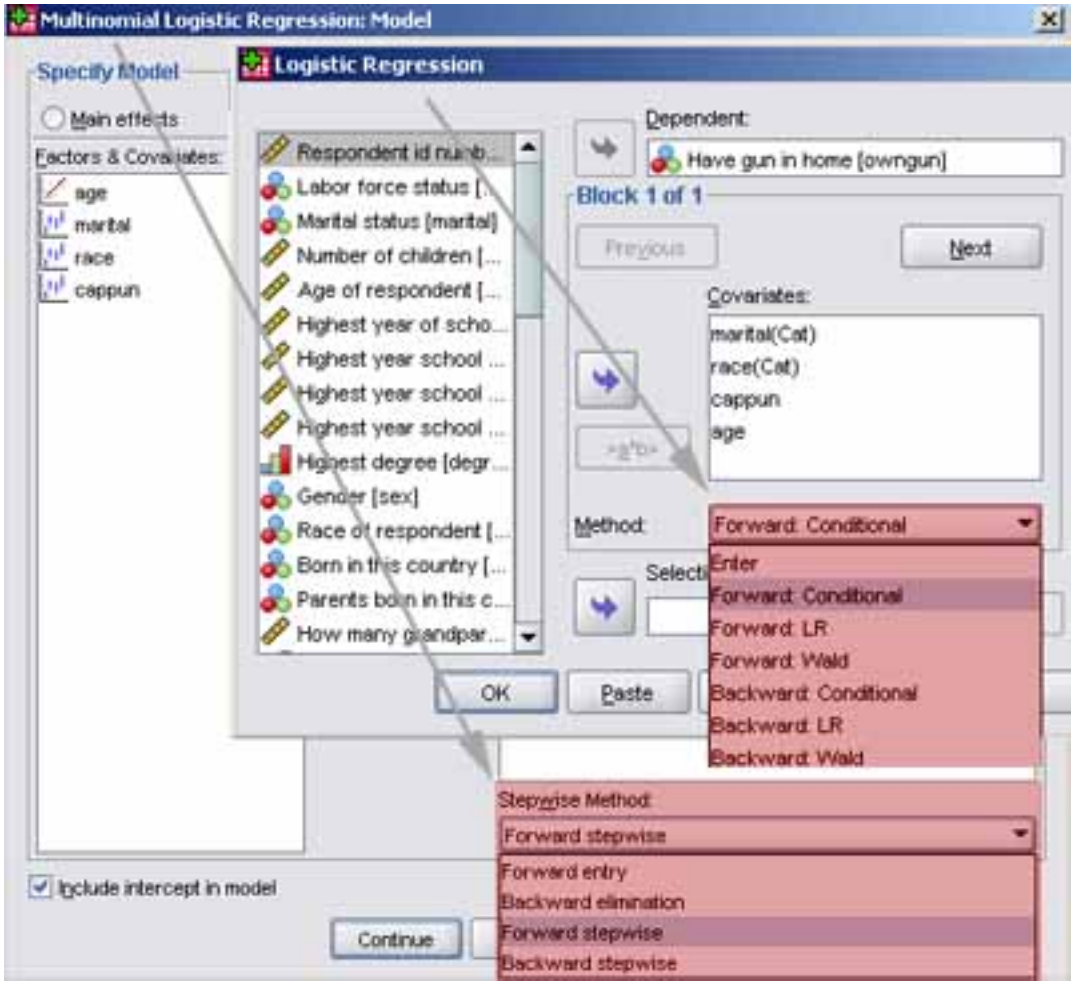
Also called *standardized effect coefficients* or *beta weights,* standardized logistic coefficients correspond to beta (standardized regression) coefficients. While they might be used to compare the relative strength of the independents. This is deprecated for reasons given below and consequently SPSS does not output standardized logistic coefficients. Odds ratios are preferred for this purpose, since when using standardized logit coefficients one is discussing relative importance of the independent variables in terms of effect on the dependent variable's logged odds, which is less intuitive than relative to the actual odds of the dependent variable, which is the referent when odds ratios are used.

As in OLS regression, in logistic (and probit) regression, unstandardized logistic coefficients are used for comparing coefficients for the same independent variable between samples. Unlike OLS regression, however, standardized coefficients (beta weights in OLS regression) are generally not used in logistic regression even for comparing coefficients of different predictor variables within the same sample. Rather, odds ratios are used. Comparison of standardized coefficients between samples (groups) is inappropriate because variances of variables differ between groups. (Recall standardization involves dividing by the standard deviation, which is the square root of the variance). As a further problem of using standardized logistic coefficients, different authors have proposed different algorithms for "standardization," and these result in different values. See further discussion below in the FAQ section and also see Allison (1999).

Stepwise logistic regression

Overview

The forward or backward stepwise logistic regression methods, available in both binary and multinomial regression in SPSS, determine automatically which variables to add or drop from the model. As data-driven methods, stepwise procedures run the risk of modeling noise in the data and are considered useful only for exploratory purposes. Selecting model variables on a theoretic basis and using the "Enter" method is preferred. Both binary and multinomial logistic regression offer both forward and backward stepwise logistic regression, though will slight differences illustrated below.



Binary logistic regression in SPSS offers these variants in the Method area of the main binary logistic regression dialog: forward conditional, forward LR, forward Wald, backward conditional, backward LR, or backward Wald. The conditional options uses a computationally faster version of the likelihood ratio test, LR options utilize the likelihood ratio test (chi-square difference), and the Wald options use the Wald test. The LR option is most often preferred. The likelihood ratio test computes -2LL for the current model, then re-estimates -2LL with the target variable removed. The conditional option is preferred when LR estimation proves too computationally time-consuming. The conditional statistic is considered not as accurate as the likelihood ratio test but more so than the third possible criterion, the Wald test. Stepwise procedures are selected in the Method drop-down list of the binary logistic regression dialog.

Multinomial logistic regression offers these variants under the Model button if a Custom model is specified: forward stepwise, backward stepwise, forward entry, and backward elimination. These four options are described in the FAQ section below. All are based on maximum likelihood estimation (ML), with forward methods using the likelihood ratio or score statistic and backward methods using the likelihood ratio or Wald's statistic. LR is the default, but score and Wald alternatives are available under the Options button. Forward entry adds terms to the model until no omitted variable would contribute significantly to the model. Forward stepwise determines the forward entry model and then alternates between backward elimination and forward entry until all variables not in the model fail to meet entry or removal criteria. Backward elimination and backward stepwise are similar, but begin with all terms in the model and work backward. with backward elimination stopping when the model contains only terms which are significant and with backward stepwise taking this result and further alternating between forward entry and backward elimination until no omitted variable would contribute significantly to the model.

Forward selection vs. backward elimination

Forward selection is the usual option, starting with the constant-only model and adding variables one at a time in the order they are best by some criterion (see below) until some cutoff level is reached (ex., until the step at which all variables not in the model have a significance higher than .05). Backward selection starts with all variables and deletes one at a time, in the order they are worst by some criterion. In the illustration below, forward stepwise modeling of a binary

dependent, which was having a gun in the home or not (Gunown), as predicted by the categorical variable marital status (marital, with four categories), race (three categories), and attitude on the death penalty (binary). The forward stepwise procedure adds Marital first, then Race, then Gunown.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B) Lower	Upper
Step 1[a]	marital			77.292	4	.000			
	marital(1)	-1.139	.140	66.300	1	.000	.320	.243	.421
	marital(2)	-.284	.215	1.733	1	.188	.753	.494	1.149
	marital(3)	-.628	.177	12.593	1	.000	.534	.377	.755
	marital(4)	-.574	.305	3.544	1	.060	.563	.310	1.024
	Constant	1.293	.121	114.127	1	.000	3.644		
Step 2[b]	marital			60.942	4	.000			
	marital(1)	-1.014	.143	50.529	1	.000	.363	.274	.480
	marital(2)	-.168	.218	.594	1	.441	.845	.551	1.297
	marital(3)	-.582	.180	10.428	1	.001	.559	.393	.796
	marital(4)	-.563	.311	3.275	1	.070	.570	.310	1.048
	race			49.463	2	.000			
	race(1)	-1.638	.314	27.254	1	.000	.194	.105	.359
	race(2)	-.757	.350	4.666	1	.031	.469	.236	.932
	Constant	2.669	.325	67.370	1	.000	14.421		
Step 3[c]	marital			58.239	4	.000			
	marital(1)	-1.003	.144	48.772	1	.000	.367	.277	.486
	marital(2)	-.188	.220	.727	1	.394	.829	.539	1.276
	marital(3)	-.555	.181	9.347	1	.002	.574	.402	.819
	marital(4)	-.659	.315	4.370	1	.037	.517	.279	.960
	race			37.258	2	.000			
	race(1)	-1.571	.316	24.792	1	.000	.208	.112	.386
	race(2)	-.879	.353	6.186	1	.013	.415	.208	.830
	cappun	.690	.134	26.521	1	.000	1.994	1.533	2.593
	Constant	1.773	.366	23.484	1	.000	5.889		

a. Variable(s) entered on step 1: marital.

b. Variable(s) entered on step 2: race.

c. Variable(s) entered on step 3: cappun.

Cross-validation

When using stepwise methods, problems of overfitting the model to noise in the current data may be mitigated by cross-validation, fitting the model to a test subset of the data and validating the model using a hold-out validation subset.

Rao's efficient score as a variable entry criterion for forward selection

Rao's efficient score statistic has also been used as the forward selection criterion for adding variables to a model. It is similar but not identical to a likelihood ratio test of the coefficient for an individual explanatory variable. Score appears in the "score" column of the "Variables not in the equation" table of SPSS output for binary logistic regression. Score may be selected as a variable entry criterion in multinomial regression, but there is no corresponding table (its use is just noted in a footnote in the "Step Summary" table). In the example below, having a gun in the home or not is predicted from marital status, race, age, and attitudes toward capital punishment. Race, with the highest score in Step 1, gets added to the initial predictor, marital status, in Step 2. The variable with the highest Score in Step 2, attitude toward capital punishment (Cappun) gets added in Step 3. Age, with a non-significant Score, is never added as the stepwise procedure stops at Step 3.

Variables not in the Equation

			Score	df	Sig.
Step 1	Variables	race	55.060	2	.000
		race(1)	51.850	1	.000
		race(2)	21.038	1	.000
		cappun	41.025	1	.000
		age	6.887	1	.009
	Overall Statistics		83.630	4	.000
Step 2	Variables	cappun	27.096	1	.000
		age	2.753	1	.097
	Overall Statistics		30.144	2	.000
Step 3	Variables	age	3.091	1	.079
	Overall Statistics		3.091	1	.079

Score statistic

Rao's efficient score, labeled simply "score" in SPSS output, is test for whether the logistic regression coefficient for a given explanatory variable is zero. It is mainly used as the criterion for variable inclusion in forward stepwise logistic regression (discussed above), because of its advantage of being a non-iterative and therefore computationally fast method of testing individual parameters compared to the

likelihood ratio test. In essence, the score statistic is similar to the first iteration of the likelihood ratio method, where LR typically goes on to three or four more iterations to refine its estimate. In addition to testing the significance of each variable, the score procedure generates an "Overall statistics" significance test for the model as a whole. A finding of nonsignificance (ex., p>.05) on the score statistic leads to acceptance of the null hypothesis that coefficients are zero and the variable may be dropped. SPSS continues by this method until no remaining predictor variables have a score statistic significance of .05 or better.

Which step is the best model?

Stepwise methods do not necessarily identify "best models" at all as they work by fitting an automated model to the current dataset, raising the danger of overfitting to noise in the particular dataset at hand. However, there are three possible methods of selecting the "final model" that emerges from the stepwise procedure. binary stepwise logistic regression offers only the first, but multinomial stepwise logistic regression offers two more, based on the information theory measures AIC and BIC, discussed below.

1. *Last step*. The final model is the last step model, where adding another variable would not improve the model significantly. In the figure below, a general happiness survey variable (3 levels) is predicted from age, education, and a categorical survey item on whether life is exciting or dull (3 levels). By the customary last step rule, the best model is model 3.

2. *Lowest AIC*. The "Step Summary" table will print the Akaike Information Criterion (AIC) for each step. AIC is commonly used to compare models, where the lower the AIC, the better. The step with the lowest AIC thus becomes the "final model." By the lowest AIC criterion, the best model would be model 1.

3. *Lowest BIC*. The "Step Summary" table will print the Bayesian Information Criterion (BIC) for each step. BIC is also used to compare models, again where the lower the BIC, the better. The step with the lowest BIC thus becomes the "final model." Often BIC will point to a more parsimonious model than will AIC as its formula factors in degrees of freedom, which is related to number of variables. By the lowest BIC criterion, the best model would be model 1.

Step Summary

Model	Action	Effect(s)	Model Fitting Criteria			Effect Selection Tests		
			AIC	BIC	-2 Log Likelihood	Chi-Square[a]	df	Sig.
0	Entered	Intercept	1.468E3	1.478E3	1.464E3			
1	Entered	life	1.326E3	1.356E3	1.314E3	149.817	4	.000
2	Entered	age	1.417E3	2.128E3	1.125E3	189.567	140	.003
3	Entered	educ	1.434E3	2.321E3	1.070E3	55.059	36	.022

Stepwise Method: Forward Entry

a. The chi-square for entry is based on the likelihood ratio test.

Click here for further discussion of stepwise methods.

Contrast Analysis

Repeated contrasts

"Repeated contrasts" is an SPSS option (called *profile contrasts* in SAS) which computes the logit coefficient for each category of the independent (except the "reference" category, which is the last one by default). Contrasts are used when one has a categorical independent variable and wants to understand the effects of various levels of that variable. Specifically, a "contrast" is a set of coefficients that sum to 0 over the levels of the independent categorical variable. SPSS automatically creates K-1 internal dummy variables when a covariate is declared to be categorical with K values (by default, SPSS leaves out the last category, making it the reference category). The user can choose various ways of assigning values to these internal variables, including *indicator contrasts*, *deviation contrasts*, or *simple contrasts*. In SPSS, indicator contrasts are now the default (old versions used deviation contrasts as default).

Indicator contrasts

Indicator contrasts produce estimates comparing each other group to the reference group. David Nichols, senior statistician at SPSS, gives this example of indicator coding output:

Parameter codings for indicator contrasts

	Value	Freq	Parameter Coding (1)	(2)
GROUP				
	1	106	1.000	.000
	2	116	.000	1.000
	3	107	.000	.000

This example shows a three-level categorical independent (labeled GROUP), with category values of 1, 2, and 3. The predictor here is called simply GROUP. It takes on the values 1-3, with frequencies listed in the "Freq" column. The two "Coding" columns are the internal values (parameter codings) assigned by SPSS under indicator coding. There are two columns of codings because two dummy variables are created for the three-level variable GROUP. For the first variable, which is Coding (1), cases with a value of 1 for GROUP get a 1, while all other cases get a 0. For the second, cases with a 2 for GROUP get a 1, with all other cases getting a 0.

- *Simple contrasts* compare each group to a reference category (like indicator contrasts). The contrasts estimated for simple contrasts are the same as for indicator contrasts, but the intercept for simple contrasts is an unweighted average of all levels rather than the value for the reference group. That is, with one categorical independent in the model, simple contrast coding means that the intercept is the log odds of a response for an unweighted average over the categories.

- *Deviation contrasts* compare each group other than the excluded group to the unweighted average of all groups. The value for the omitted group is then equal to the negative of the sum of the parameter estimates.

Contrasts and ordinality

For nominal variables, the pattern of contrast coefficients for a given independent should be random and nonsystematic, indicating the nonlinear, nonmonotonic pattern characteristic of a true nominal variable. Contrasts can thus be used as a

method of empirically differentiating categorical independents into nominal and ordinal classes.

Analysis of residuals

Overview

Residuals may be plotted to detect outliers visually. Residual analysis may lead to development of separate models for different types of cases. For logistic regression, it is usual to use the standardized difference between the observed and expected probabilities. SPSS calls this the "standardized residual (ZResid)," while SAS calls this the "chi residual," while Menard (1995) and at other times (including by SPSS in the table of "Observed and Predicted Frequencies" in multinomial logistic output) it is called the "Pearson residual." In a model which fits in every cell formed by the independents, no absolute standardized residual will be > 1.96. Cells which do not meet this criterion signal combinations of independent variables for which the model is not working well.

Residual analysis in binary logistic regression

Outliers

Outliers beyond a certain number of standard deviations (default = 2) can be requested under Options in SPSS, as discussed above.

The DfBeta statistic

DfBeta is available under the Save button to indicate cases which are poorly fitted by the model. DfBeta measures the change in the logit coefficients for a given variable when a case is dropped. There is a DfBeta statistic for each case for each explanatory variable and for the constant. An arbitrary cutoff criterion for cases to be considered outliers is those with dbeta > 1.0 on critical variables in the model. The DfBeta statistic can be saved as DFB0_1 for the constant, DFB1_1 for the first independent, DFB1_2 for the second independent, etc.

The leverage statistic

The *leverage statistic, h,* is available is available under the Save button to identify cases which influence the logistic regression model more than others. The leverage statistic varies from 0 (no influence on the model) to 1 (completely determines the model). By common rule of thumb, leverage is considered high if greater than 2p/n, where p is the number of predictors (number of continuous predictors plus number of non-reference levels of categorical predictors) and n is sample size. This cutoff is equivalent to twice the average leverage, which equals p/n. By a different rule of thumb, cases with leverage under .2 are not a problem but if a case has leverage over .5, the case has undue leverage and should be examined for the possibility of measurement error or the need to model such cases separately. Minimum, maximum, and mean leverage may be displayed. Note that influential cases may nonetheless have small leverage values when predicted probabilities are .9. Leverage is an option in SPSS, in which a plot of leverage by case id will quickly identify cases with unusual impact.

Cook's distance

Cook's distance, D, is a third measure of the influence of a case, also available under the Save button. Its value is a function of the case's leverage and of the magnitude of its standardized residual. It is a measure of how much deleting a given case affects residuals for all cases. An approximation to Cook's distance is an option in SPSS logistic regression.

Other. The Save button in the SPSS binary logistic dialog will also save the standardized residual as ZRE_1. One can also save predictions as PRE_1.

Residual analysis in multinomial logistic regression

SPSS NOMREG has little support for residual analysis, but the Save button in multinomial logistic regression supports saving predicted and actual category probabilities. One can then compute the difference for purposes of residual analysis.

Assumptions

Logistic regression is popular in part because it enables the researcher to overcome many of the restrictive assumptions of OLS regression. Logistic regression does not assume a linear relationship between the dependents and the independents. It may handle nonlinear effects even when exponential and polynomial terms are not explicitly added as additional independents because the logit link function on the left-hand side of the logistic regression equation is non-linear. However, it is also possible and permitted to add explicit interaction and power terms as variables on the right-hand side of the logistic equation, as in OLS regression.

- The dependent variable need not be normally distributed (but does assume its distribution is within the range of the exponential family of distributions, such as normal, Poisson, binomial, gamma). Solutions may be more stable if predictors have a multivariate normal distribution.
- The dependent variable need not be homoscedastic for each level of the independents; that is, there is no homogeneity of variance assumption: variances need not be the same within categories.
- Normally distributed error terms are not assumed.
- Logistic regression does not require that the independents be interval.
- Logistic regression does not require that the independents be unbounded.

However, other assumptions still apply, as discussed below.

Data level

A dichotomous or polytomous dependent variable is assumed for binary or multinomial logistic regression respectively. Reducing a continuous variable to a binary or categorical one loses information and attenuates effect sizes, reducing the power of the logistic procedure compared to OLS regression. That is, OLS regression is preferred to logistic regression for continuous variables when OLS assumptions are met and is superior to categorizing a continuous dependent for purposes of running a logistic regression. Likewise, discriminant function analysis is more powerful than binary logistic regression for a binary dependent variable, when the assumptions of the former are met. If the categories of the dependent variable are ordinal, ordinal regression will be more powerful and is preferred.

Independent variables may be interval or categorical, but if categorical, it is assumed that they are dummy or indicator coded. Categories must be mutually exclusive, such that a given case will be in only one group of any given factor. SPSS gives an option to recode categorical variables automatically. If all predictor variables are categorical, loglinear analysis is an alternative procedure.

Meaningful coding

Logistic coefficients will be difficult to interpret if not coded meaningfully. The convention for binary logistic regression is to code the dependent class of greatest interest as 1 ("the event occurring") and the other class as 0 ("the event not occurring), and to code its expected correlates also as +1 to assure positive correlation. For multinomial logistic regression, the class of greatest interest should be the last class. Logistic regression is predicting the log odds of being in the class of greatest interest. The "1" group or class of interest is sometimes called the target group or the response group.

Proper specification of the model

Proper specification is particularly crucial; parameters may change magnitude and even direction when variables are added to or removed from the model.

- *Inclusion of all relevant variables in the regression model*: If relevant variables are omitted, the common variance they share with included variables may be wrongly attributed to those variables, or the error term may be inflated.

- *Exclusion of all irrelevant variables*: If causally irrelevant variables are included in the model, the common variance they share with included variables may be wrongly attributed to the irrelevant variables. The more the correlation of the irrelevant variable(s) with other independents, the greater the standard errors of the regression coefficients for these independents.

Independence of irrelevant alternatives

In multinomial logistic regression, where the dependent represents choices or alternatives, it is assumed that adding or removing alternatives does not affect

the odds associated with the remaining alternatives. When the IIA assumption is violated, alternative-specific multinomial probit regression is recommended, discussed in the separate "blue book" volume on loglinear analysis.

Error terms are assumed to be independent (independent sampling)

Violations of this assumption can have serious effects. Violations will occur, for instance, in correlated samples and repeated measures designs, such as before-after or matched-pairs studies, cluster sampling, or time-series data. That is, subjects cannot provide multiple observations at different time points. Conditional logit models in Cox regression and logistic models for matched pairs in multinomial logistic regression are available to adapt logistic models to handle non-independent data.

Low error in the explanatory variables

Ideally assumes low measurement error and no missing cases. See here for further discussion of measurement error in GLM models.

Linearity

Logistic regression does not require linear relationships between the independent factor or covariates and the dependent variable, as does OLS regression, but it does assume a linear relationship between the continuous independents and the log odds (logit) of the dependent. When the assumption of linearity in the logits is violated, then logistic regression will underestimate the degree of relationship of the independents to the dependent and will lack power (generating Type II errors, thinking there is no relationship when there actually is). The Box-Tidwell test of linearity in the logit is used primarily with interval-level covariates. The logit step test is used primarily with ordinal-level covariates.

When a continuous or ordinal covariate is shown to lack linearity in the logit, the researcher may divide the covariate into categories and use it as a factor, thereby getting separate parameter estimates for various levels of the variable. For ordinal covariates sometimes linearity in the logit can be achieved by combining categories

1. *Box-Tidwell Transformation (Test)*: Add to the logistic model interaction terms which are the crossproduct of each independent times its natural logarithm [(X)ln(X)]. If these terms are significant, then there is nonlinearity in the logit. This method is not sensitive to small nonlinearities.

2. *Orthogonal polynomial contrasts*: This option treats a categorical independent as a categorical variable with categories assumed to be equally spaced. The logit (effect) coefficients for each category of the categorical explanatory variable should not change over the contrasts. This method is not appropriate when the independent has a large number of values, inflating the standard errors of the contrasts. Select polynomial from the contrast list after clicking the Categorical button in the SPSS logistic regression dialog.

3. *Logit step tests*. Another simple method of checking for linearity between an ordinal or interval independent variable and the logit of the dependent variable is to (1) create a new variable which divides the existing independent variable into categories of equal intervals, then (2) run a logistic regression with the same dependent but using the newly categorized version of the independent as a categorical variable with the default indicator coding. If there is linearity with the logit, the b coefficients for each class of the newly categorized explanatory variable should increase (or decrease) in roughly linear steps.

4. *The SPSS Visual Bander* can be used to create a categorical variable based on an existing interval or ordinal one. It is invoked from the SPSS menu by selecting Transform, Visual Bander. In the Visual Bander dialog, enter the variable to categorize and click Continue. In the next dialog, select the variable just entered and also give a name for the new variable to be created; click the Make Cutpoints button and select Equal Intervals (the default), then enter the starting point (ex., 0) and the number of categories (ex., 5) and tab to the Width box, which will be filled in with a default value. Click Apply to return to the Visual Bander dialog, then click OK. A new categorized variable of the name provided will be created at the end of the Data Editor spreadsheet.

5. *A logit graph* can be constructed to visually display linearity in the logit, or lack thereof, for a banded (categorical) version of an underlying continuous variable. After a banded version of the variable is created and substituted into the logistic model, separate b coefficients will be output in the "Parameter Estimates" table. Once can go to this table in SPSS output, double-click it, highlight the b coefficients for each band, right-click, and select Create Graph, Line. An ordinal covariate that is consistent with linearity in the logit will display a consistent pattern of stepped increase or decrease without reversals in direction in the line.

Additivity

Like OLS regression, logistic regression does not account for interaction effects except when interaction terms (usually products of standardized independents) are created as additional variables in the analysis. This is done by using the categorical covariates option in SPSS's logistic procedure.

Absence of perfect separation

If groups of the dependent are perfectly separated by an independent variable or set of variables, implausibly large b coefficients and effect sizes may be computed for the independent variable(s).

Absence of perfect multicollinearity

If one variable is a perfect linear function of another in the model, standard errors become infinite and the solution to the model becomes indeterminate. In many computer software programs a fatal error message will be issued. SPSS will warn "The parameter covariance matrix cannot be computed. Remaining statistics will be omitted," and will output a 'Model Summary Table' which will show a zero -2 log likelihood statistic with a note similar to "Estimation terminated at iteration number 20 because a perfect fit is detected. This solution is not unique."

Absence of high multicollinearity

To the extent that one independent is a near but not perfect linear function of another independent, the problem of multicollinearity will occur in logistic regression, as it does in OLS regression. As the independents increase in

correlation with each other, the standard errors of the logit (effect) coefficients will become inflated. Multicollinearity does not change the estimates of the coefficients, only their reliability. High standard errors flag possible multicollinearity. Multicollinearity and its handling is discussed more extensively in the separate "blue book" volume on multiple regression.

Centered variables

As in OLS regression, centering may be necessary either to reduce multicollinearity or to make interpretation of coefficients meaningful. Centering is almost always recommended for independent variables which are components of interaction terms in a logistic model. See the full discussion in the separate "blue book" volume on multiple regression.

No outliers

As in OLS regression, outliers can affect results significantly. The researcher should analyze standardized residuals for outliers and consider removing them or modeling them separately. Standardized residuals >2.58 are outliers at the .01 level, which is the customary level (standardized residuals > 1.96 are outliers at the less-used .05 level). Standardized residuals are requested under the "Save" button in the binary logistic regression dialog box in SPSS. For multinomial logistic regression, checking "Cell Probabilities" under the "Statistics" button will generate actual, observed, and residual values.

Sample size

Also, unlike OLS regression, logistic regression uses maximum likelihood estimation (ML) rather than ordinary least squares (OLS) to derive parameters. ML relies on large-sample asymptotic normality which means that reliability of estimates declines when there are few cases for each observed combination of independent variables. That is, in small samples may lead to high standard errors. In the extreme, if there are too few cases in relation to the number of variables, it may be impossible to converge on a solution. Very high parameter estimates (logistic coefficients) may signal inadequate sample size.

In general, the sample size should be larger than for the corresponding OLS regression. One rule of thumb is that the number of cases in the smaller of the

two binary outcomes in binary logistic regression divided by the number of predictor variables should be at least 20 (Harrell, 2001), where "number of predictor variables" refers to all variables modeled (ex., in stepwise procedures), not just variables in the final model. Peduzzi et al. (1996) recommended that the smaller of the classes of the dependent variable have at least 10 events per parameter in the model. Hosmer & Lemeshow (1989) recommended a minimum of 10 cases per independent variable. Pedhazur (1997) recommended sample size be at least 30 times the number of parameters. Yet another rule of thumb calls for 50 cases per independent variable.

Sampling adequacy

Goodness of fit measures like model chi-square assume that for cells formed by the categorical independents, all cell frequencies are >=1 and 80% or more of cells are > 5. Researchers should run crosstabs to assure this requirement is met. Sometimes one can compensate for small samples by combining categories of categorical independents or by deleting independents altogether. The presence of small or empty cells may cause the logistic model to become unstable, reporting implausibly large b coefficients and odds ratios for independent variables.

Expected dispersion

Logistic regression does not assume homogeneity of variance: each group formed by the independent categorical variables does not need to have the same dispersion or variance on the dependent variable.

In logistic regression the expected variance of the dependent can be compared to the observed variance, and discrepancies may be considered under- or overdispersion. If there is moderate discrepancy, standard errors will be over-optimistic and one should use adjusted standard error. Adjusted standard error will make the confidence intervals wider. However, if there are large discrepancies, this indicates a need to respecify the model, or that the sample was not random, or other serious design problems. The expected variance is ybar*(1 - ybar), where ybar is the mean of the fitted (estimated) y. This can be compared with the actual variance in observed y to assess under- or overdispersion. Adjusted SE equals SE * SQRT(D/df), where D is the scaled deviance, which for logistic regression is -2LL, which is -2Log Likelihood in SPSS logistic regression output.

Frequently Asked Questions

How should logistic regression results be reported?

Considering the increasing use of logistic regression, there is surprisingly little consensus on exactly what should be reported in journal articles based on this procedure. Peng, Lee, & Ingersoll (2002) recommend reporting in table form the b parameter, its standard error, the Wald statistic, degrees of freedom, p significance level, and the odds ratio (exp(b)) for the constant and each predictor in the model. In addition, the reference category for the dependent should always be displayed in a table footnote. A table should report all overall model fit tests (likelihood ratio, score, Wald, and Hosmer & Lemeshow tests) with their associated chi-square, p significance levels, and degrees of freedom. These authors also recommend that in a footnote to the table, all pseudo-R^2 and monotone association measures be reported; and in an additional table, the classification table should be reproduced, and in a footnote to it, the researcher should list its associated terms (hit rate, sensitivity, specificity, false positive rate, false negative rate). Beyond this, some authors also always report the confidence limits on the odds ratio.

Example

"A binary logistic regression analysis was conducted to predict attitudes in favor or against the death penalty, using gender, race, and educational level as categorical predictor variables. The model was found to be statistically different from the null (constant-only) model at the 0.000 level (chi-square = 89.011, df = 7). All three categorical predictor variables were found to be significant at the 0.000 level, meaning that at least one of their levels was significantly related to attitude toward capital punishment.

Using odds ratios to interpret the significant levels of the predictor variables, we may say that education level has a higher effect on attitude toward capital punishment than does race, and race has a higher effect than does gender. Specifically, we may say that:

- Being female rather than male multiplies (reduces) the odds of opposing rather than favoring the death penalty by a factor of .542.
- Being white rather than "other race" multiplies (reduces) the odds of opposing rather than favoring the death penalty by a factor of .457.
- Having a less than a high school education rather than a graduate education multiplies (reduces) the odds of opposing rather than favoring the death penalty by a factor of .424.
- Having a high school education rather than a graduate education multiplies (reduces) the odds of opposing rather than favoring the death penalty by a factor of .364.
- Having a junior college education rather than a graduate education multiplies (reduces) the odds of opposing rather than favoring the death penalty by a factor of .371.

Though significant, Nagelkerke's R-square of 0.095 indicated a relatively weak relationship between the predictor variables and attitude toward capital punishment. A weak relationship was also suggested by a table of successful and unsuccessful classification of cases, showing that although the model classified 77.5% of cases correctly, this was only 0.1% more that classifying by chance based on always selecting the most numerous category. Both Nagelkerke's R-square and the classification table results suggest the need for model respecification using different or additional variables.

Why not just use regression with dichotomous dependents?

Use of a dichotomous dependent variable in OLS regression violates the assumptions of normality and homoscedasticity because a normal distribution is impossible with only two values. Likewise, when the values can only be 0 or 1, residuals (error) will be low for the portions of the regression line near Y=0 and Y=1, but high in the middle -- hence the error term will violate the assumption of homoscedasticity (equal variances) when a dichotomy is used as a dependent. Even with large samples, standard errors and significance tests will be in error because of lack of homoscedasticity. Also, for a dependent variable which assumes values of 0 and 1, the regression model will allow out-of-range estimates below 0 and above 1. Also, multiple linear regression does not handle non-linear relationships, whereas log-linear methods do. These objections to the use of

regression with dichotomous dependents apply to polytomous dependents used in OLS regression also.

How does OLS regression compare to logistic regression?

Today it is considered unacceptable to use OLS regression with a binary dependent variable since OLS assumptions of normal distribution cannot be met. Nonetheless, OLS may be used on such dependent variables on an exploratory basis. Allison (2012: 10) notes, "as an approximate method, OLS regression does a surprisingly good job with dichotomous variables, despite clear-cut violations of assumptions." If used for exploratory purposes, the split of the binary dependent variable should not be extreme (not 90:10 or worse). OLS regression in most cases will not return dramatically different substantive results compared to binary logistic regression.

If data are dichotomized or binned into categories for purposes of binary or multinomial logistic regression, information is lost, leading to increased probability of attenuation of effect size. OLS regression with continuous dependent variables will have more power (fewer Type II errors = fewer false negatives) than logistic models.

OLS regression for a multinomial dependent variable with unordered levels is never acceptable and any results are arbitrary. For an ordered multinomial variable, ordinal regression is preferred over either multinomial or OLS regression.

When uncertain about meeting the assumptions of OLS regression, logistic regression is preferred.

When is discriminant analysis preferred over logistic regression?

With a binary dependent variable when all assumptions of discriminant function analysis are met (e.g., multivariate normality and equal variance-covariance matrices), discriminant analysis usually will have more power than logistic regression as well as yield a quicker solution. That is, there will be fewer Type II errors (false negatives).

What is the SPSS syntax for logistic regression?

With SPSS, logistic regression is found under Analyze > Regression > Binary Logistic or Multinomial Logistic.

```
LOGISTIC REGRESSION /VARIABLES income WITH age SES gender
opinion1 opinion2 region
    /CATEGORICAL=gender, opinion1, opinion2, region
    /CONTRAST(region)=INDICATOR(4)
    /METHOD FSTEP(LR)
    /CLASSPLOT
```

SPSS syntax is shown above in simplified form.

- The dependent variable is the variable immediately after the VARIABLES term. The categorical independent variables are those immediately after the WITH term. The continuous predictor variables are those immediately after the BY term.
- The CATEGORICAL command specifies any categorical variables. Note these must also be listed in the VARIABLES statement.
- The CONTRAST command tells SPSS which category of a categorical predictor variable is to be the reference category, by default the highest-coded category.
- The METHOD subcommand sets the method of computation, above specified as FSTEP to indicate forward stepwise logistic regression. Alternatives are BSTEP (backward stepwise logistic regression) and ENTER (enter terms as listed, usually because their order is set by theories which the researcher is testing). ENTER is the default method. The (LR) term following FSTEP specifies that likelihood ratio criteria are to be used in the stepwise addition of variables to the model.
- The CLASSPLOT option specifies a histogram of predicted probabilities is to output (see above).

The full binary logistic regression syntax is below:

```
LOGISTIC REGRESSION VARIABLES = dependent var
        [WITH independent varlist [BY var [BY var] ... ]]
[/CATEGORICAL = var1, var2, ... ]
[/CONTRAST (categorical var) = [{INDICATOR [(refcat)]    }]]
                               {DEVIATION [(refcat)]    }
                               {SIMPLE [(refcat)]       }
                               {DIFFERENCE              }
```

```
                                {HELMERT                }
                                {REPEATED               }
                                {POLYNOMIAL[({1,2,3...})]}
                                             {metric  }
                                {SPECIAL (matrix)       }
 [/METHOD = {ENTER**       }   [{ALL    }]]
            {BSTEP [{COND}]}    {varlist}
                    {LR  }
                    {WALD}
            {FSTEP [{COND}]}
                    {LR  }
                    {WALD}
 [/SELECT = {ALL**                 }]
            {varname relation value}
 [/{NOORIGIN**}]
   {ORIGIN    }
 [/ID = [variable]]
 [/PRINT = [DEFAULT**] [SUMMARY] [CORR] [ALL] [ITER [({1})]] [GOODFIT]]
                                                      {n}
           [CI(level)]
 [/CRITERIA = [BCON ({0.001**})] [ITERATE({20**})] [LCON({0**   })]
                     {value  }           {n   }         {value }
              [PIN({0.05**})] [POUT({0.10**})] [EPS({.00000001**})]]
                   {value }         {value }        {value      }
              [CUT[{O.5**  }]]
                   [value  }
 [/CLASSPLOT]
 [/MISSING = {EXCLUDE **}]
             {INCLUDE   }
 [/CASEWISE = [tempvarlist]  [OUTLIER({2    })]]
                                  {value}
 [/SAVE = tempvar[(newname)] tempvar[(newname)]...]
 [/OUTFILE = [{MODEL    }(filename)]]
              {PARAMETER}
 [/EXTERNAL]
```

**Default if the subcommand or keyword is omitted.

The syntax for multinomial logistic regression is:

```
NOMREG dependent varname [(BASE = {FIRST } ORDER = {ASCENDING**})] [BY factor list]
                                  {LAST**}         {DATA       }
                                  {value }         {DESCENDING }
       [WITH covariate list]
 [/CRITERIA = [CIN({95**})] [DELTA({0**})] [MXITER({100**})] [MXSTEP({5**})]
                   {n   }          {n  }           {n    }           {n  }
              [LCONVERGE({0**})] [PCONVERGE({1.0E-6**})] [SINGULAR({1E-8**})]
                        {n  }             {n       }             {n     }
              [BIAS({0**})] [CHKSEP({20**})]  ]
                   {n  }           {n   }
 [/FULLFACTORIAL]
 [/INTERCEPT = {EXCLUDE   }]
               {INCLUDE** }
 [/MISSING = {EXCLUDE**}]
             {INCLUDE  }
 [/MODEL = {[effect effect ...]} [| {BACKWARD} = { effect effect ...}]]
                                   {FORWARD }
                                   {BSTEP   }
                                   {FSTEP   }
 [/STEPWISE =[RULE({SINGLE**   })][MINEFFECT({0**  })][MAXEFFECT(n)]]
```

```
                  {SFACTOR    }             {value}
                  {CONTAINMENT}
                  {NONE       }
            [PIN({0.05**})]  [POUT({0.10**})]
                 {value }          {value }
            [ENTRYMETHOD({LR** })] [REMOVALMETHOD({LR**})]
                        {SCORE}                  {WALD}
[/OUTFILE = [{MODEL    }(filename)]]
             {PARAMETER}
[/PRINT = [CELLPROB] [CLASSTABLE] [CORB] [HISTORY({1**})] [IC] ]
                                                {n  }
          [SUMMARY ] [PARAMETER ] [COVB] [FIT] [LRT] [KERNEL]
    [ASSOCIATION] [CPS**] [STEP**] [MFI**] [NONE]
[/SAVE = [ACPROB[(newname)]] [ESTPROB[(rootname[:{25**}])] ]
                                               {n   }
         [PCPROB[(newname)]] [PREDCAT[(newname)]]
[/SCALE = {1**     }]
          {n       }
          {DEVIANCE}
          {PEARSON }
[/SUBPOP = varlist]
[/TEST[(valuelist)] = {['label'] effect valuelist effect valuelist...;}]
                      {['label'] ALL list;                            }
                      {['label'] ALL list                              }
```

** Default if the subcommand is omitted.

Apart from indicator coding, what are the other types of contrasts?

Here are the common types of coding:

- *Indicator* (dummy or reference coding): If the presence of the case in the given category is contrasted with absence of membership in the category, indicator coding is being used. The researcher sets first or last as the reference category. This is the most common type.

Indicator coding, last as reference

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)	(4)
Marital status	Married	1259	1.000	.000	.000	.000
	Widowed	242	.000	1.000	.000	.000
	Divorced	412	.000	.000	1.000	.000
	Separated	79	.000	.000	.000	1.000
	Never married	607	.000	.000	.000	.000

- *Simple:* If each category of the factor is compared to the reference category, simple coding is being used. The researcher sets first or last as the reference category.

Simple coding, last as reference

Categorical Variables Codings

			Parameter coding			
		Frequency	(1)	(2)	(3)	(4)
Marital status	Married	1259	.800	-.200	-.200	-.200
	Widowed	242	-.200	.800	-.200	-.200
	Divorced	412	-.200	-.200	.800	-.200
	Separated	79	-.200	-.200	-.200	.800
	Never married	607	-.200	-.200	-.200	-.200

- *Deviation (effect coding)*: If the reference value is the overall effect (grand mean) of all categories with which others are contrasted, deviation coding is being used. The researcher sets first or last as the reference category.

Deviation coding, last as reference

Categorical Variables Codings

			Parameter coding			
		Frequency	(1)	(2)	(3)	(4)
Marital status	Married	1259	1.000	.000	.000	.000
	Widowed	242	.000	1.000	.000	.000
	Divorced	412	.000	.000	1.000	.000
	Separated	79	.000	.000	.000	1.000
	Never married	607	-1.000	-1.000	-1.000	-1.000

- *Helmert*: If the reference value is the mean of subsequent categories (except for the last category, which has no higher categories) with which others are contrasted, Helmert coding is being used. The researcher does not select a reference category.

Helmert coding

Categorical Variables Codings

			Parameter coding			
		Frequency	(1)	(2)	(3)	(4)
Marital status	Married	1259	.800	.000	.000	.000
	Widowed	242	-.200	.750	.000	.000
	Divorced	412	-.200	-.250	.667	.000
	Separated	79	-.200	-.250	-.333	.500
	Never married	607	-.200	-.250	-.333	-.500

- *Difference:* If the reference value is the mean of previous categories (except for the first category, which has no lower categories) with which others are contrasted, difference coding is being used. This type of coding is also called "reverse Helmert". The researcher does not select a reference category.

Difference coding

Categorical Variables Codings

			Parameter coding			
		Frequency	(1)	(2)	(3)	(4)
Marital status	Married	1259	-.500	-.333	-.250	-.200
	Widowed	242	.500	-.333	-.250	-.200
	Divorced	412	.000	.667	-.250	-.200
	Separated	79	.000	.000	.750	-.200
	Never married	607	.000	.000	.000	.800

- *Repeated*: If the reference category is the previous category (except for the first category) with which others are contrasted, repeated coding is being used.

Repeated coding

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)	(4)
Marital status	Married	1259	.800	.600	.400	.200
	Widowed	242	-.200	.600	.400	.200
	Divorced	412	-.200	-.400	.400	.200
	Separated	79	-.200	-.400	-.600	.200
	Never married	607	-.200	-.400	-.600	-.800

- *Polynomial*. With orthogonal polynomial contrasts categories are assumed to be equally spaced numeric categories. This coding is used to check linearity. For factor "X", output will list X(1) for the linear effect of X, X(2) for the quadratic effect, etc.

Polynomial coding

Categorical Variables Codings

		Frequency	Parameter coding (1)	(2)	(3)	(4)
Marital status	Married	1259	-.632	.535	-.316	.120
	Widowed	242	-.316	-.267	.632	-.478
	Divorced	412	.000	-.535	.000	.717
	Separated	79	.316	-.267	-.632	-.478
	Never married	607	.632	.535	.316	.120

Can I create interaction terms in my logistic model, as with OLS regression?

Yes. As in OLS regression, interaction terms are constructed as crossproducts of the two interacting variables.

Will SPSS's binary logistic regression procedure handle my categorical variables automatically?

No. You must declare your categorical variables categorical if they have more than two values. This is done by clicking on the "Categorical" button in the Logistic

Regression dialog box. After this, SPSS will automatically create dummy variables based on the categorical variable.

Can I handle missing cases the same in logistic regression as in OLS regression?

In SPSS and some statistical packages for OLS regression, the researcher may choose to estimate missing values based on OLS regression of the variable with missing cases, based on non-missing data. However, the nonlinear model assumed by logistic regression requires a full set of data. Therefore SPSS and most statistical packages for logistic regression provide only for LISTWISE deletion of cases with missing data, using the remaining full dataset to calculate logistic parameters. The alternative is to use multiple imputation or some other missing values method to estimate missing data, then use the imputed dataset for logistic regression.

Explain the error message I am getting about unexpected singularities in the Hessian matrix.

Even though "normal" output will still appear in SPSS and some other statistical packages, this error message means that the printed solution is unreliable, may not be valid, and should not be reported. A singularity is a perfect correlation (perfect multicollinearity). Even if the researcher's independent variables are no collinear, it is possible that one of the predictors is a constant for one of the categories of the dependent variable and that constitutes a singularity also.

To diagnose this problem, the researcher should look at the final "Parameter Estimates" table. When the intercept is large in one direction and the coefficient for one of the independent variables is very large in the opposite direction, a problem is indicated.

If there is only one very large logit coefficient of opposite sign from a very large intercept, the problem is in the category of the dependent variable used in the numerator of that logit and the researcher may need to merge that category of the dependent variable with an adjacent one. It also may be possible to resolve the problem by dropping an offending predictor variable from the model, or by merging some of its categories if it is a categorical variable.

If all logit coefficients are large and opposite in sign to the intercept, the problem is in the reference category of the dependent variable used in the denominator in for all coefficients and the researcher may need to merge that category of the dependent variable with an adjacent one.

Explain the error message I am getting in SPSS about cells with zero frequencies.

The researcher may see an error message that looks like this: "Warning. There are 742 (65.8%) cells (i.e., dependent variable levels by subpopulations) with zero frequencies." If there are only factors in the model, the researcher should strive to minimize 0-count cells by eliminating categories or recoding to combine categories. However, if there are continuous covariates in the logistic model, it is virtually impossible for the cells formed by values of the covariates by values of other predictors to be well populated and large numbers of such cells (subpopulations) will be zero. This is normal when covariates are in the model and the warning may be disregarded if cell count is adequate for the categorical variables in the model.

Is it true for logistic regression, as it is for OLS regression, that the beta weight (standardized logit coefficient) for a given independent reflects its explanatory power controlling for other variables in the equation, and that the betas will change if variables are added or dropped from the equation?

Yes, the same basic logic applies. This is why it is best in either form of regression to compare two or more models for their relative fit to the data rather than simply to show the data are not inconsistent with a single model. The model, of course, dictates which variables are entered and one uses the ENTER method in SPSS, which is the default method.

What is the coefficient in logistic regression which corresponds to R-Square in **multiple regression?**

There is no exactly analogous coefficient. See the discussion of R-squared, above. *Cox and Snell's R-Square* is an attempt to imitate the interpretation of multiple R-

Square, and *Nagelkerke's R-Square* is a further modification of the Cox and Snell coefficient to assure that it can vary from 0 to 1.

Is multicollinearity a problem for logistic regression the way it is for multiple linear regression?

Yes, it is a problem. The discussion in separate Statistical Associates "Blue Book" volume on "Multiple Regression" discusses multicollinearity at greater length and is also relevant to logistic regression.

What is the logistic equivalent to the VIF test for multicollinearity in OLS regression? Can odds ratios be used?

Multicollinearity is a problem when high in either logistic or OLS regression because in either case standard errors of the b coefficients will be high and interpretations of the relative importance of the independent variables will be unreliable. In an OLS regression context, recall that VIF is the reciprocal of tolerance, which is 1 - R-squared. When there is high multicollinearity, R-squared will be high also, so tolerance will be low, and thus VIF will be high. When VIF is high, the b and beta weights are unreliable and subject to misinterpretation. For typical social science research, multicollinearity is considered not to be a problem if VIF <= 4, a level which corresponds to doubling the standard error of the b coefficient.

As there is no direct counterpart to R-squared in logistic regression, VIF cannot be computed. Applying VIF logic to various pseudo-R-squared measures has exploratory value at best but this test would be less reliable than in OLS regression and is recommended. Also, a high odds ratio in logistic regression would not be sufficient evidence of multicollinearity in itself.

To the extent that one independent variable is linearly or nonlinearly related to another independent variable, multicollinearity could be a problem in logistic regression. Some authors use the VIF test in OLS regression to screen for multicollinearity prior to running logistic regression, but this is sufficient only if nonlinearity can be ruled out. Evidence of absence of nonlinearity may be (1) eta-square not significantly higher than R-square, or (2) in logistic regression, the Box-Tidwell transformation and orthogonal polynomial contrasts may be used to test linearity among the independents (see above).

How can one use estimated variance of residuals to test for model misspecification?

The misspecification problem may be assessed by comparing expected variance of residuals with observed variance. Since logistic regression assumes binomial errors, the estimated variance (y) = m(1 - m), where m = estimated mean residual. "Overdispersion" is when the observed variance of the residuals is greater than the expected variance. Overdispersion indicates misspecification of the model, non-random sampling, or an unexpected distribution of the variables. If misspecification is involved, one must respecify the model. If that is not the case, then the computed standard error will be over-optimistic (confidence intervals will be too wide). One suggested remedy is to use adjusted SE = SE*SQRT(s), where s = D/df, where D = dispersion and df=degrees of freedom in the model.

How are interaction effects handled in logistic regression?

The same as in OLS regression. The researcher must add interaction terms to the model as cross-products of the standardized independent variables and/or dummy independent variables. Some statistics programs will allow the researcher to specify the pairs of interacting variables and will do all the computation automatically. In SPSS, use the categorical covariates option: highlight two variables, then click on the button that shows >a*b> to put them in the Covariates box .The significance of an interaction effect is the same as for any other variable, except in the case of a set of dummy variables representing a single ordinal variable.

When an ordinal variable has been entered as a set of dummy variables, the interaction of another variable with the ordinal variable will involve multiple interaction terms. In this case the significance of the interaction of the two variables is the significance of the change of R-square of the equation with the interaction terms and the equation without the set of terms associated with the ordinal variable. Computing the significance of the difference of two R-squares is discussed in the separate Statistical Associates “Blue Book” volume on “Multiple Regression”.

Does stepwise logistic regression exist, as it does for OLS regression?

Yes, it exists, but it is not supported by all statistics packages. It is supported by SPSS, in both binary and multinomial logistic regression as described above. Stepwise regression is used in the exploratory phase of research or for purposes of pure prediction, not theory testing. In the theory testing stage the researcher should base selection of the variables on theory, not on a computer algorithm. Menard (1995: 54) writes, "there appears to be general agreement that the use of computer-controlled stepwise procedures to select variables is inappropriate for theory testing because it capitalizes on random variations in the data and produces results that tend to be idiosyncratic and difficult to replicate in any sample other than the sample in which they were originally obtained." Those who use this procedure often focus on *step chi-square* output in SPSS, which represents the change in the likelihood ratio test (model chi-square test; see above) at each step. See the discussion of stepwise logistic regression above.

What are the stepwise options in multinomial logistic regression in SPSS?

As described in the SPSS manual, the four options are forward entry, forward stepwise, backward elimination, and backward stepwise:

FORWARD ENTRY

1. Estimate the parameter and likelihood function for the initial model and let it be our current model.

2. Based on the ML estimates of the current model, calculate the score or LR statistic for every variable eligible for inclusion and find its significance.

3. Choose the variable with the smallest significance. If that significance is less than the probability for a variable to enter, then go to step 4; otherwise, stop FORWARD.

4. Update the current model by adding a new variable. If there are no more eligible variable left, stop FORWARD; otherwise, go to step 2.

FORWARD STEPWISE

1. Estimate the parameter and likelihood function for the initial model and let it be our current model.

2. Based on the ML estimates of the current model, calculate the score statistic or likelihood ratio statistic for every variable eligible for inclusion and find its significance.

3. Choose the variable with the smallest significance (p-value). If that significance is less than the probability for a variable to enter, then go to step 4; otherwise, stop FSTEP.

4. Update the current model by adding a new variable. If this results in a model which has already been evaluated, stop FSTEP.

5. Calculate the significance for each variable in the current model using LR or Wald's test.

6. Choose the variable with the largest significance. If its significance is less than the probability for variable removal, then go back to step 2. If the current model with the variable deleted is the same as a previous model, stop FSTEP; otherwise go to the next step.

7. Modify the current model by removing the variable with the largest significance from the previous model. Estimate the parameters for the modified model and go back to step 5.

BACKWARD ELIMINATION

1. Estimate the parameters for the full model that includes all eligible variables. Let the current model be the full model.

2. Based on the ML estimates of the current model, calculate the LR or Wald's statistic for all variables eligible for removal and find its significance.

3. Choose the variable with the largest significance. If that significance is less than the probability for a variable removal, then stop BACKWARD; otherwise, go to the next step.

4. Modify the current model by removing the variable with the largest significance from the model.

5. Estimate the parameters for the modified model. If all the variables in the BACKWARD list are removed then stop BACKWARD; otherwise, go back to step 2.

BACKWARD STEPWISE

1. Estimate the parameters for the full model that includes the final model from previous method and all eligible variables. Only variables listed on the BSTEP variable list are eligible for entry and removal. Let current model be the full model.

2. Based on the ML estimates of the current model, calculate the LR or Wald's statistic for every variable in the BSTEP list and find its significance.

3. Choose the variable with the largest significance. If that significance is less than the probability for a variable removal, then go to step 5. If the current model without the variable with the largest significance is the same as the previous model, stop BSTEP; otherwise go to the next step.

4. Modify the current model by removing the variable with the largest significance from the model.

5. Estimate the parameters for the modified model and go back to step 2.

6. Check to see any eligible variable is not in the model. If there is none, stop BSTEP; otherwise, go to the next step.

7. Based on the ML estimates of the current model, calculate LR statistic or score statistic for every variable not in the model and find its significance.

8. Choose the variable with the smallest significance. If that significance is less than the probability for the variable entry, then go to the next step; otherwise, stop BSTEP.

9. Add the variable with the smallest significance to the current model. If the model is not the same as any previous models, estimate the parameters for the new model and go back to step 2; otherwise, stop BSTEP.

May I use the multinomial logistic option when my dependent variable is binary?

Binary dependent variables can be fitted in both the binary and multinomial logistic regression options of SPSS, with different options and output. This can be done but the multinomial procedure will aggregate the data, yielding different goodness of fit tests. The SPSS online help manual notes, " An important theoretical distinction is that the Logistic Regression procedure produces all predictions, residuals, influence statistics, and goodness-of-fit tests using data at the individual case level, regardless of how the data are entered and whether or not the number of covariate patterns is smaller than the total number of cases, while the Multinomial Logistic Regression procedure internally aggregates cases to form subpopulations with identical covariate patterns for the predictors, producing predictions, residuals, and goodness-of-fit tests based on these subpopulations. If all predictors are categorical or any continuous predictors take on only a limited number of values (so that there are several cases at each distinct covariate pattern) the subpopulation approach can produce valid goodness-of-fit tests and informative residuals, while the individual case level approach cannot."

What is nonparametric logistic regression and how is it more nonlinear?

In general, nonparametric regression as discussed in the separate Statistical Associates "Blue Book" volume on “Multiple Regression”, can be extended to the case of GLM regression models like logistic regression. See Fox (2000: 58-73).

GLM nonparametric regression allows the logit of the dependent variable to be a nonlinear function of the parameter estimates of the independent variables. While GLM techniques like logistic regression are nonlinear in that they employ a transform (for logistic regression, the natural log of the odds of a dependent

variable) which is nonlinear, in traditional form the result of that transform (the logit of the dependent variable) is a linear function of the terms on the right-hand side of the equation. GLM non-parametric regression relaxes the linearity assumption to allow nonlinear relations over and beyond those of the link function (logit) transformation.

Generalized nonparametric regression is a GLM equivalent to OLS local regression (local polynomial nonparametric regression), which makes the dependent variable a single nonlinear function of the independent variables. The same problems noted for OLS local regression still exist, notably difficulty of interpretation as independent variables increase.

Generalized additive regression is the GLM equivalent to OLS additive regression, which allow the dependent variable to be the additive sum of nonlinear functions which are different for each of the independent variables. Fox (2000: 74-77) argues that generalized additive regression can reveal nonlinear relationships under certain circumstances where they are obscured using partial residual plots alone, notably when a strong nonlinear relationship among independents exists alongside a strong nonlinear relationship between an independent and a dependent.

How many independent variables can I have?

There is no precise answer to this question, but the more independent variables, the more likelihood of multicollinearity. In general, there should be significantly fewer independent variables than in OLS regression as logistic dependent variables, being categorized, have lower information content. Also, if you have 20 independent variables, at the .05 level of significance you would expect one to be found to be significant just by chance. A rule of thumb is that there should be no more than 1 independent for each 10 cases in the sample. In applying this rule of thumb, keep in mind that if there are categorical independent variables, such as dichotomies, the number of cases should be considered to be the lesser of the groups (ex., in a dichotomy with 480 0's and 20 1's, effective size would be 20), and by the 1:10 rule of thumb, the number of independent variables should be the smaller group size divided by 10 (in the example, 20/10 = 2 independent variables maximum).

How do I express the logistic regression equation if one or more of my independent variables is categorical?

When a covariate is categorical, SPSS will print out "parameter codings," which are the internal-to-SPSS values which SPSS assigns to the levels of each categorical variable. These parameter codings are the X values which are multiplied by the logit (effect) coefficients to obtain the predicted values. Similar codings tables are available in SAS and Stata.

How do I compare logit coefficients across groups formed by a categorical independent variable?

There are two strategies, the subgroup approach and the indicator approach.

The first strategy is to separate the sample into subgroups, then perform otherwise identical logistic regression for each. The researcher then computes the p value for a Wald chi-square test of the significance of the differences between the corresponding coefficients. The formula for this test, for the case of two subgroup parameter estimates, is Wald chi-square = $[(b_1 - b_2)^2]/\{[se(b_1)]^2 + [se(b_2)]^2$, where the b's are the logit coefficients for groups 1 and 2 and the se terms are their corresponding standard errors. This chi-square value is read from a table of the chi-square distribution with 1 degree of freedom.

The second strategy is to create an indicator (dummy) variable or set of variables which reflects membership/non-membership in the group, and also to have interaction terms between the indicator dummies and other independent variables, such that the significant interactions are interpreted as indicating significant differences across groups for the corresponding independent variables. When an indicator variable has been entered as a set of dummy variables, its interaction with another variable will involve multiple interaction terms. In this case the significance of the interaction of the indicator variable and another independent variable is the significance of the change of R-square of the equation with the interaction terms and the equation without the set of terms associated with the ordinal variable.

Allison (1999: 186) has shown that "Both methods may lead to invalid conclusions if residual variation differs across groups." Unequal residual variation across groups will occur, for instance, whenever an unobserved variable (whose effect is

incorporated in the disturbance term) has different impacts on the dependent variable depending on the group. Allison suggests that, as a rule of thumb, if "one group has coefficients that are consistently higher or lower than those in another group, it is a good indication of a potential problem ..." (p. 199). Allison explicated a new test to adjust for unequal residual variation, presenting the code for computation of this test in SAS, LIMDEP, BMDP, and STATA. The test is not implemented directly by SPSS or SAS. Note Allison's test is conservative in that it will always yield a chi-square which is smaller than the conventional test, making it harder to prove the existence of cross-group differences.

How do I compute the confidence interval for the unstandardized logit (effect) coefficients?

To obtain the upper confidence limit at the 95% level, where b is the unstandardized logit coefficient, se is the standard error, and e is the natural logarithm, take e to the power of (b + 1.96*se). Subtract to get the lower CI.

Acknowledgments

Appreciation is expressed to the peer reviewer of an earlier draft of this manuscript for the generous donation of time and effort, by Prof. Ankit Mahajan of AnalyticsTraining.com.

Bibliography

Agresti, Alan (1996). *An introduction to categorical data analysis*. NY: John Wiley. An excellent, accessible introduction.

Allison, Paul D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods and Research*, 28(2): 186-208.

Allison, Paul D. (2012). *Logistic regression using SAS: Theory and application, Second ed.* Cary, NC: SAS Institute.

Breslow, N. E. & Day. N. E. (1980) *Statistical methods in cancer research. Volume I - The analysis of case-control studies.* Lyon, France: International Agency for Research on Cancer. IARC Scientific Publications No. 32.

Cleves, Mario A. (2002). From the help desk: Comparing areas under receiver operating characteristic curves from two or more probit or logit models, *The Stata Journal* 2(3): 301-313.

Cox, D.R. and E. J. Snell (1989). Analysis of binary data (2nd edition). London: Chapman & Hall.

DeMaris, Alfred (1992). *Logit modeling: Practical applications*. Thousand Oaks, CA: Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 106.

Estrella, A. (1998). A new measure of fit for equations with dichotomous dependent variables. *Journal of Business and Economic Statistics* 16(2): 198-205. Discusses proposed measures for an analogy to R^2.

Fox, John (2000). Multiple and generalized nonparametric regression. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences Series No.131. Covers nonparametric regression models for GLM techniques like logistic regression. Nonparametric regression allows the logit of the dependent to be a nonlinear function of the logits of the independent variables.

Greenland, Sander ; Schwartzbaum, Judith A.; & Finkle, William D. (2000). Problems due to small samples and sparse data in conditional logistic regression. *American Journal of Epidemiology* 151:531-539.

Hair J. F.; Anderson Jr., R.; Tatham, R. L. (1987). *Multivariate data analysis with readings. 2nd edition*. NY: Macmillan Publishing Company.

Hand, D.; Mannila, H.; & Smyth, P. (2001) *Principles of data mining.* Cambridge, MA: MIT Press.

Hanley, James A. & McNeil, Barbara J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148(3): 839-843.

Harrell, Frank (2001). *Regression modeling strategies*. NY: Springer.

Hosmer, David and Stanley Lemeshow (1989, 2000). *Applied Logistic Regression*. 2nd ed., 2000. NY: Wiley & Sons. A much-cited treatment utilized in SPSS routines.

Jaccard, James (2001). *Interaction effects in logistic regression*. Thousand Oaks, CA: Sage Publications. Quantitative Applications in the Social Sciences Series, No. 135.

Jennings, D. E. (1986). Outliers and residual distributions in logistic regression. *Journal of the American Statistical Association* (81), 987-990.

Kleinbaum, D. G. (1994). *Logistic regression: A self-learning text*. New York: Springer-Verlag. What it says.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models, 2nd ed.* London: Chapman & Hall. Recommended by the SPSS multinomial logistic tutorial.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In: *Frontiers in Economics*, P. Zarembka, eds. NY: Academic Press.

McKelvey, Richard and William Zavoina (1994). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4: 103-120. Discusses polytomous and ordinal logits.

Menard, Scott (2002). *Applied logistic regression analysis, 2nd Edition*. Thousand Oaks, CA: Sage Publications. Series: Quantitative Applications in the Social Sciences, No. 106. First ed., 1995.

Menard, Scott (2010). *Logistic regression: From introductory to advanced concepts and applications*. Thousand Oaks, CA: Sage Publications.

Meyers, Lawrence S.; Gamst, Glenn; & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation*. Thousand Oaks, CA: Sage Publications.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, Vol. 78, No. 3: 691-692. Covers the two measures of R-square for logistic regression which are found in SPSS output.

O'Connell, Ann A. (2005). *Logistic regression models for ordinal response variables*. Thousand Oaks, CA: Sage Publications. Quantittive Applications in the Social Sciences, Volume 146.

Osborne, Jason W. (2014). *Best practices in logistic regression*. Thousand Oaks, CA: Sage Publications.

Pampel, Fred C. (2000). *Logistic regression: A primer*. Sage Quantitative Applications in the Social Sciences Series #132. Thousand Oaks, CA: Sage Publications. Pp. 35-38 provide an example with commented SPSS output.

Peduzzi, P., J. Concato, E. Kemper, T. R. Holford, and A. Feinstein (1996). A simulation of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology* 99: 1373-1379.

Pedhazur, E. J. (1997). Multiple regression in behavioral research, 3rd ed. Orlando, FL: Harcourt Brace.

Peng, Chao-Ying Joann; Lee, Kuk Lida; & Ingersoll, Gary M. (2002). An introduction to logistic regression analysis and reporting. Journal of Educational Research 96(1): 3-13.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. NY: Oxford University Press.

Press, S. J. and S. Wilson (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, Vol. 73: 699-705. The authors make the case for the superiority of logistic regression for situations where the assumptions of multivariate normality are not met (ex., when dummy variables are used), though discriminant analysis is held to be better when they are. They conclude that logistic and discriminant analyses will usually yield the same conclusions, except in the case when there are independents which result in predictions very close to 0 and 1 in logistic analysis. This can be revealed by examining a 'plot of observed groups and predicted probabilities' in the SPSS logistic regression output.

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden, ed., *Sociological Methodology 1995*: 111-163. London: Tavistock. Presents BIC criterion for evaluating logits.

Rice, J. C. (1994). "Logistic regression: An introduction". In B. Thompson, ed., *Advances in social science methodology*, Vol. 3: 191-245. Greenwich, CT: JAI Press. Popular introduction.

Tabachnick, B.G., and L. S. Fidell (1996). *Using multivariate statistics*, 3rd ed. New York: Harper Collins. Has clear chapter on logistic regression.

Wright, R.E. (1995). "Logistic regression". In L.G. Grimm & P.R. Yarnold, eds., *Reading and understanding multivariate statistics*. Washington, DC: American Psychological Association. A widely used recent treatment.

Statistical Associates Publishing
Blue Book Series

Association, Measures of
Case Studies
Cluster Analysis
Content Analysis
Correlation
Correlation, Partial
Correspondence Analysis
Cox Regression
Creating Simulated Datasets
Crosstabulation
Curve Estimation & Nonlinear Regression
Delphi Method in Quantitative Research
Discriminant Function Analysis
Ethnographic Research
Evaluation Research
Factor Analysis
Focus Group Research
Game Theory
Generalized Linear Models/Generalized Estimating Equations
GLM Multivariate, MANOVA, and Canonical Correlation
GLM (Univariate), ANOVA, and ANCOVA
Grounded Theory
Life Tables & Kaplan-Meier Survival Analysis
Literature Review in Research and Dissertation Writing
Logistic Regression: Binary & Multinomial
Log-linear Models,
Longitudinal Analysis
Missing Values & Data Imputation
Multidimensional Scaling
Multiple Regression
Narrative Analysis
Network Analysis
Neural Network Models

Nonlinear Regression
Ordinal Regression
Parametric Survival Analysis
Partial Correlation
Partial Least Squares Regression
Participant Observation
Path Analysis
Power Analysis
Probability
Probit and Logit Response Models
Research Design
Scales and Measures
Significance Testing
Social Science Theory in Research and Dissertation Writing
Structural Equation Modeling
Survey Research & Sampling
Testing Statistical Assumptions
Two-Stage Least Squares Regression
Validity & Reliability
Variance Components Analysis
Weighted Least Squares Regression

Statistical Associates Publishing
http://www.statisticalassociates.com
sa.publishers@gmail.com

