Binomial Logistic Regression

Introduction

A binomial logistic regression attempts to predict the probability that an observation falls into one of two categories of a dichotomous dependent variable based on one or more independent variables that can be either continuous or categorical.

In many ways, binomial logistic regression is similar to linear regression, with the exception of the measurement type of the dependent variable (i.e., linear regression uses a continuous dependent variable rather than a dichotomous one). However, unlike linear regression, you are not attempting to determine the predicted value of the dependent variable, but the probability of being in a particular category of the dependent variable given the independent variables. An observation is assigned to whichever category is predicted as most likely. As with other types of regression, binomial logistic regression can also use interactions between independent variables to predict the dependent variable.

Note: Binomial logistic regression is often referred to as just logistic regression.

For example, you could use binomial logistic regression to predict whether students will pass or fail an exam based on the amount of time they spend revising, whether English is their first language and their pre-exam stress levels. Here, your dichotomous dependent variable would be "exam performance", which has two categories – "pass" and "fail" – and you would have three independent variables: the continuous variable, "time spent revising", measured in hours, the dichotomous independent variable, "English as a first language", which has two categories – "yes" and "no" – and the ordinal independent variable, "pre-exam stress levels", which has three levels: "low stress", medium stress" and "high stress".

Basic requirements of a binomial logistic regression

In order to run a binomial logistic regression, there are seven assumptions that need to be considered. The first four assumptions relate to your choice of study design and the measurements you chose to make, whilst the other three assumptions relate to how your data fits the binomial logistic regression model. These assumptions are:

- **Assumption #1:** You have **one dependent variable** that is **dichotomous** (i.e., a **nominal variable** with two outcomes). Examples of **dichotomous variables** include gender (two outcomes:

"males" or "females"), presence of heart disease (two outcomes: "yes" or "no"), employment status (two outcomes: "employed" or "unemployed"), transport type (two outcomes: "bus" or "car").

Note 1: The dependent variable can also be referred to as the "outcome", "target" or "criterion" variable. It does not matter which of these you use, but we will continue to use "dependent variable" for consistency.
Note 2: We refer to the dependent variable as being a nominal variable with two "outcomes", but it is also common to use the word "categories" (i.e., a variable such as "gender" would have two categories: "males" or "females"). Again, it does not matter which of these you use.

- **Assumption #2**: You have **one or more independent variables** that are measured on either a **continuous** or **nominal** scale. Examples of **continuous variables** include height (measured in metres and centimetres), temperature (measured in °C), salary (measured in US dollars), revision time (measured in hours), intelligence (measured using IQ score), firm size (measured in terms of the number of employees), age (measured in years), reaction time (measured in milliseconds), grip strength (measured in kg), power output (measured in watts), test performance (measured from 0 to 100), sales (measured in number of transactions per month) and academic achievement (measured in terms of GMAT score). Examples of **nominal variables** include gender (e.g., two categories: male and male), ethnicity (e.g., three categories: Caucasian, African American and Hispanic) and profession (e.g., five categories: surgeon, doctor, nurse, dentist, therapist).

Note: The "categories" of the independent variable are also referred to as "groups" or "levels", but the term "levels" is usually reserved for the categories of an ordinal variable (e.g., an ordinal variable such as "fitness level", which has three levels: "low", "moderate" and "high"). However, these three terms – "categories", "groups" and "levels" – can be used interchangeably. We refer to them as categories in this guide.
Important: If one of your independent variables was measured at the ordinal level, it can still be entered in a binomial logistic regression, but it must be treated as either a continuous or nominal variable. It cannot be entered as an ordinal variable. Examples of ordinal variables include Likert items (e.g., a 7-point scale from strongly agree through to strongly disagree), physical activity level (e.g., 4 groups: sedentary, low, moderate and high), customer liking a product (ranging from "Not very much", to "It is OK", to "Yes, a lot"), and so forth.

- **Assumption #3**: You should have **independence of observations** and the **categories of the dichotomous dependent variable** and **all your nominal independent variables** should be **mutually exclusive and exhaustive**.

Independence of observations means that there is no relationship between the observations in each category of the dependent variable or the observations in each category of any nominal independent variables. In addition, there is no relationship between the categories. Indeed, an important distinction is made in statistics when comparing values from either different individuals or from the same individuals.

To illustrate this, consider again the example from the Introduction where binomial logistic regression could be used to predict whether students will pass or fail an exam based on the amount of time they spend revising, whether English is their first language and their pre-exam stress levels. Here, your dichotomous dependent variable would be "exam performance", which has two categories – "pass" and "fail" – and you would have three independent variables: the continuous variable, "time spent revising", measured in hours, the dichotomous independent variable, "English as a first language", which has two categories – "yes" and "no" – and the ordinal independent variable, "pre-exam stress levels", which has three levels: "low stress", medium stress" and "high stress".

In this scenario, independence of observations means that a student could **either** "pass" or "fail" the exam. They **could not** pass "and" fail the exam. As such, the student has to be placed into one of the two categories of the dependent variable. The student cannot be placed into both categories. Similar, take the dichotomous independent variable, "English as a first language". The correct answer for the purposes of a binomial logistic regression is **either** "yes" or "no". A student cannot be entered into both categories.

Independence of observations is largely a study design issue rather than something you can test for using SPSS Statistics, but it is an important assumption of binomial logistic regression. If there is a relationship between the categories of any variables or between the categories themselves, this means that the observations are **related**. Therefore, if your study fails this assumption, you will need to use another statistical test instead of binomial logistic regression; possibly **linear mixed models** or **Generalized Estimating Equations (GEE)**.

- **Assumption #4**: You should have a **bare minimum of 15 cases per independent variable**, although some recommend as high as **50 cases per independent variable**. As with other multivariate techniques, such as multiple regression, there are a number of recommendations regarding minimum **sample size**. Indeed, binomial logistic regression relies on maximum likelihood estimation (MLE) and the reliability of estimates declines significantly for combinations of cases where there are few cases.

- **Assumptions #5, #6 and #7**: A binomial logistic regression must also meet three assumptions that relate to how your data fits the binomial logistic regression model in order to provide a valid result: (a) there should be a linear relationship between the continuous independent variables and the logit transformation of the dependent variable; (b) there should be no multicollinearity; and (c) there should be no significant outliers, leverage or influential points.

Assuming that you are confident your study design meets assumptions #1, #2, #3 and #4 above, we explain the main characteristics of binomial logistic regression analysis in the section that follows.

Fitting a binomial logistic regression model

Binomial logistic regression is part of a larger statistical group of tests called Generalized Linear Models (GzLM). These tests are an extension of Linear Models (e.g., multiple regression) to incorporate dependent variables that are not just continuous, but may be measured on other types of measurement scale (e.g., dichotomous or ordinal measurement scales).

Like multiple regression, binomial logistic regression allows for a relationship to be modelled between multiple independent variables and a single dependent variable where the independent variables are being used to predict the dependent variable. However, in the case of a binomial logistic regression, the dependent variable is dichotomous. In addition, a transformation is applied so that instead of predicting the category of the binomial logistic regression directly, the logit of the dependent variable is predicted instead.

For example, if we consider four independent variables to be "X1" through "X4" and the dependent variable to be "Y", a binomial logistic regression models the following:

$$\text{logit}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon.$$

Where β_0 is the intercept (also known as the constant), β_1 is the slope parameter (also known as the slope coefficient) for X_1, and so forth, and ε represents the errors. This represents the population model, but it can be estimated as follows:

$$\text{logit}(Y) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + b_4 X_4 + e$$

In the formula above, b_0 is the sample intercept (aka constant) and estimates β_0, b_1 is the sample slope parameter for X_1 and estimates β_1, and so forth, e represents the sample errors/residuals and estimates ε.

A *logit* is the natural log of the odds of an event occurring. It has little direct meaning. However, by applying an anti-log it can have a much more interpretative meaning. In addition, through further calculations you can ascertain other useful properties of the predictive power of your binomial logistic regression model, such as the percentage of correctly classified cases.

Example used in this guide

A health researcher wants to be able to predict whether the incidence of heart disease can be predicted based on age, weight, gender and maximal aerobic capacity (VO_2max) (an indicator of fitness and health). To this end, the researcher recruited 100 participants to perform a maximum VO_2max test as well as recording their age, weight and gender. The participants were also evaluated for the presence of heart disease. A logistic regression was then run to determine whether the presence of heart disease could be predicted from their VO_2max, age, weight and gender. *Note:* this data is fictitious.

Download the SPSS Statistics data file for the example used throughout this guide from here.

http://pc.cd/BqS7

NOTE: You need to have the SPSS Statistics software on your computer for this file to open.

Setting up your data

For a binomial logistic regression you will have at least two variables – one dependent variable and one independent variable – but you will typically have two or more independent variables. In addition, you may also choose to include a case identifier, as discussed below. In this example, we have the following six variables:

1) The dependent variable, `heart_disease`, which is whether the participant has heart disease;

and

2) The independent variable, `age`, which is the participant's age in years;
3) The independent variable, `weight`, which is the participant's weight (technically, it is their 'mass');
4) The independent variable, `gender`, which has two categories: "Male" and "Female";
5) The independent variable, `VO2max`, which is the maximal aerobic capacity.

and

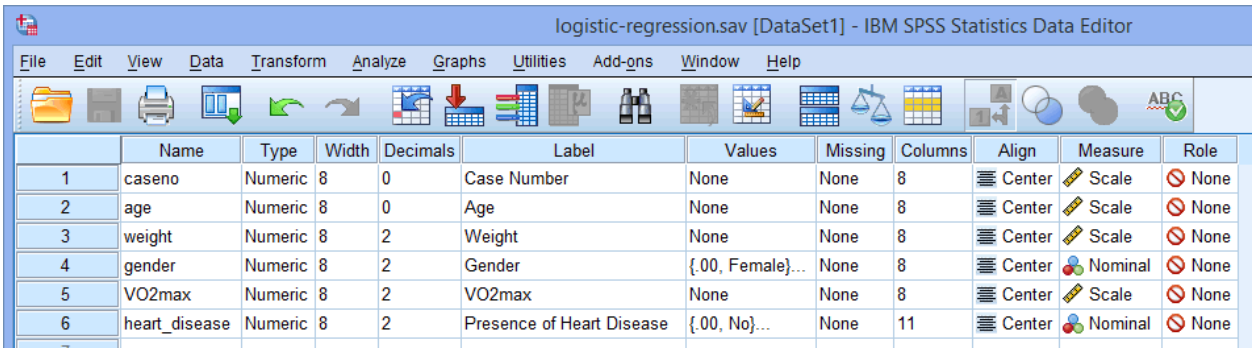
6) The case identifier, `caseno`, which is used for easy elimination of cases (e.g., participants) that might occur when checking assumptions.

> **Note:** The case identifier is not used directly in calculations for a logistic regression analysis. Therefore, it is not essential and you can choose not to include it. That said, we do find that it is very useful.

To set up these variables, SPSS Statistics has a **Variable View** where you define the types of variables you are analysing and a **Data View** where you enter your data for these variables. First, we show you how to set up your variables – the case identifier, independent variables and dependent variable in the **Variable View** window of SPSS Statistics. Finally, we show you how to enter your data into the **Data View** window.

The Variable View in SPSS Statistics

At the end of the setup process, your **Variable View** window will look like the one below, which illustrates the setup for both the case identifier, independent variables and dependent variable:

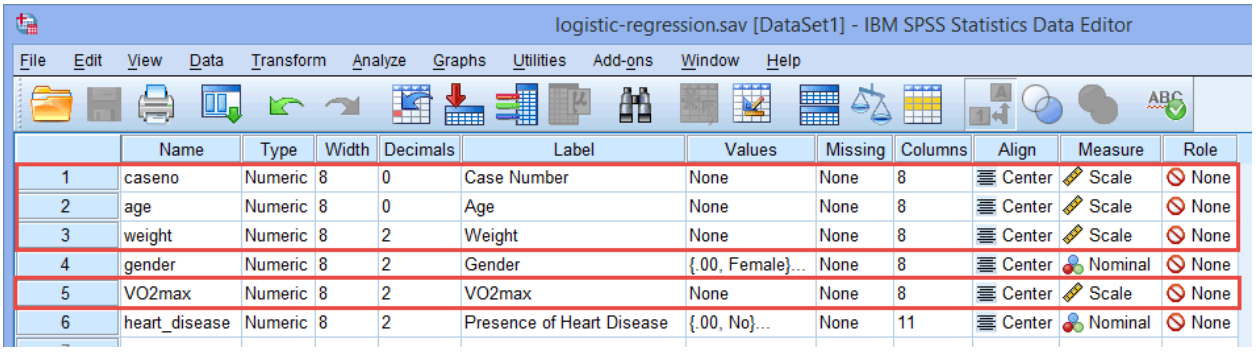


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	caseno	Numeric	8	0	Case Number	None	None	8	Center	Scale	None
2	age	Numeric	8	0	Age	None	None	8	Center	Scale	None
3	weight	Numeric	8	2	Weight	None	None	8	Center	Scale	None
4	gender	Numeric	8	2	Gender	{.00, Female}...	None	8	Center	Nominal	None
5	VO2max	Numeric	8	2	VO2max	None	None	8	Center	Scale	None
6	heart_disease	Numeric	8	2	Presence of Heart Disease	{.00, No}...	None	11	Center	Nominal	None

In the **Variable View** window above, you will have entered all your variables: one on each row. For our example, we have put the case identifier, caseno, on row 1, the four independent variables – age, weight, gender and VO2max – on rows 2, 3, 4 and 5, respectively, and the dependent variable, heart_disease, on row 6.

> **Note:** The order that you enter your variables into the **Variable View** is irrelevant. It will simply determine the order of the columns in the **Data View**.

First, look at the continuous independent variables, age, weight and VO2max – on rows 2, 3 and 5 respectively, as well as the case identifier, caseno, on row 1, as shown below:



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	caseno	Numeric	8	0	Case Number	None	None	8	Center	Scale	None
2	age	Numeric	8	0	Age	None	None	8	Center	Scale	None
3	weight	Numeric	8	2	Weight	None	None	8	Center	Scale	None
4	gender	Numeric	8	2	Gender	{.00, Female}...	None	8	Center	Nominal	None
5	VO2max	Numeric	8	2	VO2max	None	None	8	Center	Scale	None
6	heart_disease	Numeric	8	2	Presence of Heart Disease	{.00, No}...	None	11	Center	Nominal	None

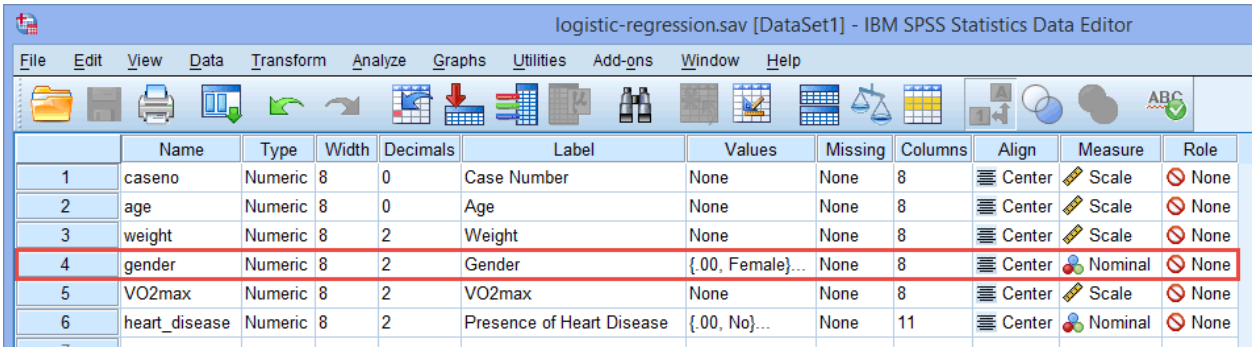
Start by entering the name of each of the continuous independent variables and the case identifier in the cell under the Name column (e.g., "caseno" on row 1 and "age" on row 2). There are certain "illegal" characters that cannot be entered into the Name cell. Therefore, if you get an error message, you can learn what these are in our general data setup guide here. For your own clarity, you can also provide a label for your variables in the Label column (e.g., the label we entered for "caseno" was "Case number").

The cell under the Measure column should show Scale, indicating that these variables were measured on a continuous scale, whilst the cell under the Role column should show None.

Note 1: The case identifier, caseno, can be labelled as Scale or Nominal. However, all continuous independent variables must be labelled as Scale. The term "Scale" is used in SPSS Statistics to refer to variables that are measured on a continuous scale.

Note 2: We suggest changing the cell under the Role column from Input to None for all your variables, but you do not have to make this change. We suggest that you do because there are certain analyses in SPSS Statistics where the Input setting results in your variables being automatically transferred into certain fields in the dialogue boxes you may use to carry out your analysis. Since you may not want to transfer these variable, we suggest changing the Input setting to None so that this does not happen automatically.

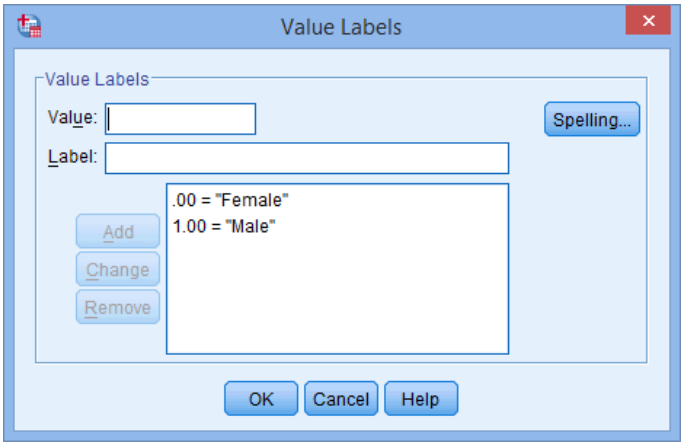
Next, look at the nominal independent variable, gender, on row 4 below:



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	caseno	Numeric	8	0	Case Number	None	None	8	Center	Scale	None
2	age	Numeric	8	0	Age	None	None	8	Center	Scale	None
3	weight	Numeric	8	2	Weight	None	None	8	Center	Scale	None
4	gender	Numeric	8	2	Gender	{.00, Female}...	None	8	Center	Nominal	None
5	VO2max	Numeric	8	2	VO2max	None	None	8	Center	Scale	None
6	heart_disease	Numeric	8	2	Presence of Heart Disease	{.00, No}...	None	11	Center	Nominal	None

Enter the name of your nominal independent variable(s) in the cell under the Name column (e.g., "gender" in our example). The cell under the Measure column should show Nominal, indicating that you have a nominal independent variable, whilst the cell under the Role column should show None

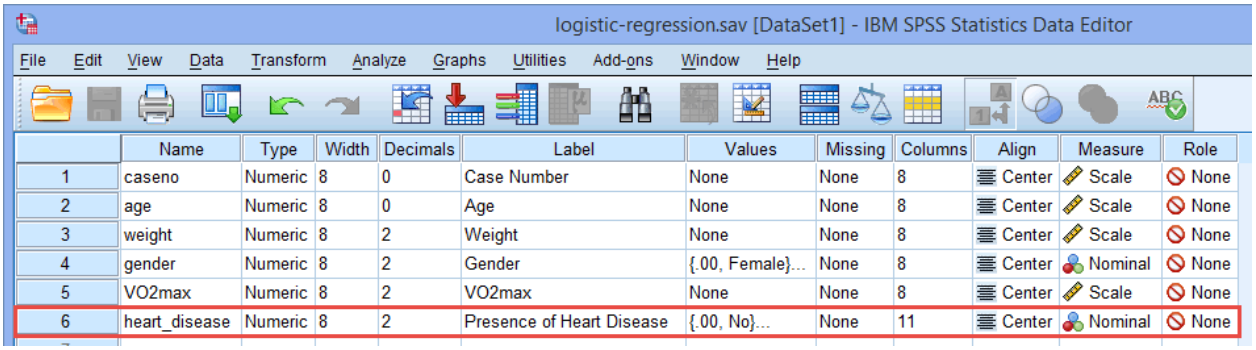
Next, the cell under the Values column should contain the information about the groups/levels of your nominal independent variables (e.g., "Male" and "Female" for gender). To enter this information, click into the cell under the Values column for one of your independent variables. The ... button will appear in the cell. Click on this button and the **Value Labels** dialogue box will appear. You now need to give each group/level of your nominal independent variable a "value", which you enter into the Value: box (e.g., "**1**"), as well as a "label", which you enter into the Label: box (e.g., "**Male**"). By clicking the Add button the coding will appear in the main box (e.g., **"1.00 = "Male"** for gender). The setup for our nominal independent variable, gender, is shown below:



Note: You will typically enter an integer (e.g., "**1**") into the Value: box to represent the group/level of your independent variable and not a decimal (e.g., "**1.00**"). However, SPSS Statistics adds the 2 decimal places by default when you click on the Add button (e.g., **1.00 = "Male"**), as shown in the **Value Labels** box above. Therefore, do not think that you have done anything wrong.

Important: In a binomial logistic regression your independent variables will be either Scale variables or Nominal variables. If any of your independent variables were measured on an ordinal scale (i.e., they are Ordinal variables), you need to decide whether to enter these as Scale (i.e., continuous) or Nominal variables. They cannot be entered as ordinal variables.

Finally, look at the dichotomous dependent variable, heart_disease, on row 6 below:

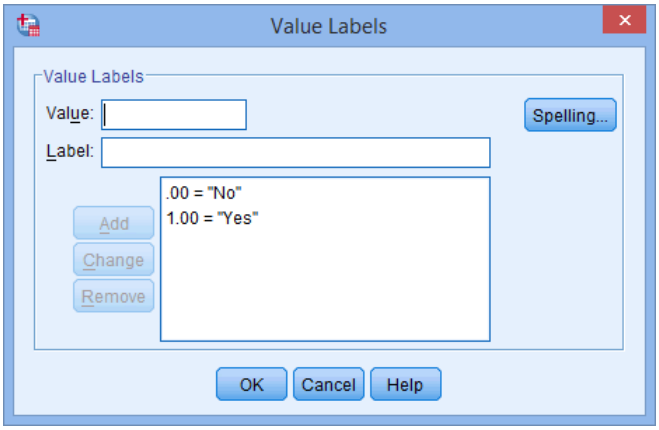


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	caseno	Numeric	8	0	Case Number	None	None	8	Center	Scale	None
2	age	Numeric	8	0	Age	None	None	8	Center	Scale	None
3	weight	Numeric	8	2	Weight	None	None	8	Center	Scale	None
4	gender	Numeric	8	2	Gender	{.00, Female}...	None	8	Center	Nominal	None
5	VO2max	Numeric	8	2	VO2max	None	None	8	Center	Scale	None
6	heart_disease	Numeric	8	2	Presence of Heart Disease	{.00, No}...	None	11	Center	Nominal	None

The setup of your dichotomous dependent variable will be the same as any nominal independent variables you have since a dichotomous variable is simply a nominal variable with just two groups/levels.

Therefore, enter the name of your dichotomous dependent variable in the cell under the Name column (e.g., "heart_disease" in our example). The cell under the Measure column should show Nominal, indicating that you have a nominal dependent variable (i.e., a dichotomous variable), whilst the cell under the Role column should show None .

Again, the cell under the Values column should contain the information about the two levels of your dichotomous dependent variable (e.g., "No" and "Yes" for heart_disease to indicate the presence or absence of heart disease amongst participants). Therefore, the setup for heart_disease is as follows:

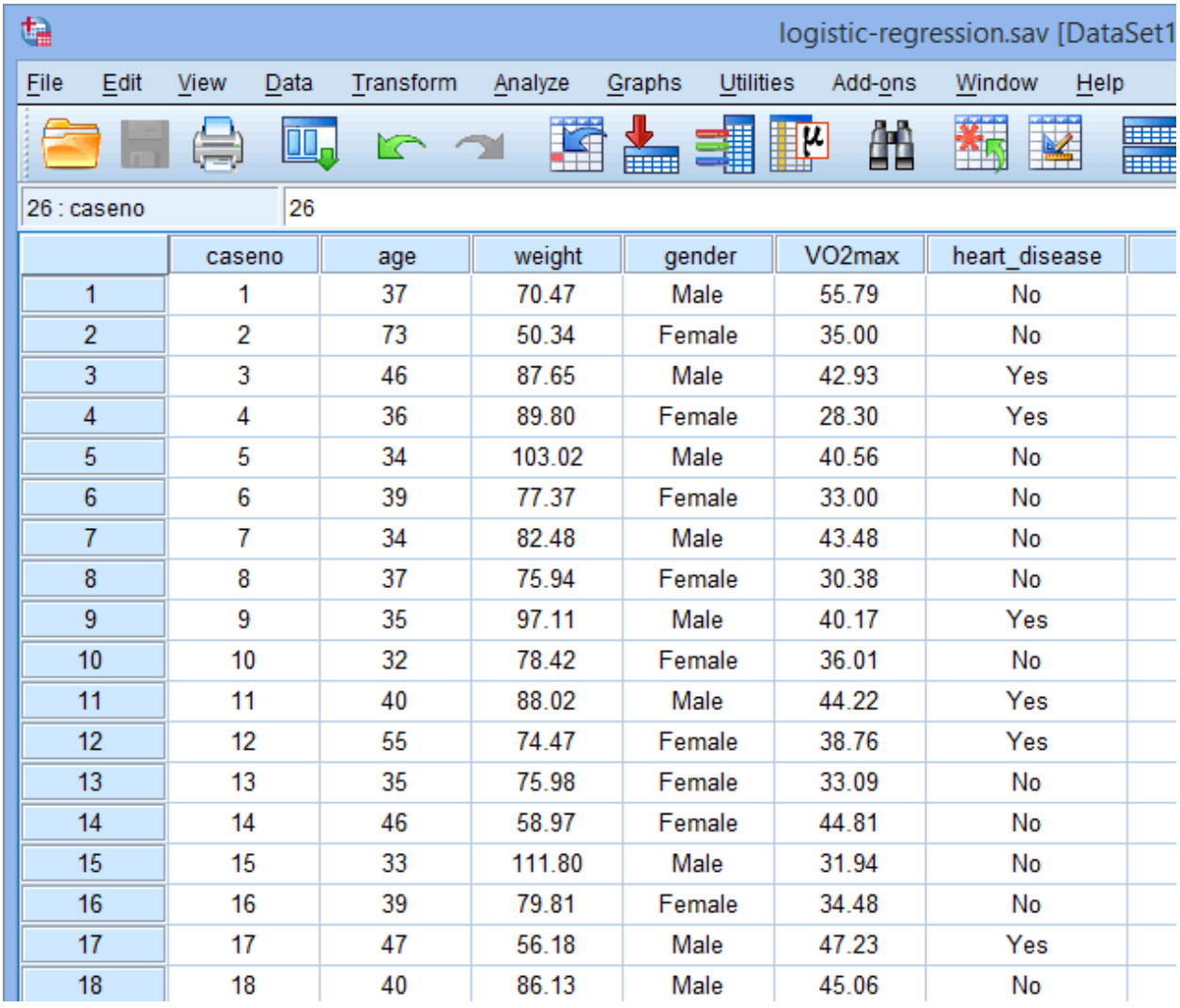


Note: You should code your dichotomous dependent variable as "0" for the "negative" response (i.e., no presence of heart disease) and "1" for the "positive" response (i.e., presence of heart disease).

You have now successfully entered all the information that SPSS Statistics needs to know about your variables into the **Variable View** window. In the next section, we show you how to enter your data into the **Data View** window.

The Data View in SPSS Statistics

Based on the file setup for the variables in the **Variable View** above, the **Data View** window should look as follows:



	caseno	age	weight	gender	VO2max	heart_disease
1	1	37	70.47	Male	55.79	No
2	2	73	50.34	Female	35.00	No
3	3	46	87.65	Male	42.93	Yes
4	4	36	89.80	Female	28.30	Yes
5	5	34	103.02	Male	40.56	No
6	6	39	77.37	Female	33.00	No
7	7	34	82.48	Male	43.48	No
8	8	37	75.94	Female	30.38	No
9	9	35	97.11	Male	40.17	Yes
10	10	32	78.42	Female	36.01	No
11	11	40	88.02	Male	44.22	Yes
12	12	55	74.47	Female	38.76	Yes
13	13	35	75.98	Female	33.09	No
14	14	46	58.97	Female	44.81	No
15	15	33	111.80	Male	31.94	No
16	16	39	79.81	Female	34.48	No
17	17	47	56.18	Male	47.23	Yes
18	18	40	86.13	Male	45.06	No

Your variables will be displayed in the columns based on the order you entered them into the **Variable View** window. Therefore, in our example, we first entered the case identifier, caseno, so this appears in the first column, entitled caseno, followed by the scores for age in the second column, entitled age, and so forth from left to right.

Now, you simply have to enter your data into the cells under each column. Remember that "each row" represents "one case" (e.g., a case could be a single participant). Therefore, in row 1 of our example, the first participant was a 37 years old male, weighing 70.47 kg, having a VO2max of 55.79 ml/min/kg, with no presence of heart disease.

Since these cells will initially be empty, you need to click into the cells to enter your data. You'll notice that when you click into the cells under your dichotomous dependent variable and any nominal independent variables, SPSS Statistics will give you a drop-down option with the groups/levels of the variables already populated (e.g., the cells under the gender column will include the two groups of our nominal independent variable, gender, namely "Male" and "Female"). However, for your continuous independent variables and case identifier, you simply need to enter the values.

Assumptions I

The assumptions of a binomial logistic regression will allow you to: (a) provide information on the accuracy of your predictions; (b) test how well the regression model fits your data; (c) determine the variation in your dependent variable explained by your independent variables; and (d) test hypotheses on your regression equation. If these assumptions are violated, you need to make corrections and re-test these assumptions. If they still do not pass, you must find alternative statistical tests.

The first four assumptions of a binomial logistic regression relate to your study design: (a) you have a dichotomous dependent variable; (b) you have one or more independent variables, which can be either continuous variables (i.e., an interval or ratio variable) or nominal variables; (c) there should be independence of observations; (d) the categories of the dichotomous dependent variable and all your nominal independent variables should be mutually exclusive and exhaustive; and (e) there should be a bare minimum of 15 cases per independent variable (although some recommend as high as 50 cases per independent variable). If your study design does not meet these four assumptions, a binomial logistic regression is the incorrect statistical test to use to analyse your data. However, there will be other tests you can use instead.

The other three assumptions relate to the nature of your data and can be tested using SPSS Statistics. Since it is not uncommon for the data you have collected to violate (i.e., fail) one or more of these three

assumptions, we show you different ways to proceed. This could include (a) making corrections to your data so that it no longer violates the assumptions, (b) using an alternative statistical test, or (c) proceeding with your analysis even when your data violates certain assumptions.

- **Assumption #5 There needs to be a linear relationship between the continuous independent variables and the logit transformation of the dependent variable.**

  The assumption of linearity in a binomial logistic regression requires that there is a linear relationship between the continuous independent variables, `age`, `weight` and `VO2max`, and the logit transformation of the dependent variable, `heart_disease`.

  There are a number of methods to test for a linear relationship between the continuous independent variables and the logit of the dependent variable. In this guide, we use the Box-Tidwell approach, which adds an interaction terms between the continuous independent variables and their natural logs to the regression equation. Therefore, we show you how to: (a) use the **Binary Logistic** procedure in SPSS Statistics to test this assumption; (b) interpret and report the results from this test; and (c) proceed with your analysis depending on whether you have met or violated this assumption.

- **Assumption #6 Your data must not show multicollinearity**

  Multicollinearity occurs when you have two or more independent variables that are highly correlated with each other. This leads to problems with understanding which independent variable contributes to the variance explained in the dependent variable, as well as technical issues in calculating a binomial logistic regression model. You can detect for multicollinearity through an inspection of **correlation coefficients** and **Tolerance/VIF values**, which will inform you whether your data meets or violates this assumption.

- **Assumption #7 There should be no significant outliers, high leverage points or highly influential points**

  Outliers, **leverage** and **influential points** are different terms used to represent observations in your data set that are in some way unusual when you wish to perform a binomial logistic regression analysis. These different classifications of **unusual points** reflect the different impact they have on the regression line. An observation can be classified as more than one type of unusual point. However, all these points can have a very negative effect on the regression equation that is used to predict the value of the dependent variable based on the independent variables. This can change

the output that SPSS Statistics produces and reduce the predictive accuracy of your results as well as the statistical significance. Fortunately, when using SPSS Statistics to run binomial logistic regression on your data, you can detect possible outliers, high leverage points and highly influential points.

Testing for linearity

The most important assumption in logistic regression is that the model is correctly specified, a component of which is the assumption of **linearity**, which is often expressed as "**linearity in the logit**" (e.g., Hilbe, 2016; Menard, 2002). It is similar to the linearity assumption in multiple regression, only with respect to the **log odds transformation (logit) of the dependent variable** rather than to the **dependent variable itself**.

The linearity assumption states that for every **one-unit increase** in a continuous independent variable, the **value** of the **log odds (logit) of the dependent variable increases by a constant amount**. For example, in our analysis, for every one-year increase in age, the log odds (logit) of heart disease should increase by a constant amount (e.g., 0.085). Similarly, for every one kilogram increase in weight, the log odds (logit) of heart disease should also increase by a constant amount (e.g., 0.006). These constants are the values of the slope coefficients and will be **different** for each continuous independent variable. This constant increase applies over the **full range of values of the continuous independent variable**. In other words, if the log odds (logit) of heart disease increases by 0.085 from 22 to 23 years of age, for example, then it will also increase by 0.085 from 67 to 68 years of age, or any other one-year difference.

One method that can be used to check the assumption of "linearity in the logit" is the **Box-Tidwell procedure** (Box & Tidwell, 1962), which was developed for linear regression, but is also appropriate for logistic regression models (Fox, 2016; Guerrero & Johnson, 1982). The procedure is simple to use and can be carried out in various statistical packages, including SPSS Statistics. It is one of several methods recommended to assess whether a continuous independent variable is linearly related to the logit of the dependent variable (e.g., Hosmer & Lemeshow, 1989; Menard 2002, 2010).

Although simple to use, two concerns have been raised with regard to the use of the Box-Tidwell procedure. First, it is **insensitive to small departures from linearity**; that is, it has **low power** to small departures from linearity (Hosmer & Lemeshow, 1989). In other words, if a relationship is not quite linear, but not by much, the Box-Tidwell procedure might not be able to detect this. Hosmer and Lemeshow (1989) considered this a disadvantage of the procedure. However, this lack of sensitivity to small departures from

linearity has also been put forward as an advantage of the procedure (Menard, 2000), usually because of a concern of **overfitting** the data/model (Menard, 2010; Osborne, 2015).

Note: Overfitting a model is when the model developed is based too closely on the **sample data**, resulting in the model fitting the sample data **well**, but modelling the **population poorly**. Overfitting is a **serious concern** in statistics/model building and should be **avoided at all costs**.

Second, the Box-Tidwell procedure does **not** inform you of the **type** (i.e. shape) of **nonlinearity** (e.g., cubic, exponential, etc.) (Hosmer & Lemeshow, 1989; Menard, 2002). However, it is simple to use the results of the Box-Tidwell procedure to estimate an appropriate transformation to correct for some types of nonlinearity (Menard, 2010; Osborne, 2015) and we will be adding a section to this guide shortly to explain how to do this.

Note: The Box-Tidwell procedure is a specific case of the more general procedure of **fractional polynomials** (Royston & Altman, 1994; Thompson, Xie & White, 2003). These and other advanced methods (e.g., **Generalized Additive Linear Models**, **partial residual plots**) can be used to determine the form/type of the relationship between a continuous independent variable and the logit of the dependent variable (e.g., Hilbe, 2009, 2016; Steyerberg, 2009). Indeed, Hosmer and Lemeshow's 1989 book, *Applied Logistic Regression*, recommended the Box-Tidwell procedure (and arguably made it popular), but later editions (2000; 2013) do not reference the Box-Tidwell procedure and recommend the more advanced methods. Unfortunately, none of these more advanced methods are easily carried out in SPSS Statistics. However, other statistical packages such as Stata or R can run some or all of these advanced methods.

Therefore, for a binomial logistic regression to be valid, the continuous independent variables need to be linearly related to the logit of the dependent variable, which can be tested in SPSS Statistics using the Box-Tidwell (1962) procedure. This requires two procedures in SPSS Statistics:

- The first part of the Box-Tidwell (1962) procedure requires that **all continuous independent variables are first transformed into their natural logs**. In our example, this means that we need to perform **natural log transformations** on our three continuous independent variables: age, weight and VO2max. This will generate three new variables – ln_age, ln_weight and ln_VO2max – which are the natural log transformations for age, weight and VO2max respectively. In the section, Procedure to create natural log transformations, we set out the **Compute Variable** procedure in SPSS Statistics that you need to use to create these three new natural log transformed variables.

> **Important:** You only need to test the assumption of linearity for continuous independent variables. You do not need to do this for categorical independent variables (i.e., nominal or ordinal independent variables).

- The second part of the Box-Tidwell (1962) procedure requires that you **create interaction terms for each of your continuous independent variables and their respective natural log transformed variables**. Since we have three continuous independent variables in our example, this means that we have to create three interaction terms: ln_age*age (i.e., the product of ln_age by age), ln_weight*weight (i.e., the product of ln_weight by weight) and ln_VO2max*VO2max (i.e., the product of ln_VO2max by VO2max). These three interaction terms – ln_age*age, ln_weight*weight and ln_VO2max*VO2max – then need to be entered into the binomial logistic regression procedure, together with the dichotomous dependent variable, heart_disease, all three continuous variables – age, weight and VO2max – and the categorical independent variable, gender, in order to run the Box-Tidwell (1962) procedure. Therefore, in the section, Box-Tidwell (1962) procedure to test for linearity, we show you how to use the **Binary Logistic** procedure in SPSS Statistics to produce the results for the Box-Tidwell (1962) test.

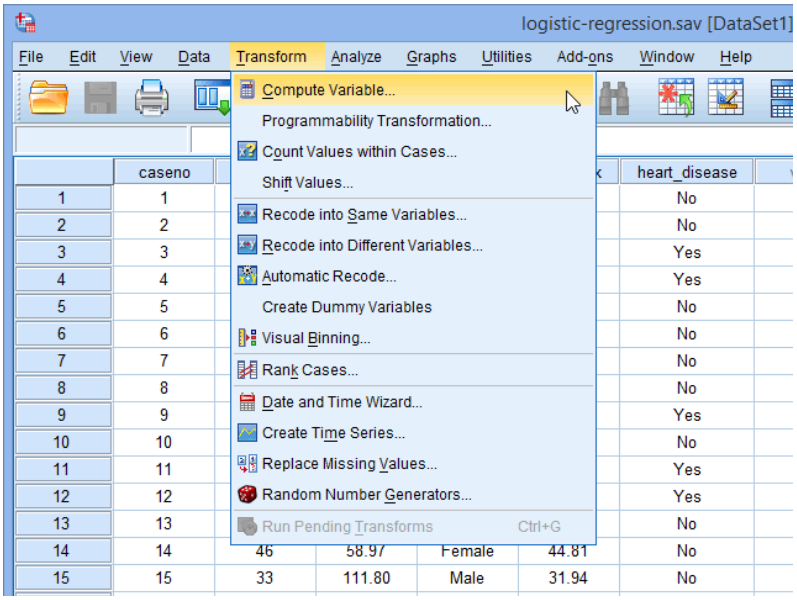
We show you how to determine if your continuous independent variables are linearly related to the logit of the dependent variable in the section: Interpreting the linearity assumption, which follows the two procedures above. If any of your continuous independent variables are not linearly related to the logit of the dependent variable, these variables have failed the assumption of linearity. However, there are ways to overcome this problem. If one or more of your continuous independent variables are not linearly related to the logit of the dependent variable, you can apply a transformation to the independent variable(s) that violated this assumption to determine if that corrects the problem

There are a few things to consider when doing this: (a) the transformations are applied to the original variable (e.g., age); (b) you only need to transform the continuous independent variable(s) that fail this assumption. You do not need to transform all of your continuous independent variables or any of your categorical independent variables (i.e., any nominal or ordinal variables) since this assumption does not apply to categorical independent variables; (c) if the transformation is successful, you then need to re-run the Box-Tidwell procedure – but now using the transformed version of the original variable that violated this assumption (i.e., the transformed version of age) (you will need to calculate a new natural log transformed variable and interaction term); before finally (d) interpreting the results to determine whether your continuous independent variable now meets the assumption of linearity. If this does not solve the problem, you might have to split the variable into ordinal categories and not consider the variable as continuous.

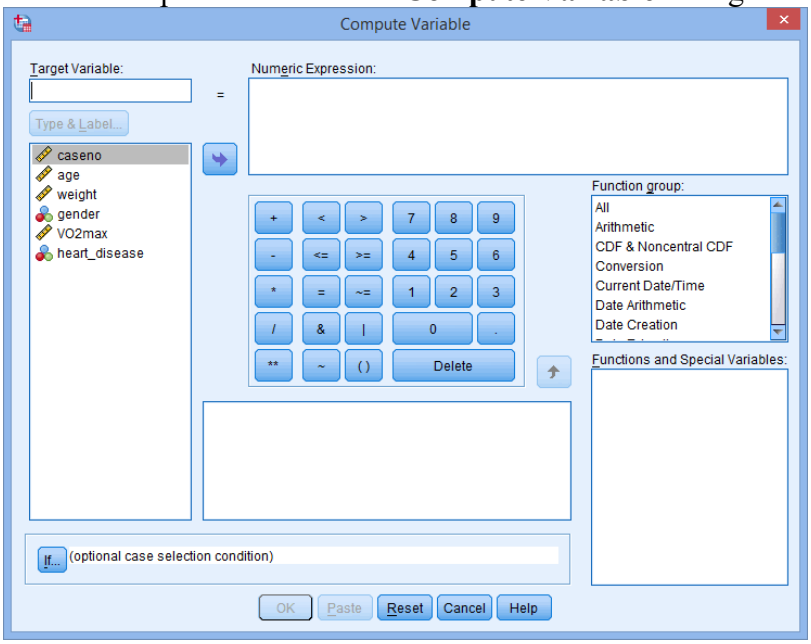
Procedure to create natural log transformations

The following **Compute Variable** procedure in SPSS Statistics sets out how to create natural log transformations for the three continuous independent variables in our example: age, weight and VO2max. Therefore, after completing this procedure, you will have created the following three natural log transformed variables: ln_age, ln_weight and ln_VO2max.

1 Click **Transform** > **Compute Variable**... on the main menu, as shown below:



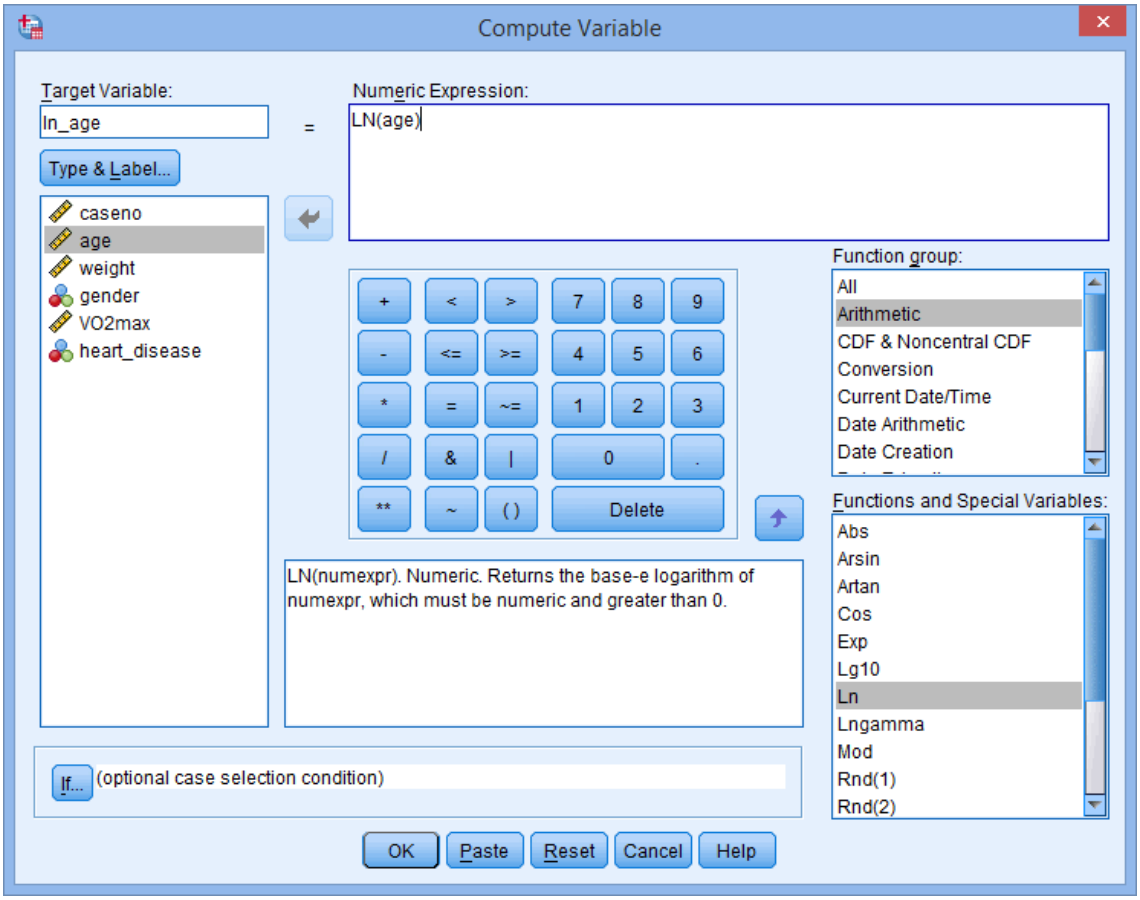
You will be presented with the **Compute Variable** dialogue box, as shown below:



2 To create a natural log transformation of your continuous independent variables, you have to repeat the following two steps for each of your continuous independent variables. We start by illustrating the process for the continuous independent variable, age.

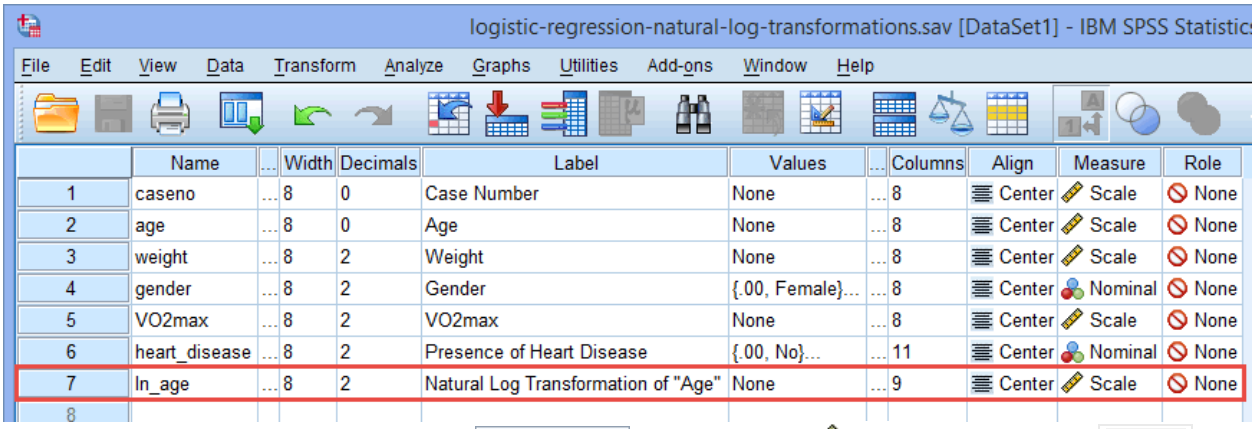
Enter a name for the natural log transformed variable you want to create into the Target Variable: box. For our continuous independent variable, age, we entered the name "**ln_age**" to highlight we are creating the natural log transformation (i.e., the "**ln**" part) of the variable, age (i.e., the "**age**" part.

Next, there are two methods to create the natural log transformation for our continuous independent variable, age. You can can either: (a) directly type in "**LN(age)**" into the Numeric Expression: box; or (b) select "Arithmetic" from the Function group: menu, followed by selecting "LN" from the Functions and special variables menu. Then, double-click on "LN" or use the button, either of which will transfer this function into the Numeric Expression: box. Finally, double-click on age, which will transfer this variable into the "LN()" function. Both methods will give you the same result, as shown below:



3 Click the OK button to compute the new variable, ln_age (i.e., the natural log transformation of age).

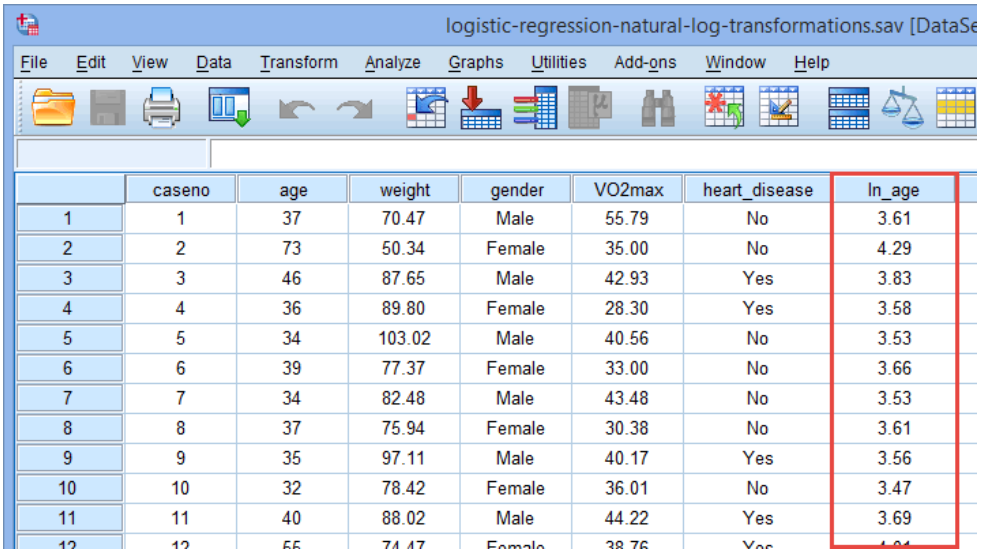
Note: When you create each of the natural log transformations of your independent variables, the new variables will appear in both the **Variable View** and **Data View** of SPSS Statistics, as highlighted for the **Variable View** below:



	Name	Width	Decimals	Label	Values	Columns	Align	Measure	Role
1	caseno	8	0	Case Number	None	8	Center	Scale	None
2	age	8	0	Age	None	8	Center	Scale	None
3	weight	8	2	Weight	None	8	Center	Scale	None
4	gender	8	2	Gender	{.00, Female}...	8	Center	Nominal	None
5	VO2max	8	2	VO2max	None	8	Center	Scale	None
6	heart_disease	8	2	Presence of Heart Disease	{.00, No}...	11	Center	Nominal	None
7	ln_age	8	2	Natural Log Transformation of "Age"	None	9	Center	Scale	None
8									

You will notice that the cell under the Measure column reads Scale, indicating that ln_age, the natural log transformation of age, is also a continuous variable. We have entered the label, "**Natural Log Transformation of "Age"**" into the Label column for clarity (SPSS Statistics does not do this for you), as well as changing the cell under the Role column from Input to None.

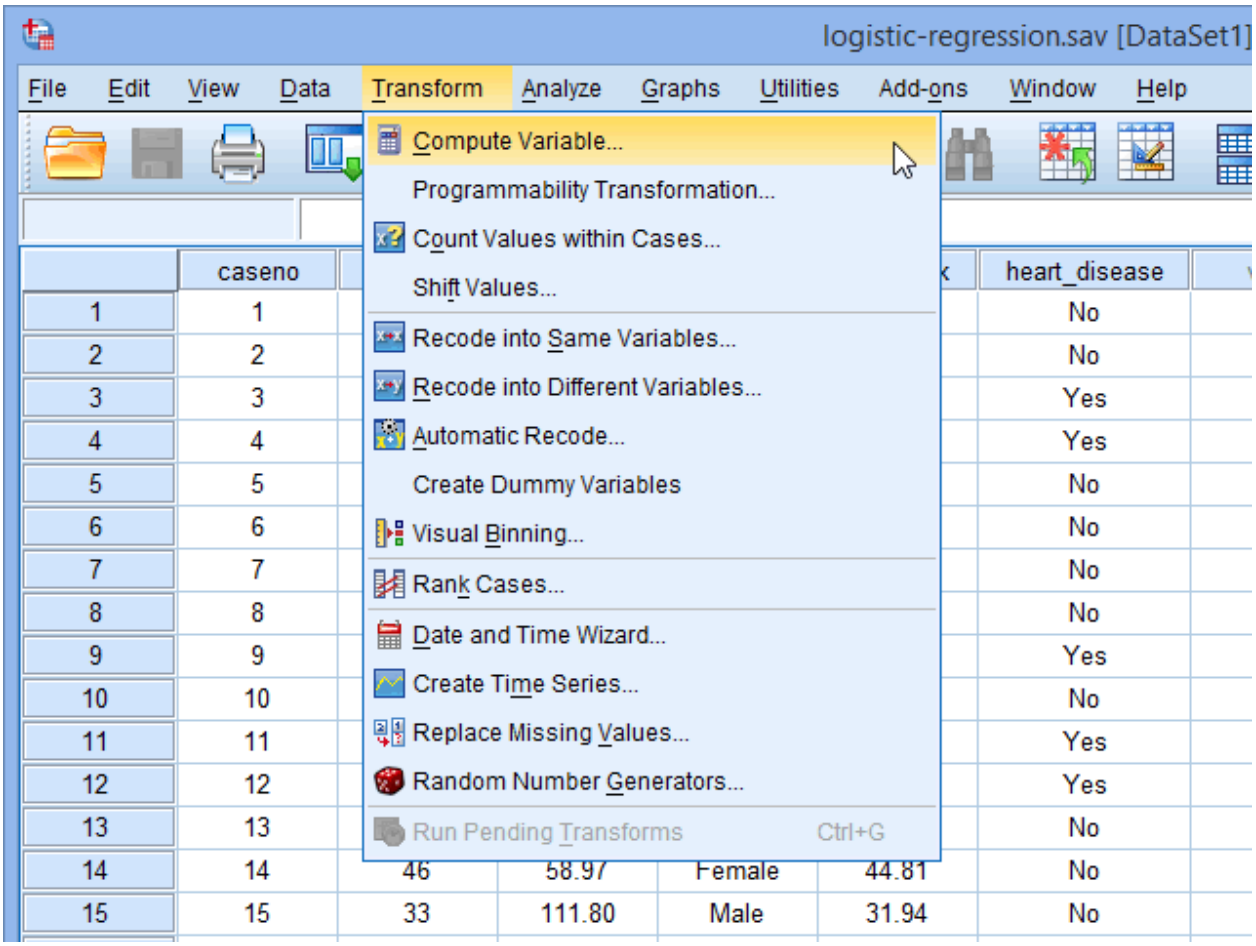
The values for the new natural log transformed variable, ln_age, will also now appear in the **Data View** of SPSS Statistics, as highlighted below under the ln_age column:



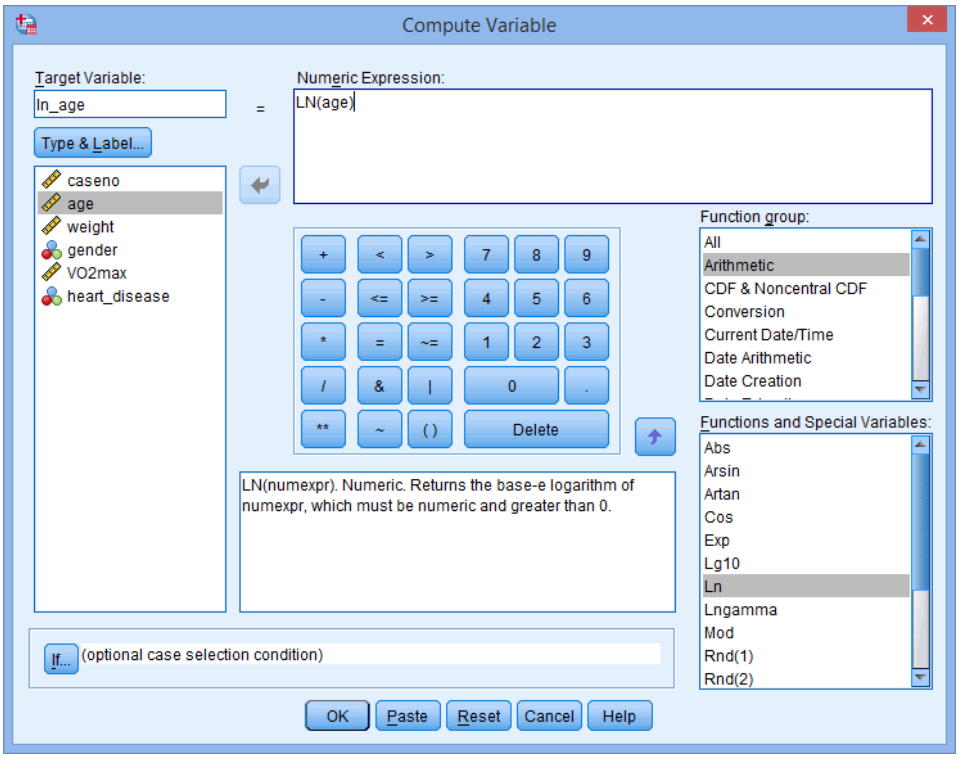
	caseno	age	weight	gender	VO2max	heart_disease	ln_age
1	1	37	70.47	Male	55.79	No	3.61
2	2	73	50.34	Female	35.00	No	4.29
3	3	46	87.65	Male	42.93	Yes	3.83
4	4	36	89.80	Female	28.30	Yes	3.58
5	5	34	103.02	Male	40.56	No	3.53
6	6	39	77.37	Female	33.00	No	3.66
7	7	34	82.48	Male	43.48	No	3.53
8	8	37	75.94	Female	30.38	No	3.61
9	9	35	97.11	Male	40.17	Yes	3.56
10	10	32	78.42	Female	36.01	No	3.47
11	11	40	88.02	Male	44.22	Yes	3.69
12	12	55	74.47	Female	38.76	Yes	4.01

4 You now have to repeated the process above for all your other continuous independent variables. If you feel confident how to do this, you can skip to the next section, Box-Tidwell (1962) procedure to test for linearity, which shows you how to test for the assumption of linearity using the new natural log transformed variables you have just created. However, if you are following this guide step-by-step, we illustrate the process for creating natural log transformations of all our continuous independent variables in the steps that follow.

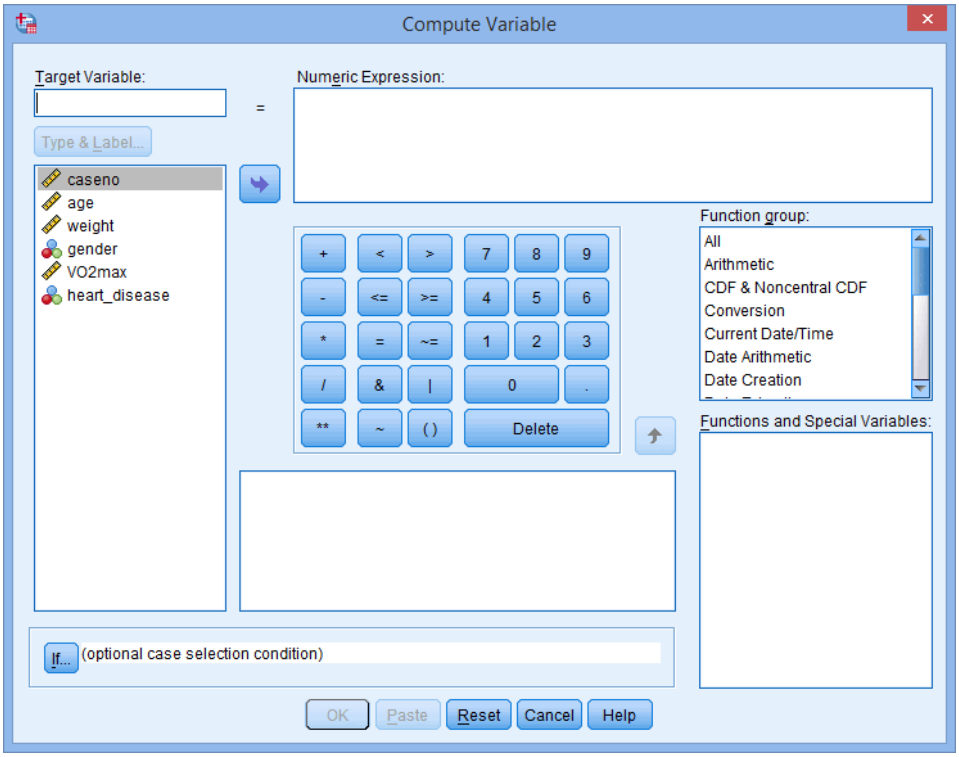
- Click **Transform > Compute Variable...** on the main menu, as shown below:



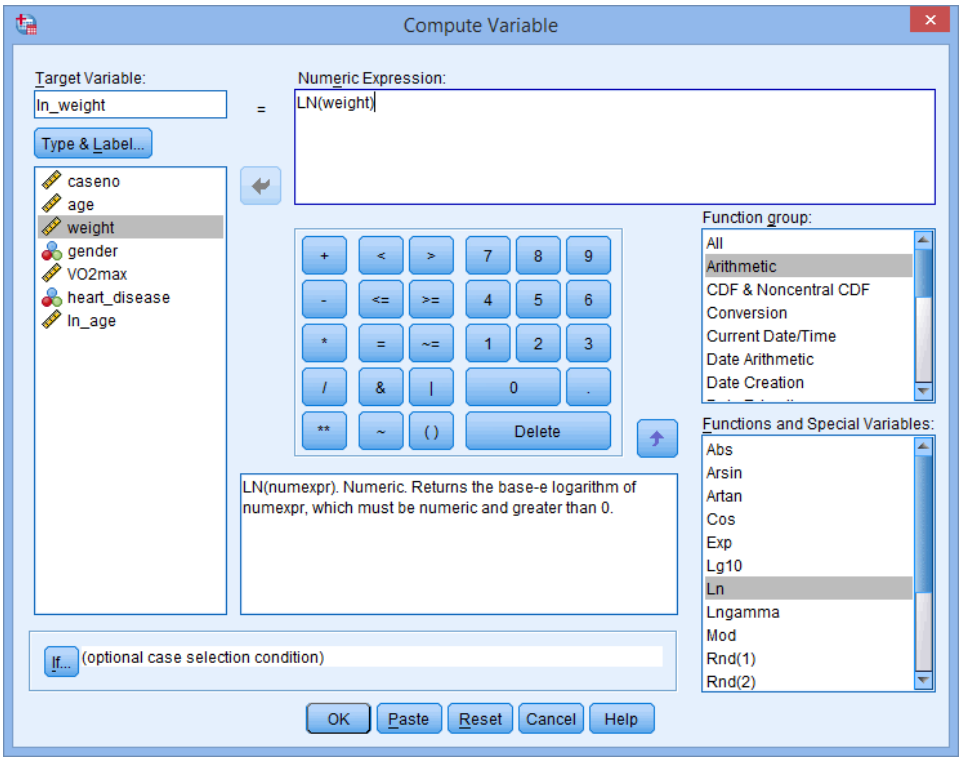
You will be presented with the **Compute Variable** dialogue box, but rather than being blank as in the example above, the fields will be populated with the options you selected when creating the ln_age variable in the previous three steps, as shown below:



5 Click the Reset button to clear all the fields in the **Compute Variable** dialogue box. You will be presented with the following screen:



6 Type the name "**ln_weight**" into the Target Variable: box. Next, type "**LN(weight)**" into the Numeric Expression: box, as shown below:

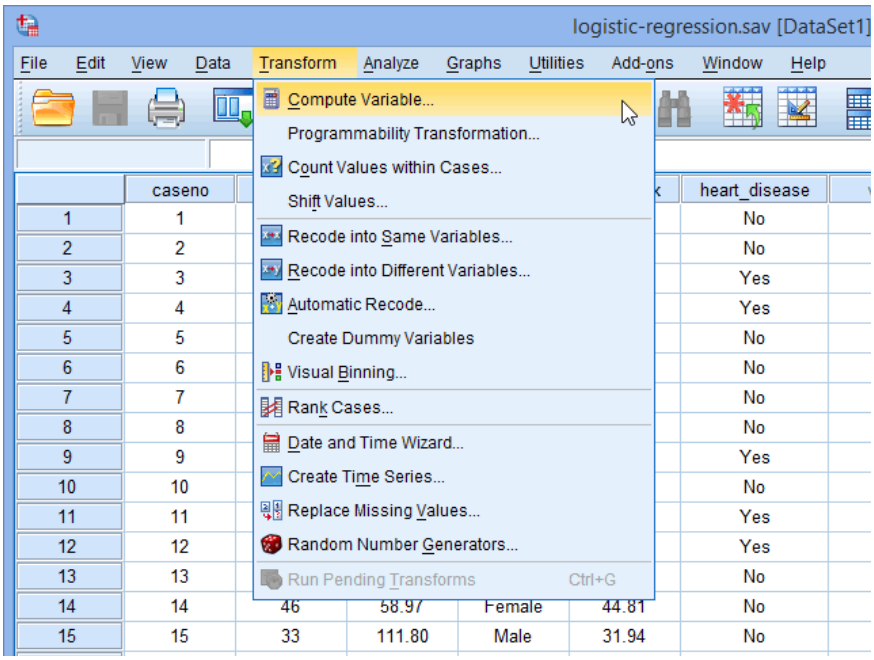


Note: Just to reiterate, the name we entered into the Target Variable: box, "**ln_weight**", highlights that we are creating the natural log transformation (i.e., the "**ln**" part) of the variable, weight (i.e., the "**weight**" part). Also, rather than just typing in "**LN(weight)**" into the Numeric Expression: box, you could have chosen the other method to do this, as we explained earlier; that is, you could have selected "Arithmetic" from the Function group: menu, and then selected "LN" from the Functions and special variables menu. Next, you would have double-clicked on "LN" or used the button, either of which would have transferred this function into the Numeric Expression: box. Finally, you would have double-clicked on weight, which would have transferred this variable into the "LN()" function. Both methods would have given you the same result.

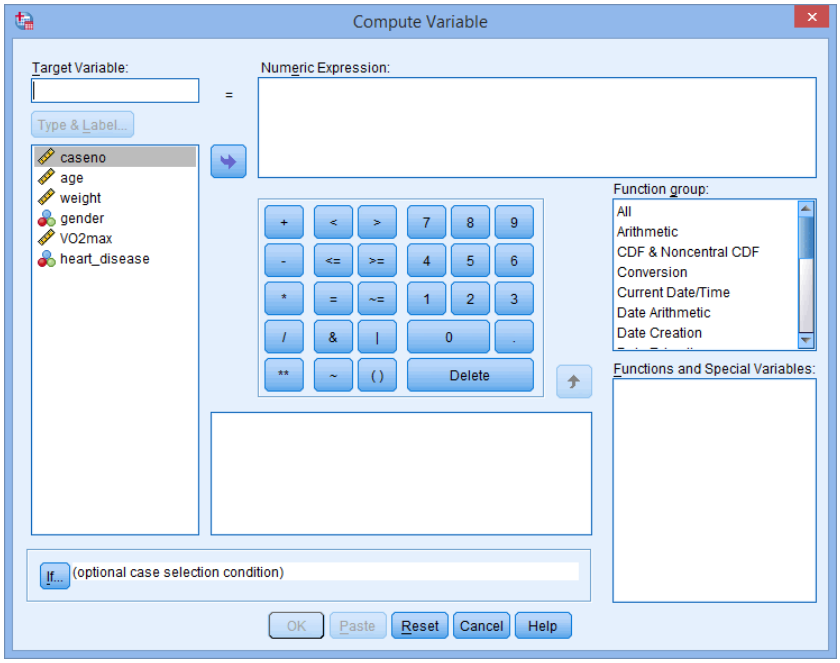
7 Click the OK button to compute the new variable, ln_weight (i.e., the natural log transformation of weight).

Note: The new variable, ln_weight, will appear in both the **Variable View** and **Data View** of SPSS Statistics, as explained above.

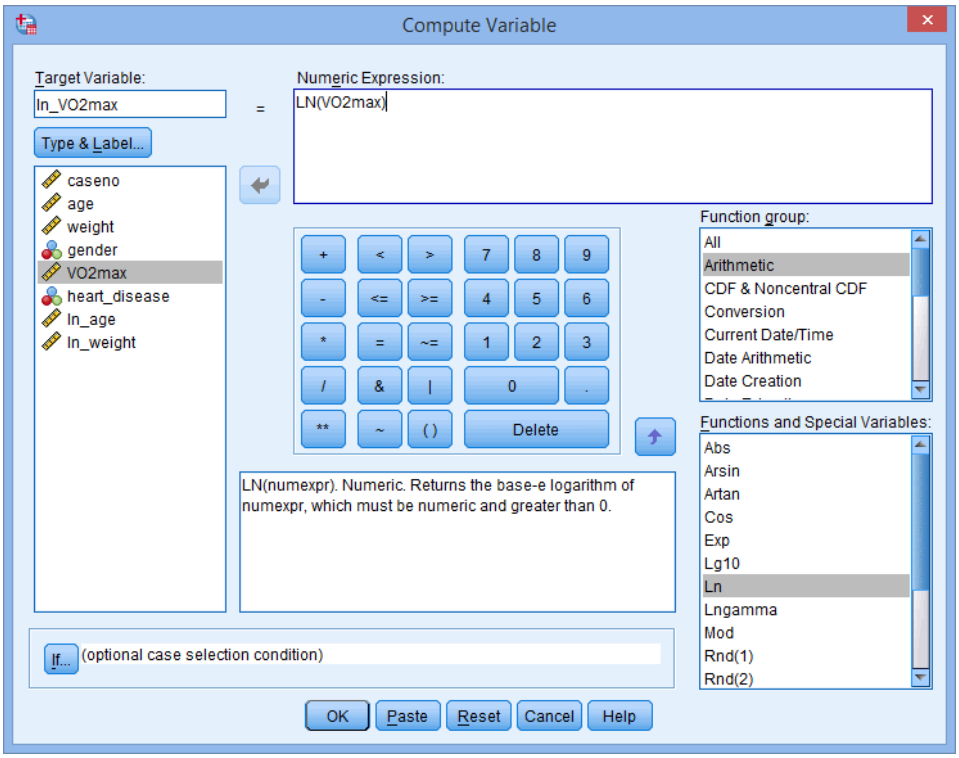
8 To create the natural log transformation for our third and last continuous independent variable, VO2max, click **Transform > Compute Variable...** on the main menu, as shown below:



You will be presented with the **Compute Variable** dialogue box. Again, the fields will be populated with the options you selected when creating the ln_weight variable in the previous three steps, so click the Reset button to clear all the fields, which will give you the following blank **Compute Variable** dialogue box:



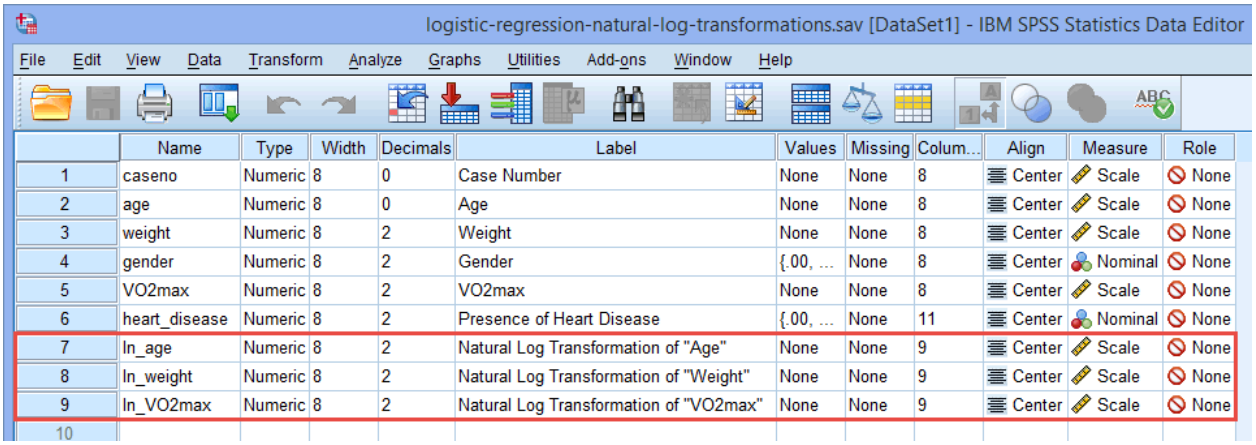
9 Type the name "**ln_VO2max**" into the Target Variable: box. Next, type "**LN(VO2max)**" into the Numeric Expression: box, as shown below:



Note: Just to explain one more time, the name we entered into the Target Variable: box, "**ln_VO2max**", highlights that we are creating the natural log transformation (i.e., the "**ln**" part) of the variable, VO2max (i.e., the "**VO2max**" part). Also, rather than just typing in "**LN(VO2max)**" into the Numeric Expression: box, you could have chosen the other method to do this, as we explained earlier; that is, you could have selected "Arithmetic" from the Function group: menu, and then selected "LN" from the Functions and special variables menu. Next, you would have double-clicked on "LN" or used the button, either of which would have transferred this function into the Numeric Expression: box. Finally, you would have double-clicked on VO2max, which would have transferred this variable into the "LN()" function. Both methods would have given you the same result.

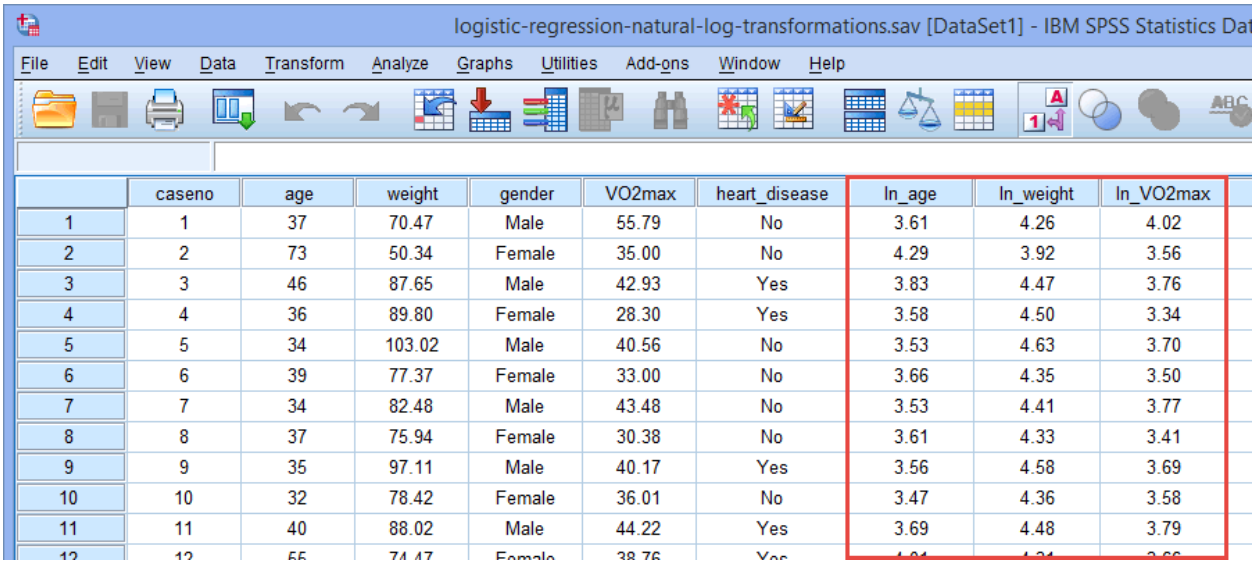
10 Click the OK button to compute the new variable, ln_VO2max (i.e., the natural log transformation of VO2max).

You have now created the natural log transformations for all three continuous independent variables in our example – ln_age, ln_weight and ln_VO2max – as highlighted in the **Variable View** of SPSS Statistics below:



	Name	Type	Width	Decimals	Label	Values	Missing	Colum...	Align	Measure	Role
1	caseno	Numeric	8	0	Case Number	None	None	8	Center	Scale	None
2	age	Numeric	8	0	Age	None	None	8	Center	Scale	None
3	weight	Numeric	8	2	Weight	None	None	8	Center	Scale	None
4	gender	Numeric	8	2	Gender	{.00, ...	None	8	Center	Nominal	None
5	VO2max	Numeric	8	2	VO2max	None	None	8	Center	Scale	None
6	heart_disease	Numeric	8	2	Presence of Heart Disease	{.00, ...	None	11	Center	Nominal	None
7	ln_age	Numeric	8	2	Natural Log Transformation of "Age"	None	None	9	Center	Scale	None
8	ln_weight	Numeric	8	2	Natural Log Transformation of "Weight"	None	None	9	Center	Scale	None
9	ln_VO2max	Numeric	8	2	Natural Log Transformation of "VO2max"	None	None	9	Center	Scale	None
10											

The values for these three natural log transformed variables will also now appear in the **Data View** of SPSS Statistics, as highlighted in the ln_age, ln_weight and ln_VO2max columns below:



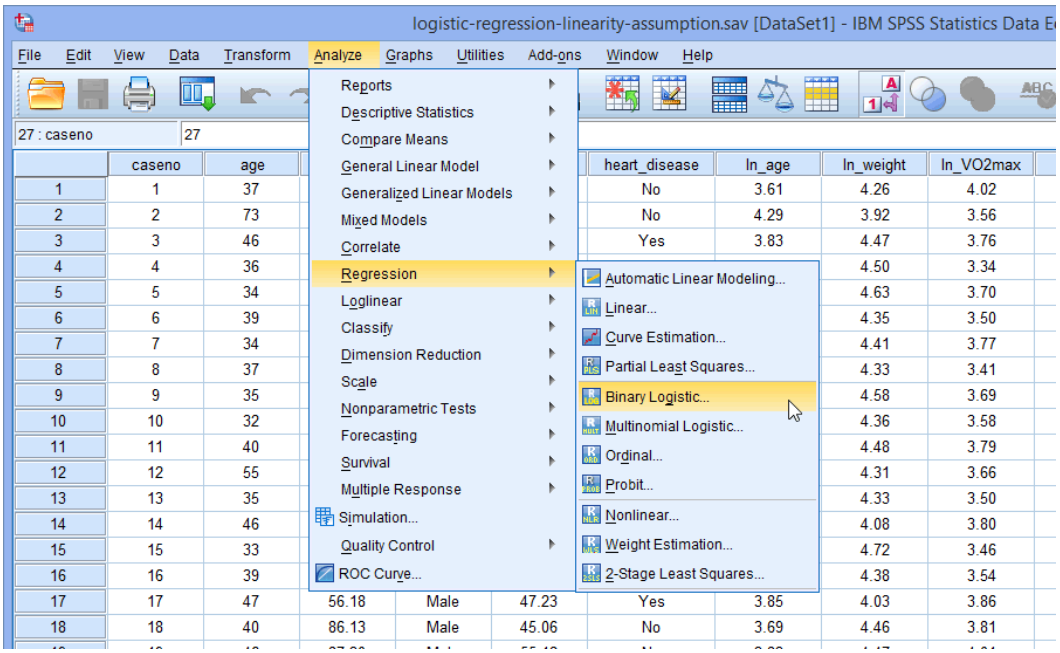
	caseno	age	weight	gender	VO2max	heart_disease	ln_age	ln_weight	ln_VO2max
1	1	37	70.47	Male	55.79	No	3.61	4.26	4.02
2	2	73	50.34	Female	35.00	No	4.29	3.92	3.56
3	3	46	87.65	Male	42.93	Yes	3.83	4.47	3.76
4	4	36	89.80	Female	28.30	Yes	3.58	4.50	3.34
5	5	34	103.02	Male	40.56	No	3.53	4.63	3.70
6	6	39	77.37	Female	33.00	No	3.66	4.35	3.50
7	7	34	82.48	Male	43.48	No	3.53	4.41	3.77
8	8	37	75.94	Female	30.38	No	3.61	4.33	3.41
9	9	35	97.11	Male	40.17	Yes	3.56	4.58	3.69
10	10	32	78.42	Female	36.01	No	3.47	4.36	3.58
11	11	40	88.02	Male	44.22	Yes	3.69	4.48	3.79
12	12	55	74.47	Female	38.76	Yes			

Now that you have created the ln_age, ln_weight and ln_VO2max, you can carry out the **Binary Logistic** procedure in SPSS Statistics to test for the assumption of linearity, which is set out in the next section.

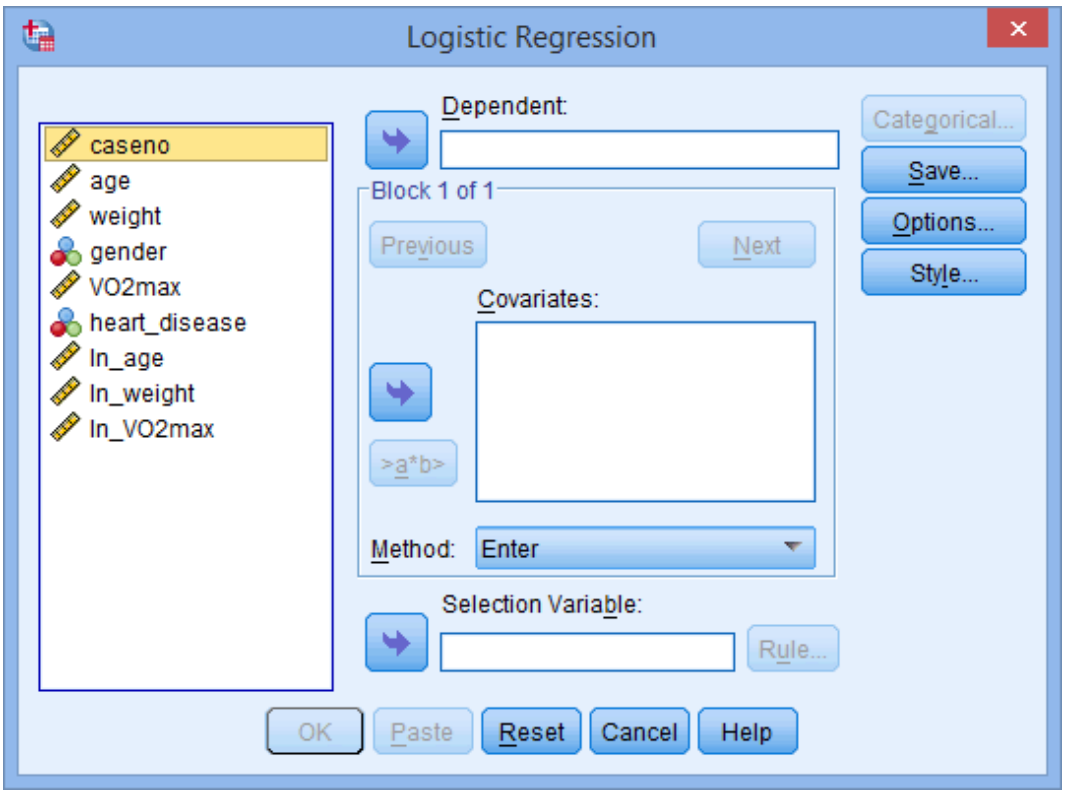
Box-Tidwell (1962) procedure to test for linearity

The following **Binary Logistic** procedure in SPSS Statistics sets out how to test whether the continuous independent variables in our example are linearly related to the logit of the dependent variable (i.e., the assumption of linearity):

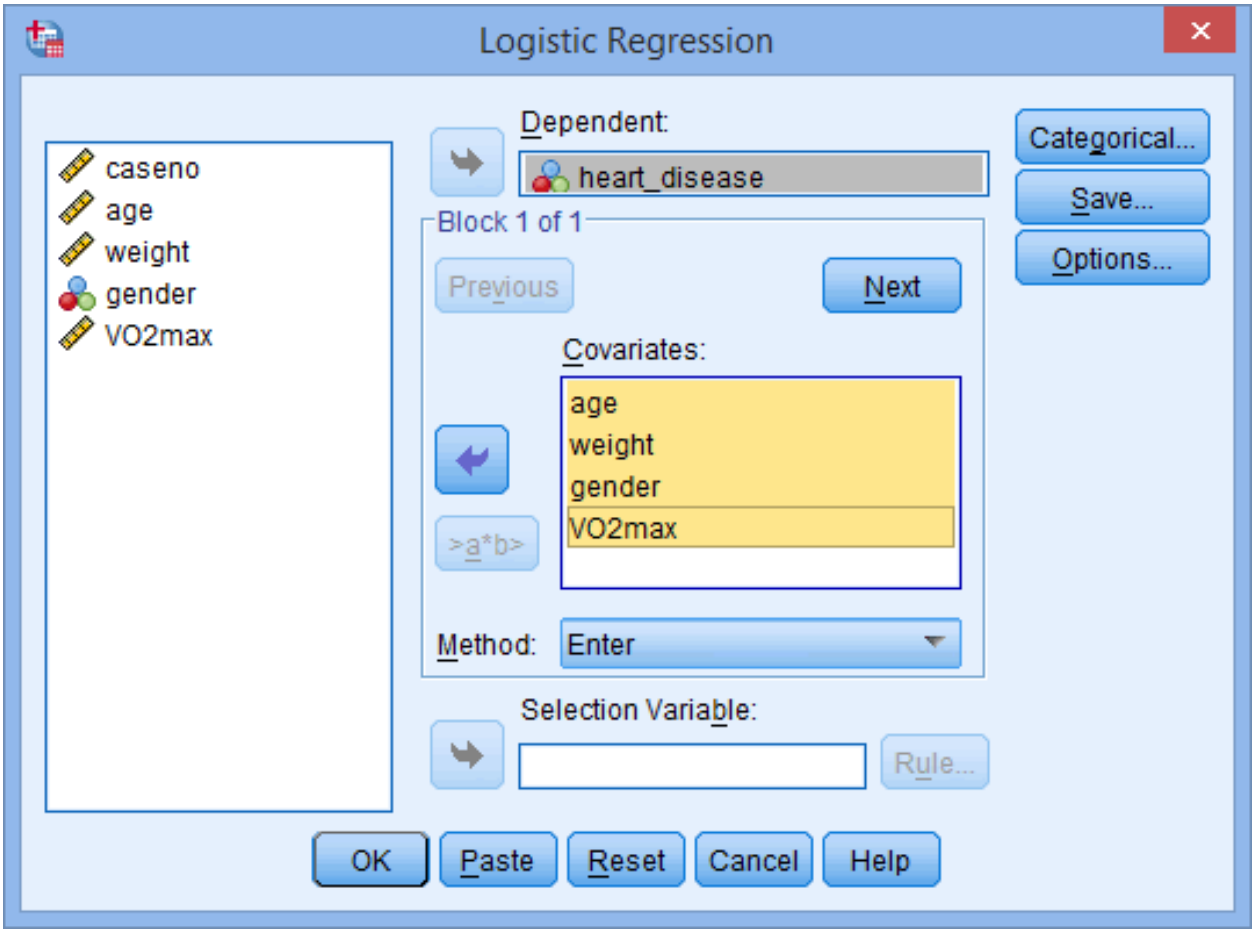
1 Click **Analyze > Regression > Binary Logistic...** on the main menu, as shown below:



You will be presented with the **Logistic Regression** dialogue box, as shown below:

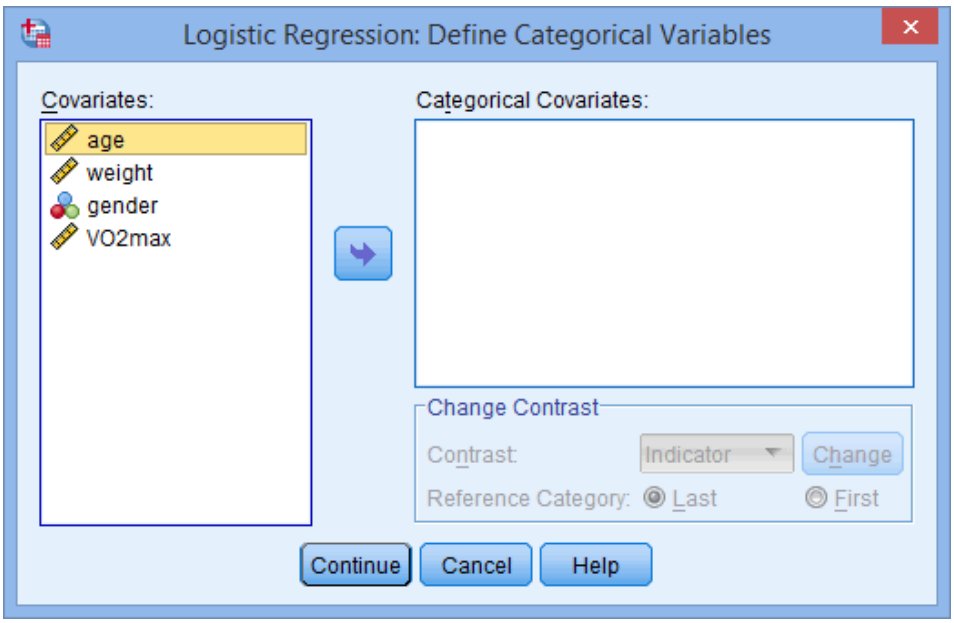


2 Transfer the dependent variable, heart_disease, into the Dependent: box, and the independent variables, age, weight, gender and VO2max into the Covariates: box, using the buttons, as shown below:



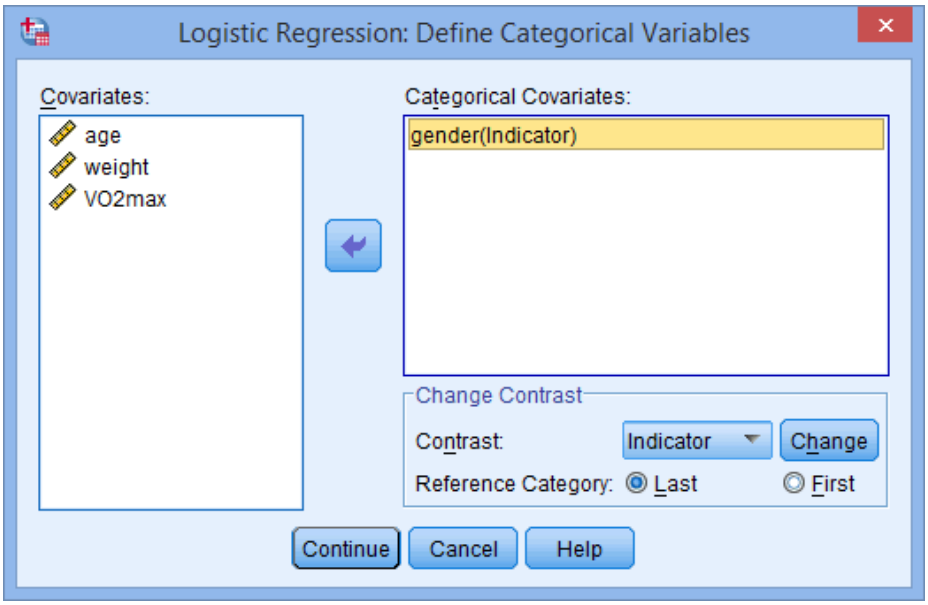
Note: For a standard logistic regression you should ignore Previous and Next buttons as they are for sequential (hierarchical) logistic regression. The Method: option needs to be kept at the default value, which is "Enter". If, for whatever reason, "Enter" is not selected, you need to change Method: back to "Enter". The "Enter" method is the name given by SPSS Statistics to standard regression analysis.

3 Click the Categorical... button. You will be presented with the **Logistic Regression: Define Categorical Variables** dialogue box, as shown below:

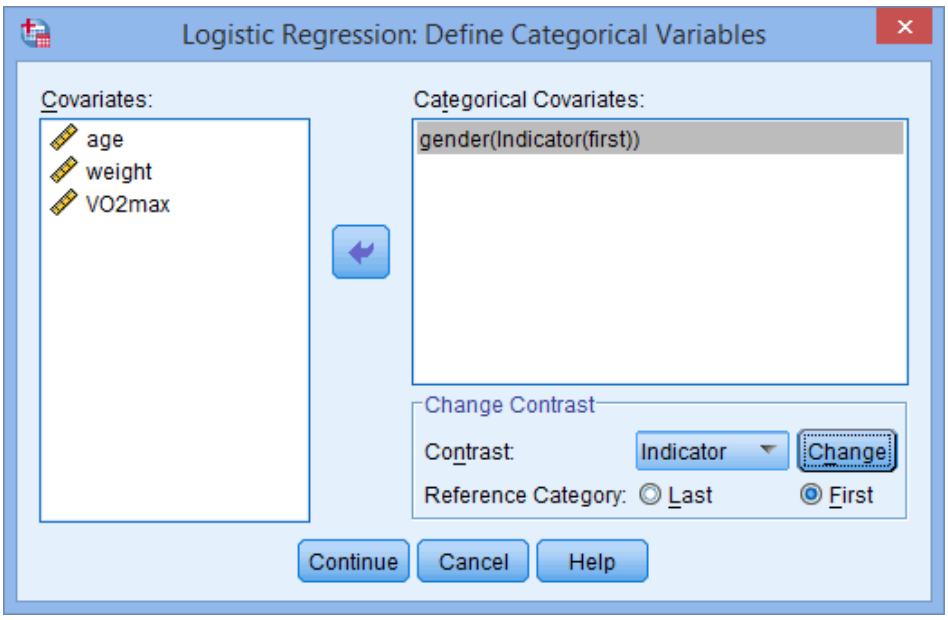


Note: SPSS Statistics requires you to define all the categorical predictor values in the logistic regression model. It does not do this automatically.

4 Transfer the categorical independent variable, gender, from the Covariates: box to the Categorical Covariates: box, as shown below:

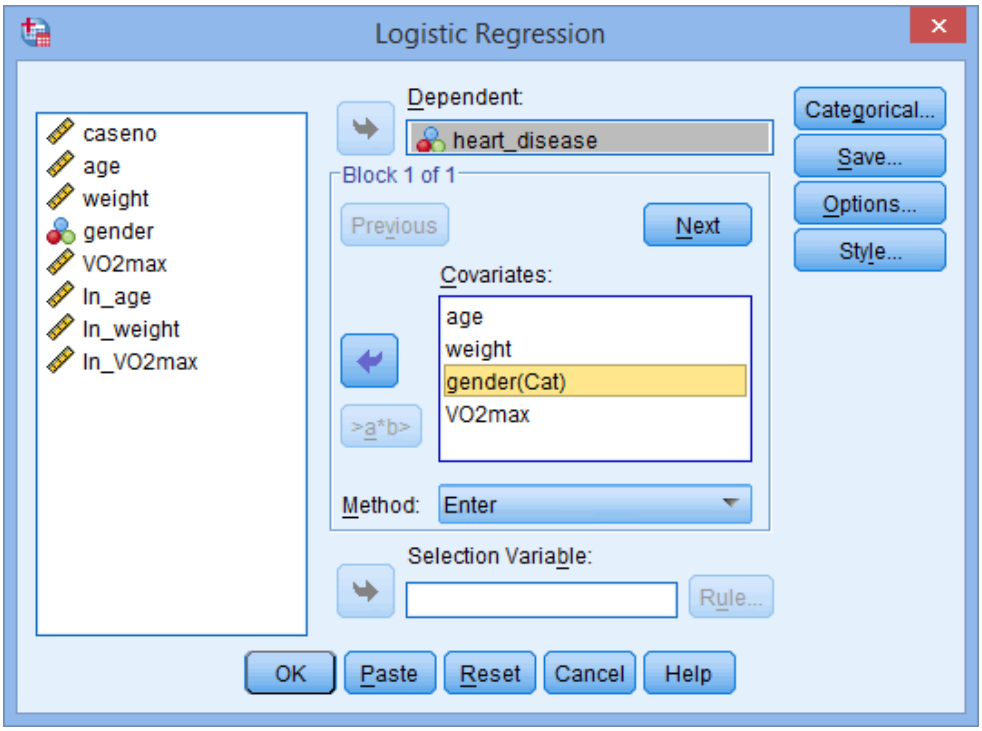


5 In the –Change Contrast– area, change the Reference Category: from the Last option to the First option. Then, click the Change button, as shown below:



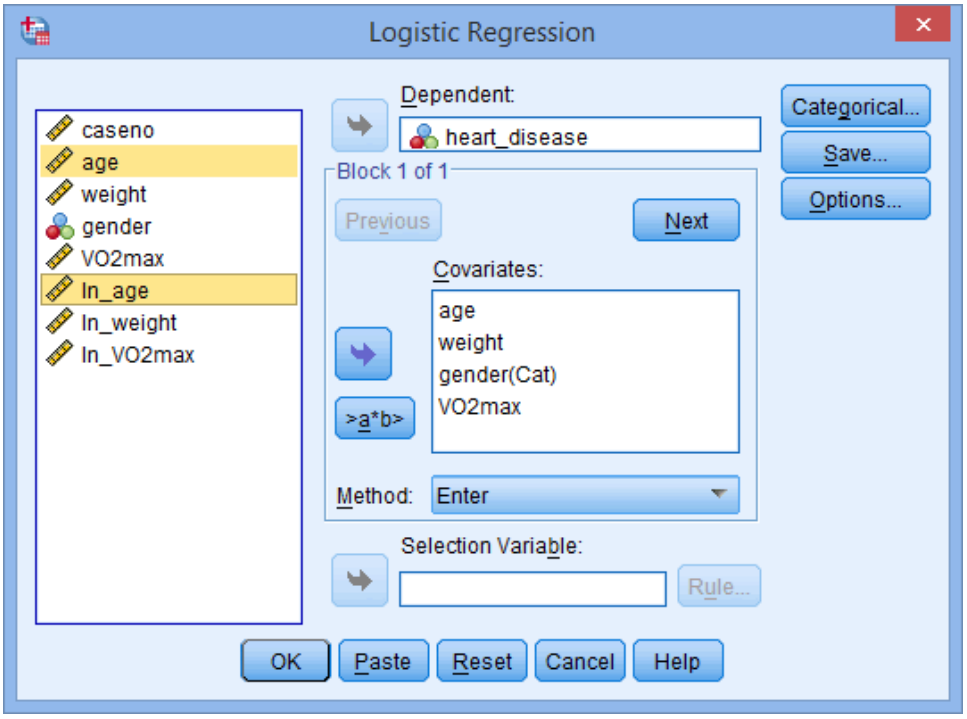
Note: Whether you choose Last or First will depend on how you set up your data. In this example, males are to be compared to females, with females acting as the reference category (who were coded "0"). Therefore, First is chosen.

6 Click the Continue button. You will be returned to the **Logistic Regression** dialogue box where the categorical independent variable, gender, will now be labelled "**gender(Cat)**", as shown below:

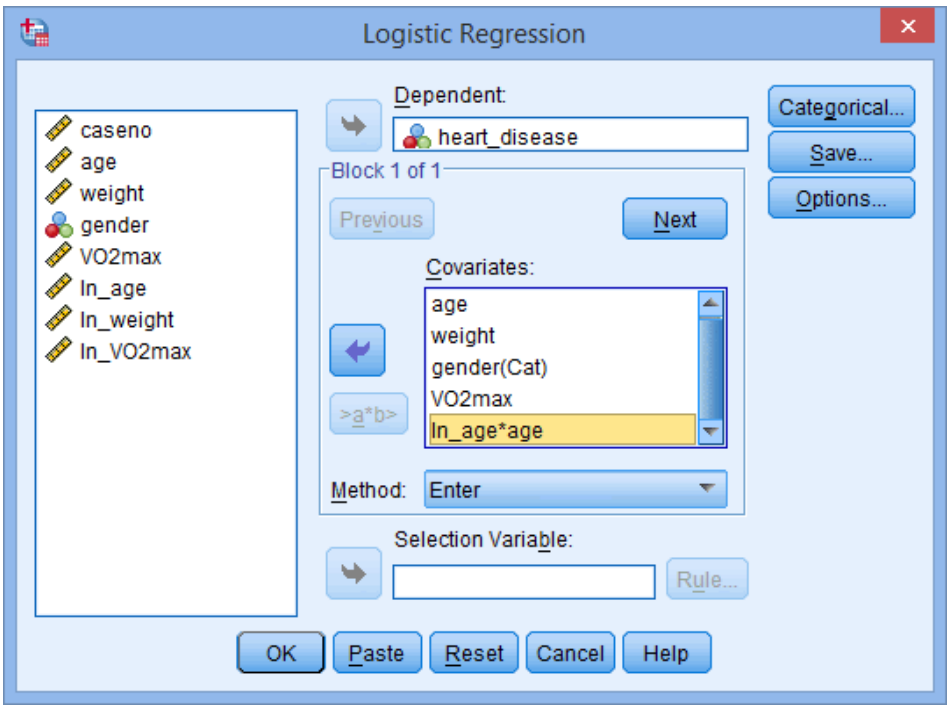


Note: You will know that your categorical independent variable has been correctly defined because the label of the variable will now include "**(Cat)**" at the end. Therefore, in our example, "**(Cat)**" was added to the end of "**gender**" to create "**gender(Cat)**".

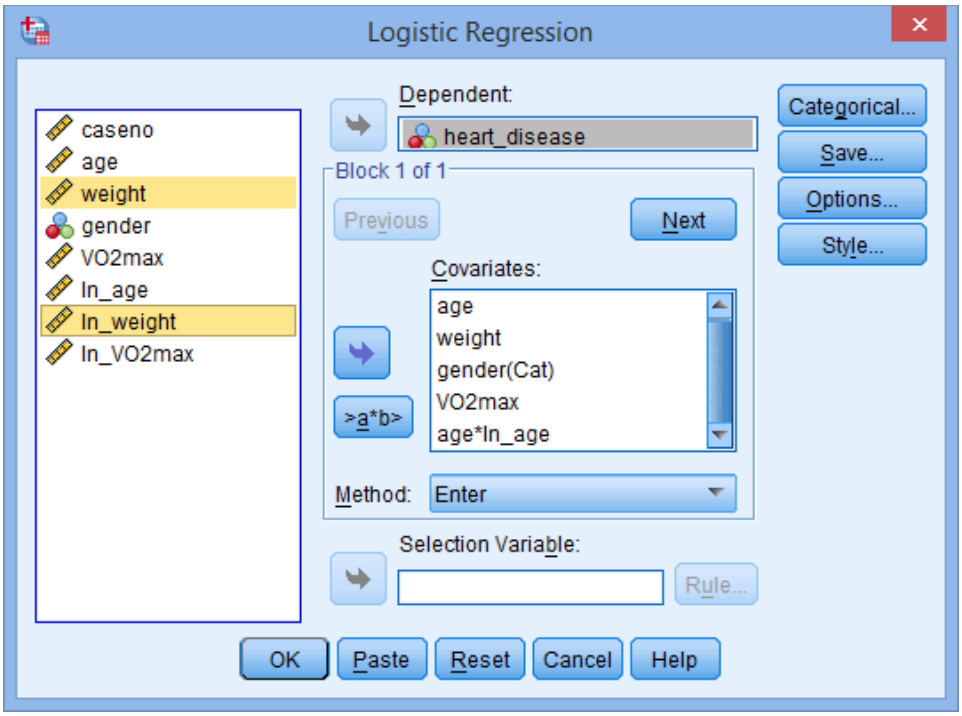
7 Now that you have set up the main elements required to carry out a binomial logistic regression, you can start to create the three interaction terms: ln_age*age, ln_weight*weight and ln_VO2max*VO2max. We start by illustrating the process for the interaction term ln_age*age. Therefore, highlight the first continuous independent variable, age, and its natural log transformation, ln_age. This will activate the >a*b> button (i.e., it changes from >a*b> to >a*b>), as shown below:



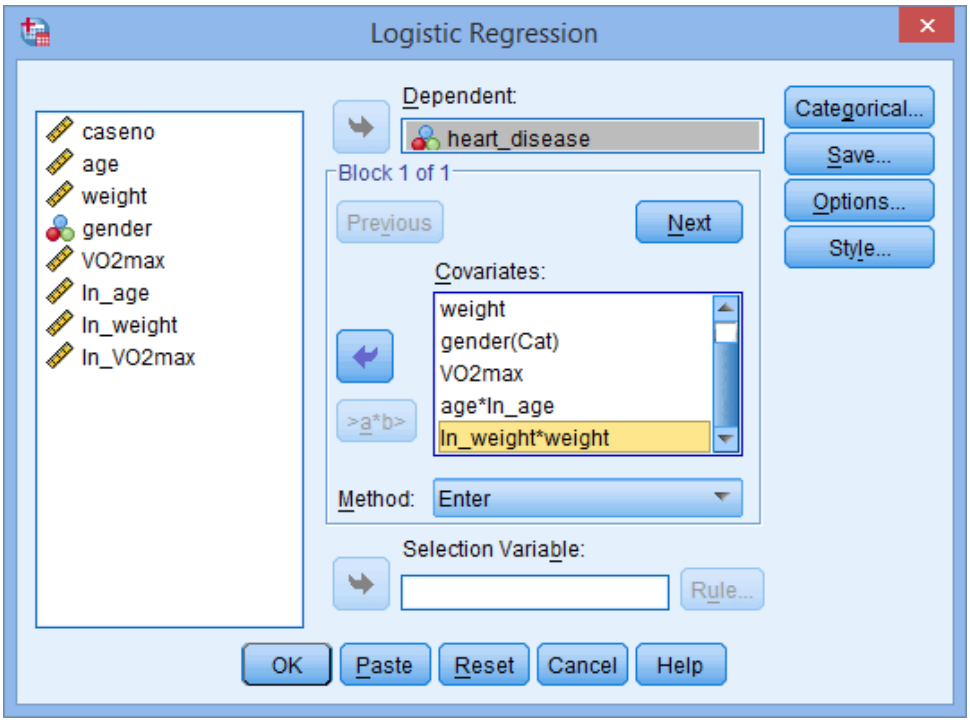
8 Click the >a*b> button. This will add the interaction term, "**ln_age*age**" (i.e., ln_age*age), into the Covariates: box, as shown below:



9 Highlight the second continuous independent variable, weight, and its natural log transformation, ln_weight. This will activate the >a*b> button (i.e., it changes from >a*b> to >a*b>), as shown below:

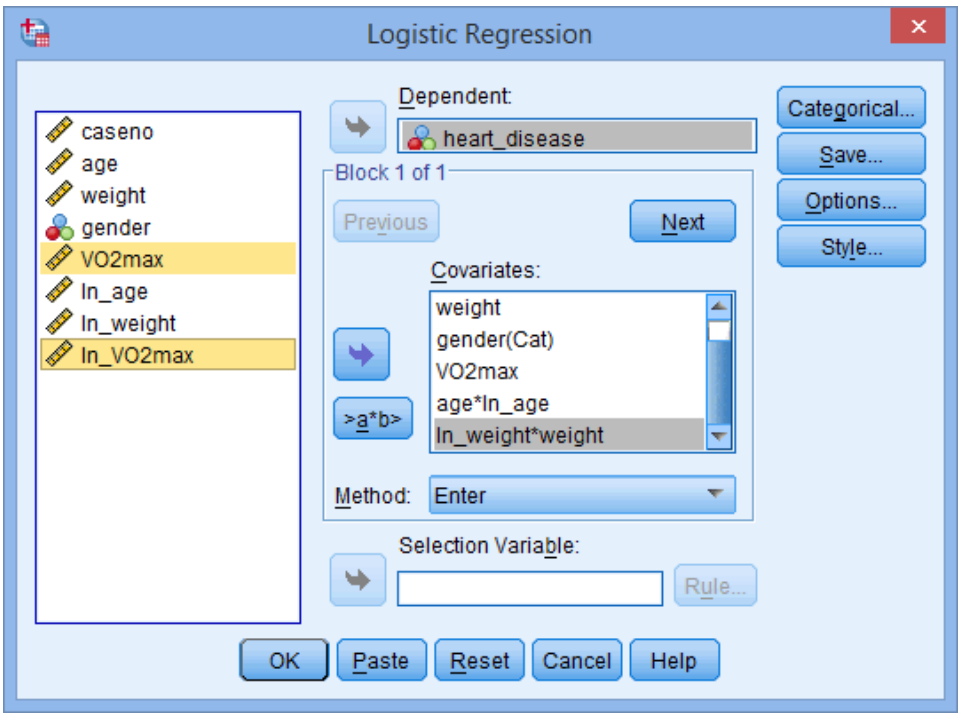


10 Click the >a*b> button. This will add the interaction term, "**ln_weight*weight**" (i.e., ln_weight*weight), into the Covariates: box, as shown below:

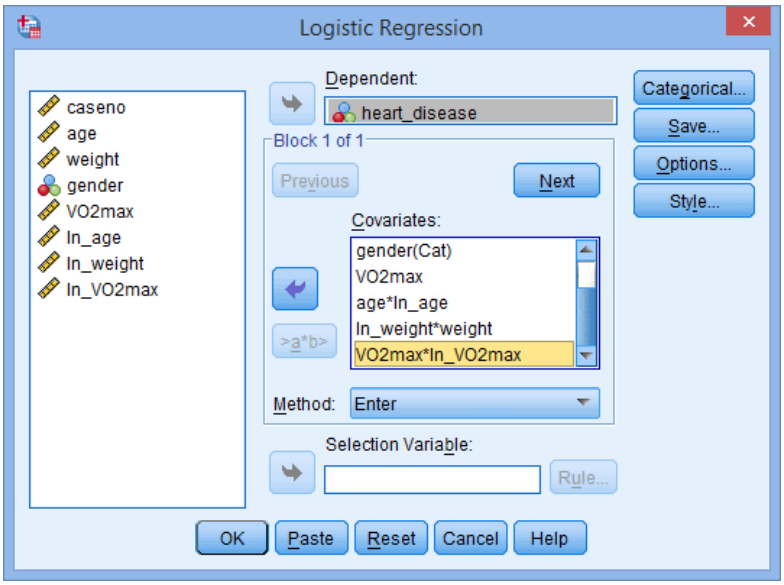


Note: You will have to scroll down the variables listed in the Covariates: box in order to see the "**ln_weight*weight**" interaction term. This will be the case whenever you have more than five variables in the Covariates: box, but you can resize the **Logistic Regression** dialogue box to bring more of your variables into view.

11 Highlight the third and final continuous independent variable, VO2max , and its natural log transformation, ln_VO2max . This will activate the >a*b> button (i.e., it changes from >a*b> to >a*b>), as shown below:



12 Click the >a*b> button. This will add the interaction term, "**ln_VO2max*VO2max**" (i.e., ln_VO2max*VO2max), into the Covariates: box, as shown below:



13 Click the OK button.

The steps above will have run the Box-Tidwell (1962) procedure to determine whether the continuous independent variables are linearly related to the logit of the dependent variable.

Interpreting the linearity assumption

Since you had to run the main elements of the binomial logistic regression procedure when creating interaction terms in the previous section, a lot of SPSS Statistics output that was generated is not required at this stage. In fact, you only need to consult the **Variables in the Equation** table, as shown below:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age	2.725	1.103	6.100	1	.014	15.258	1.755	132.650
	weight	.144	.783	.034	1	.854	1.155	.249	5.360
	gender(1)	1.854	.922	4.047	1	.044	6.384	1.049	38.857
	VO2max	1.320	1.823	.524	1	.469	3.745	.105	133.452
	age by ln_age	-.543	.227	5.754	1	.016	.581	.373	.905
	ln_weight by weight	-.027	.146	.033	1	.855	.974	.731	1.297
	VO2max by ln_VO2max	-.301	.382	.620	1	.431	.740	.350	1.566
	Constant	-40.585	21.707	3.496	1	.062	.000		

a. Variable(s) entered on step 1: age, weight, gender, VO2max, age * ln_age , ln_weight * weight , VO2max * ln_VO2max .

Specifically, you need to look for the rows that contain the interaction terms (i.e., the "**age by ln_age**", "**ln_weight by weight**" and "**VO2max by ln_VO2max**" rows) and then examine the values in the "**Sig.**" column for these rows. In our example there are three interaction terms and the values in the "**Sig.**" column for these interaction terms are highlighted below:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age	2.725	1.103	6.100	1	.014	15.258	1.755	132.650
	weight	.144	.783	.034	1	.854	1.155	.249	5.360
	gender(1)	1.854	.922	4.047	1	.044	6.384	1.049	38.857
	VO2max	1.320	1.823	.524	1	.469	3.745	.105	133.452
	age by ln_age	-.543	.227	5.754	1	.016	.581	.373	.905
	ln_weight by weight	-.027	.146	.033	1	.855	.974	.731	1.297
	VO2max by ln_VO2max	-.301	.382	.620	1	.431	.740	.350	1.566
	Constant	-40.585	21.707	3.496	1	.062	.000		

a. Variable(s) entered on step 1: age, weight, gender, VO2max, age * ln_age , ln_weight * weight , VO2max * ln_VO2max .

If the interaction term is statistically significant, the original continuous independent variable is not linearly related to the logit of the dependent variable (i.e., it has failed the assumption of linearity). Now although it is common practice to not correct for multiple comparisons when interpreting terms in regression, it has been recommended as sensible to apply a Bonferroni correction based on all terms (including the intercept) in the model when assessing this linearity assumption (Tabachnick & Fidell, 2014).

In our example, there are 8 terms in this model. These are highlighted in the table above and include the three continuous independent variables – age, weight and VO2max – the categorical independent variable, gender, the three interaction terms – age*ln_age, ln_weight*weight and VO2max*ln_VO2max – and the intercept (called the "**Constant**" in the SPSS Statistics, as highlighted in the final row of the table above).

Since there are 8 terms in this model, we divide the *p*-value at which statistical significance is accepted – that is, $p < 0.05$ – by the number of terms in the model. As such, the new level at which statistical significance would be accepted when $p < .00625$ (i.e., .05 ÷ 8).

Based on this new level of acceptance of statistical significance, we can see that all continuous independent variables are linearly related to the logit of the dependent variable. We know this because all *p*-values are above .00625 (i.e., they are **.016**, **.855** and **.431**).

To calculate the new alpha (α) level (i.e., *p*-value) for your own research, you need to divided the alpha level ($p < .05$) by the number of terms in your model. Formulaically, this is:

adjusted alpha level = original alpha level ÷ number of comparisons

For example, if you had four terms in your model, you would need to change the alpha level to .0125 (i.e., .05 ÷ 4 = .0125). Adjusted alpha values for up to eight terms to 6 decimal places (the maximum number of decimal place allowed in SPSS Statistics for alpha level values) are shown in the table below:

# of contrasts	Original alpha (α) level	New alpha (α) level
1	.05	.05
2	.05	.025
3	.05	.016667
4	.05	.0125
5	.05	.01
6	.05	.008333
7	.05	.007143
8	.05	.00625

Table:**Bonferroni-corrected alpha (α) levels**

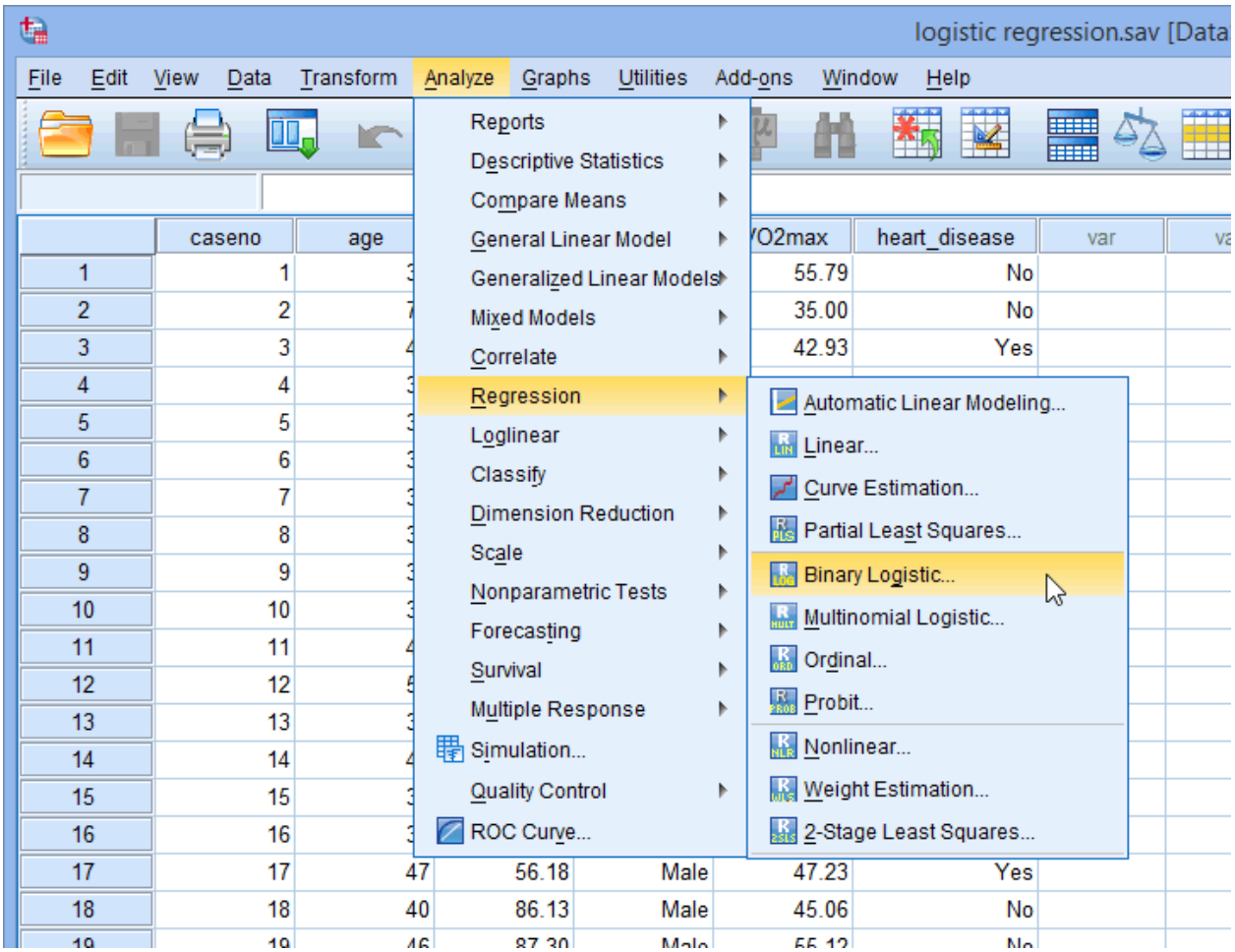
You could report the finding of the linearity assumption as follows:

> Linearity of the continuous variables with respect to the logit of the dependent variable was assessed via the Box-Tidwell (1962) procedure. A Bonferroni correction was applied using all eight terms in the model resulting in statistical significance being accepted when $p < .00625$ (Tabachnick & Fidell, 2014). Based on this assessment, all continuous independent variables were found to be linearly related to the logit of the dependent variable.

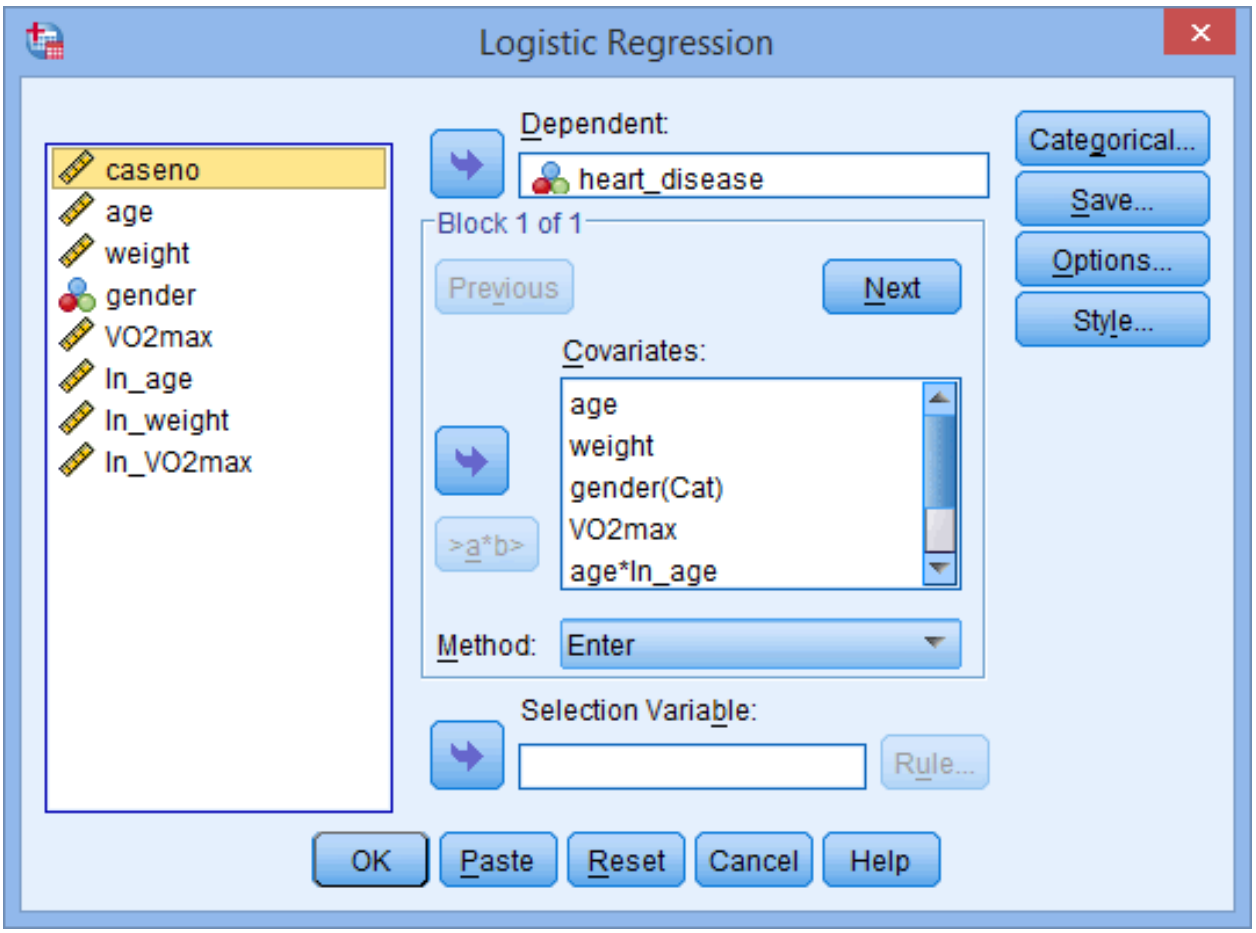
Procedure

A binomial logistic regression can be run using the **Binary Logistic...**procedure in SPSS Statistics. The instructions below will show you how to build the regression model in SPSS Statistics and which options to select in order to test whether you have any outliers (i.e., one of the assumptions of binomial logistic regression).

1 Click **Analyze > Regression > Binary Logistic...** on the main menu, as shown below:

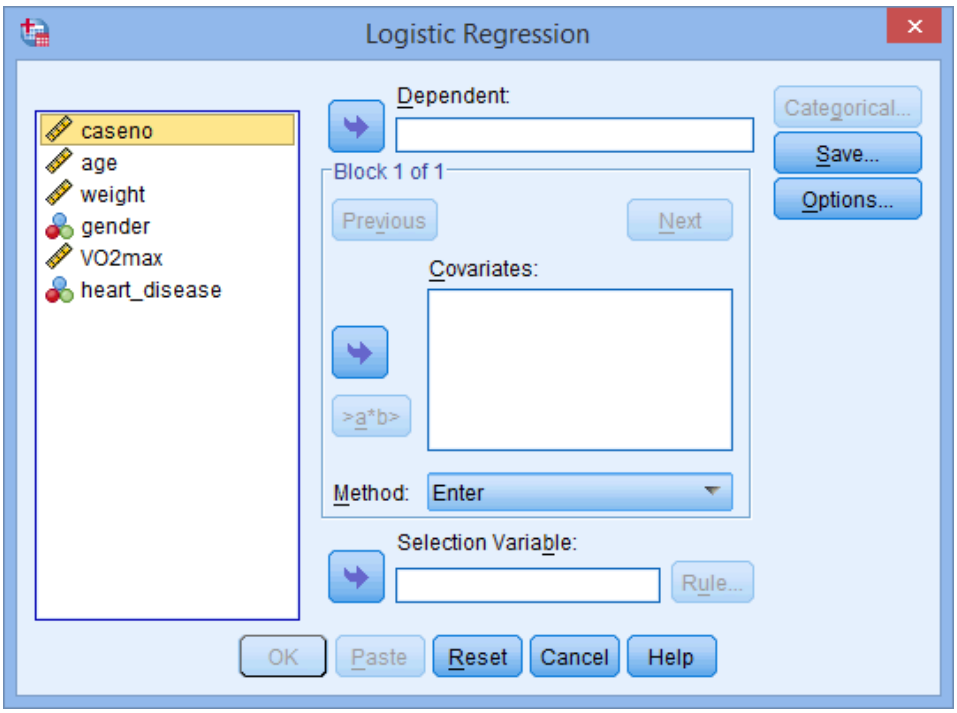


You will be presented with the **Logistic Regression** dialogue box, as shown below:



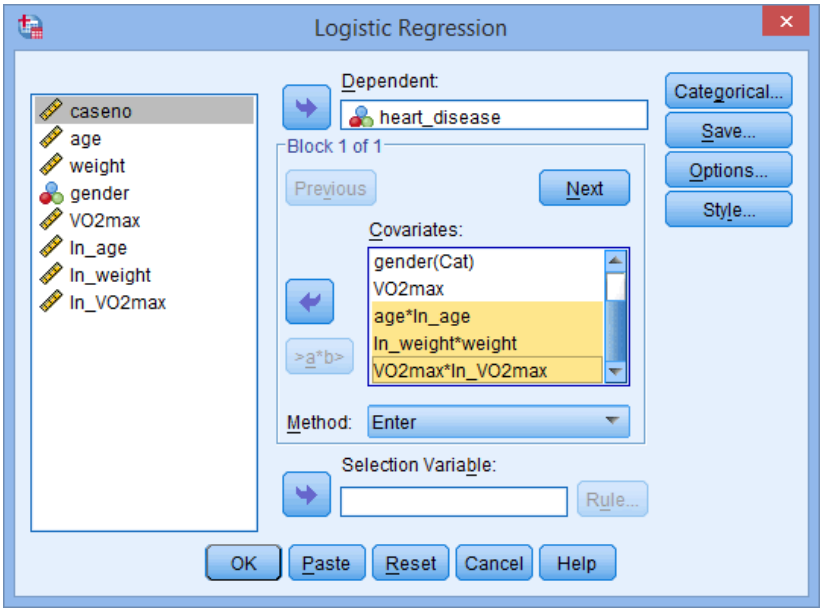
Note: If you have been following the steps throughout this guide, the **Logistic Regression** dialogue box above will already be populated with the dependent variable, heart_disease, in the Dependent: box, and the four independent variables – age, weight, gender and VO2max – and three interactions terms – ln_age*age, ln_weight*weight and ln_VO2max*VO2max – in the Covariates: box. You will have to scroll down the variables listed in the Covariates: box in order to see the "**ln_weight*weight**" and "**ln_VO2max*VO2max**" interaction terms because the default size of this box only shows five variables.

The **Logistic Regression** dialogue box was set up this way when testing for the assumption of linearity on the previous page. If you did not work through the previous page, the **Logistic Regression** dialogue box would look like the following:



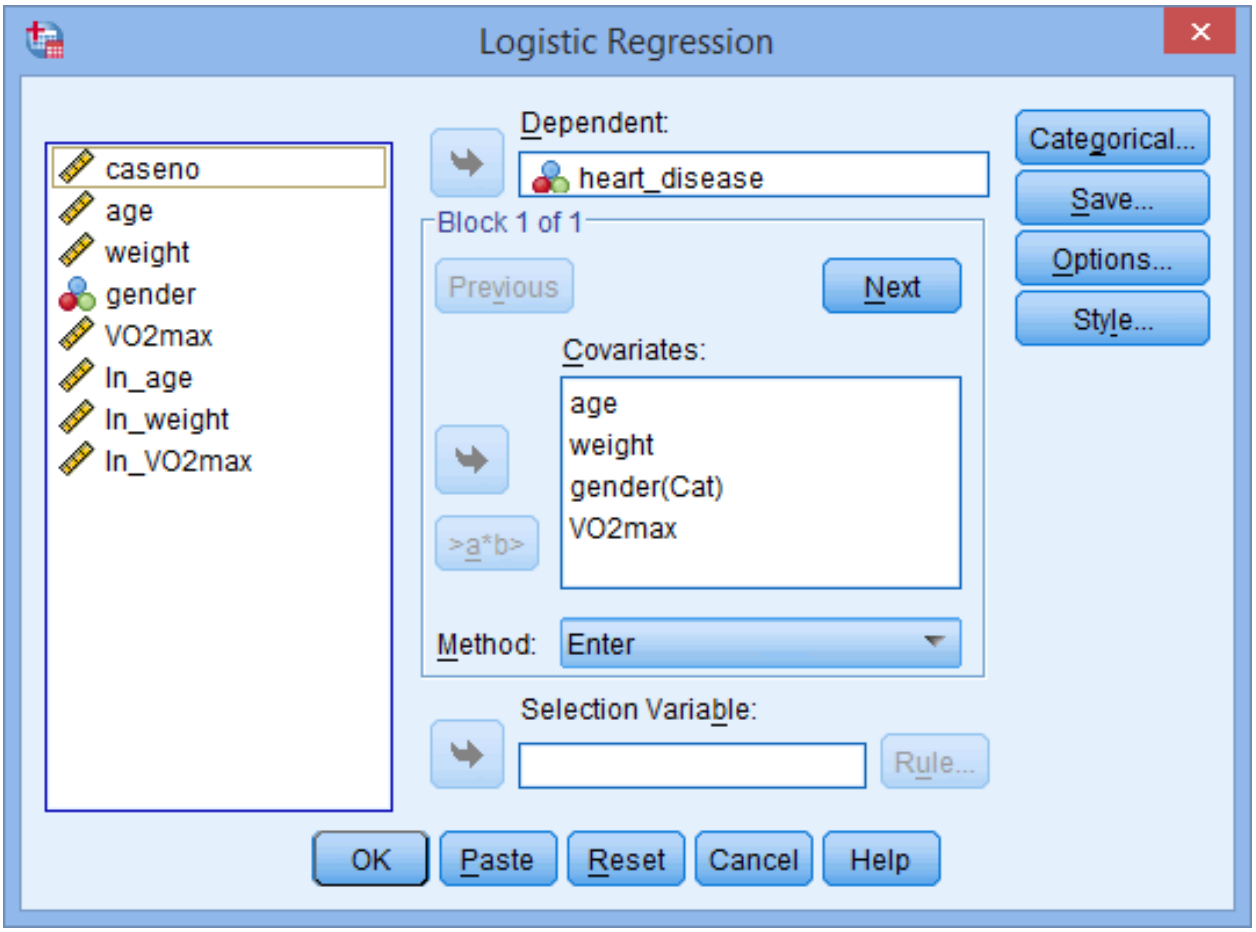
Therefore, you now need to go to the section, Box-Tidwell (1962) procedure to test for linearity, on the previous page, and run through **Steps 1 to 6**. Then, rather than moving on to Step 7 on the previous page, come back to this page and move onto **Step 3**, where you will be instructed to continue your analysis by clicking Options... button.

2 Remove the three interaction terms – "**ln_age*age**", "**ln_weight*weight**" and "**ln_VO2max*VO2max**" – from the Covariates: box by highlighting them and clicking the button, as shown below:

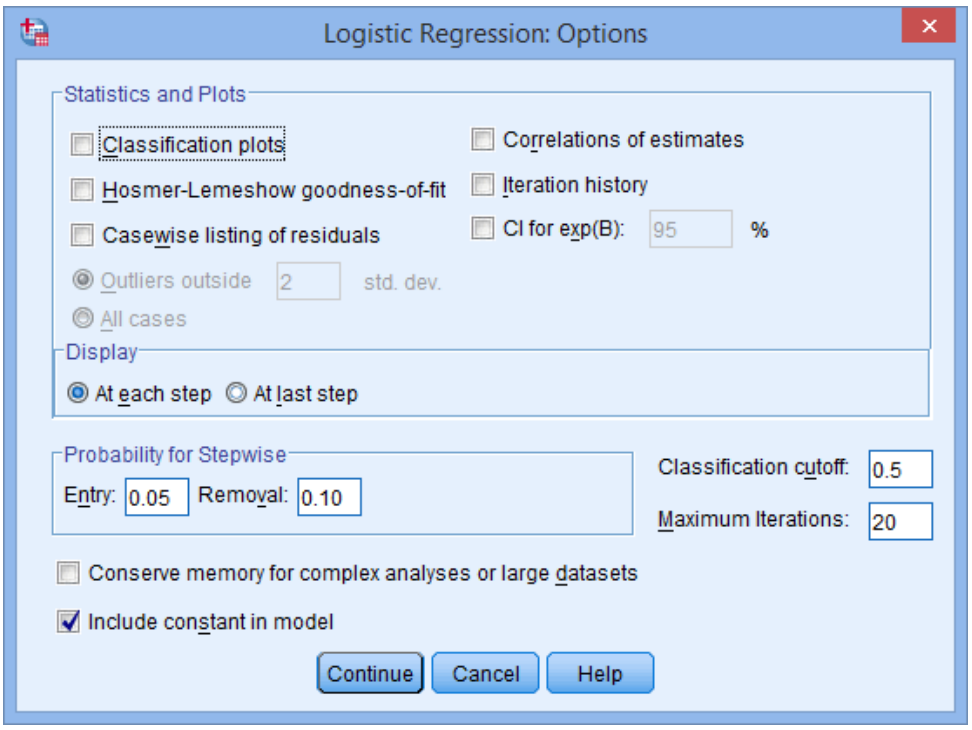


Note: Just to reiterate from the previous note, you will have to scroll down the variables listed in the Covariates: box in order to see the "**ln_weight*weight**" and "**ln_VO2max*VO2max**" interaction terms because the default size of this box only shows five variables.

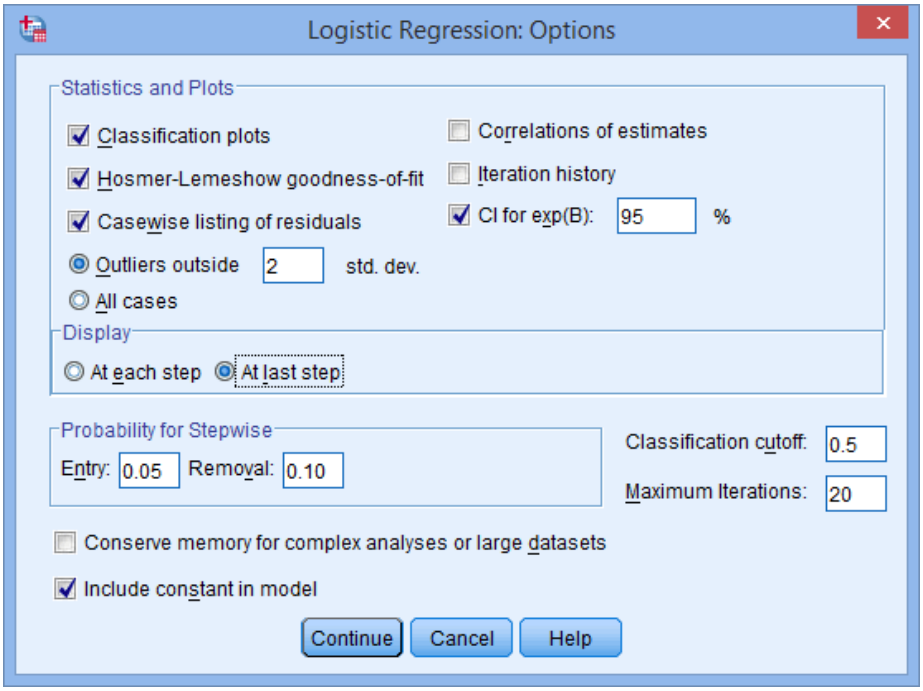
You will be presented with the **Logistic Regression** dialogue box where the three interactions have been removed from the Covariates: box, leaving just the four independent variables – "**age**", "**weight**", "**gender(Cat)**" and "**VO2max**" – as shown below:



3 Click the Options... button. You will be presented with the **Logistic Regression: Options** dialogue box, as shown below:

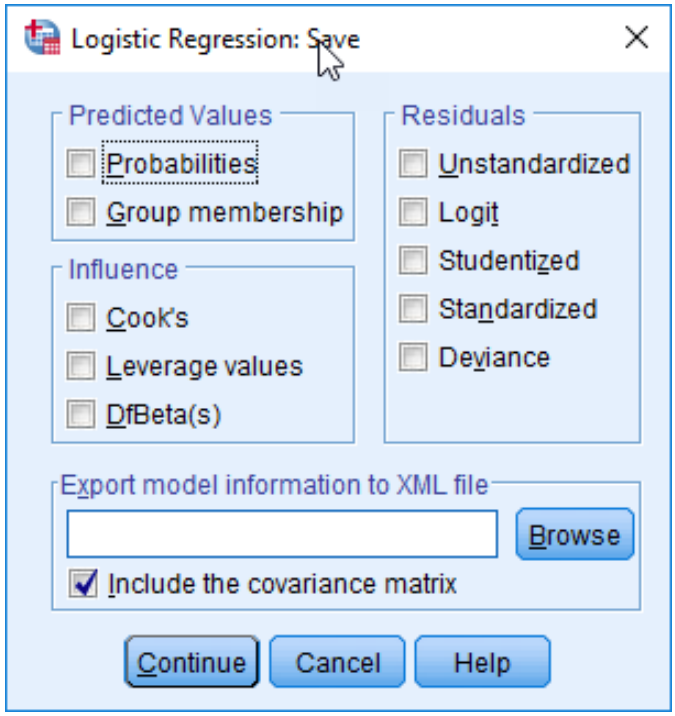


4 In the –Statistics and Plots– area, click the Classification plots, Hosmer-Lemeshow goodness-of-fit, Casewise listing of residuals and CI for exp(B): options, and in the –Display– area, click the At last step option. You will end up with a screen similar to below:

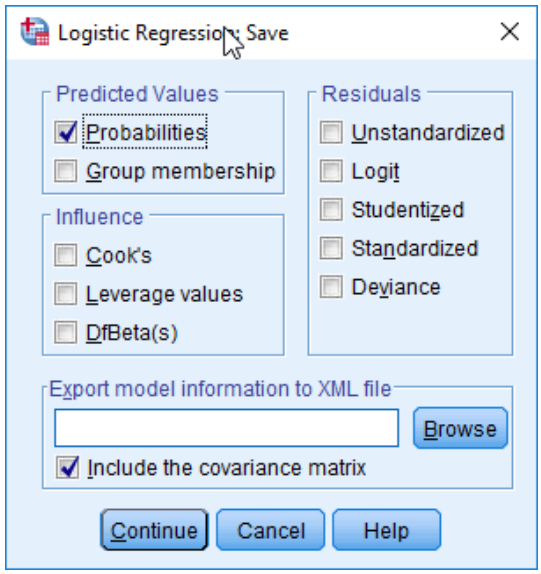


5 Click the Continue button. You will be returned to the **Logistic Regression** dialogue box.

6 Click the Save... button. You will be presented with the **Logistic Regression: Save** dialogue box, as shown below:



7 In the –Predicted Values– area, select the Probabilities checkbox. You will end up with a screen similar to below:



8 Click the Continue button. You will be returned to the **Logistic Regression** dialogue box.

9 Click the OK button. This will generate the output.

Having run the procedure above, you will first have to determine whether the assumption of outliers has been met before you interpret the binomial logistic regression results.

Assumptions II

Having run the binomial logistic regression in the Procedure section on the previous page, you can now interpret the results for the assumption of outliers (i.e., Assumption #7 that was presented in the Assumptions I). After interpreting this assumption, you will be in a position to start interpreting the results from the binomial logistic regression in the Interpreting Results section later.

Testing for outliers using case diagnostics

Like ordinary multiple regression, you need to check the binomial logistic regression results for cases which do not fit the model very well (e.g., outliers). The **Casewise List** table highlights cases where the cases (e.g., participants) with standardized residuals (i.e., the "**ZResid**" column) greater than ±2 standard deviations. Cases with standardized residual values greater than 2.5 should be inspected in further detail to determine why these cases are outliers and to remove them from the analysis if this is deemed necessary. You will have a table called **Casewise Diagnostics** that contains the relevant information, as shown below:

Casewise List[b]

Case	Selected Status[a]	Observed: heart_disease	Predicted	Predicted Group	Temporary Variable: Resid	Temporary Variable: ZResid
59	S	Y**	.142	N	.858	2.455
70	S	Y**	.082	N	.918	3.349

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

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

Note 1: If all your cases have standardized residuals less than ±2, this table will not be produced as part of the SPSS Statistics output. If you successfully remove an outlier, the **Casewise Diagnostics** table produced the first time you ran the analysis might not then be produced the second time around.

Note 2: There are different types of residuals that can be used to detect outliers: standardized residuals, studentized residuals or studentized deleted residuals. The default option that SPSS Statistics produces when you selected the Casewise diagnostics option in the Procedure section on the previous page is to use standardized residuals.

In the example in this guide, there is only one residual that conforms to this criteria and this is case number 70 (i.e., the "**Case**" column), which has a standardized residual of 3.349. This potential outlier is highlighted below:

Casewise List[b]

Case	Selected Status[a]	Observed: heart_disease	Predicted	Predicted Group	Temporary Variable: Resid	Temporary Variable: ZResid
59	S	Y**	.142	N	.858	2.455
70	S	Y**	.082	N	.918	3.349

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

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

Note: The "**Case Number**" column refers to SPSS's intrinsic numbering system (i.e., the numbers in the numbered far-left blue column in the **Data View** window), not the self-assigned variable, caseno.

You should check to see why this particular case is unusual and whether you should remove it from the model.

You could report this finding as follows:

There was one standardized residual with a value of 3.349 standard deviations, which was kept in the analysis.

If you do have outliers that are of concern, you need to consider whether you should remove them. Alternatively, you might find that a transformation solves the issue. If you decide at this stage to remove any outliers from the analysis, you will need to filter out these outliers and re-run the regression analysis. How to filter cases (e.g., outliers) is presented in our Selecting cases guide. If you make any transformations, you will need to re-run the regression analysis starting with the linearity assumption.

Interpreting Results

After running the binomial logistic regression procedures and testing that your data meets the assumptions of a binomial logistic regression in the previous sections, SPSS Statistics will have generated a number of tables that contain all the information you need to report the results of your binomial logistic regression. On the following four pages, we show you how to interpret these results.

There are two main objectives that you can achieve with the output from a binomial logistic regression: (a) determine which of your independent variables (if any) have a statistically significant effect on your dependent variable; and (b) determine how well your binomial logistic regression model predicts the dependent variable. Both of these objectives will be answered in the following sections:

- Data coding: You can start your analysis by inspecting your variables and data, including: (a) checking if any cases are missing and whether you have the number of cases you expect (the "**Case Processing Summary**" table); (b) making sure that the correct coding was used for the dependent variable (the "**Dependent Variable Encoding**" table); and (c) determining whether there are any categories amongst your categorical independent variables with very low counts – a situation that is undesirable for binomial logistic regression (the "**Categorical Variables Codings**" table). This is highlighted in the Data coding section.
- Baseline analysis: Next, you can consult the "**Classification Table**", "**Variables in the Equation**" and "**Variables not in the Equation**" tables. These all relate to the situation where no independent variables have been added to the model and the model just includes the constant. As such, you are interested in this information only as a comparison to the model with all the independent variables added. This Baseline analysis section provides a basis against which the main binomial logistic regression analysis with all independent variables added to the equation can be evaluated, which takes place.
- Binomial logistic regression results: In evaluating the main logistic regression results, you can start by determining the overall statistical significance of the model (namely, how well the model predicts categories compared to no independent variables). You can also assess the adequacy of the model by analysing how poor the model is at predicting the categorical outcomes using the **Hosmer and Lemeshow goodness of fit test**. This is explained in the Model fit section. Next, you can consult the **Cox & Snell R Square**and **Nagelkerke R Square** values to understand how much variation in the dependent variable can be explained by the model (i.e., these are two methods of calculating the explained variation), but it is preferable to report the Nagelkerke R^2 value. This is illustrated in the Variance explained section.

- Category prediction: After determining model fit and explained variation, it is very common to use binomial logistic regression to predict whether cases can be correctly classified (i.e., predicted) from the independent variables. Logistic regression estimates the probability of an event (in this case, having heart disease) occurring. If the estimated probability of the event occurring is greater than or equal to 0.5 (better than even chance), SPSS Statistics classifies the event as occurring (e.g., heart disease being present). If the probability is less than 0.5, SPSS Statistics classifies the event as not occurring (e.g., no heart disease). We explained how to interpret category prediction based on the observed and predicted classifications.
- Variables in the equation: Finally, you can assess the contribution of each independent variable to the model and its statistical significance using the **Variables in the Equation** table. You will also be able to use the odds ratios of each of the independent variables (along with their confidence intervals) to understand the change in the odds ratio for each increase in one unit of the independent variable. Using these odds ratios you will be able to, for example, make statements such as: "the odds of having heart disease is 7.026 times greater for males as opposed to females". You can make such predictions for categorical and continuous independent variables.

Data coding

The first several tables describe some aspects of the setup of the variables used in the analysis and provide an opportunity to make sure that the analysis is set up in the manner you intended. The first table is the "**Case Processing Summary**" table, as shown below:

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	100	100.0
	Missing Cases	0	.0
	Total	100	100.0
Unselected Cases		0	.0
Total		100	100.0

a. If weight is in effect, see classification table for the total number of cases.

This table is useful for determining if any cases are missing and whether you have the number of cases you expect. The next table, "**Dependent Variable Encoding**" (see below), displays the coding that you applied to the dependent variable, heart_disease.

Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

The table below, "**Categorical Variables Codings**", shows the coding only for the predictor/independent variables that are categorical. You can use this table to determine whether there are any categories with very low counts - a situation that is undesirable for binomial logistic regression. In this example, neither category has very low numbers of counts.

Categorical Variables Codings

		Frequency	Parameter coding (1)
gender	Female	37	.000
	Male	63	1.000

Baseline analysis

The next three tables headed under the main title, "**Block 0: Beginning Block**", all relate to the situation where no independent variables have been added to the model and the model just includes the constant. As such, you are interested in this information only as a comparison to the model with all the independent variables added. The table below, "**Classification Table**", shows that without any independent variables, the 'best guess' is to simply assume that all participants did not have heart disease. If you assume this, you will overall correctly classify 65% of cases (the "**Overall Percentage**" row), as shown below:

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			heart_disease		Percentage Correct
			No	Yes	
Step 0	heart_disease	No	65	0	100.0
		Yes	35	0	.0
	Overall Percentage				65.0

a. Constant is included in the model.

b. The cut value is .500

The table below, "**Variables in the Equation**", simply shows you that only the constant was included in this particular model:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-.619	.210	8.718	1	.003	.538

And the table below, "**Variables not in the Equation**", highlights the independent variables left out of the model:

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	age	11.872	1	.001
		weight	8.065	1	.005
		gender(1)	4.620	1	.032
		VO2max	4.150	1	.042
	Overall Statistics		24.338	4	.000

Binomial logistic regression results

All the next tables come after the heading "**Block 1: Method = Enter**" and represent the results of the main logistic regression analysis with all independent variables added to the equation.

Model fit

The first table, "**Omnibus Tests of Model Coefficients**", provides the overall statistical significance of the model (namely, how well the model predicts categories compared to no independent variables), as shown below:

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	27.402	4	.000
	Block	27.402	4	.000
	Model	27.402	4	.000

For this type of binomial logistic regression, you can reference the "**Model**" row. From the table above, you can see that the model is statistically significant ($p < .0005$; "**Sig.**" column). Another way of assessing the adequacy of the model is to analyse how poor the model is at predicting the categorical outcomes. This is tested using the **Hosmer and Lemeshow goodness of fit test** as found in the similarly titled table, as shown below:

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.843	8	.871

For this test, you do not want the result to be statistically significant because this would indicate that you have a poor fitting model. In this example, the Hosmer and Lemeshow test is not statistically significant (p = .871; "**Sig.**" column), indicating that the model is not a poor fit.

Variance explained

In order to understand how much variation in the dependent variable can be explained by the model (the equivalent of R^2 in multiple regression), you can consult the table below, "**Model Summary**":

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	102.088[a]	.240	.330

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

This table contains the **Cox & Snell R Square** and **Nagelkerke R Square** values, which are both methods of calculating the explained variation (it is not as straightforward to do this as compared to multiple regression). These values are sometimes referred to as *pseudo* R^2values and will have lower values than in multiple regression. However, they are interpreted in the same manner, but with more caution.

Therefore, the explained variation in the dependent variable based on our model ranges from 24.0% to 33.0%, depending on whether you reference the Cox & Snell R^2 or Nagelkerke R^2 methods, respectively. Nagelkerke R^2 is a modification of Cox & Snell R^2, the latter of which cannot achieve a value of 1. For this reason, it is preferable to report the Nagelkerke R^2 value.

Category prediction

Binomial logistic regression estimates the probability of an event (in this case, having heart disease) occurring. If the estimated probability of the event occurring is greater than or equal to 0.5 (better than even chance), SPSS Statistics classifies the event as occurring (e.g., heart disease being present). If the probability is less than 0.5, SPSS Statistics classifies the event as not occurring (e.g., no heart disease). It is very common to use logistic regression to predict whether cases can be correctly classified (i.e., predicted) from the independent variables. Therefore, it becomes necessary to have a method to assess

the effectiveness of the predicted classification against the actual classification. There are many methods to assess this with their usefulness often depending on the nature of the study conducted. However, all methods revolve around the observed and predicted classifications, which are presented in the **Classification Table**, as shown below:

Classification Table[a]

Observed			Predicted		
			heart_disease		Percentage Correct
			No	Yes	
Step 1	heart_disease	No	55	10	84.6
		Yes	19	16	45.7
	Overall Percentage				71.0

a. The cut value is .500

Firstly, notice that the table has a subscript which states, "The cut value is .500". This means that if the probability of a case being classified into the "yes" category is greater than .500, then that particular case is classified into the "yes" category. Otherwise, the case is classified as in the "no" category (as mentioned previously). The classification table from earlier – which did not include any independent variables – showed that 65.0% of cases overall could be correctly classified by simply assuming that all cases were classified as "no" heart disease. However, with the independent variables added, the model now correctly classifies 71.0% of cases overall (see "**Overall Percentage**" row). That is, the addition of the independent variables improves the overall prediction of cases into their observed categories of the dependent variable. This particular measure is referred to as the ***percentage accuracy in classification (PAC)***.

Another measure is the ***sensitivity***, which is the percentage of cases that had the observed characteristic (e.g., "yes" for heart disease) which were correctly predicted by the model (i.e., true positives). In this case, 45.7% of participants who had heart disease were also predicted by the model to have heart disease (see the "**Percentage Correct**" column in the "**Yes**" row of the observed categories).

Specificity is the percentage of cases that did not have the observed characteristic (e.g., "no" for heart disease) and were also correctly predicted as not having the observed characteristic (i.e., true negatives). In this case, 84.6% of participants who did not have heart disease were correctly predicted by the model not to have heart disease (see the "**Percentage Correct**" column in the "**No**" row of the observed categories).

The ***positive predictive value*** is the percentage of correctly predicted cases with the observed characteristic compared to the total number of cases predicted as having the characteristic. In our case, this is 100 x (16 ÷ (10 + 16)) which is 61.5%. That is, of all cases predicted as having heart disease, 61.5% were correctly predicted.

The ***negative predictive value*** is the percentage of correctly predicted cases without the observed characteristic compared to the total number of cases predicted as not having the characteristic. In our case, this is 100 x (55 ÷ (55 + 19)) which is 74.3%. That is, of all cases predicted as not having heart disease, 74.3% were correctly predicted.

ROC Curve

In the previous section you calculated five measures – such as sensitivity and specificity – that assess the ability of a binomial logistic regression model to **correctly classify cases** (i.e., to **discriminate**). All these measures were calculated based on a **cut-off point** of **0.5** (**50%**), meaning that a case (e.g., participant) with a **predicted probability of the event** (e.g., heart disease) that is **greater than or equal to 0.5**would be classified as **having** the event (e.g., having heart disease), and all participants with predicted probabilities **lower than 0.5** would be classified as **not having** the event (e.g., not having heart disease). However, instead of concentrating on **one** cut-off point only, you can consider **all possible** cut-off points in your data, and how **each cut-off point changes** the **specificity** and **sensitivity** of the test. For example, a **higher** cut-off point will **increase specificity**, **but lower sensitivity**. That is, a higher cut-off point makes it "harder" for participants to be classified as having the event of interest, but "easier" to be classified as not having the event of interest. A visual representation of this is presented in a plot called the **Receiver Operating Characteristic (ROC) curve**, which is a **plot** of **sensitivity versus 1 minus specificity** (Hilbe, 2009). The ROC curve can also be used to calculate an **overall measure of discrimination**, but this will be discussed later.

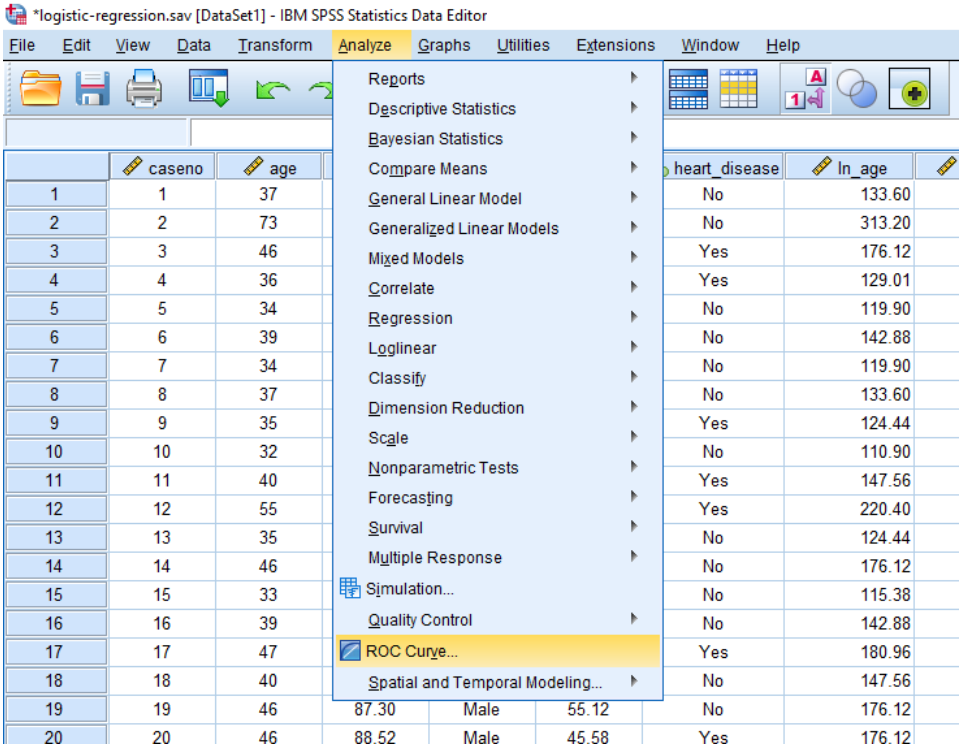
Explanation: Discrimination is the ability of a binomial logistic regression model to discriminate (i.e., discern) between those participants with and without the event of interest (i.e., to be able to predict who has, or has not, the event of interest).

Note: You should ask is whether you should report the area under the ROC curve at all. The area under the ROC curve is a measure of the overall discriminatory ability of the binomial logistic regression model, but if you not interested in this aspect of the model (i.e., the ability to classify individuals into the two groups of the dichotomous dependent variable) you will probably not want to report this measure (and certainly not as a measure of the model fit). Nonetheless, it is a popular measure to report following a binomial logistic regression analysis, and although your reasons for running a binomial logistic regression analysis might not have been to understand discrimination, readers of your study might want to know themselves. Therefore, we include it in this guide. However, if you take this latter approach (i.e., reporting the area under the ROC curve when it is not relevant to your analysis aims), this could be challenged by reviewers.

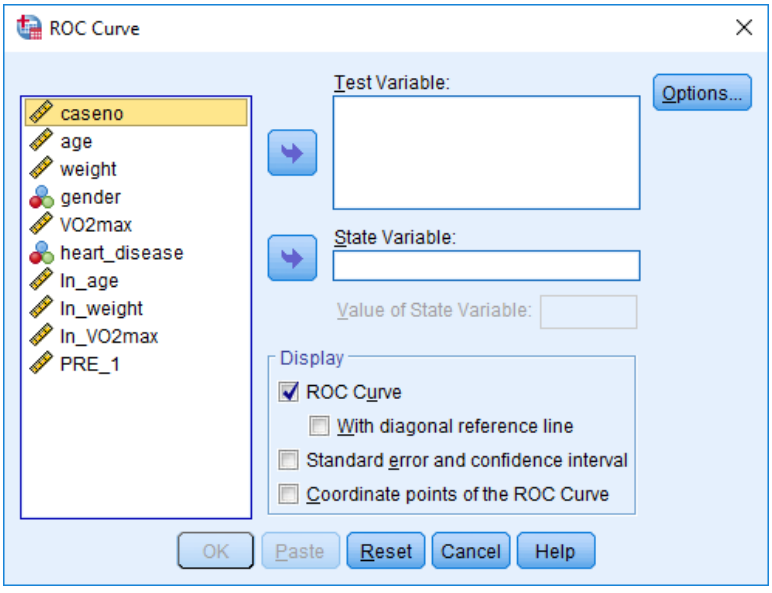
ROC curve procedure

How to generate the ROC curve is shown in the procedure below:

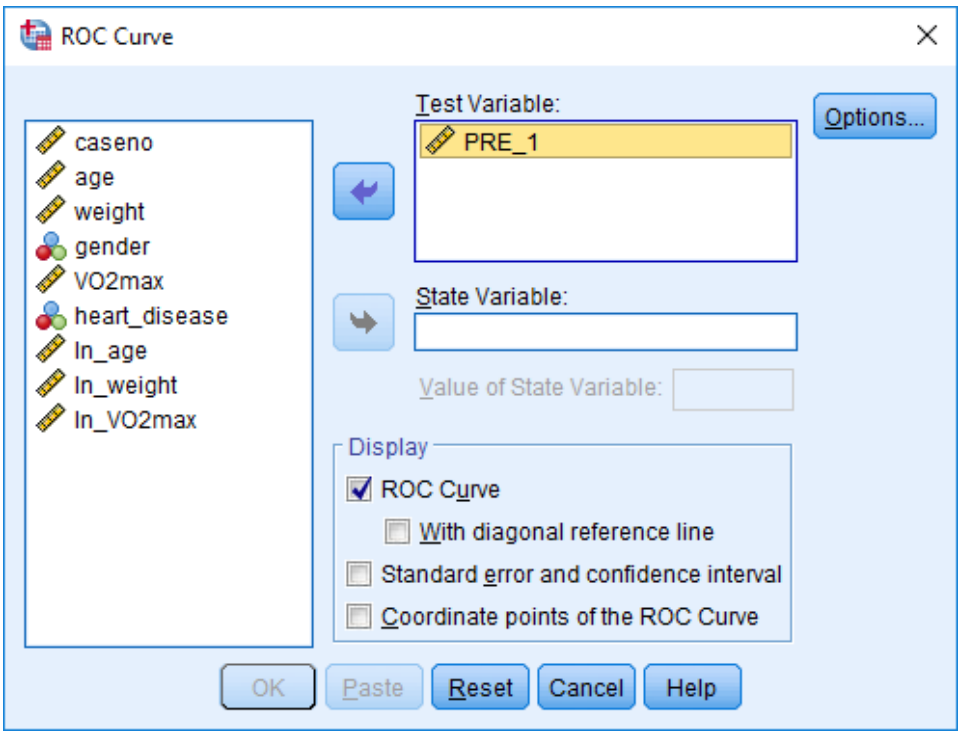
1 Click **Analyze > ROC Curve...** on the main menu, as shown below:



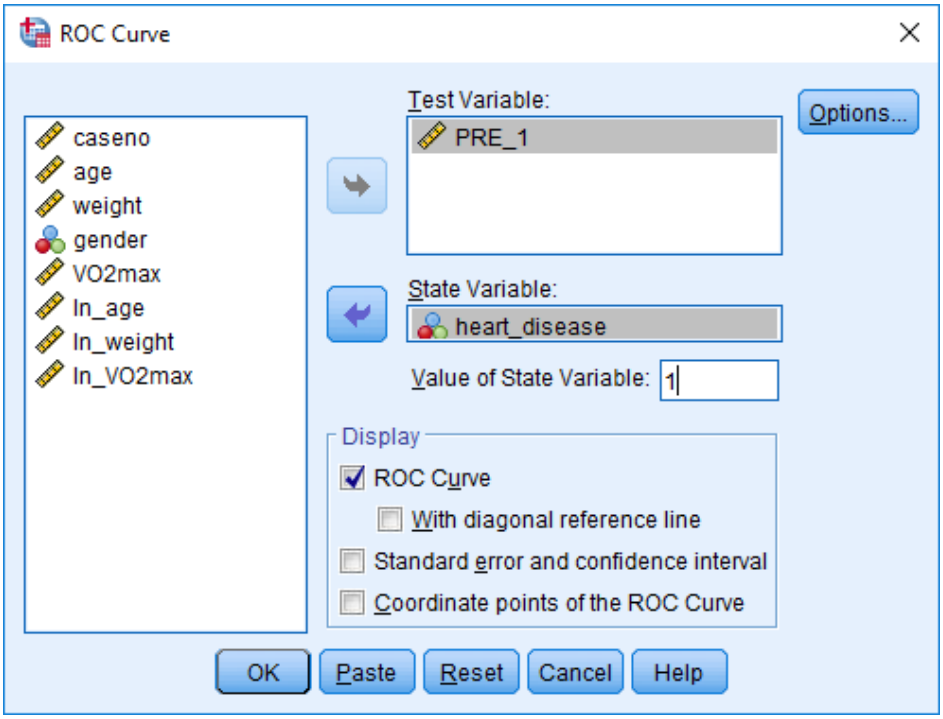
You will be presented with the **ROC Curve** dialogue box, as shown below:



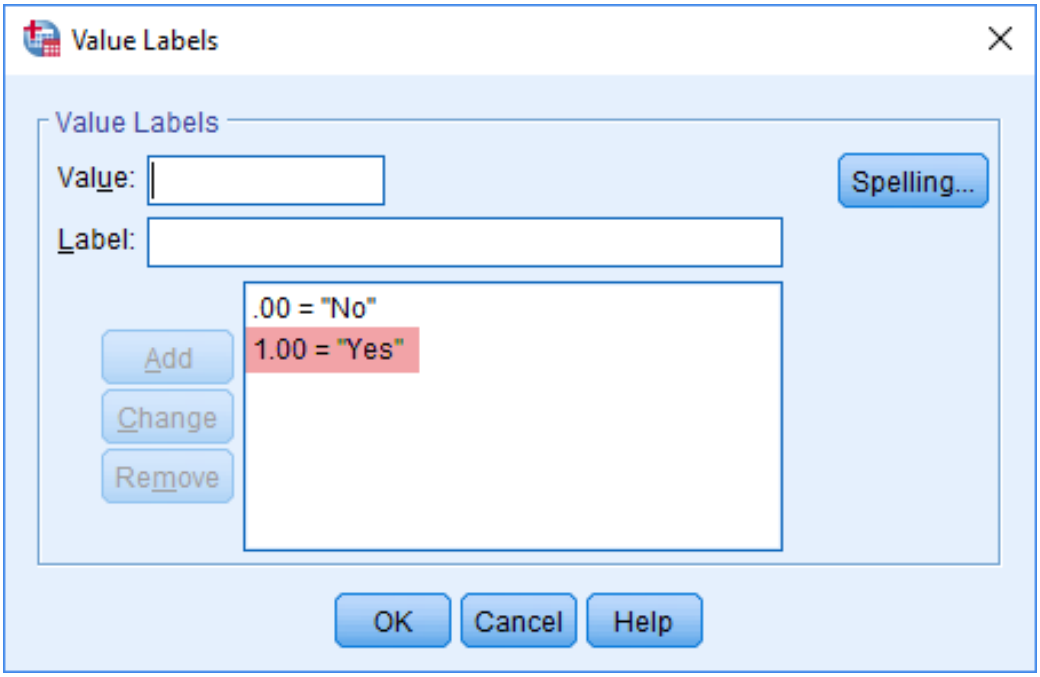
2 Transfer the predicted value variable, PRE_1, into the Test Variable: box by highlighting it and clicking on the button, as shown below:



3 Transfer the dependent variable, heart_disease, into the State Variable: box by highlighting it and clicking on the button. Then, type "**1**" (without the quotes) into the Value of State Variable: box, which reflects the "**success**" or "**event**" category of our dependent variable (i.e., where "**1**" = "**Yes**" to having heart disease). You will end up with a screen similar to below:

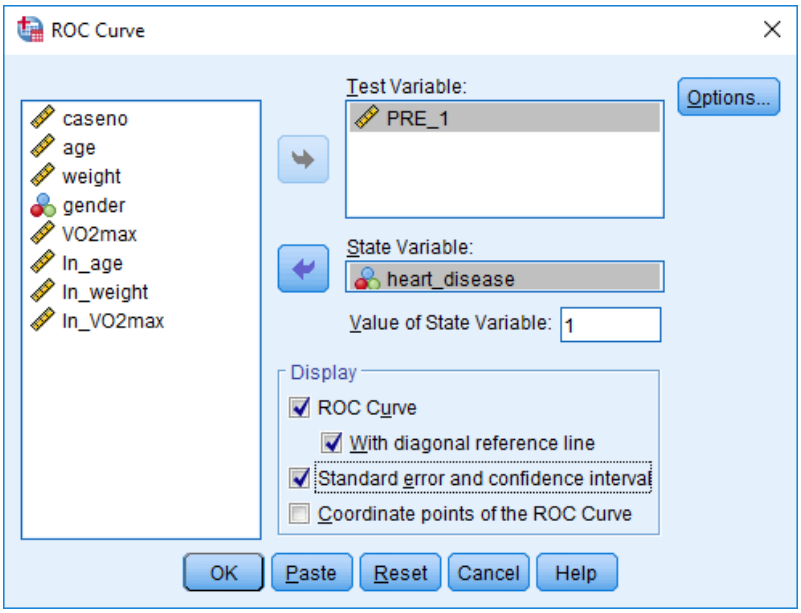


Explanation: What **value** you enter into the Value of State Variable: box will depend on how you have **coded** the category of your dependent variable that you consider to be the "**success**" or "**event**" category in the "**Value Labels**" dialogue box when setting up your data in the **Variable View** of SPSS Statistics (i.e., in the Example & Data Setup section on page 4), as highlighted for our example below:



As highlighted above, we coded "**Yes**" (i.e., having heart disease, which is our "**success/event**" category) as "**1**" in the **Value Labels** dialogue box. Therefore, we entered "**1**" into the Value of State Variable: box above.

4 In the –Display– area, click the With diagonal reference line and Standard error and confidence interval, as shown below:



5 Click on the OK button to generate the SPSS Statistics output for the ROC curve.

Interpreting the ROC curve

Before interpreting the ROC curve results, we suggest consulting the "**a.**" sub-note at the bottom of the **Case Processing Summary** table, as highlighted below, to make sure that you correctly coded the event of interest (i.e., your "**success/event**" category) in Step 3 of the ROC curve procedure in the previous section:

Case Processing Summary

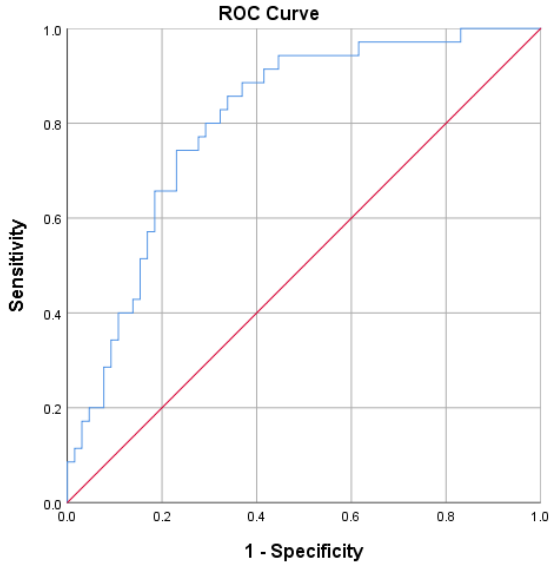
heart_disease	Valid N (listwise)
Positive[a]	35
Negative	65

Larger values of the test result variable(s) indicate stronger evidence for a positive actual state.

a. The positive actual state is 1.00 Yes.

You can see in the sub-note highlighted above that the **positive actual state** is "**1.00 Yes**", indicating that we have correctly stated the event (i.e., the event of interest in this example is having heart disease, which was coded as "**1 = Yes**"). Whatever category represents your event of interest should be reported in this sub-note. If not, you need to go back to Step 3 of the ROC procedure above and change the coding you have entered accordingly.

Now that you know you have entered the correct information in the ROC curve procedure, you can consider the ROC curve results. As such, the ROC curve is presented under the heading, **ROC Curve**, as shown below:



The **further** the blue line is **above** the straight line, the **better** the **discrimination**. The **area under** the ROC curve is equivalent to the **concordance probability** (Gönen, 2007), which can also be reported via SPSS Statistics' NOMREG procedure (i.e., its multinomial logistic regression procedure). The **concordance (c) statistic** is the most common measure of the **ability** of a generalized linear model (GzLM) to **discriminate**, of which binomial logistic regression is a GzLM (Steyerberg, 2009). It is equivalent to the area under the ROC curve for a dichotomous dependent variable (i.e., for binomial (or binary) logistic regressions) (Gönen, 2007; Steyerberg, 2009). You can find the **value**for this area and, therefore, the concordance statistic, by consulting the "**Area**" column in the **Area Under the Curve** table, as highlighted below:

Area Under the Curve

Test Result Variable(s): PRE_1

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.804	.044	.000	.718	.891

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

You can see that the **area under the ROC curve** is **.804**. The area can range from **0.5** to **1.0** with **higher values** representing **better discrimination**. According to Hosmer et al. (2013) a value of .804 puts the discrimination of this model at the lower border of **excellent discrimination**. The general rules of thumb of Hosmer et al. (2003) are presented below:

AUC	**Classification**
0.5	This suggests no discrimination, so we might as well flip a coin.
0.5 < AUC < 0.7	We consider this poor discrimination, not much better than a coin toss.
0.7 ≤ AUC < 0.8	We consider this acceptable discrimination.
0.8 ≤ AUC < 0.9	We consider this excellent discrimination.
AUC ≥ 0.9	We consider this outstanding discrimination.

Table: **Rules of thumb for the area under the ROC curve (AUC) according to Hosmer et al. (2013).**

It is also possible to provide a **95% confidence interval (CI)** for the area under the ROC curve. These are presented in the "**Lower Bound**" and "**Upper Bound**" columns under the "**Asymptotic 95% Confidence Interval**" column in the "**Area Under the Curve**" table, as highlighted below:

Area Under the Curve

Test Result Variable(s): PRE_1

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.804	.044	.000	.718	.891

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

You can see that the 95% confidence interval (CI) is from **.718** to **.891**. That is, we can be **95% confident** that the **population value** of the area under the ROC curve is between .718 and .891.

You could report your results as follows:

> The area under the ROC curve was .804 (95% CI, .718 to .891), which is an excellent level of discrimination according to Hosmer et al. (2013).

If you have space in your report, you should also present the ROC curve itself (as recommended by Hosmer et al., 2003).

> **Note:** The recommendation by Hosmer et al. (2003) and Royston & Altman (2010) to report the ROC curve is based on at least one of your study aims being to understand the ability of the binomial logistic regression model to discriminate individuals with and without the event of interest.

Now that you have used binomial logistic regression to predict whether cases can be correctly classified (i.e., predicted) from the independent variables, you can assess the contribution of each independent variable to the model and its statistical significance.

Variables in the equation

The **Variables in the Equation** table shows the contribution of each independent variable to the model and its statistical significance. This table is shown below:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age	.085	.028	9.132	1	.003	1.089	1.030	1.151
	weight	.006	.022	.065	1	.799	1.006	.962	1.051
	gender(1)	1.950	.842	5.356	1	.021	7.026	1.348	36.625
	VO2max	-.099	.048	4.266	1	.039	.906	.824	.995
	Constant	-1.676	3.336	.253	1	.615	.187		

a. Variable(s) entered on step 1: age, weight, gender, VO2max.

The Wald test ("**Wald**" column) is used to determine statistical significance for each of the independent variables. The statistical significance of the test is found in the "**Sig.**" column. From these results you can see that age (p = .003), gender (p = .021) and VO2max (p = .039) added significantly to the model/prediction, but weight (p = .799) did not add significantly to the model.

The *B* coefficients ("**B**" column) are used in the equation to predict the probability of an event occurring, but not in an immediately intuitive manner. The coefficients do, in fact, show the change in the log odds that occur for a one-unit change in an independent variable when all other independent variables are kept constant. So, for example, the log odds change for gender is 1.950, which is the increase in log odds (as *B* is positive) for males (as females were coded "0" and males as "1"). However, this is not often the most intuitive method of understanding your results. Luckily, SPSS Statistics also includes the odds ratios of each of the independent variables in the "**Exp(B)**" column along with their confidence intervals ("**95%**

C.I. for EXP(B)" column). This informs you of the change in the odds for each increase in one unit of the independent variable. For example, for gender, an increase in one unit (i.e., being male) increases the odds by 7.026. What this means is that the odds of having heart disease ("yes" category) is 7.026 times greater for males as opposed to females. Values less than 1.000 indicate a decreased odds for an increase in one unit of the independent variable. Sometimes, for clarity, the odds ratio is inverted (e.g., 1 / .906 = 1.10, for VO2max). Thus, you would state that for each unit *reduction* in the independent variable, VO2max, the odds of having heart disease increases by a factor of 1.10. Remember to invert the confidence intervals as well if you take this latter approach.

Reporting

When you report your results, you can simply focus on the main findings. However, it is good practice to report the results from the assumptions tests you carried out. Finally, we illustrate how to report your results in a table.

Reporting your main findings

You could write up the results as follows:

> A binomial logistic regression was performed to ascertain the effects of age, weight, gender and VO_2max on the likelihood that participants have heart disease. The logistic regression model was statistically significant, $\chi^2(4) = 27.402$, $p < .0005$. The model explained 33.0% (Nagelkerke R^2) of the variance in heart disease and correctly classified 71.0% of cases. Sensitivity was 45.7%, specificity was 84.6%, positive predictive value was 61.5% and negative predictive value was 74.3%. Of the five predictor variables only three were statistically significant: age, gender and VO_2max (as shown in Table 1). Males had 7.02 times higher odds to exhibit heart disease than females. Increasing age was associated with an increased likelihood of exhibiting heart disease, but increasing VO_2max was associated with a reduction in the likelihood of exhibiting heart disease.

Including assumption testing in your reporting

Adding in the information about the tests and assumptions ran, you have:

> A binomial logistic regression was performed to ascertain the effects of age, weight, gender and VO_2max on the likelihood that participants have heart disease. Linearity of the continuous variables with respect to the logit of the dependent variable was assessed via the Box-Tidwell (1962) procedure. A Bonferroni correction was applied using all eight terms in the model resulting in statistical significance being accepted when $p < .00625$ (Tabachnick & Fidell, 2014). Based on this assessment, all continuous independent

variables were found to be linearly related to the logit of the dependent variable. There was one standardized residual with a value of 3.349 standard deviations, which was kept in the analysis. The logistic regression model was statistically significant, $\chi^2(4) = 27.402$, $p < .0005$. The model explained 33.0% (Nagelkerke R^2) of the variance in heart disease and correctly classified 71.0% of cases. Sensitivity was 45.7%, specificity was 84.6%, positive predictive value was 61.5% and negative predictive value was 74.3%. Of the five predictor variables only three were statistically significant: age, gender and VO_2max (as shown in Table 1). Males had 7.02 times higher odds to exhibit heart disease than females. Increasing age was associated with an increased likelihood of exhibiting heart disease, but increasing VO_2max was associated with a reduction in the likelihood of exhibiting heart disease.

Tabulating the results of a binomial logistic regression

You can present the results from the binomial logistic regression analysis in a simple table, as shown below:

Table 1

Logistic Regression Predicting Likelihood of Heart Disease based on Age, Weight, Gender and VO_2max.

	B	SE	Wald	*df*	*p*	Odds Ratio	95% CI for Odds Ratio	
							Lower	Upper
Age	.09	.03	9.13	1	.003	1.09	1.03	1.15
Weight	.01	.02	.07	1	.799	1.01	.96	1.05
Gender	1.95	.84	5.36	1	.021	7.03	1.35	36.63
VO_2max	-.099	.05	4.27	1	.039	.91	.82	1.00
Constant	-1.68	3.34	.25	1	.615	.19		

Note: Gender is for males compared to females.

Disclosure : THIS ARTICLE is copied from

Laerd Statistics (2017). Binomial logistic regression using SPSS Statistics. *Statistical tutorials and software guides*. Retrieved from https://statistics.laerd.com/

References

1. Box, G. E. P., & Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics*, *4*, 531-550.
2. Collett, D. (2003). *Modelling binary data* (2nd ed.). Boca Raton, FL: CRC Press.
3. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Alternative Regression Models: Logistic, Poisson Regression, and the Generalized Linear Model. *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). London: Lawrence Erlbaum Associates.
4. Fox, J. (2016). *Applied regression analysis and generalized linear models* (3rd ed.). Thousand Oaks, CA: Sage.

5. Gönen, M. (2007). *Analyzing receiver operating characteristic curves with SAS*. Cary, NC: SAS Institute Inc.
6. Guerrero, V. M., & Johnson, R. A. (1982). Use of the Box-Cox transformation with binary response models. *Biometrika*, *69*(2), 309-314.
7. Hilbe, J. M. (2009). *Logistic regression models*. CRC Press: Kindle Edition.
8. Hilbe, J. M. (2016). *Practical guide to logistic regression*. Boca Raton, FL: CRC Press.
9. Hosmer, D. W., Jr. & Lemeshow, S. (1989). *Applied logistic regression*. New York, NY: John Wiley & Sons.
10. Hosmer, D. W., Jr., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Hoboken, NJ: Wiley.
11. Jaccard, J. (2001). *Interaction effects in logistic regression*. London: SAGE Publications.
12. Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Thousand Oaks, CA: Sage.
13. Menard, S. (2010). *Logistic regression: From introductory to advanced concepts and applications*. London: SAGE Publications.
14. Osborne, J. W. (2015). *Best practices in logistic regression*. Thousand Oaks, CA: Sage.
15. Royston, P., & Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling. *Applied Statistics*, *43*, 429-467.
16. Royston, P., & Altman, D. G. (2010). Visualizing and assessing discrimination in the logistic regression model. *Statistics in Medicine*, *29*, 2508-2520.
17. Steyerberg, E. W (2009). *Clinical prediction models*. New York, NY: Springer.
18. Tabachnick, B. G., & Fidell, L. S. (2014). *Using multivariate statistics* (6th ed.). Essex, UK: Pearson.
19. Thompson, W. K., Xie, M., & White, H. R. (2003). Transformations of covariates for longitudinal data. *Biostatistics*, *4*(3), 353-364.