

## Binary Logistic Regression

A binomial logistic regression (often referred to simply as logistic regression), predicts 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.

For example, you could use binomial logistic regression to understand whether exam performance can be predicted based on revision time, test anxiety and lecture attendance (i.e., where the dependent variable is "exam performance", measured on a dichotomous scale – "passed" or "failed" – and you have three independent variables: "revision time", "test anxiety" and "lecture attendance"). Alternately, you could use binomial logistic regression to understand whether drug use can be predicted based on prior criminal convictions, drug use amongst friends, income, age and gender (i.e., where the dependent variable is "drug use", measured on a dichotomous scale – "yes" or "no" – and you have five independent variables: "prior criminal convictions", "drug use amongst friends", "income", "age" and "gender").

### Assumptions

When you choose to analyse your data using binomial logistic regression, part of the process involves checking to make sure that the data you want to analyse can actually be analysed using a binomial logistic regression. You need to do this because it is only appropriate to use a binomial logistic regression if your data "passes" seven assumptions that are required for binomial logistic regression to give you a valid result. In practice, checking for these seven assumptions just adds a little bit more time to your analysis, requiring you to click a few more buttons in SPSS Statistics when performing your analysis, as well as think a little bit more about your data, but it is not a difficult task.

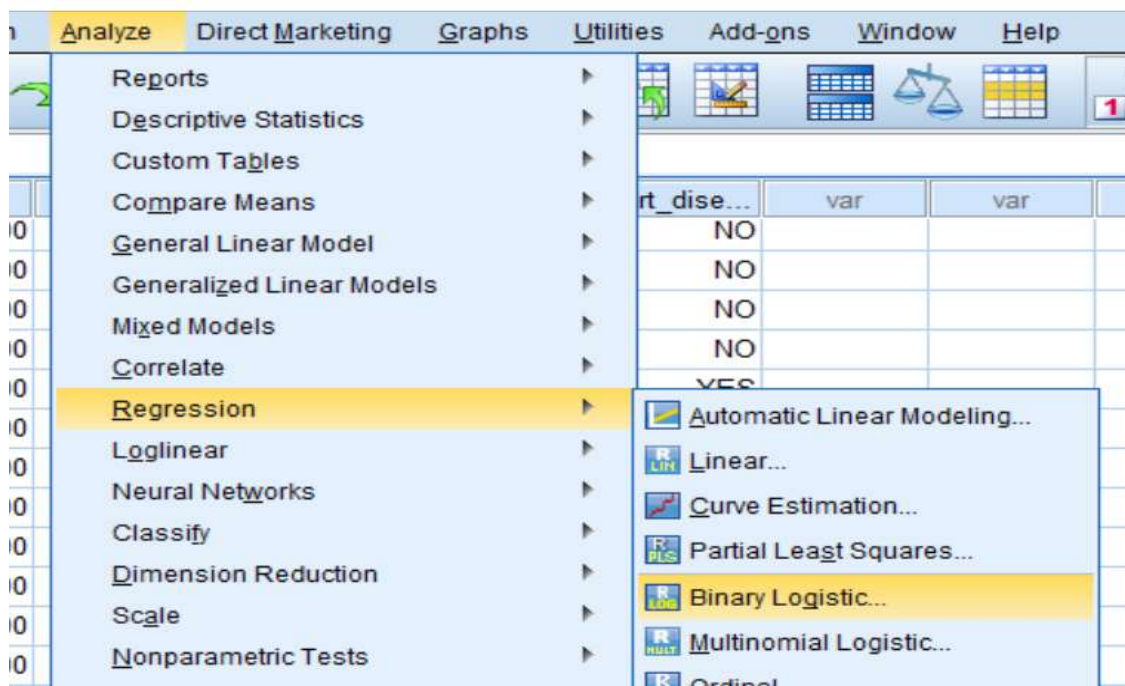
- Assumption #1: Your dependent variable should be measured on a dichotomous scale. Examples of dichotomous variables include gender (two groups: "males" and "females"), presence of heart disease (two groups: "yes" and "no"), personality type (two groups: "introversion" or "extroversion"), body composition (two groups: "obese" or "not obese"), and so forth. However, if your dependent variable was not measured on a dichotomous scale, but a continuous scale instead, you will need to carry out multiple regression, whereas if your dependent variable was measured on an ordinal scale, ordinal logistic regression would be a more appropriate starting point.
- Assumption #2: You have one or more independent variables, which can be either continuous (i.e., an interval or ratio variable) or categorical (i.e., an ordinal or nominal variable). Examples of continuous variables include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth. Examples of ordinal variables include Likert items (e.g., a 7-point scale from "strongly agree" through to "strongly disagree"), amongst other ways of ranking categories (e.g., a 3-point scale explaining how much a customer liked a product, ranging from "Not very much" to "Yes, a lot"). Examples of nominal variables include gender (e.g., 2 groups: male and female), ethnicity (e.g., 3 groups: Caucasian, African American and Hispanic), profession (e.g., 5 groups: surgeon, doctor, nurse, dentist, therapist), and so forth.
- Assumption #3: You should have independence of observations and the dependent variable should have mutually exclusive and exhaustive categories.
- Assumption #4: There needs to be a linear relationship between any continuous independent variables and the logit transformation of the dependent variable.


**Example:** A health researcher wants to be able to predict whether the "incidence of heart disease" can be predicted based on "age", "weight", "gender" and "VO2max" (i.e., where VO2max refers to maximal aerobic capacity, an indicator of fitness and health). To this end, the researcher recruited 40 participants to perform a maximum VO2max test as well as recording their age, weight and gender. The participants were also evaluated for the presence of heart disease. A binomial logistic regression was then run to determine whether the presence of heart disease could be predicted from their VO2max, age, weight and gender.

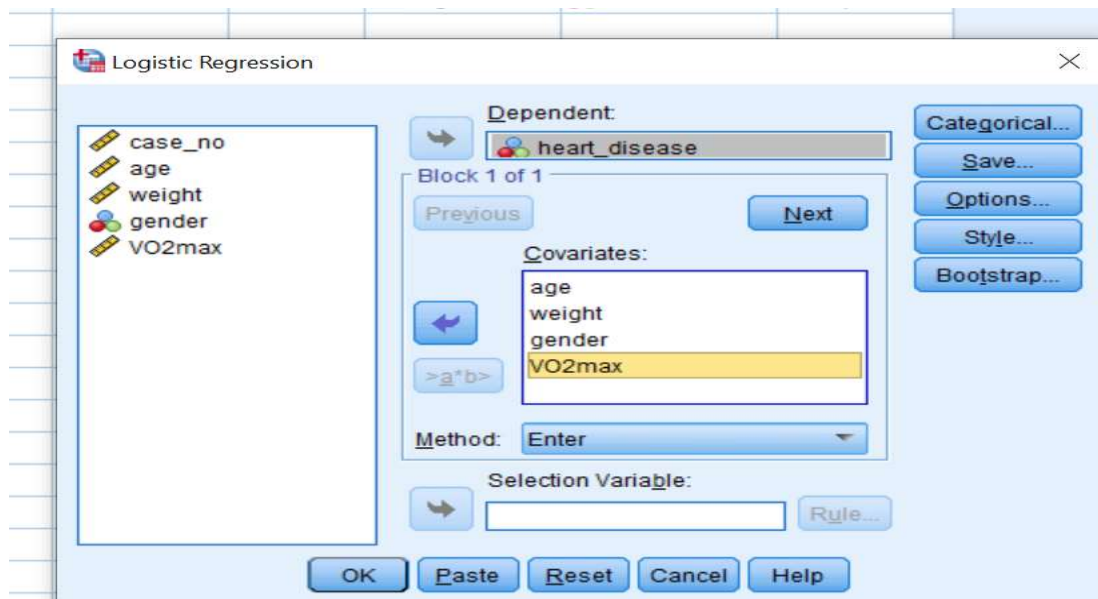
<u>case</u>	<u>Age</u>	<u>Wght</u>	<u>Gndr</u>	<u>VO2</u>	<u>heartd</u>	<u>case</u>	<u>Age</u>	<u>Wght</u>	<u>Gndr</u>	<u>VO2</u>	<u>heartd</u>
1,00	35,00	47,25	female	41,00	No	21,00	38,00	66,20	female	44,20	No
2,00	71,00	85,00	male	43,00	Yes	22,00	38,00	58,80	female	45,00	No
3,00	47,00	83,00	male	42,93	Yes	23,00	51,00	72,00	male	52,10	Yes
4,00	31,00	58,25	female	43,90	No	24,00	35,00	75,80	male	51,20	No
5,00	32,00	68,14	male	44,70	No	25,00	42,00	84,00	male	43,90	No
6,00	35,00	75,48	male	45,00	Yes	26,00	47,00	88,00	female	44,20	No
7,00	48,00	48,50	female	46,20	Yes	27,00	48,00	45,90	female	43,17	No
8,00	33,00	49,20	female	43,18	No	28,00	44,00	55,00	male	48,00	Yes
9,00	35,00	59,00	female	45,00	No	29,00	41,00	57,50	female	52,00	Yes
10,00	36,00	66,00	female	45,50	No	30,00	36,00	62,50	male	53,00	Yes
11,00	55,00	69,00	female	48,00	Yes	31,00	30,00	68,00	male	47,25	No
12,00	51,00	65,00	male	52,00	Yes	32,00	32,00	77,00	male	49,00	No
13,00	38,00	63,00	male	55,40	Yes	33,00	47,00	71,00	female	54,12	Yes
14,00	46,00	70,00	male	43,80	No	34,00	44,00	68,00	female	43,12	Yes
15,00	38,00	75,00	male	44,30	No	35,00	45,00	49,00	female	46,25	No
16,00	48,00	77,50	male	44,00	No	36,00	55,00	55,00	male	52,10	Yes
17,00	35,00	56,18	male	42,23	Yes	37,00	33,00	88,00	male	50,00	Yes
18,00	44,00	86,13	male	45,06	No	38,00	36,00	95,00	male	48,00	No
19,00	46,00	87,30	male	55,12	Yes	39,00	37,00	89,00	male	44,12	No
20,00	55,00	88,25	male	43,00	No	40,00	38,00	87,50	male	45,00	No

case_no	age	weight	gender	VO2max	heart_dise...	var	
13,00	51,00	63,00	FEMALE	55,40	NO		
14,00	38,00	70,00	MALE	43,80	NO		
15,00	46,00	75,00	MALE	44,30	NO		
16,00	38,00	77,50	MALE	44,00	NO		
17,00	31,00	56,18	MALE	42,23	YES		
18,00	47,00	86,13	MALE	45,06	NO		
19,00	40,00	87,30	MALE	55,12	NO		
20,00	46,00	88,25	MALE	43,00	NO		
21,00	52,00	66,20	FEMALE	44,20	NO		
22,00	38,00	58,80	FEMALE	45,00	NO		
23,00	36,00	72,00	MALE	52,10	YES		
24,00	35,00	75,80	MALE	51,20	NO		
25,00	42,00	84,00	MALE	43,90	NO		
26,00	47,00	88,00	FEMALE	44,20	NO		
27,00	48,00	45,90	FEMALE	43,17	NO		
28,00	44,00	55,00	MALE	48,00	YES		
29,00	41,00	57,50	FEMALE	52,00	YES		
30,00	36,00	62,50	MALE	53,00	NO		
31,00	30,00	68,00	MALE	47,25	NO		
32,00	32,00	77,00	MALE	49,00	NO		

[Click Analyze > Regression > Binary Logistic... on the main menu, as shown below:](#)

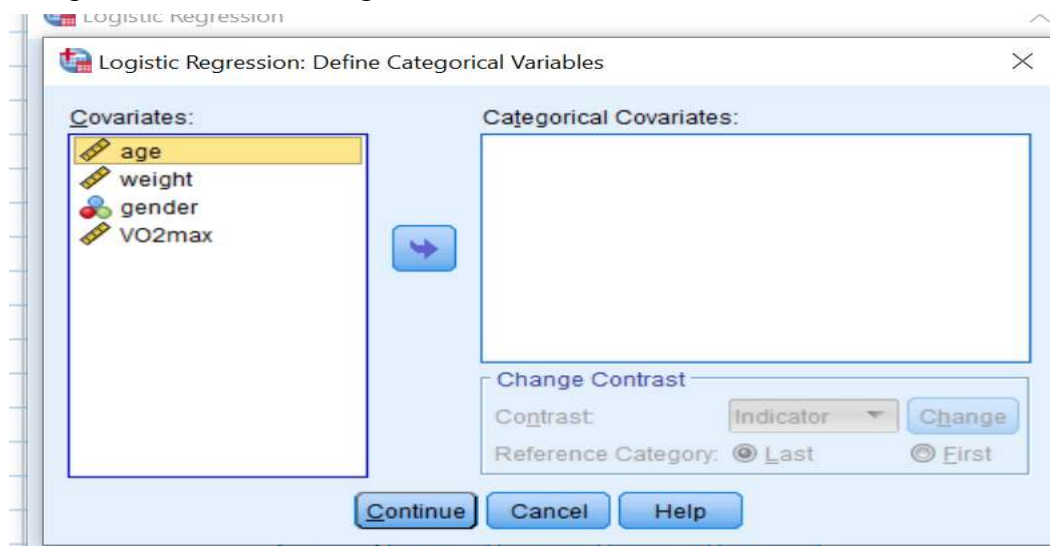


Transfer the dependent variable, heart\_disease, into the Dependent: box, and the independent-variables, age, weight, gender and VO2max into the Covariates: box, using the  buttons, as shown below:



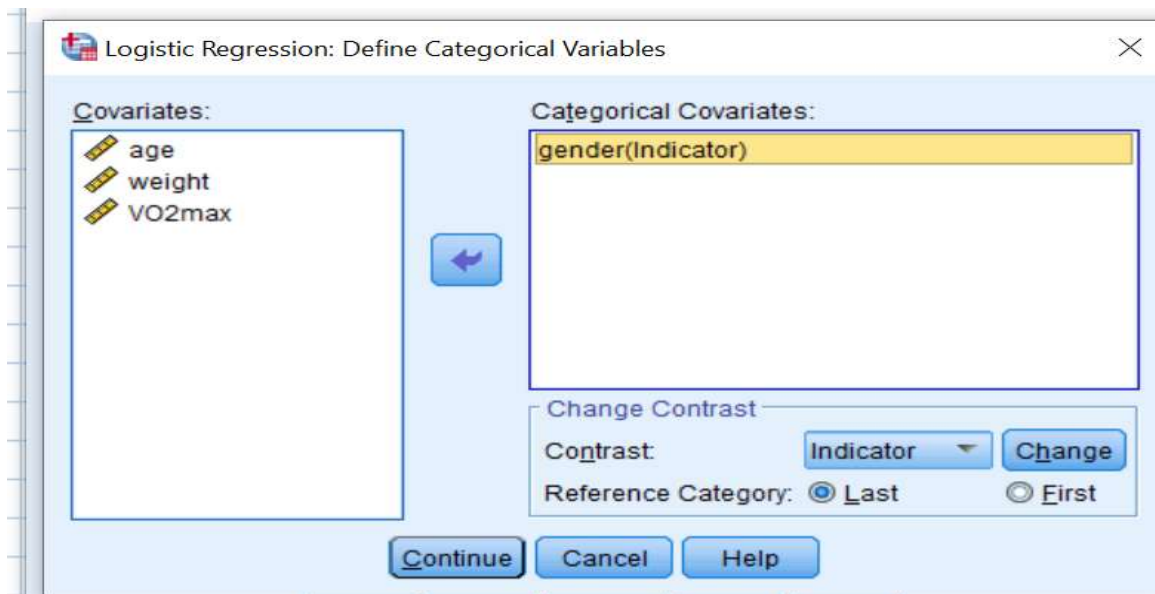
**Note:** For a standard logistic regression you should ignore the **Previous** and **Next** buttons because they are for sequential (hierarchical) logistic regression. The **Method:** option needs to be kept at the default value, which is **Enter**. If, for whatever reason, **Enter** is not selected,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

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



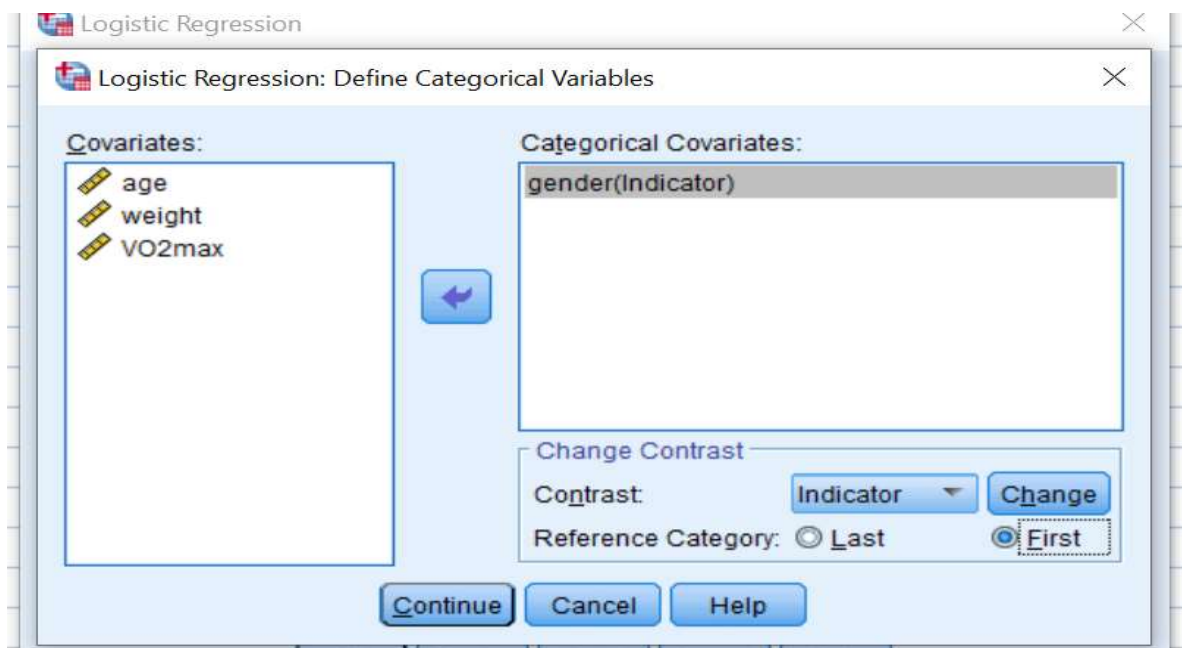
Note: SPSS Statistics requires you to define all the categorical predictor values in the logistic regression model. It does not do this automatically.

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



In the –Change Contrast– area, change the Reference Category: from the Last option to the First option. Then, click on 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.



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.

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

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 the one below:



**Logistic Regression: Options**

**Statistics and Plots**

☒ Classification plots ☐ Correlations of estimates

☒ Hosmer-Lemeshow goodness-of-fit ☐ Iteration history

☒ Casewise listing of residuals ☒ CI for exp(B) 95 %

☒ Outliers outside 2 std. dev. ☐ All cases

**Display**

☒ At each step ☐ At last step

**Probability for Stepwise**

Entry: 0,05 Removal: 0,10

Classification cutoff: 0,5

Maximum iterations: 20

☐ Conserve memory for complex analyses or large datasets

☒ Include constant in model

**Buttons:** Continue Cancel Help

## Variance explained

In order to understand how much variation in the dependent variable can be explained by the model (the equivalent of R<sup>2</sup> 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	30,025 <sup>a</sup>	,458	,616

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than ,001.

**Hosmer and Lemeshow Test**

Step	Chi-square	df	Sig.
1	10,574	8	,227

This table contains the Cox & Snell R Square and Nagelkerke R Square values, which are both methods of calculating the explained variation. These values are sometimes referred to as pseudo R<sup>2</sup> values (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 45% to 61%, depending on whether you reference the Cox & Snell R<sup>2</sup> or Nagelkerke R<sup>2</sup> methods, respectively. Nagelkerke R<sup>2</sup> is a modification of Cox & Snell R<sup>2</sup>, the latter of which cannot achieve a value of 1. For this reason, it is preferable to report the Nagelkerke R<sup>2</sup> value.

Accurate classification percentage and Hosmer-Lemeshow criteria are used to examine the goodness of fit of the model. This statistic tests the logistic regression model in general.

It is desirable that the model be NOT significant ( $p > 0.05$ ). It means that model fits the data.

## 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 binomial 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<sup>a</sup>**

Observed			Predicted		
			heartdisease		Percentage Correct
			No	Yes	
Step 1	heartdisease	No	20	3	87,0
		Yes	6	11	64,7
	Overall Percentage				77,5

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). Whilst the classification table appears to be very simple, it actually provides a lot of important information about your binomial logistic regression result, including:

- A. The percentage accuracy in classification (PAC), which reflects the percentage of cases that can be correctly classified as "no" heart disease with the independent variables added (not just the overall model).
- B. 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).
- C. Specificity, which 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).
- D. The positive predictive value, which is the percentage of correctly predicted cases "with" the observed characteristic compared to the total number of cases predicted as having the characteristic.
- E. The negative predictive value, which is the percentage of correctly predicted cases "without" the observed characteristic compared to the total number of cases predicted as not having the characteristic.

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 <sup>a</sup>	age	,225	,096	5,438	1	,020	1,252	1,037	1,513
	weight	-,023	,039	,339	1	,560	,978	,906	1,055
	gender(1)	-,736	1,150	,410	1	,522	,479	,050	4,559
	VO2	,448	,183	6,006	1	,014	1,565	1,094	2,240
	Constant	-28,445	11,787	5,824	1	,016	,000		

a. Variable(s) entered on step 1: age, weight, gender, VO2.

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 ( $p = .020$ ) and VO2max ( $p = .014$ ) added significantly to the model/prediction, but weight ( $p = .560$ ) and gender ( $p = .522$ ) did not add significantly to the model. You can use the information in the "Variables in the Equation" table to predict the probability of an event occurring based on a one unit change in an independent variable when all other independent variables are kept constant. For example, the table shows that the odds of having heart disease ("yes" category) is 0,479 times greater for males as opposed to females.

Based on the results above, we could report the results of the study as follows

A logistic regression was performed to ascertain the effects of age, weight, gender and VO2max on the likelihood that participants have heart disease. The model explained 61.0% (Nagelkerke R<sup>2</sup>) of the variance in heart disease and correctly classified 77.5% of cases. Males were 0,479 times more likely to exhibit heart disease than females. Increasing age and increasing VO2 were associated with an increased likelihood of exhibiting heart disease.