

# Logistic regression with XLSTAT

[demoLOG.xls](#)

## Logistic regression

Logistic regression, and related methods such as Probit analysis, are very useful techniques when one wants to understand or to predict the effect of a series of variables on a binary response variable (a variable which can take only two values, 0/1 or Yes/no, for example). Logistic regression can be helpful to model the effect of doses in medicine or agriculture, or to anticipate the likelihood of customers responding to a direct mail, or to evaluate the risk for a bank that a client will not pay back a loan.

With XLSTAT you can either run the logistic regression on raw data (the response is given as 0s and 1s) or on aggregated data (the response is a sum of "successes" or ones, and the number of repetitions must also be available).

ID	Temp	Response
1	41	0
2	42	0
3	43	0
4	44	0
5	45	1
6	46	0
7	47	1
8	48	1
9	49	1
10	50	1

Example of raw data - (effect of temperature on the resistance of a chip)

OBS	Log-DOSE	TESTED	KILLED
1	1.691	59	6
2	1.724	60	13
3	1.755	62	18
4	1.784	56	28
5	1.811	63	52
6	1.837	59	53
7	1.861	62	61
8	1.884	60	60

Example of aggregated data - (effect of an insecticide on a specific species of insects)

Note that Addinsoft has developed a specific module for dose analysis. This module is called XLSTAT-Dose and can be ordered separately.

The methodology of logistic regression aims at modeling the probability of success depending on the values of the explanatory variables, which can be categorical or numerical variables.

## Dataset for logistic regression

The example treated here is a marketing case where we want to detect if customers are likely to renew their subscription for an online sports information service. An Excel sheet with both the data and the results can be downloaded by clicking [here](#).

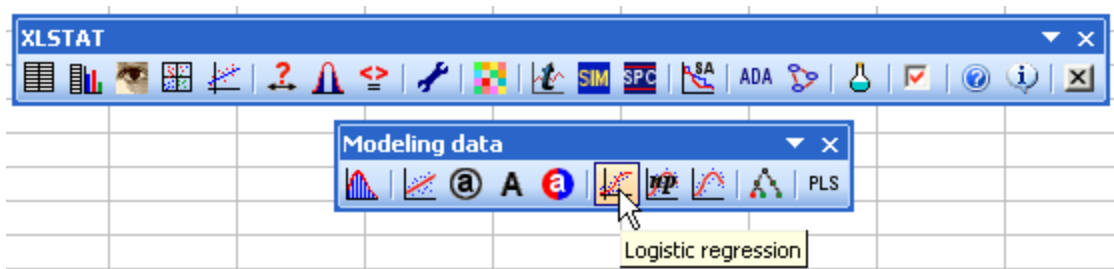
The data show a sample of 60 users, with their age category, the average number of pages viewed per week, and the number of pages viewed during the previous week. These users have been offered the possibility of renewing their subscription, which is expiring in two weeks. Our goal is to understand why some renewed their subscription and others did not.

## Goal of the logistic regression

Using a Logistic regression model, we want to explain the results we obtained, and then use the model on the whole population in order to identify the users who might not renew the subscription. These users could be targeted and offered an incentive (an added service, for example) to renew.

## Setting up a logistic regression

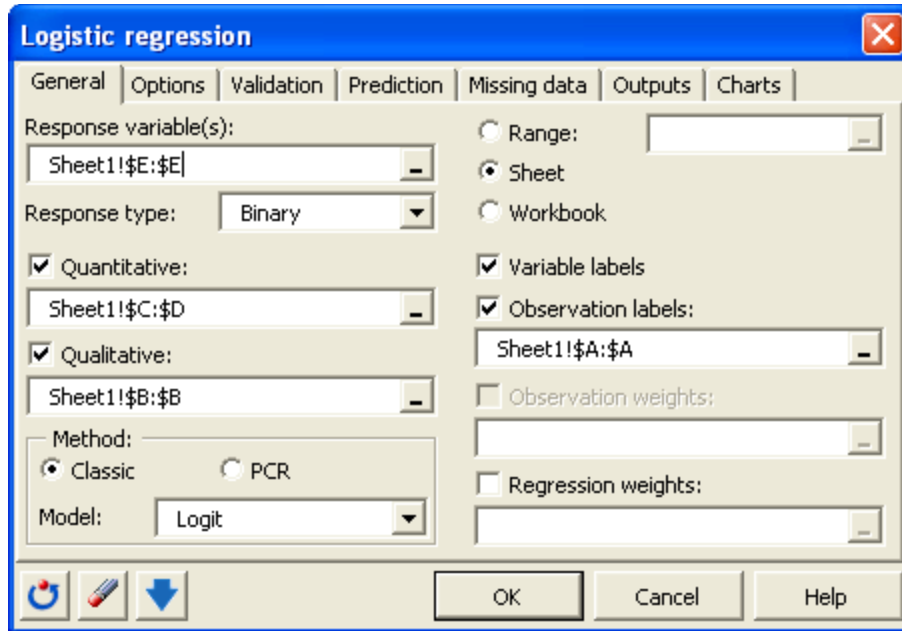
To activate the Logistic regression dialog box, start XLSTAT, then select the **XLSTAT / Modeling data / Logistic regression** data command, or click on the corresponding button of the **Modeling Data** toolbar (see below).



When you click on the button, the **Logistic regression** dialog box appears. Select the data on the Excel sheet.

The **Response** corresponds to the column where the binary variable or the counts of positive cases are stored (NB: when using aggregated data the **Weights** must be selected). In this particular case we have three explanatory variables, one categorical - the age group - and two numerical: the pages counts.

As we selected the column titles of all variables, we have selected the option **Column labels included**.



The computations begin once you have clicked on the **OK** button.

## Interpretation of the logistic regression

The following table gives details on the model. This table is helpful in understanding the effect of the various variables and the relative effects of the age categories.

Model parameters (Variable Renewed):

Source	Value	Standard error	Wald Chi-Square	Pr > Chi <sup>2</sup>	Wald Lower bound (5%)	Wald Upper bound (5%)
Intercept	-2,357	1,246	3,579	0,059	-2,435	-2,277
AvPages/Week	0,023	0,026	0,826	0,363	0,022	0,024
Pages/Week	0,089	0,036	6,274	0,012	0,087	0,091
Age-18-24	0,000	0,000				
Age-25-29	0,688	1,206	0,325	0,569	0,612	0,764
Age-30-39	0,963	1,276	0,570	0,450	0,883	1,043
Age-40-49	-2,983	1,377	4,690	0,030	-3,069	-2,897
Age-50-59	1,086	1,168	0,864	0,352	1,012	1,159
Age-60+	0,309	1,264	0,060	0,807	0,230	0,388

On this table we can see from looking at the probability of the Chi-squares that the variable most influencing renewal is the number of pages viewed the previous month. The intercept is significant, and the fact that the customer's age is between 40-49 also has a strong negative influence on subscription renewal. This last point needs to be interpreted by marketing people so that the right action can be taken towards this specific population.

The next table gives several indicators of the quality of the model (or goodness of fit). These results are equivalent to the  $R^2$  and to the analysis of variance table in linear regression and ANOVA. The most important value to look at is the probability of Chi-square test on the log ratio. This is equivalent to the Fisher's F test: we try to evaluate if the variables bring significant information by comparing the model as it is defined with a simpler model with only one constant. In this case, as the probability is lower than 0.0001, we can conclude that significant information is brought by the variables.

Goodness of fit statistics (Variable Renewed):			
Statistic	Independent	Full	
Observations	60	60	
Sum of weights	60,000	60,000	
DF	59	52	
-2 Log(Likelihood)	80,761	46,943	
R <sup>2</sup> (McFadden)	0,000	0,419	
R <sup>2</sup> (Cox and Snell)	0,000	0,431	
R <sup>2</sup> (Nagelkerke)	0,000	0,582	
AIC	84,761	62,943	
SBC	88,950	79,698	
Iterations	0	7	
Test of the null hypothesis H0: Y=0,600 (Variable Renewed)			
Statistic	DF	Chi-square	Pr > Chi <sup>2</sup>
-2 Log(Likelihood)	7	33,819	< 0,0001
Score	7	23,816	0,001
Wald	7	12,623	0,082

The last step is the application of the model on the whole population. In this case the model writes:

$Y = \text{Exp}(L(x)) / [1 + \text{Exp}(L(x))]$ , where  $L(x) = -2.3567 + 0.0235 \cdot \text{AvPage/Week} + 0.0893 \cdot \text{Page/Week} + \text{Factor}$  and Factor takes the value of the parameter corresponding to the age group to which a customer belongs.

When we applied the model to the 600 customers who needed to renew their subscription the following month, we found that only 40% would renew. By taking the right marketing actions, we were able to boost the result to 85%!