



Some Basics of Logistic Regression

by Katrien Declercq

Academic

Suppose a company that sells products via a catalog wants to identify those customers to whom advertising efforts should be directed. It has been decided that customers who spend 100 dollars or more are the target group. Based on orders received over the last six months, the company wants to characterize this group of customers. For each order, the data available are the purchase price (recoded to 1 if purchase \$100 or more, 0 if purchase under \$100), the age of the customer in years, the gender of the customer and the annual income (Low, Middle, High). So, to help the company to answer their question, an analysis has to be done to identify what factors (age, gender, income) determine whether a customer purchases for \$100 or more (i.e. show the response 1).

In general, regression analysis enables to characterize the relationship between a response variable and one or more predictor variables. In linear regression, the response variable is continuous (cfr. previous newsletter) and in logistic regression, the response variable is categorical, so that's the type of analysis needed here). The logistic regression model uses the predictor variables (categorical or continuous) to predict the probability of specific outcomes. To ensure that the estimated probabilities are between 0 and 1 and since the typical relationship between the probability of the outcome and a predictor variable is nonlinear (S-shaped) rather than linear, logistic regression transforms the probabilities (p) with the logit transformation ($\text{logit}(p) = \log(p/(1-p))$) (where log is the natural logarithm). In SAS, logistic regression can be performed using the LOGISTIC procedure or the GENMOD procedure.

```
data b_sales;
  input purchase age gender $ income $ @@;
  datalines;
0 41 F L 0 47 F L 1 41 F L 1 39 F L 0 32 F L 0 32 F L 0 33 F L
0 45 F L 0 43 F L 0 40 F L 0 39 F L 1 46 F L 0 38 F L 0 38 F L
1 44 F L 0 39 F L 0 26 F L 0 45 F L 0 40 F L 0 35 F L 1 30 F L
0 44 F L 1 47 F L 1 35 F L 0 36 F L 0 34 F L 1 38 F L 0 38 F L
0 47 F L 0 41 F L 0 45 F L 1 37 F L 1 38 F L 1 29 F L 1 40 F L
0 35 F L 0 44 F L 0 39 F L 1 50 F L 1 41 F L 0 41 F L 0 40 F L
1 38 F L 1 36 F L 1 37 F L 0 39 F L 0 41 F L 1 33 F L 1 51 F L
0 31 F L 0 31 F L 0 35 F L 0 46 F L 1 39 F L 0 47 F L 0 40 F L
1 56 F L 0 36 F L 0 37 F L 0 36 F L 0 44 F L 1 36 F L 0 40 F L
1 38 F L 0 35 F L 1 35 F L 0 45 F L 0 41 F L 0 42 F L 1 37 F L
0 55 F L 1 41 F L 1 33 F L 1 36 F L 0 33 F L 0 38 F L 0 38 F L
0 55 F L 0 50 F L 1 31 F L 1 37 F L 1 37 F L 0 38 F L 0 41 F L
0 42 F L 1 34 F L 1 31 F L 0 28 F L 1 35 F L 1 39 F M 0 39 F M
1 34 F M 1 45 F M 0 41 F M 0 42 F M 0 46 F M 0 42 F M 0 33 F M
1 47 F M 1 39 F M 0 41 F M 0 33 F M 0 43 F M 1 31 F M 1 43 F M
0 45 F M 1 48 F M 0 40 F M 1 35 F M 0 33 F M 1 34 F M 1 43 F M
0 35 F M 1 31 F M 0 48 F M 1 36 F M 0 39 F M 0 37 F M 0 40 F M
0 39 F M 1 44 F M 0 35 F M 0 33 F M 1 34 F M 0 50 F M 1 44 F M
0 40 F M 0 36 F M 1 41 F M 1 37 F M 1 38 F M 0 38 F M 1 41 F M
0 36 F M 0 42 F M 1 43 F M 0 43 F M 0 50 F M 0 42 F M 0 42 F M
0 38 F M 0 33 F M 0 39 F M 1 41 F M 1 44 F M 0 42 F M 1 33 F M
1 49 F M 0 40 F M 1 37 F M 0 38 F M 0 36 F M 0 39 F M 1 34 F M
0 44 F M 1 34 F M 0 39 F M 1 38 F M 1 34 F M 1 52 F M 1 41 F H
0 41 F H 0 37 F H 0 45 F H 0 44 F H 1 35 F H 0 30 F H 0 28 F H
1 40 F H 0 33 F H 1 44 F H 0 39 F H 0 42 F H 0 31 F H 0 29 F H
0 49 F H 1 33 F H 1 35 F H 0 34 F H 1 40 F H 1 40 F H 0 32 F H
1 38 F H 1 23 F H 1 49 F H 1 39 F H 0 32 F H 0 43 F H 1 38 F H
```

```

0 34 F H 0 38 F H 0 35 F H 0 45 F H 1 28 F H 1 35 F H 0 26 F H
1 38 F H 1 34 F H 1 46 F H 0 35 F H 0 36 F H 0 37 F H 0 48 F H
0 41 F H 0 40 F H 1 51 F H 1 45 F H 0 33 F H 1 37 F H 1 36 F H
0 42 F H 1 51 F H 1 51 F H 1 41 F H 1 26 F H 1 49 F H 0 46 F H
0 41 F H 0 40 F H 1 44 F H 0 35 F H 0 40 F H 0 37 F H 1 32 F H
1 45 F H 1 29 F H 1 41 F H 1 39 F H 0 26 F H 0 30 F H 1 38 F H
0 35 F H 0 39 F H 0 24 F H 0 32 F H 0 46 F H 1 40 F H 0 38 F H
1 42 F H 0 43 F H 0 29 M L 0 58 M L 0 40 M L 0 35 M L 0 47 M L
0 33 M L 0 40 M L 0 37 M L 0 37 M L 0 40 M L 0 36 M L 0 32 M L
0 37 M L 1 49 M L 0 38 M L 1 36 M L 0 42 M L 0 44 M L 0 38 M L
0 40 M L 0 41 M L 0 40 M L 0 47 M L 1 39 M L 0 38 M L 0 43 M L
0 46 M L 1 48 M L 1 38 M L 0 35 M L 0 38 M L 0 37 M L 0 32 M L
0 29 M L 1 41 M L 1 39 M L 0 40 M L 1 48 M L 0 37 M L 0 40 M L
0 42 M L 0 42 M L 0 38 M L 0 46 M M 0 45 M M 0 33 M M 0 34 M M
1 33 M M 0 42 M M 1 47 M M 0 43 M M 0 35 M M 0 43 M M 1 33 M M
0 26 M M 0 34 M M 0 43 M M 0 41 M M 0 38 M M 0 25 M M 1 45 M M
0 37 M M 0 34 M M 0 39 M M 0 32 M M 0 44 M M 0 39 M M 0 30 M M
0 36 M M 0 45 M M 0 39 M M 0 44 M M 0 47 M M 0 46 M M 1 34 M M
1 38 M M 0 31 M M 0 34 M M 1 33 M M 0 38 M M 1 43 M M 0 43 M M
1 47 M M 0 41 M M 1 37 M M 1 38 M M 0 38 M M 1 35 M M 0 34 M M
0 33 M M 0 40 M M 0 39 M M 0 41 M M 1 44 M M 0 37 M M 0 41 M M
0 45 M M 0 29 M M 0 40 M M 0 40 M M 0 35 M M 0 49 M M 0 40 M M
0 38 M M 0 41 M M 0 35 M M 0 33 M M 0 47 M M 0 36 M M 0 37 M M
0 40 M M 1 44 M M 0 48 M M 1 43 M M 0 41 M M 0 30 M M 0 42 M H
0 33 M H 1 41 M H 0 33 M H 0 37 M H 1 42 M H 0 25 M H 1 39 M H
0 39 M H 0 36 M H 0 43 M H 0 41 M H 0 34 M H 0 37 M H 1 46 M H
1 42 M H 0 40 M H 0 41 M H 1 39 M H 1 34 M H 1 35 M H 1 35 M H
0 44 M H 1 45 M H 1 38 M H 0 39 M H 0 38 M H 0 43 M H 0 31 M H
1 45 M H 1 36 M H 1 33 M H 1 49 M H 1 41 M H 1 41 M H 1 31 M H
0 34 M H 1 39 M H 0 44 M H 1 46 M H 0 32 M H 1 52 M H 0 41 M H
1 48 M H 0 50 M H 1 38 M H 0 36 M H 1 34 M H 1 37 M H 0 38 M H
0 42 M H 1 33 M H 1 35 M H 0 30 M H 0 48 M H 1 35 M H 0 44 M H
0 41 M H 0 37 M H 1 35 M H 0 46 M H 0 37 M H 1 47 M H 1 39 M H
1 47 M H 1 45 M H 1 37 M H 1 37 M H 1 33 M H 0 40 M H 1 43 M H
0 31 M H 1 32 M H 0 34 M H 0 34 M H
;;;

```

```

options nonumber nodate;
PROC LOGISTIC DATA=b_sales descending;
  CLASS gender (param=ref ref='M');
  MODEL purchase = gender;
  TITLE 'LOGISTIC REGRESSION MODEL: Purchase = Gender';
  title2 'Income in <H>igh, <M>edium or <L>ow';
RUN;

```

The descending option will model the probability that a customer places an order of \$100 or more (response 1). Otherwise, by default, the response 0 would be modeled. The param option specifies the parametrization of the model that will be used, which in this example is reference cell coding, i.e. the females will be compared to the males (reference group because of ref='Male').

The first part of the output shows information on the model, response values and predictor variable values (design variables). The next section shows information on the model convergence by giving three tests that allow to compare one model to another; lower values indicate a more desirable model. The lower part of the output shows a test of the global null hypothesis that all parameters of the model are 0. If the model shows to be significant here, it makes sense to look into more detail at the tests for the individual effects in the model (next part of the output). In the model we specified, only one predictor variable was present, of course income and age can also be added to the model (and eventually interactions). PROC LOGISTIC has some model selection techniques available for selecting the "best" model amongst a series of possible models.

Much more about this in the Online Doc.