Contents

What is Logistic Regression?

Logistic regression is a discriminitive probabilistic classification model that operates over real-valued vector inputs.

The dimensions of the input vectors being classified are called "features" and there is no restriction against them being correlated.

Logistic regression is one of the best probabilistic classifiers, measured in both log loss and first-best classification accuracy across a number of tasks.

Logistic regression requires extensive tuning in the form of feature selection and implementation to achieve state-of-the-art classification performance.

1 Logistic Regression

Logistic regression determines the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories.

1.1 The purpose of logistic regression

The crucial limitation of linear regression is that it cannot deal with Dependent Variabless that are ***dichotomous*** and categorical. Many interesting variables in the business world are dichotomous: for example, consumers make a decision to buy or not buy (***Buy/Don't Buy***), a product may pass or fail quality control (***Pass/Fail***), there are good or poor credit risks (***Good/Poor***), an employee may be promoted or not (***Promote/Don't Promote***).

A range of regression techniques have been developed for analysing data with categorical dependent variables, including logistic regression and discriminant analysis (Hence referred to as DA, which is the next section of course).

Logistical regression is regularly used rather than discriminant analysis when there are only two categories for the dependent variable. Logistic regression is also easier to use with SPSS than DA when there is a mixture of numerical and categorical Independent Variabless, because it includes procedures for generating the necessary dummy variables automatically, requires fewer assumptions, and is more statistically robust. DA strictly requires the continuous independent variables (though dummy variables can be used as in multiple regression). Thus, in instances where the independent variables are categorical, or a mix of continuous and categorical, and the DV is categorical, logistic regression is necessary.

1.2 Use of Binomial Probability Theory

Since the dependent variable is dichotomous we cannot predict a numerical value for it using logistic regression, so the usual regression least squares deviations criteria for best fit approach of minimizing error around the line of best fit is inappropriate.

Instead, logistic regression employs binomial probability theory in which there are only two values to predict: that probability (p) is 1 rather than 0, i.e. the event/person belongs to one group rather than the other. Logistic regression forms a best fitting equation or function using the maximum likelihood method (not part of course), which maximizes the probability of classifying the observed data into the appropriate category given the regression coefficients.

1.3 Variable Selection

Like ordinary regression, logistic regression provides a coefficient **b** estimates, which measures each IVs partial contribution to variations in the response variables. The goal is to correctly predict the category of outcome for individual cases using the most parsimonious model.

To accomplish this goal, a model (i.e. an equation) is created that includes all predictor variables that are useful in predicting the response variable. Variables can, if necessary, be entered into the model in the order specified by the researcher in a stepwise fashion like regression.

There are two main uses of logistic regression:

- The first is the prediction of group membership. Since logistic regression calculates the probability of success over the probability of failure, the results of the analysis are in the form of an **odds ratio**.
- Logistic regression also provides knowledge of the relationships and strengths among the variables (e.g. playing golf with the boss puts you at a higher probability for job promotion than undertaking five hours unpaid overtime each week).

1.4 Assumptions of logistic regression

- Logistic regression does not assume a linear relationship between the dependent and independent variables.
- The dependent variable must be a dichotomy (2 categories). (Remark: Dichotomous refers to two outcomes. Multichotomous refers to more than two outcomes).
- The independent variables need not be interval, nor normally distributed, nor linearly related, nor of equal variance within each group.
- The categories (groups) must be mutually exclusive and exhaustive; a case can only be in one group and every case must be a member of one of the groups.
- Larger samples are needed than for linear regression because maximum likelihood coefficients are large sample estimates. A minimum of 50 cases per predictor is recommended.

1.5 Odds and Odds Ratio

Logistic regression calculates changes in the log odds of the dependent, not changes in the dependent value as OLS regression does. For a dichotomous variable the odds of membership of the target group are equal to the probability of membership in the target group divided by the probability of membership in the other group. Odds value can range from 0 to infinity and tell you how much more likely it is that an observation is a member of the target group rather than a member of the other group. If the probability is 0.80, the odds are 4 to 1 or 0.80/0.20; if the probability is 0.25, the odds are .33 (0.25/0.75).

If the probability of membership in the target group is 0.50, the odds are 1 to 1 (0.50/0.50), as in coin tossing when both outcomes are equally likely.

Another important concept is the odds ratio (OR), which estimates the change in the odds of membership in the target group for a one unit increase in the predictor. It is calculated by using the regression coefficient of the predictor as the exponent. Suppose we were predicting

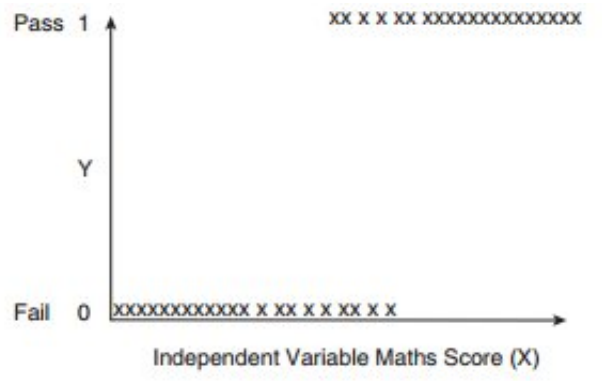

Figure 1: Accountancy Exam Results

exam success by a maths competency predictor with an estimate b = 2.69. Thus the odds ratio is exp(2.69) or 14.73. Therefore the odds of passing are 14.73 times greater for a student, for example, who had a pre-test score of 5, than for a student whose pre-test score was 4.

1.6 The Sigmoid Graph

While logistic regression gives each predictor (IV) a coefficient **b** which measures its independent contribution to variations in the dependent variable, the dependent variable can only take on one of the two values: 0 or 1.

What we want to predict from a knowledge of relevant independent variables and coefficients is therefore not a numerical value of a dependent variable as in linear regression, but rather the probability (p) that it is 1 rather than 0 (belonging to one group rather than the other). But even to use probability as the dependent variable is unsound, mainly because numerical predictors may be unlimited in range. If we expressed p as a linear function of investment, we might then find ourselves predicting that p is greater than 1 (which cannot be true, as probabilities can only take values between 0 and 1). Additionally, because logistic regression has only two y values in the category or not in the category a straight line best fit (as in linear regression) is not possible to draw.

1.6.1 Hypothetical Example

Consider the following hypothetical example: 200 accountancy first year students are graded on a pass-fail dichotomy on the end of the semester accountancy exam. At the start of the course, they all took a maths pre-test with results reported in interval data ranging from 050 the higher the pretest score the more competency in maths. Logistic regression is applied to determine the relationship between maths pretest score (IV or predictor) and whether a student passed the course (DV). Students who passed the accountancy course are coded 1 while those who failed are coded 0.

We can see from Figure 1 of the plotted xs that there is somewhat greater likelihood that those who obtained above average to high score on the maths test passed the accountancy course, while below average to low scorers tended to fail. There is also an overlap in the middle

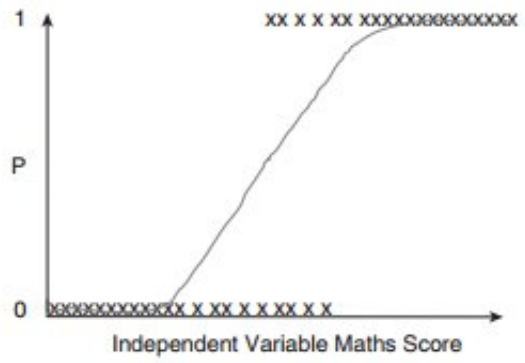

Figure 2: Accountancy Exam Results - Fitted Curve

area. But if we tried to draw a straight (best fitting) line, as with linear regression, it just would not work, as intersections of the maths results and pass/fail accountancy results form two lines of xs, as in Figure 1.

The solution is to convert or transform these results into probabilities. We might compute the average of the Y values at each point on the X axis. We could then plot the probabilities of Y at each value of X and it would look something like the wavy graph line superimposed on the original data in Figure 2. This is a smoother curve, and it is easy to see that the probability of passing the accountancy course (Y axis) increases as values of X increase. What we have just done is transform the scores so that the curve now fits a cumulative probability curve, i.e. adding each new probability to the existing total. As you can see, this curve is not a straight line; it is more of an s-shaped curve (A sigmoid curve).

Predicted values are interpreted as probabilities and are now not just two conditions with a value of either 0 or 1 but continuous data that can take any value from 0 to 1. The slope of the curve in Figure 2 is low at the lower and upper extremes of the independent variable and greatest in the middle where it is most sensitive. In the middle, of course, are a number of cases that are out of order, in the sense that there is an overlap with average maths scores in both accountancy pass and fail categories, while at the extremes are cases which are almost universally allocated to their correct group. The outcome is not a prediction of a Y value, as in linear regression, but a probability of belonging to one of two conditions of Y, which can take on any value between 0 and 1 rather than just 0 and 1 in Figure 1.

1.6.2 Log transformation

Unfortunately a further mathematical transformation a log transformation is needed to normalize the distribution. Transformations, such as log transformations and square root transformations transform non-normal/skewed distributions closer to normality.

This log transformation of the p values to a log distribution enables us to create a link with the normal regression equation. The log distribution (or logistic transformation of p) is also called the logit of p or ***logit(p)***.

1.7 The Logit

The convention for binomial logistic regression is to code the dependent class of greatest interest as 1 and the other class as 0, because the coding will affect the odds ratios and slope estimates.

The logit(p) is the log (to base e) of the odds ratio or likelihood ratio that the dependent variable is 1. In symbols it is defined as:

$$logit(p) = ln\left(\frac{p}{(1-p)}\right)$$

Whereas p can only range from 0 to 1, logit(p) scale ranges from negative infinity to positive infinity and is symmetrical around the logit of 0.5 (which is zero)

1.8 The Logistic Regression Equation

The form of the logistic regression equation is:

$$logit[p(x)] = log\frac{p(x)}{1-p(x)} = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots$$

This looks just like a linear regression and although logistic regression finds a best fitting equation, just as linear regression does, the principles on which it does so are rather different. Instead of using a least-squared deviations criterion for the best fit, it uses a maximum likelihood method, which maximizes the probability of getting the observed results given the fitted regression coefficients. A consequence of this is that the goodness of fit and overall significance statistics used in logistic regression are different from those used in linear regression.

The probability that a case is in a particular category,p, can be calculated with the following formula (which is simply another rearrangement of the previous formula).

$$p = \frac{exp(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots)}{1 + exp(b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \ldots)}$$

1.9 Exercise Data Set

The exercise data set comes from a survey of home owners conducted by an electricity company about an offer of roof solar panels with a 50% subsidy from the state government as part of the states environmental policy. The variables involve household income measured in units of a thousand dollars, age, monthly mortgage, size of family household, and as the dependent variable, whether the householder would take or decline the offer. The purpose of the exercise is to conduct a logistic regression to determine whether family size and monthly mortgage will predict taking or declining the offer.

For the first demonstration, we will use 'family size and 'mortgage only. For the options, select Classification Plots, Hosmer-Lemeshow Goodness Of Fit, Casewise Listing Of Residuals and select Outliers Outside 2sd. Retain default entries for probability of stepwise, classification cutoff and maximum iterations.

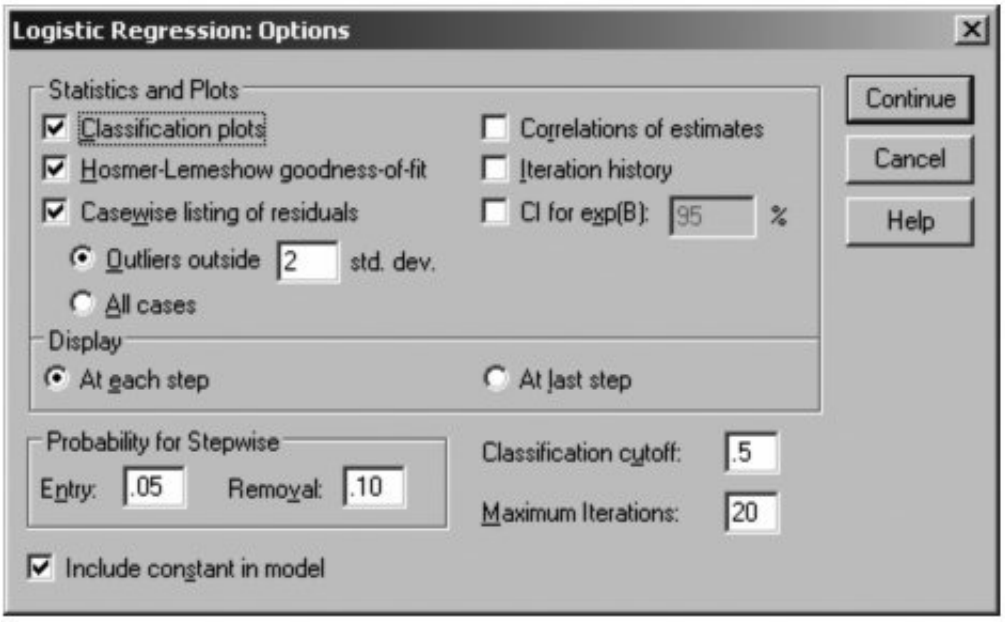

Figure 3: Selected Options for Exercises

We are not using any categorical variables this time. If there are categorical variables, use the ***categorical*** option. For most situations, choose the indicator coding scheme (it is the default).

1.10 SPSS Outout - Block 0: Beginning Block.

Block 0 presents the results with only the constant included before any coefficients (i.e. those relating to family size and mortgage) are entered into the equation. Logistic regression compares this model with a model including all the predictors (family size and mortgage) to determine whether the latter model is more appropriate. The table suggests that if we knew nothing about our variables and guessed that a person would not take the offer we would be correct 53.3% of the time. The variables not in the equation table tells us whether each IV improves the model. The answer is yes for both variables, with family size slightly better than mortgage

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			take solar panel offer		
			decline offer	take offer	Percentage Correct
Step 0	take solar panel offer	decline offer	0	14	.0
		take offer	0	16	100.0
	Overall Percentage				53.3

a. Constant is included in the model.
b. The cut value is .500

Figure 4: Classification table

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.134	.366	.133	1	.715	1.143

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Mortgage	6.520	1	.011
		Famsize	14.632	1	.000
	Overall Statistics		15.085	2	.001

Figure 5: Variables in / not in the equation

size, as both are significant and if included would add to the predictive power of the model. If they had not been significant and able to contribute to the prediction, then termination of the analysis would obviously occur at this point

This presents the results when the predictors family size and mortgage are included. Later SPSS prints a classification table which shows how the classification error rate has changed from the original 53.3we can now predict with 90% accuracy (see Classification Table later). The model appears good, but we need to evaluate model fit and significance as well. SPSS will offer you a variety of statistical tests for model fit and whether each of the independent variables included make a significant contribution to the model.

1.11 Omnibus Test for Model Coefficients

The overall significance is tested using what SPSS calls the ***Model Chi-square***, which is derived from the likelihood of observing the actual data under the assumption that the model that has been fitted is accurate. There are two hypotheses to test in relation to the overall fit of the model:

H_0 The model is a good fitting model.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	24.096	2	.000
	Block	24.096	2	.000
	Model	24.096	2	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	17.359[a]	.552	.737

a. Estimation terminated at iteration number 8 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.378	8	.605

Figure 6: Test Outcomes

H_1 The model is not a good fitting model (i.e. the predictors have a significant effect).

In our case model chi square has 2 degrees of freedom, a value of 24.096 and a probability of $p < 0.000$.

Thus, the indication is that the model has a poor fit, with the model containing only the constant indicating that the predictors do have a significant effect and create essentially a different model. So we need to look closely at the predictors and from later tables determine if one or both are significant predictors.

This table has 1 step. This is because we are entering both variables and at the same time providing only one model to compare with the constant model. In stepwise logistic regression there are a number of steps listed in the table as each variable is added or removed, creating different models. The step is a measure of the improvement in the predictive power of the model since the previous step. (I will revert to this next class).

1.12 Model Summary Table

The likelihood function can be thought of as a measure of how well a candidate model fits the data (although that is a very simplistic definition). The AIC criterion is based on the Likelihood function. The likelihood function of a fitted model is commonly re-expressed as -2LL (i.e. The log of the likelihood times minus 2). The 2LL value from the Model Summary table below is 17.359.

Although there is no close analogous statistic in logistic regression to the coefficient of determination R^2 the Model Summary Table provides some approximations. Cox and Snells R-Square attempts to imitate multiple R-Square based on likelihood, but its maximum can be (and usually is) less than 1.0, making it difficult to interpret. Here it is indicating that 55.2% of the variation in the DV is explained by the logistic model. The Nagelkerke modification

that does range from 0 to 1 is a more reliable measure of the relationship. Nagelkerkes R^2 will normally be higher than the Cox and Snell measure. Nagelkerkes R^2 is part of SPSS output in the Model Summary table and is the most-reported of the R-squared estimates. In our case it is 0.737, indicating a moderately strong relationship of 73.7% between the predictors and the prediction.

1.13 Hosmer and Lemeshow Statistic

An alternative to model chi square is the Hosmer and Lemeshow test which divides subjects into 10 ordered groups of subjects and then compares the number actually in the each group (observed) to the number predicted by the logistic regression model (predicted). The 10 ordered groups are created based on their estimated probability; those with estimated probability below .1 form one group, and so on, up to those with probability .9 to 1.0.

Each of these categories is further divided into two groups based on the actual observed outcome variable (success, failure). The expected frequencies for each of the cells are obtained from the model. A probability (p) value is computed from the chi-square distribution with 8 degrees of freedom to test the fit of the logistic model.

If the H-L goodness-of-fit test statistic is greater than .05, as we want for well-fitting models, we fail to reject the null hypothesis that there is no difference between observed and model-predicted values, implying that the models estimates fit the data at an acceptable level. That is, well-fitting models show non-significance on the H-L goodness-of-fit test. This desirable outcome of non-significance indicates that the model prediction does not significantly differ from the observed.

The H-L statistic assumes sampling adequacy, with a rule of thumb being enough cases so that 95% of cells (typically, 10 decile groups times 2 outcome categories = 20 cells) have an expected frequency > 5. Our H-L statistic has a significance of .605 which means that it is not statistically significant and therefore our model is quite a good fit.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6.378	8	.605

Figure 7: Hosmer and Lemeshow Statistic

Contingency Table for Hosmer and Lemeshow Test

		take solar panel offer = decline offer		take solar panel offer = take offer		
		Observed	Expected	Observed	Expected	Total
Step 1	1	3	2.996	0	.004	3
	2	3	2.957	0	.043	3
	3	3	2.629	0	.371	3
	4	2	2.136	1	.864	3
	5	2	1.881	1	1.119	3
	6	0	.833	3	2.167	3
	7	0	.376	3	2.624	3
	8	1	.171	2	2.829	3
	9	0	.019	3	2.981	3
	10	0	.000	3	3.000	3

Figure 8: Hosmer and Lemeshow Table

1.14 Classification Table

Rather than using a goodness-of-fit statistic, we often want to look at the proportion of cases we have managed to classify correctly. For this we need to look at the classification table printed out by SPSS, which tells us how many of the cases where the observed values of the dependent variable were 1 or 0 respectively have been correctly predicted.

In the Classification table, the columns are the two predicted values of the dependent, while the rows are the two observed (actual) values of the dependent. In a perfect model, all cases will be on the diagonal and the overall percent correct will be 100%. In this study, 87.5% were correctly classified for the take offer group and 92.9% for the decline offer group. Overall 90% were correctly classified. This is a considerable improvement on the 53.3% correct classification with the constant model so we know that the model with predictors is a significantly better mode.

Classification Table[a]

			Predicted		
			Take solar panel offer		
	Observed		Decline offer	Take offer	Percentage correct
Step 1	take solar panel offer	decline offer	13	1	92.9
		take offer	2	14	87.5
	Overall Percentage				90.0

[a] The cut value is .500.

Figure 9: Classification Table

1.15 Variables in the Equation

The Variables in the Equation table has several important elements. The Wald statistic and associated probabilities provide an index of the significance of each predictor in the equation. The simplest way to assess Wald is to take the significance values and if less than 0.05 reject the null hypothesis as the variable does make a significant contribution. In this case, we note that family size contributed significantly to the prediction (p = .013) but mortgage did not (p = .075). The researcher may well want to drop independents from the model when their effect is not significant by the Wald statistic (in this case mortgage).

The ***Exp(B)*** column in the table presents the extent to which raising the corresponding measure by one unit influences the odds ratio. We can interpret ***Exp(B)***) in terms of the

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Mortgage	.005	.003	3.176	1	.075	1.005
	Famsize	2.399	.962	6.215	1	.013	11.007
	Constant	-18.627	8.654	4.633	1	.031	.000

a. Variable(s) entered on step 1: Mortgage, Famsize.

Figure 10: Variables in the Equation

$$\text{Probability of a case} = \frac{e\{(2.399 \times \text{family size}) + (.005 \times \text{mortgage}) - 18.627\}}{1 + e\{(2.399 \times \text{family size}) + (.005 \times \text{mortgage}) - 18.627\}}$$

Figure 11: Logistic Regression Equation

$$\begin{aligned}\text{Probability of a case taking offer} &= \frac{e\{(2.399 \times 7) + (.005 \times 2500) - 18.627\}}{1 + e\{(2.399 \times 7) + (.005 \times 2500) - 18.627\}} \\ &= \frac{e^{10.66}}{1 + e^{10.66}} \\ &= 0.99\end{aligned}$$

Figure 12: Logistic Regression Equation : Example

change in odds. If the value exceeds 1 then the odds of an outcome occurring increase; if the figure is less than 1, any increase in the predictor leads to a drop in the odds of the outcome occurring. For example, the ***Exp(B)*** value associated with family size is 11.007. Hence when family size is raised by one unit (one person) the odds ratio is 11 times as large and therefore householders are 11 more times likely to belong to the take offer group.

The ***B*** values are the logistic coefficients that can be used to create a predictive equation (similar to the b values in linear regression) formula seen previously.

Here is an example of the use of the predictive equation for a new case. Imagine a householder whose household size including themselves was seven and paying a monthly mortgage of $2,500$ euros. Would they take up the offer, i.e. belong to category 1? Substituting in we get:

Therefore, the probability that a householder with seven in the household and a mortgage of 2,500 p.m. will take up the offer is 99%, or 99% of such individuals will be expected to take up the offer. Note that, given the non-significance of the mortgage variable, you could be justified in leaving it out of the equation. As you can imagine, multiplying a mortgage value by B adds a negligible amount to the prediction as its B value is so small (.005).