1 Binary Classification

1.1 The Logistic Regression Equation

The form of the logistic regression equation is:

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

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).

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

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

1.2 Category Prediction Table

- 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 oftening 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:
- 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.

1.3 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.

1.4 Classification Plot

The classification plot or histogram of predicted probabilities provides a visual demonstration of the correct and incorrect predictions. Also called the '`classplot`' or the '`plot of observed groups and predicted probabilities`?,it is another very useful piece of information from

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 1: Classification Table

the SPSS output when one chooses `Classification plots`' under the Options button in the Logistic Regression dialogue box.

1.5 Interpreting the Classifcation Table

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.

1.6 SPSS Output

The variable Vote2005 is a binary variable describing turnout at a general election. The predictor variables are gender and age.

$$\text{logit}(\text{vote2005}) = -.779 + .077\text{gender}(1) + .037\text{age}$$

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	gender(1)	.077	.074	1.087	1	.297	1.080	.935	1.248
	age	.037	.002	267.015	1	.000	1.038	1.033	1.042
	Constant	-.779	.118	43.942	1	.000	.459		

a. Variable(s) entered on step 1: age.

Figure 2: General Election 2005

The age coefficient is statistically significant. Exp(B) for age is 1.038, which means for each year different in age, the person is 1.038 times more likely to turn out to vote, having allowed for gender in the model. Eg. a 21 year old is 1.038 times as likely to turn out to vote than a 20 year old. This might not seem much of a difference but a 20 year difference leads to a person being $1.038^20 = 2.11$ times more likely to turn out to vote. E.g. a 40 year old is 2.11 times more likely to turn out to vote than a 20 year old, having allowed for gender in the model.

The gender coefficient is not statistically significant.