Contents

1 Assumptions of logistic regression

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.

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).

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.

- 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.

Types of Variables (Revision)

- 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.

South Africa Heart Disease Data Example

A retrospective sample of males in a heart-disease high-risk region of the Western Cape, South Africa. There are roughly two controls per case of CHD. Many of the CHD positive men have undergone blood pressure reduction treatment and other programs to reduce their risk factors after their CHD event. In some cases the measurements were made after these treatments. These data are taken from a larger dataset, described in Rousseauw et al, 1983, South African Medical Journal.

Exercise
Fit a logistic regression model with

- *Coronary Heart Disease* (`chd`) as the dependent variable
- *age at onset, current alcohol consumption, obesity levels, cumulative tabacco, type-A behavior*, and *low density lipoprotein cholesterol* as predictor variables.

```
> head(SAheart)
sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52   1
2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46   0
4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58   1
5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49   1
6 132    6.20 6.47     36.21 Present    62   30.77   14.14  45   0
...
...
```

Calculate the misclassification rate for your model using this model

2 Review of Logistic Regression

Logistic Regression: Logit Transformation

The logit transformation is given by the following formula:

$$\eta_i = \text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

The inverse of the logit transformation is given by the following formula:

$$\pi_i = \text{logit}^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

2.1 Dummy variables

When an explanatory variable is categorical we can use **dummy variables** to contrast the different categories. For each variable we choose a baseline category and then contrast all remaining categories with the base line. If an explanatory variable has k categories, we need k-1 dummy variables to investigate all the differences in the categories with respect to the dependent variable.

For example suppose the explanatory variable was ***housing*** coded like this:

1: Owner occupier

2: renting from a private landlord

3: renting from the local authority

We would therefore need to choose a baseline category and create two dummy variables. For example if we chose owner occupier as the baseline category we would code the dummy variables (House1 and House2) like this

2.2 Log Likelihood

A "likelihood" is a probability, specifically the probability that the observed values of the dependent may be predicted from the observed values of the independents.

Like any probability, the likelihood varies from 0 to 1. The log likelihood (LL) is its log and varies from 0 to minus infinity (it is negative because the log of any number less than 1 is negative). LL is calculated through iteration, using maximum likelihood estimation (MLE).

2.3 Maximum Likelihood Estimation

- Maximum likelihood estimation, MLE, is the method used to calculate the logit coefficients. This contrasts to the use of ordinary least squares (OLS) estimation of coefficients in regression. OLS seeks to minimize the sum of squared distances of the data points to the regression line.
- MLE seeks to maximize the log likelihood, LL, which reflects how likely it is (the odds) that the observed values of the dependent may be predicted from the observed values of the independents. (Equivalently MLE seeks to minimize the -2LL value.)

- MLE is an iterative algorithm which starts with an initial arbitrary "guesstimate" of what the logit coefficients should be, the MLE algorithm determines the direction and size change in the logit coefficients which will increase LL.
- After this initial function is estimated, the residuals are tested and a re-estimate is made with an improved function, and the process is repeated (usually about a half-dozen times) until convergence is reached (that is, until LL does not change significantly). There are several alternative convergence criteria.

Wald Test

- The Wald test is a way of testing the signicance of particular explanatory variables in a statistical model. In logistic regression we have a binary outcome variable and one or more explanatory variables. For each explanatory variable in the model there will be an associated parameter.
- The Wald test, described by Polit (1996) and Agresti (1990), is one of a number of ways of testing whether the parameters associated with a group of explanatory variables are zero.
- If for a particular explanatory variable, or group of explanatory variables, the Wald test is significant, then we would conclude that the parameters associated with these variables are not zero, so that the variables should be included in the model. If the Wald test is not significant then these explanatory variables can be omitted from the model. When considering a single explanatory variable, Altman (1991) uses a t-test to check whether the parameter is significant.
- For a single parameter the Wald statistic is just the square of the t-statistic and so will give exactly equivalent results. An alternative and widely used approach to testing the significance of a number of explanatory variables is to use the likelihood ratio test. This is appropriate for a variety of types of statistical models. Agresti (1990) argues that the likelihood ratio test is better, particularly if the sample size is small or the parameters are large.

The Wald Test is a statistical test used to determine whether an effect exists or not,

It tests whether an independent variable has a statistically significant relationship with a dependent variable.

It is used in a great variety of different models including models for dichotomous variables and model for continuous variables.

- $\hat{\theta}$ Maximum likelihood estimate of the parameter of interest θ
- θ_o Proposed value. This is an assumption of the fact that the differences between $\hat{\theta}$ and θ_o is normal.

Univariate case

$$\frac{(\hat{\theta} - \theta_o)^2}{\text{var}(\hat{\theta})} \sim \chi^2$$

$$\frac{(\hat{\theta} - \theta_o)^2}{\text{s.e.}(\hat{\theta})} \sim \text{Normal}$$

The likelihood ratio test is also used to determine whether an effect exists.]

3 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.

3.1 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.

3.2 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.

4 Introduction to Logistic Regression

- Logistic regression or logit regression is a type of probabilistic statistical classification model.

- It is also used to predict a binary response from a binary predictor, used for predicting the outcome of a categorical dependent variable (i.e., a class label) based on one or more predictor variables (features).
- That is, it is used in estimating empirical values of the parameters in a qualitative response model. The probabilities describing the possible outcomes of a single trial are modeled, as a function of the explanatory (predictor) variables, using a logistic function.
- Logistic regression, also called a logit model, is used to model **dichotomous (i.e. Binary) outcome variables**. In the logit model the log odds of the outcome is modeled as a linear combination of the predictor variables.
- Binary Logistic regression is used to determine the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories.
- Logistic regression determines the impact of multiple independent variables presented simultaneously to predict membership of one or other of the two dependent variable categories.
- 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 regression** would be a more appropriate starting point.

Introduction to Logistic Regression

The term ***generalized linear model*** is used to describe a procedure for transforming the dependent variable so that the right hand side of the model equation can be interpreted as a ***linear combination*** of the explanatory variables. In logistic regression, the logit may be computed in a manner similar to linear regression:

$$\eta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

In situations where the dependent (y) variable is continuous and can be reasonably assumed to have a normal distribution we do not transform the y variable at all and we can simply run a multiple linear regression analysis.

Otherwise some sort of transformation is applied.

4.1 Binomial 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.

Binomial logistic regression estimates the probability of an event (as an example, having heart disease) occurring.

- If the estimated probability of the event occurring is greater than or equal to 0.5 (better than even chance), the procedure classifies the event as occurring (e.g., heart disease being present).

- If the probability is less than 0.5, Logistic regression classifies the event as not occurring (e.g., no heart disease).

4.2 Examples of Logistic Regression

Example 1: Suppose that we are interested in the factors that influence whether a political candidate wins an election. The outcome (response) variable is binary (0/1); *win or lose*. The predictor variables of interest are the amount of money spent on the campaign, the amount of time spent campaigning negatively and whether or not the candidate is an incumbent.

Example 2: A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, *admit/don't admit*, is a binary variable.

4.3 Assumptions of logistic regression

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.

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).

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.

- 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.

Types of Variables (Revision)

- 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.

Logistic function

The logistic function, with $\beta_0 + \beta_1 x$ on the horizontal axis and $\pi(x)$ on the vertical axis An explanation of logistic regression begins with an explanation of the logistic function, which always takes on values between zero and one:

$$F(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}},$$

and viewing t as a linear function of an explanatory variable x (or of a linear combination of explanatory variables), the logistic function can be written as:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x)}}{e^{(\beta_0+\beta_1 x)} + 1} = \frac{1}{1 + e^{-(\beta_0+\beta_1 x)}}.$$

This will be interpreted as the probability of the dependent variable equalling a "success" or "case" rather than a failure or non-case. We also define the inverse of the logistic function, the logit:

$$g(x) = \log \frac{\pi(x)}{1 - \pi(x)} = \beta_0 + \beta_1 x,$$

and equivalently:

$$\frac{\pi(x)}{1 - \pi(x)} = e^{(\beta_0+\beta_1 x)}.$$

5 Logistic Regression

Logistic regression, also called a logit model, is used to model **dichotomous outcome** variables. In the logit model the **log odds** of the outcome is modeled as a linear combination of the predictor variables.

In logistic regression theory, the predicted dependent variable is a function of the probability that a particular subject will be in one of the categories (for example, the probability that a patient has the disease, given his or her set of scores on the predictor variables).

5.1 Dummy variables

When an explanatory variable is categorical we can use **dummy variables** to contrast the different categories. For each variable we choose a baseline category and then contrast all remaining categories with the base line. If an explanatory variable has k categories, we need k-1 dummy variables to investigate all the differences in the categories with respect to the dependent variable.

For example suppose the explanatory variable was ***housing*** coded like this:

1: Owner occupier

2: renting from a private landlord

3: renting from the local authority

We would therefore need to choose a baseline category and create two dummy variables. For example if we chose owner occupier as the baseline category we would code the dummy variables (House1 and House2) like this

5.2 Log Likelihood

A "likelihood" is a probability, specifically the probability that the observed values of the dependent may be predicted from the observed values of the independents.

Like any probability, the likelihood varies from 0 to 1. The log likelihood (LL) is its log and varies from 0 to minus infinity (it is negative because the log of any number less than 1 is negative). LL is calculated through iteration, using maximum likelihood estimation (MLE).

5.3 Maximum Likelihood Estimation

- Maximum likelihood estimation, MLE, is the method used to calculate the logit coefficients. This contrasts to the use of ordinary least squares (OLS) estimation of coefficients in regression. OLS seeks to minimize the sum of squared distances of the data points to the regression line.
- MLE seeks to maximize the log likelihood, LL, which reflects how likely it is (the odds) that the observed values of the dependent may be predicted from the observed values of the independents. (Equivalently MLE seeks to minimize the -2LL value.)
- MLE is an iterative algorithm which starts with an initial arbitrary "guesstimate" of what the logit coefficients should be, the MLE algorithm determines the direction and size change in the logit coefficients which will increase LL.
- After this initial function is estimated, the residuals are tested and a re-estimate is made with an improved function, and the process is repeated (usually about a half-dozen times) until convergence is reached (that is, until LL does not change significantly). There are several alternative convergence criteria.

Wald Test

- The Wald test is a way of testing the signicance of particular explanatory variables in a statistical model. In logistic regression we have a binary outcome variable and one or more explanatory variables. For each explanatory variable in the model there will be an associated parameter.
- The Wald test, described by Polit (1996) and Agresti (1990), is one of a number of ways of testing whether the parameters associated with a group of explanatory variables are zero.
- If for a particular explanatory variable, or group of explanatory variables, the Wald test is significant, then we would conclude that the parameters associated with these variables are not zero, so that the variables should be included in the model. If the Wald test is not significant then these explanatory variables can be omitted from the model. When considering a single explanatory variable, Altman (1991) uses a t-test to check whether the parameter is significant.
- For a single parameter the Wald statistic is just the square of the t-statistic and so will give exactly equivalent results. An alternative and widely used approach to testing the significance of a number of explanatory variables is to use the likelihood ratio test. This is appropriate for a variety of types of statistical models. Agresti (1990) argues that the likelihood ratio test is better, particularly if the sample size is small or the parameters are large.

The Wald Test is a statistical test used to determine whether an effect exists or not,

It tests whether an independent variable has a statistically significant relationship with a dependent variable.

It is used in a great variety of different models including models for dichotomous variables and model for continuous variables.

- $\hat{\theta}$ Maximum likelihood estimate of the parameter of interest θ
- θ_o Proposed value. This is an assumption of the fact that the differences between $\hat{\theta}$ and θ_o is normal.

  Univariate case

$$\frac{(\hat{\theta}-\theta_o)^2}{\text{var}(\hat{\theta})} \sim \chi^2$$

$$\frac{(\hat{\theta}-\theta_o)^2}{\text{s.e.}(\hat{\theta})} \sim \text{Normal}$$

The likelihood ratio test is also used to determine whether an effect exists.]

6 Wald statistic

- Alternatively, when assessing the contribution of individual predictors in a given model, one may examine the significance of the Wald statistic. The Wald statistic, analogous to the t-test in linear regression, is used to assess the significance of coefficients.

- Alternatively, when assessing the contribution of individual predictors in a given model, one may examine the significance of the Wald statistic. The Wald statistic, analogous to the t-test in linear regression, is used to assess the significance of coefficients.
- The Wald statistic is commonly used to test the significance of individual logistic regression coefficients for each independent variable (that is, to test the null hypothesis in logistic regression that a particular logit (effect) coefficient is zero).
- The Wald Statistic is the ratio of the unstandardized logit coefficient to its standard error. The Wald statistic and its corresponding p probability level is part of SPSS output in the section ***Variables in the Equation.*** This corresponds to significance testing of b coefficients in OLS regression. The researcher may well want to drop independents from the model when their effect is not significant by the Wald statistic.
- The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient and is asymptotically distributed as a chi-square distribution.

$$W_j = \frac{B_j^2}{SE_{B_j}^2}$$

- Although several statistical packages (e.g., SPSS, SAS) report the Wald statistic to assess the contribution of individual predictors, the Wald statistic has limitations.
- When the regression coefficient is large, the standard error of the regression coefficient also tends to be large increasing the probability of Type-II error.
- The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient and is asymptotically distributed as a chi-square distribution.
- The Wald statistic also tends to be biased when data are sparse.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age	.085	.028	9.132	1	.003	1.089	1.030	1.151
	weight	.006	.022	.065	1	.799	1.006	.962	1.051
	gender(1)	1.950	.842	5.356	1	.021	7.026	1.348	36.625
	VO2max	-.099	.048	4.266	1	.039	.906	.824	.995
	Constant	-1.676	3.336	.253	1	.615	.187		

a. Variable(s) entered on step 1: age, weight, gender, VO2max.

Figure 1:

6.1 The Wald Test

- The Wald test is a way of testing the significance of particular predictor variables in a statistical model.

- In logistic regression we have a binary outcome variable and one or more explanatory variables. For each predictor variable in the model there will be an associated parameter. The Wald test is one of a number of ways of testing whether the parameters associated with a group of explanatory variables are zero.
- If for a particular explanatory variable, or group of explanatory variables, the Wald test is significant, then we would conclude that the parameters associated with these variables are not zero, so that the variables should be included in the model. If the Wald test is not significant then these explanatory variables can be omitted from the model.
- When considering a single explanatory variable, Altman (1991) uses a t-test to check whether the parameter is significant. For a single parameter the Wald statistic is just the square of the t-statistic and so will give exactly equivalent results.
- An alternative and widely used approach to testing the significance of a number of explanatory variables is to use the likelihood ratio test. This is appropriate for a variety of types of statistical models.
- Agresti (1990) argues that the likelihood ratio test is better, particularly if the sample size is small or the parameters are large.

6.2 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).

6.3 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.

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).

6.4 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 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

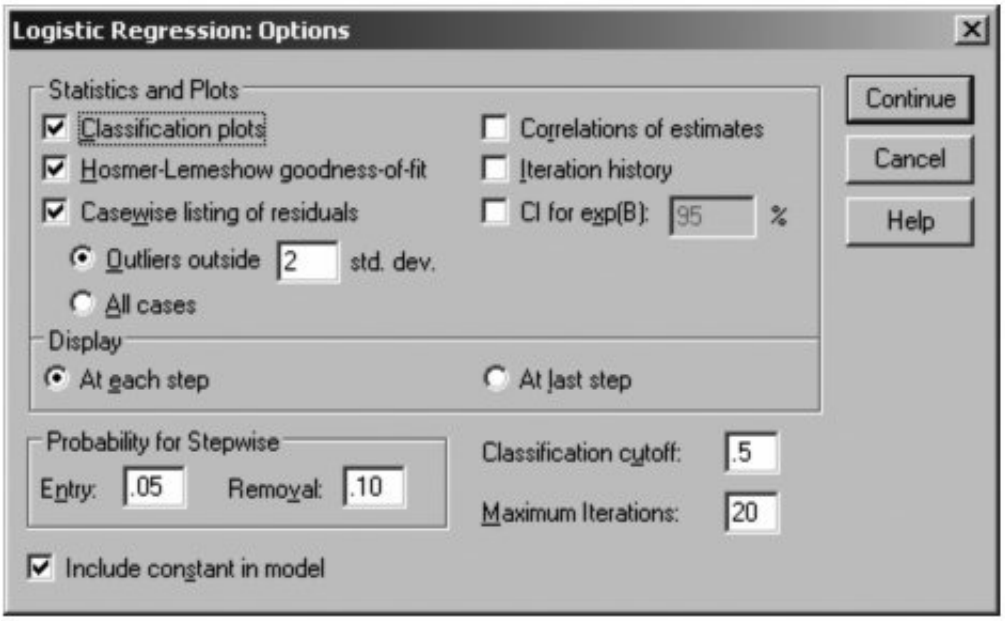

Figure 2: Selected Options for Exercises

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 3: Classification table

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.

6.5 Logistic Regression: Decision Rule

Our decision rule will take the following form: If the probability of the event is greater than or equal to some threshold, we shall predict that the event will take place. By default, SPSS sets

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 4: Variables in / not in the equation

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 5: Test Outcomes

this threshold to .5. While that seems reasonable, in many cases we may want to set it higher or lower than .5.

6.6 SPSS Output

- The variable Vote2005 is a binary variable describing turnout at a general election. The predictor variables are gender and 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 6: General Election 2005

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

- 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.

6.7 Hosmer-Lemeshow Prostate Example

We will now consider a real life example to demonstrate Logistic Regression. This example is taken from a Prostate Cancer Study from Hosmer and Lemeshow (2000). The goal of the analysis is to determine if variables measured at baseline can predict whether a tumour has penetrated the prostatic capsule. The variables are as follows:

6.8 Kasser and Bruce Infarction Data Example

We use a set of coronary data (Kasser and Bruce, 1969; Kronmal and Tarter, 1974) to see if age, history of angina pectoris (ANGINA: yes, no), history of high blood pressure (HIGHBP: yes, no), and functional class (FUNCTION: none, minimal, moderate, and more than moderate) can be used to predict the probability of past myocardial infarction (INFARCT: yes, no).

Variables from the Dataset Prostate (Hosmer and Lemeshow, 2000):

Variable	Label	Values
ID	Patient ID	1 – 380
Capsule	Tumor Penetration of Prostatic Capsule	0 = No Penetration, 1 = Penetration
Age	Age in Years	Number
Race	Race of Patient	1 = White, 2 = Black
Dpros	Results of the Digital Rectal Exam	1 = No Nodule, 2 = Left Lobe, 3 = Right Lobe, 4 = Both Lobes
Dcaps	Detection of Capsular Involvement	1 = No, 2 = Yes
PSA	Prostatic Specific Antigen Value	mg / ml
Vol	Tumor Volume Obtained from US	cm3
Gleason	Total Gleason Score	2 - 10

Figure 7: Variables

6.9 The Likelihood Ratio Test

The likelihood ratio test to test this hypothesis is based on the likelihood function. We can formally test to see whether inclusion of an explanatory variable in a model tells us more about the outcome variable than a model that does not include that variable. Suppose we have to evaluate two models.

$$\text{Model 1:} \quad \text{logit}(\pi) = \beta_0 + \beta_1 X_1$$

$$\text{Model 2:} \quad \text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Figure 8: Variables

Here, Model 1 is said to be nested within Model 2 all the explanatory variables in Model 1 (X1) are included in Model 2. We are interested in whether the additional explanatory variable in Model 2 (X_2) is required, i.e. does the simpler model (Model 1) fit the data just as well as the fuller model (Model 2). In other words, we test the null hypothesis that $\beta_2 = 0$ against the alternative hypothesis that $\beta_2 \neq 0$.

7 Wald statistic

- Alternatively, when assessing the contribution of individual predictors in a given model, one may examine the significance of the Wald statistic. The Wald statistic, analogous to the t-test in linear regression, is used to assess the significance of coefficients.
- Alternatively, when assessing the contribution of individual predictors in a given model, one may examine the significance of the Wald statistic. The Wald statistic, analogous to the t-test in linear regression, is used to assess the significance of coefficients.

- The Wald statistic is commonly used to test the significance of individual logistic regression coefficients for each independent variable (that is, to test the null hypothesis in logistic regression that a particular logit (effect) coefficient is zero).
- The Wald Statistic is the ratio of the unstandardized logit coefficient to its standard error. The Wald statistic and its corresponding p probability level is part of SPSS output in the section ***Variables in the Equation.*** This corresponds to significance testing of b coefficients in OLS regression. The researcher may well want to drop independents from the model when their effect is not significant by the Wald statistic.
- The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient and is asymptotically distributed as a chi-square distribution.

$$W_j = \frac{B_j^2}{SE_{B_j}^2}$$

- Although several statistical packages (e.g., SPSS, SAS) report the Wald statistic to assess the contribution of individual predictors, the Wald statistic has limitations.
- When the regression coefficient is large, the standard error of the regression coefficient also tends to be large increasing the probability of Type-II error.
- The Wald statistic is the ratio of the square of the regression coefficient to the square of the standard error of the coefficient and is asymptotically distributed as a chi-square distribution.
- The Wald statistic also tends to be biased when data are sparse.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age	.085	.028	9.132	1	.003	1.089	1.030	1.151
	weight	.006	.022	.065	1	.799	1.006	.962	1.051
	gender(1)	1.950	.842	5.356	1	.021	7.026	1.348	36.625
	VO2max	-.099	.048	4.266	1	.039	.906	.824	.995
	Constant	-1.676	3.336	.253	1	.615	.187		

a. Variable(s) entered on step 1: age, weight, gender, VO2max.

Figure 9:

7.1 The Wald Test

- The Wald test is a way of testing the significance of particular predictor variables in a statistical model.
- In logistic regression we have a binary outcome variable and one or more explanatory variables. For each predictor variable in the model there will be an associated parameter. The Wald test is one of a number of ways of testing whether the parameters associated with a group of explanatory variables are zero.

- If for a particular explanatory variable, or group of explanatory variables, the Wald test is significant, then we would conclude that the parameters associated with these variables are not zero, so that the variables should be included in the model. If the Wald test is not significant then these explanatory variables can be omitted from the model.
- When considering a single explanatory variable, Altman (1991) uses a t-test to check whether the parameter is significant. For a single parameter the Wald statistic is just the square of the t-statistic and so will give exactly equivalent results.
- An alternative and widely used approach to testing the significance of a number of explanatory variables is to use the likelihood ratio test. This is appropriate for a variety of types of statistical models.
- Agresti (1990) argues that the likelihood ratio test is better, particularly if the sample size is small or the parameters are large.

7.2 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).

7.3 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.

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).

7.4 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 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

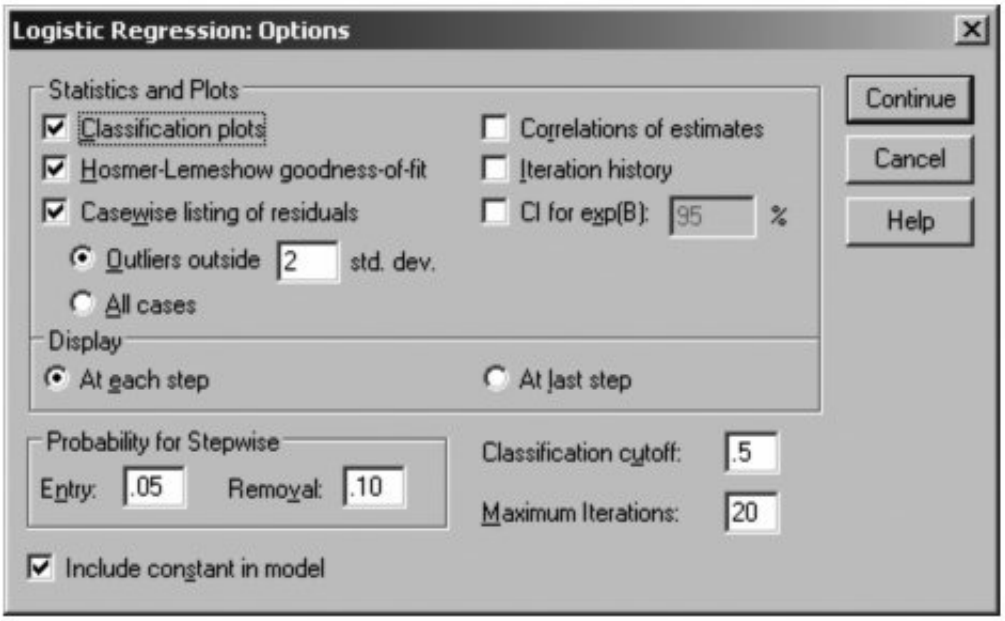

Figure 10: Selected Options for Exercises

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			take solar panel offer		
	Observed		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 11: Classification table

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.

7.5 Logistic Regression: Decision Rule

Our decision rule will take the following form: If the probability of the event is greater than or equal to some threshold, we shall predict that the event will take place. By default, SPSS sets

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 12: Variables in / not in the equation

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 13: Test Outcomes

this threshold to .5. While that seems reasonable, in many cases we may want to set it higher or lower than .5.

7.6 SPSS Output

- The variable Vote2005 is a binary variable describing turnout at a general election. The predictor variables are gender and 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 14: General Election 2005

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

- 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.

7.7 Hosmer-Lemeshow Prostate Example

We will now consider a real life example to demonstrate Logistic Regression. This example is taken from a Prostate Cancer Study from Hosmer and Lemeshow (2000). The goal of the analysis is to determine if variables measured at baseline can predict whether a tumour has penetrated the prostatic capsule. The variables are as follows:

7.8 Kasser and Bruce Infarction Data Example

We use a set of coronary data (Kasser and Bruce, 1969; Kronmal and Tarter, 1974) to see if age, history of angina pectoris (ANGINA: yes, no), history of high blood pressure (HIGHBP: yes, no), and functional class (FUNCTION: none, minimal, moderate, and more than moderate) can be used to predict the probability of past myocardial infarction (INFARCT: yes, no).

Variables from the Dataset Prostate (Hosmer and Lemeshow, 2000):

Variable	Label	Values
ID	Patient ID	1 – 380
Capsule	Tumor Penetration of Prostatic Capsule	0 = No Penetration, 1 = Penetration
Age	Age in Years	Number
Race	Race of Patient	1 = White, 2 = Black
Dpros	Results of the Digital Rectal Exam	1 = No Nodule, 2 = Left Lobe, 3 = Right Lobe, 4 = Both Lobes
Dcaps	Detection of Capsular Involvement	1 = No, 2 = Yes
PSA	Prostatic Specific Antigen Value	mg / ml
Vol	Tumor Volume Obtained from US	cm3
Gleason	Total Gleason Score	2 - 10

Figure 15: Variables

7.9 The Likelihood Ratio Test

The likelihood ratio test to test this hypothesis is based on the likelihood function. We can formally test to see whether inclusion of an explanatory variable in a model tells us more about the outcome variable than a model that does not include that variable. Suppose we have to evaluate two models.

$$\text{Model 1:} \quad \text{logit}(\pi) = \beta_0 + \beta_1 X_1$$

$$\text{Model 2:} \quad \text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Figure 16: Variables

Here, Model 1 is said to be nested within Model 2 all the explanatory variables in Model 1 (X1) are included in Model 2. We are interested in whether the additional explanatory variable in Model 2 (X_2) is required, i.e. does the simpler model (Model 1) fit the data just as well as the fuller model (Model 2). In other words, we test the null hypothesis that $\beta_2 = 0$ against the alternative hypothesis that $\beta_2 \neq 0$.

8 Summary of Logistic Regression

logistic regression or logit regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable (a dependent variable that can take on a limited number of values, whose magnitudes are not meaningful but whose ordering of magnitudes may or may not be meaningful) based on one or more predictor variables.

(1.) Logistic regression is intended for the modeling of dichotomous categorical outcomes (e.g., characterized by binary responses: buy vs Don't buy, dead vs. alive, cancer vs. none,).

(2.) We want to predict the probability of a particular response (0 to 1 scale).

(3.) For binary responses, linear regression should not be used for several reasons but the most common-sense reason is that linear regression can provide predictions NOT on a 0 to 1 scale. but rather a predicted response of some numeric value (e.g 2.4 or -800.3).

(4.) We need a way to link the probabilistic response variable to the continuous and/or categorical predictors and keep things on this 0 to 1 scale.

(5.) Logistic regression winds up transforming the probabilities to odds and then taking the natural logarithm of these odds, called logits.

(6.) Suppose a response variable is passing a test (by convention, 0=no and 1=yes). You have 1 predictor - number of days present in class over the past 30 days. Suppose the regression coefficient (often just called beta) in the output is 0.14. You would then say that, on average, as class presence increases by 1 day, the natural logarithm of the odds of passing the test increases by 0.14.

7.) For the interpretation, you can just talk about the odds. Most computer output will give you this number. Suppose the answer in odds is 1.24. Then, you just say that,on average, as class presence increases by 1 day, the odds of passing the test are multiplied by 1.24. In other words, for each additional day present, the odds of passing are 24

8.) To validate our findings, normally, we test whether the regression coefficient is equal to zero in the population. In logistic regression, the corresponding value for the odds is one (not zero). We got an odds of 1.24. Can we trust this? Or should we go with one (which would mean that the odds are the same for both passing and not passing, and hence class presence makes no difference at all)? Look at the p-value (significance). If it less than .05 (by convention), you have enough evidence to reject the notion that the odds are really one. You go ahead and support the 1.24 result.

8.1 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 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.

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 17: 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 18: Logistic Regression Equation

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).

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

Figure 19: Logistic Regression Equation : Example

9 Summary of Logistic Regression

logistic regression or logit regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable (a dependent variable that can take on a limited number of values, whose magnitudes are not meaningful but whose ordering of magnitudes may or may not be meaningful) based on one or more predictor variables.

(1.) Logistic regression is intended for the modeling of dichotomous categorical outcomes (e.g., characterized by binary responses: buy vs Don't buy, dead vs. alive, cancer vs. none,).

(2.) We want to predict the probability of a particular response (0 to 1 scale).

(3.) For binary responses, linear regression should not be used for several reasons but the most common-sense reason is that linear regression can provide predictions NOT on a 0 to 1 scale. but rather a predicted response of some numeric value (e.g 2.4 or -800.3).

(4.) We need a way to link the probabilistic response variable to the continuous and/or categorical predictors and keep things on this 0 to 1 scale.

(5.) Logistic regression winds up transforming the probabilities to odds and then taking the natural logarithm of these odds, called logits.

(6.) Suppose a response variable is passing a test (by convention, 0=no and 1=yes). You have 1 predictor - number of days present in class over the past 30 days. Suppose the regression coefficient (often just called beta) in the output is 0.14. You would then say that, on average, as class presence increases by 1 day, the natural logarithm of the odds of passing the test increases by 0.14.

7.) For the interpretation, you can just talk about the odds. Most computer output will give you this number. Suppose the answer in odds is 1.24. Then, you just say that,on average, as class presence increases by 1 day, the odds of passing the test are multiplied by 1.24. In other words, for each additional day present, the odds of passing are 24

8.) To validate our findings, normally, we test whether the regression coefficient is equal to zero in the population. In logistic regression, the corresponding value for the odds is one (not zero). We got an odds of 1.24. Can we trust this? Or should we go with one (which would mean that the odds are the same for both passing and not passing, and hence class presence makes no difference at all)? Look at the p-value (significance). If it less than .05 (by convention), you have enough evidence to reject the notion that the odds are really one. You go ahead and support the 1.24 result.

9.1 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).

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 20: 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 21: Logistic Regression Equation

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 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).

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

Figure 22: Logistic Regression Equation : Example