# Contents

# 1 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:

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

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.

### 1.1 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).

### 1.2 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.

### 1.3 Assumptions

**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.

**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 dis-

agree"), 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.

# 2 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).

## 2.1 Logistic Regression: Odds Ratio

What are odds? The odds of outcome 1 versus outcome 2 are the probability (or frequency) of outcome 1 divided by the probability (or frequency) of outcome 2.

$$\hat{Y} = \frac{\text{Odds}}{1 + \text{Odds}}$$

Loglinear models essentially define a pattern of odds ratios, apply the marginals to them, and compare the resulting table with the observed table, in pretty much the same way we apply the Pearson $\chi^2$ test for association. The big difference is the pattern we define can be much more complicated than independence.

## 2.2 Odds

The odds in favor of an event or a proposition are the ratio of the probability that an event will happen to the probability that it will not happen. 'Odds' are an expression of relative probabilities. Often 'odds' are quoted as odds against, rather than as odds in favor of, because of the possibility of confusion of the latter with the fractional probability of an event occurring.

$$\text{Odds} = \frac{p}{1-p}$$

### 2.2.1 Example

There are 5 pink marbles, 2 blue marbles, and 8 purple marbles.

- What is the probability of picking a blue marble? (Answer: 2/15).
- What are the odds in favor of picking a blue marble? (Answer: 2/13).

## 2.3 Introduction to the Odds Ratio

Let's begin with probability. Suppose that the probability of success is 0.8, thus p = 0.8 Then the probability of failure is

$$q = 1 - p = 0.2$$

The odds of success are defined as

$$odds(success) = p/q = 0.8/0.2 = 4$$

that is, the odds of success are 4 to 1. The odds of failure would be

$$odds(failure) = q/p = .2/.8 = .25$$

This looks a little strange but it is really saying that the odds of failure are 1 to 4. The odds of success and the odds of failure are just reciprocals of one another, i.e., 1/4 = .25 and 1/.25 = 4.

Next, we will add another variable to the equation so that we can compute an odds ratio.

**Another example**

Suppose that seven out of 10 males are admitted to an engineering school while three of 10 females are admitted.

- The probabilities for admitting a male are, p = 7/10 = 0.7 ( q = 1 - 0.7 = 0.3)
- Here are the same probabilities for females, p = 3/10 = 0.3 (q = 1 - 0.3 = 0.7)

Now we can use the probabilities to compute the admission odds for both males and females,

- *odds(male)* = 0.7/0.3 = 2.33333
- *odds(female)* = 0.3/0.7 = 0.42857

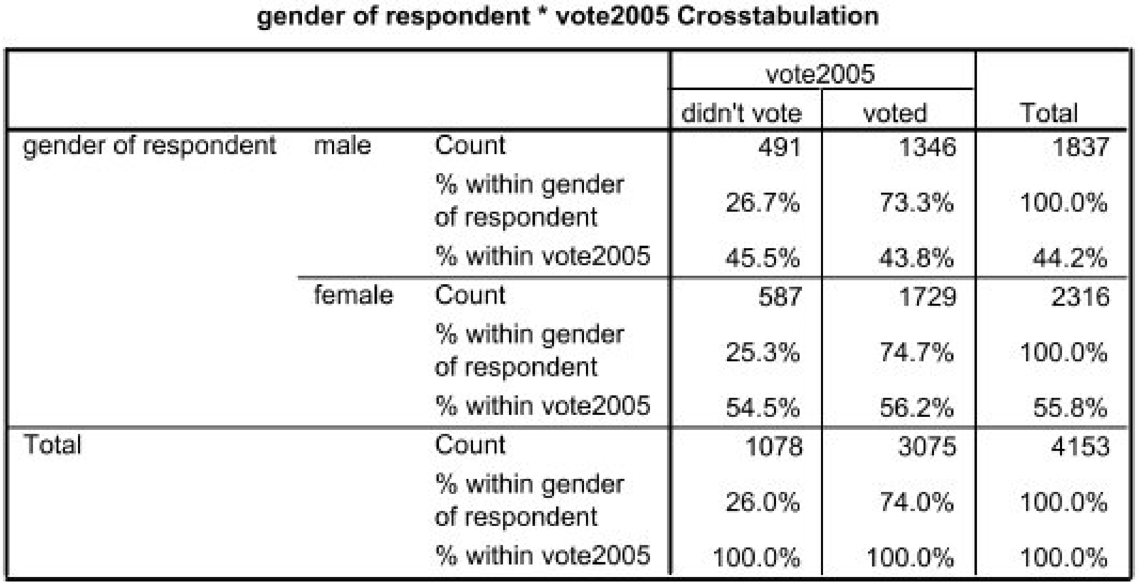
Next, we compute the odds ratio for admission,

$$OR = 2.3333/0.42857 = 5.44$$

Thus, for a male, the odds of being admitted are 5.44 times as large than the odds for a female being admitted.

### 2.4 Odds Ratio Example

These data are taken from the British Election Study 2005 pre-campaign and post-election panel data. We will consider the propensity to vote (sometimes called turnout) as the dependent variable, which has 2 categories. 0=did not turn out to vote, 1 turned out to vote.



gender of respondent * vote2005 Crosstabulation

| | | | vote2005 | | |
|---|---|---|---|---|---|
| | | | didn't vote | voted | Total |
| gender of respondent | male | Count | 491 | 1346 | 1837 |
| | | % within gender of respondent | 26.7% | 73.3% | 100.0% |
| | | % within vote2005 | 45.5% | 43.8% | 44.2% |
| | female | Count | 587 | 1729 | 2316 |
| | | % within gender of respondent | 25.3% | 74.7% | 100.0% |
| | | % within vote2005 | 54.5% | 56.2% | 55.8% |
| Total | | Count | 1078 | 3075 | 4153 |
| | | % within gender of respondent | 26.0% | 74.0% | 100.0% |
| | | % within vote2005 | 100.0% | 100.0% | 100.0% |

Figure 1: General Election 2005

The odds of a male turning out to vote are:

$$1346/491 = 2.741$$

The odds of female turning out to vote are

$$1729/587 = 2.945$$

The Odds ratio (female: male) are (1729/587) / (1346/491) = 1.074

### 2.5 Logistic Regression: Odds Ratios and Log-Odds

- Suppose that in a sample of 100 men, 90 drank wine in the previous week, while in a sample of 100 women only 20 drank wine in the same period.
- The odds of a man drinking wine are 90 to 10, or 9:1, while the odds of a woman drinking wine are only 20 to 80, or 1:4 = 0.25:1.
- The odds ratio is thus 9/0.25, or 36, showing that men are much more likely to drink wine than women.

- The detailed calculation is:

$$\frac{0.9/0.1}{0.2/0.8} = \frac{0.9 \times 0.8}{0.1 \times 0.2} = \frac{0.72}{0.02} = 36$$

- This example also shows how odds ratios are sometimes sensitive in stating relative positions: in this sample men are $90/20 = 4.5$ times more likely to have drunk wine than women, but have 36 times the odds.

- The logarithm of the odds ratio, the difference of the logits of the probabilities, tempers this effect, and also makes the measure symmetric with respect to the ordering of groups.

- For example, using natural logarithms, an odds ratio of 36/1 maps to 3.584, and an odds ratio of 1/36 maps to -3.584.

#### 2.5.1 Confidence Intervals for Odds Ratios

- Many statistical implementations of logistic regression include Confidence Intervals for the odds ratios. Odds ratios whose confidence limits exclude 1 are statistically significant.

- The odds ratio is referred to in SPSS as `Exp(B)`, the exponentiation of the B coefficient

### 2.6 Log-Odds

As an alternative to modeling the value of the outcome, logistic regression focuses instead upon the relative probability (odds) of obtaining a given result category. The natural logarithm of the odds is linear across most of its range, allowing us to continue using many of the methods developed for linear models. The result of this type of regression can be expressed as follows:

$$\text{Ln}\left[\frac{p}{1-p}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \ldots b_kx_k + e$$

### 2.7 About logits

There is a direct relationship between the coefficients produced by **logit** and the odds ratios produced by the logistic procedure. First, let's define what is meant by a logit: A logit is defined as the log base e (log) of the odds,

$$logit(p) = log(odds) = log(p/q)$$

Logistic regression is in reality ordinary regression using the logit as the response variable,

$$logit(p) = log(p/q) = b_0 + b_1X$$

This means that the coefficients in logistic regression are in terms of the log odds, that is, the coefficient 1.694596 implies that a one unit change in gender results in a 1.694596 unit change in the log of the odds.

Equation [3] can be expressed in odds by getting rid of the log. This is done by taking e to the power for both sides of the equation.

$$p/q = e^{b_0+b_1}$$

The end result of all the mathematical manipulations is that the odds ratio can be computed by raising e to the power of the logistic coefficient,

$$OR = e^b = e^1.694596 = 5.44$$

## 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:

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

and equivalently:

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

## 3 Logistic Regression: Logits

The logit transformation is given by the following formula:

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

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

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

# 4 Review of Logistic Regression

### Example 1

Given that $\pi_i = 0.2$, compute $\eta_i$.

$$\eta_i = \log\left(\frac{0.2}{1-0.2}\right) = \log\left(\frac{0.2}{0.8}\right)$$

$$\eta_i = \log(0.25) = -1.386$$

## 4.1 Example 2

Let us suppose that the probability of survival of a marine species of fauna is dependent on pollution, depth and water temperature. Suppose the logit for the logistic regression was computed as follows:

$$\eta_i = 0.14 + 0.76x_1 - 0.093x_2 + 1.2x_3$$

| Variables | case 1 | case 2 |
|---|---|---|
| Pollution($x_1$) | 6.0 | 1.9 |
| Depth ($x_2$) | 51 | 99 |
| Temp ($x_3$) | 3.0 | 2.9 |

Compute the probability of success for both case 1 and case 2.

- case $1\eta_1 = 0.14 + (0.76 \times 6) - (0.093 \times 51) + (1.2 \times 3) = 3.557$
- case $2\eta_2 = 0.14 + (0.76 \times 1.9) - (0.093 \times 99) + (1.2 \times 2.9) = -4.143$

The probabilities for success are therefore:

$$\pi_1 = \frac{e^{3.557}}{1+e^{3.557}} = \frac{35.057}{1+35.057} = 0.972$$

$$\pi_2 = \frac{e^{-4.143}}{1+e^{-4.143}} = \frac{0.0158}{1+0.0158} = 0.0156$$

**Example 2**

Given that $\eta_i = 2.3$, compute $\pi_i$.

$$\pi_i = \frac{e^{2.3}}{1+e^{2.3}} = \frac{9.974}{1+9.974} = 0.908$$

**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:

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

**Logits**

In logistic regression, the logit may be computed in a manner similar to linear regression:

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

### 4.2 Logistic function

The logistic function of any number is given by the inverse-logit:

$$\text{logit}^{-1}(\alpha) = \frac{1}{1+\exp(-\alpha)} = \frac{\exp(\alpha)}{1+\exp(\alpha)}$$

### 4.3 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

### 4.4 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).

### 4.5 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.

### 4.6 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.
- 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.

**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 $\theta$
- $theta_o$ Proposed value. This is an assumption of the fact that the differences between $\hat{\theta}$ and $theta_o$ 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. ]

**South Africa Heart Disease Data Example**

Load the South Africa Heart Disease Data and create training and test sets with the following code:

```
install.packages("ElemStatLearn")
library(ElemStatLearn)
data(SAheart)

set.seed(8484)
train = sample(1:dim(SAheart)[1],
size=dim(SAheart)[1]/2,replace=F)
trainSA = SAheart[train,]
testSA = SAheart[-train,]
```

Then fit a logistic regression model with *Coronary Heart Disease* (chd) as the outcome and *age at onset, current alcohol consumption, obesity levels, cumulative tabacco, type-A behavior*, and *low density lipoprotein cholesterol* as predictors.
Calculate the misclassification rate for your model using this function and a prediction on the ”response” scale:
What is the misclassification rate on the training set? What is the misclassification rate on the test set?

```
head(SAheart)

lr1 <- glm(chd ~ age + alcohol + obesity +
tobacco + typea + ldl, data=trainSA,
family="binomial")

lr1.train.predict <- predict(lr1, type="response")

missclass.lr1.train <- missClass(trainSA$chd,
lr1.train.predict)

lr1.test.predict <- predict(lr1, newdata=testSA,
type="response")

missclass.lr1.test <- missClass(testSA$chd,
lr1.test.predict)
```

# 5 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. 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.[9]

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

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 also tends to be biased when data are sparse.

## 5.1 The Wald Test

- The Wald test is a way of testing the significance 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 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 Model Diagnostics for Logistic Regression

## 6.1 Likelihood Ratio Test

The likelihood ratio test is a test of the difference between ?2LL for the full model with predictors and ?2LL for initial chi-square in the null model. When probability fails to reach the 5% significance level, we retain the null hypothesis that knowing the independent variables (predictors) has no increased effects (i.e. make no difference) in predicting the dependent.

## 6.2 Hosmer-Lemeshow Goodness-of-Fit

The Hosmer-Lemeshow Goodness-of-Fit test tells us whether we have constructed a valid overall model or not. If the model is a good fit to the data then the HosmerLemeshow Goodness-of-Fit test should have an associated p-value greater than 0.05.

### 6.3 The Hosmer-Lemeshow Test

The Hosmer-Lemeshow test of goodness of fit is not automatically a part of the SPSS logistic regression output. To get this output, we need to go into `options` and tick the box marked Hosmer-Lemeshow test of goodness of fit.

In our example, this gives us the following output:

| Step | Chi-square | df | Sig. |
|---|---|---|---|
| 1 | 142.032 | 6 | .000 |

Therefore, our model is significant, suggesting it does not fit the data. However, as we have a sample size of over 13,000, even very small divergencies of the model from the data would be flagged up and cause significance. Therefore, with samples of this size it is hard to find models that are parsimonious (i.e. that use the minimum amount of independent variables to explain the dependent variable) and fit the data. Therefore, other fit indices might be more appropriate.

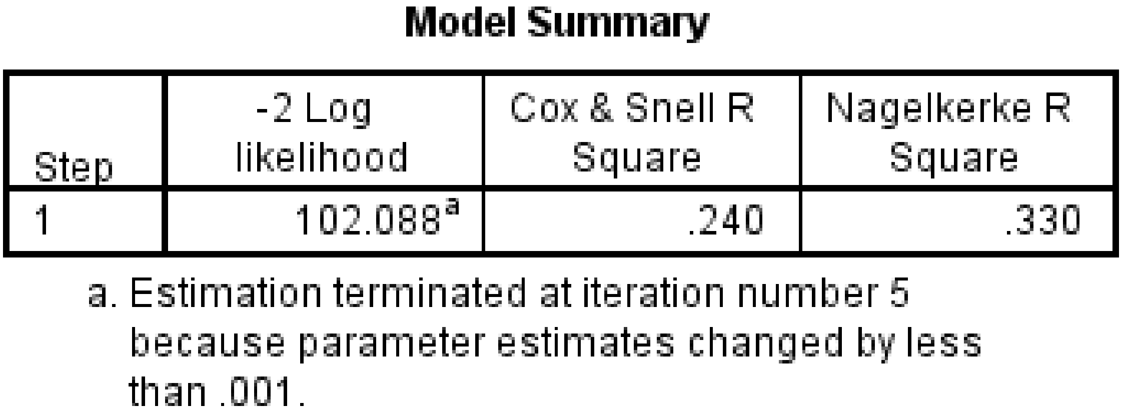
### 6.4 R Squared Diagnostics

- In order to understand how much variation in the dependent variable can be explained by the model (the equivalent of $R^2$ in multiple regression), you sho$R^2$uld consult Model Summary statistics.

- The SPSS output table below contains the *Cox & Snell R Square* and *Nagelkerke R Square* values, which are both methods of calculating the explained variation. These values are sometimes referred to as pseudo $R^2$ values (and will have lower values than in multiple regression).

- However, they are interpreted in the same manner, but with more caution. Therefore, the explained variation in the dependent variable based on our model ranges from 24.0% to 33.0%, depending on whether you reference the Cox & Snell $R^2$ or Nagelkerke $R^2$ methods, respectively.

- Nagelkerke $R^2$ is a modification of Cox & Snell $R^2$, the latter of which cannot achieve a value of 1. For this reason, it is preferable to report the Nagelkerke $R^2$ value.



**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 102.088[a] | .240 | .330 |

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Figure 2: SPSS output

### 6.5 Pseudo R-squares

Cox & Snell R Square and Nagelkerke R Square are two measures from the **pseudo R-squares** family of measures.

Logistic regression does not have an equivalent to the R-squared that is found in OLS regression; however, many researcehrs have tried to come up with one.

There are a wide variety of pseudo-R-square statistics (these are only two of them). Because this statistic does not mean what R-squared means in OLS regression (the proportion of variance explained by the predictors), we suggest interpreting this statistic with great caution.

### 6.6 Pseudo-R Squared

Cox and Snell R Square and Nagelkerke R Square - These are pseudo R-squares. Logistic regression does not have an equivalent to the R-squared that is found in OLS regression; however, many people have tried to come up with one.

### 6.6.1 Cox & Snell R Square

Cox and Snell's R-Square is an attempt to imitate the interpretation of multiple R-Square based on the likelihood, but its maximum can be (and usually is) less than 1.0, making it difficult to interpret. It is part of SPSS output.

### 6.6.2 Nagelkerke's R-Square

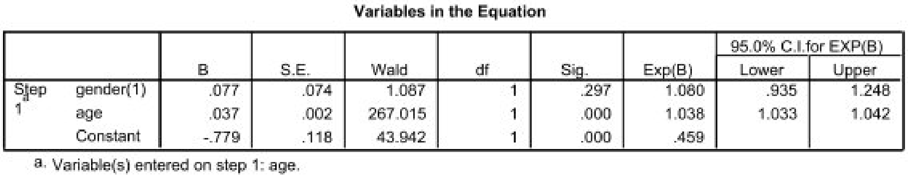
Nagelkerke's R-Square is a further modification of the Cox and Snell coefficient to assure that it can vary from 0 to 1. Nagelkerke's R-Square will normally be higher than the Cox and Snell measure. It is part of SPSS output and is the most-reported of the R-squared estimates.

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

### 6.8 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 3: 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 $1.038^{2}0 = 2.11$ 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.

### 6.9 HSB2 Example

The hsb2 dataset is taken from a national survey of high school seniors. Two hundred observation were randomly sampled from the High School and Beyond survey. Descriptive statistics and exploratory data analysis are shown below. Because we do not have a suitable dichotomous variable to use as our dependent variable, we will create one (which we will call honcomp, for honors composition) based on the continuous variable write. We do not advocate making dichotomous variables out of continuous variables; rather, we do this here only for purposes of this illustration.

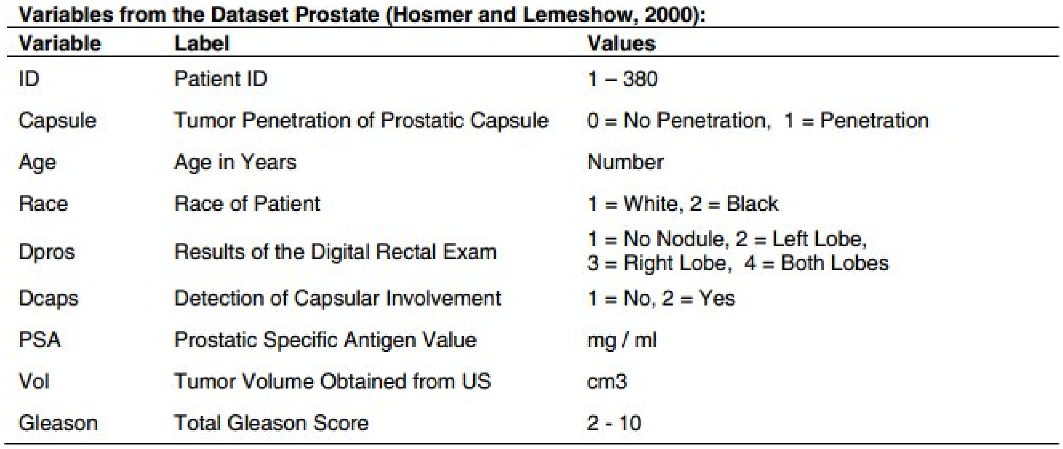
Here is the list of variables in the file.

```
  obs:            200    highschool and beyond (200 cases)
 vars:             12    28 Feb 2005 09:25
---------------------------------------------------------
              variable
variable name   type   about the variable
---------------------------------------------------------
id              scale  student id
female        nominal  (0/1)
race          nominal  ethnicity (1=hispanic 2=asian 3=africa
ses           ordinal  (1=low 2=middle 3=high)
schtyp        nominal  type of school (1=public 2=private)
prog          nominal  type of program (1=general 2=academic
read            scale  standardized reading score
write           scale  standardized writing score
```

| | | |
|---|---|---|
| math | scale | standardized math score |
| science | scale | standardized science score |
| socst | scale | standardized social studies score |
| hon | nominal | honors english (0/1) |

## 6.10 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:



**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,<br>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 |

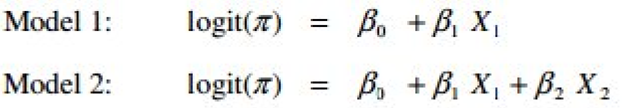
Figure 4: Variables

## 6.11 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).

### 6.12 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.



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

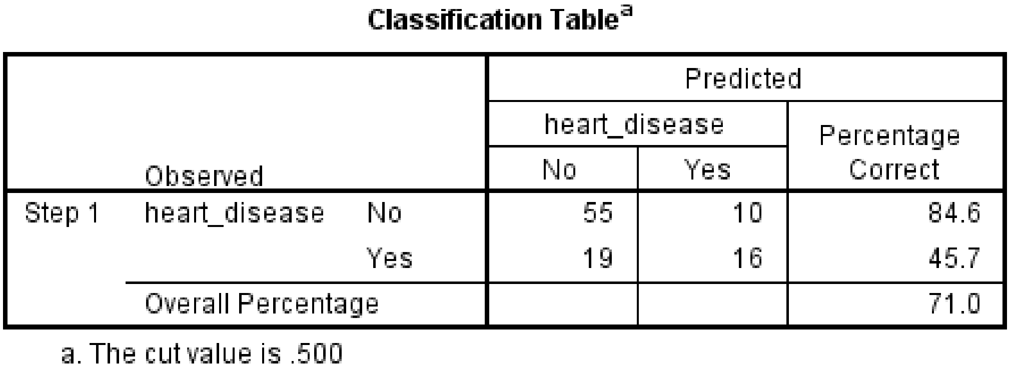
Model 2: $\text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Figure 5: 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 $\beta_2 \neq 0$.

### 6.13 Multinomial Logistic Regression

Multinomial Logistic Regression is useful for situations in which you want to be able to classify subjects based on values of a set of predictor variables. This type of regression is similar to logistic regression, but it is more general because the dependent variable is not restricted to two categories. For Example, In order to market films more effectively, movie studios want to predict what type of film a moviegoer is likely to see. By performing a Multinomial Logistic Regression, the studio can determine the strength of influence a persons age, gender, and dating status has upon the type of film they prefer. The studio can then slant the advertising campaign of a particular movie toward a group of people likely to go see it.



**Classification Table**[a]

| | | | Predicted | | |
|---|---|---|---|---|---|
| | | | heart_disease | | Percentage Correct |
| | Observed | | 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

# 7 Binary Classification

## 7.1 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.

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

### 7.3 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.

# 8 Stepwise Logistic Selection

Stepwise logistic regression involves the stepwise (or one-by-one) selection of variables, providing a fast and effective method to screen a large number of variables, and to fit multiple logistic regression equations simultaneously.

In stepwise selection, an attempt is made to remove any insignificant variables from the model before adding a significant variable to the model.

Stepwise binary logistic regression is very similar to stepwise multiple regression in terms of its advantages and disadvantages. Stepwise logistic regression is designed to find the ***most parsimonious*** set of predictors that are most effective in predicting the dependent variable.

## 8.1 Procedure for Stepwise Selection

Variables are added to the logistic regression equation one at a time, using the statistical criterion of reducing the ***-2 Log Likelihood error*** for the included variables. (Recall: The lower the -2LL value, the better the fit of the model).

After each variable is entered, each of the included variables are tested to see if the model would be better off the variable were excluded. This does not happen often.

The process of adding more variables stops when all of the available variables have been included or when it is not possible to make a statistically significant reduction in -2 Log Likelihood using any of the variables not yet included.

Categorical variables are added to the logistic regression as a group. It is possible, and often likely, that not all of the individual dummy-coded variables will have a statistically significant individual relationship with the dependent variable.

### 8.2 SPSS Implementation

SPSS provides a table of variables included in the analysis and a table of variables excluded from the analysis. It is possible that none of the variables will be included. It is possible that all of the variables will be included.

The order of entry of the variables can be used as a measure of relative importance.

Once a variable is included, its interpretation in stepwise logistic regression is the same as it would be using other methods for including variables.

### 8.3 Advantages and Disadvantages

- Stepwise logistic regression can be used when the goal is to produce a predictive model that is parsimonious and accurate because it excludes variables that do not contribute to explaining differences in the dependent variable.

- Stepwise logistic regression is less useful for testing hypotheses about statistical relationships. Its usage is recommended only for exploratory purposes, rather that as a formal procedure.

- Stepwise logistic regression can be useful in finding relationships that have not been tested before. Its findings invite one to speculate on why an unusual relationship makes sense.

- It is not legitimate to do a stepwise logistic regression and present the results as though one were testing a hypothesis that included the variables found to be significant in the stepwise logistic regression.

- Using statistical criteria to determine relationships is vulnerable to over-fitting the data set used to develop the model at the expense of generalisability.

Menard (1995: 54) writes, "there appears to be general agreement that the use of computer-controlled stepwise procedures to select variables is inappropriate for theory testing because it capitalizes on random variations in the data and produces results that tend to be idosyncratic and difficult to replicate in any sample other than the sample in which they were originally obtained."

### 8.4 Forward Selection

You can estimate models using block entry of variables or any of the following stepwise methods: forward conditional, forward LR, forward Wald, backward conditional, backward LR, or backward Wald.

Forward selection is the usual option for a stepwise regression, starting with the constant-only model and adding variables one at a time. The forward stepwise logistic regression method utilizes the likelihood ratio test which tests the change in 2LL between steps to determine automatically which variables to add or drop from the model.

Method selection allows you to specify how independent variables are entered into the analysis. Using different methods, you can construct a variety of regression models from the same set of variables.

1 Enter. A procedure for variable selection in which all variables in a block are entered in a single step.

2 Forward Selection (Conditional). Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of a likelihood-ratio statistic based on conditional parameter estimates.

3 Forward Selection (Likelihood Ratio). Stepwise selection method with entry testing based on the significance of the score statistic,

and removal testing based on the probability of a likelihood-ratio statistic based on the maximum partial likelihood estimates. (LR stands for Likelihood Ratio and is considered the criterion least prone to error.)

4 Forward Selection (Wald). Stepwise selection method with entry testing based on the significance of the score statistic, and removal testing based on the probability of the Wald statistic.

5 Backward Elimination (Conditional). Backward stepwise selection. Removal testing is based on the probability of the likelihood-ratio statistic based on conditional parameter estimates.

6 Backward Elimination (Likelihood Ratio). Backward stepwise selection. Removal testing is based on the probability of the likelihood-ratio statistic based on the maximum partial likelihood estimates.

7 Backward Elimination (Wald). Backward stepwise selection. Removal testing is based on the probability of the Wald statistic.

### 8.5 Cross Validation of Stepwise Regression

When stepwise logistic regression is used, some form of validation analysis is a necessity. We will use 75/25% cross-validation.

To do cross validation, we randomly split the data set into a 75% training sample and a 25% validation sample. We will use the training sample to develop the model, and we test its effectiveness on the validation sample to test the applicability of the model to cases not used to develop it.

In order to be successful, the follow two questions must be answers affirmatively: Did the stepwise logistic regression of the training sample produce the same subset of predictors produced by the regression model of the full data set?

If yes, compare the classification accuracy rate for the 25% validation sample to the classification accuracy rate for the 75% training sample. If the **shrinkage** (accuracy for the 75% training sample - accuracy for the 25% validation sample) is 2% (0.02) or less, we conclude that validation was successful.

Note: shrinkage may be a negative value, indicating that the accuracy rate for the validation sample is larger than the accuracy rate for the training sample. Negative shrinkage (increase in accuracy) is evidence of a successful validation analysis.

If the validation is successful, we base our interpretation on the model that included all cases.

# 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.

## 10 Multinomial Logistic Regression

Examples of multinomial logistic regression

**Example 1.** People's occupational choices might be influenced by their parents' occupations and their own education level. We can study the relationship of one's occupation choice with education level and father's occupation. The occupational choices will be the outcome variable which consists of categories of occupations.

**Example 2.** A biologist may be interested in food choices that alligators make. Adult alligators might have difference preference than young ones. The outcome variable here will be the types of food, and the predictor variables might be the length of the alligators and other environmental variables.

**Example 3.** Entering high school students make program choices among general program, vocational program and academic program. Their choice might be modeled using their writing score and their social economic status.