Part 2: Logistic Regression

1. Introduction

So far we have considered linear regression models where the response variable is continuous with a normal distribution. Often, though, we have data where the interest is in predicting or "explaining" a **binary** (or **dichotomous**) outcome variable by a set of explanatory variables. This type of outcome variable is a **two-category** variable. Examples are:

- Success/failure of a treatment, explained by dosage of medicine administered, patient's age, sex, weight and severity of condition.
- High/low cholesterol level, explained by sex, age, whether a person smokes or not, etc.
- Use/non-use of contraception, explained by gender, age, whether married, education level, religion, etc.
- Vote for/against political party, explained by age, gender, education level, region, ethnicity, etc.
- Yes/No or Agree/Disagree responses to questionnaire items in a survey.

Logistic regression is the most popular technique available for modelling dichotomous dependent variables.

2. Modelling Dichotomous Outcome Variables

Lets get technical just for a while! Going back to the simple linear regression model for a moment, what we were doing there was actually specifying a model for the mean value of our outcome variable Y. Statisticians often use the notation E[Y] for the mean of the variable Y in the population – the E stands for "expected value". Our model assumed that this mean was related to the value of the explanatory variable, x, via the linear equation

$$E[Y] = \beta_0 + \beta_1 x \ . \qquad (2.1)$$

Once we have fitted this model, the fitted line provides an estimate of the mean of the outcome variable *Y* for a "population" of subjects with explanatory variable value *x*.

So what happens if Y is dichotomous (rather than continuous)?

Now let *Y* be a dichotomous outcome variable, coded as $\boldsymbol{Y = 1}$ for the outcome of interest (denoted a "**success**"), and $\boldsymbol{Y = 0}$ for the other possible outcome (denoted a "**failure**"). We use the Greek character π to represent the probability that the "success" outcome occurs in the population. The probability of a "failure" outcome is then 1 - π. In this special case where *Y* is a binary variable, the mean of *Y* in the population is equal to π. So our model becomes a **model for the probability of a "success" outcome**

$$\pi = \beta_0 + \beta_1 x \quad , \qquad (2.2)$$

This is called the **linear probability model**.

Example

The variable *Y* denotes presence or absence of evidence of significant coronary heart disease (CHD), so that $Y = 1$ indicates CHD is present or $Y = 0$ indicates that it is not present in an individual. We are interested in the relationship between age (*X*) and the presence or absence of CHD in a population, and we have a sample of 100 subjects selected to participate in a study. See Table 1.1 on page 3 of Hosmer and Lemeshow (2000) for the data.

In a linear regression problem the first thing we would do (I hope!) to get an idea of the relationship between the two variables would be to draw a scatterplot (see Figure 2.1). This is not particularly useful here, and it does not help us to establish whether the mean of *Y* (i.e. the probability π) is linear in *x*.

To get a better picture of the relationship we could create intervals for the explanatory variable *X* and compute the mean of *Y* within each group. In this example, we create age groups 20-29, 30-34, 35-39, …, 55–59, 60-69, calculate the proportion of "successes" in each age group, and then plot the proportion of "successes" against the mid-point of the age group (see Figure 2.2). This plot shows better the dependence of π on the value of the explanatory variable, *x*. It would not appear to be linear in *x* – note the S-shaped curve. The change in π per unit change in *x* becomes progressively smaller the closer π gets to 0 or 1.

Figure 2.1: Scatterplot of CHD by Age for 100 Subjects

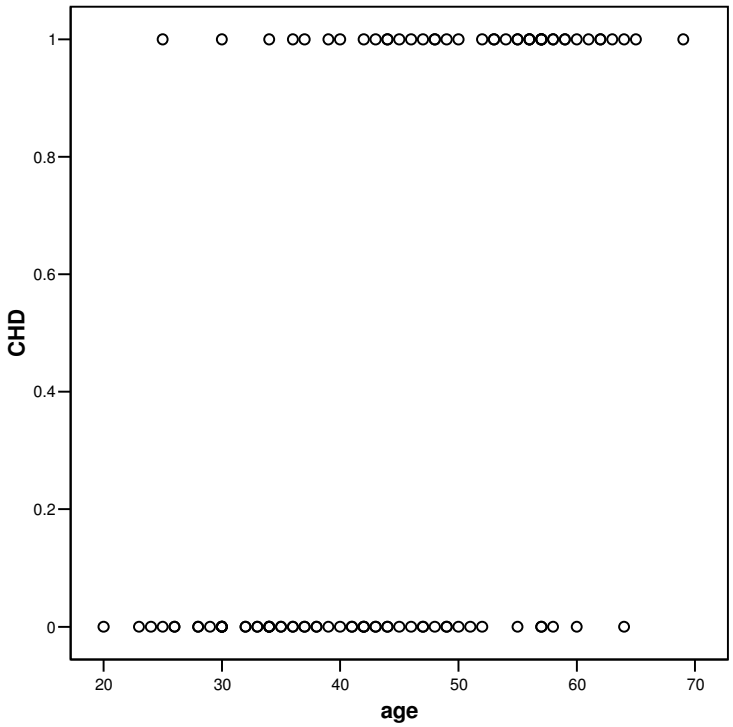
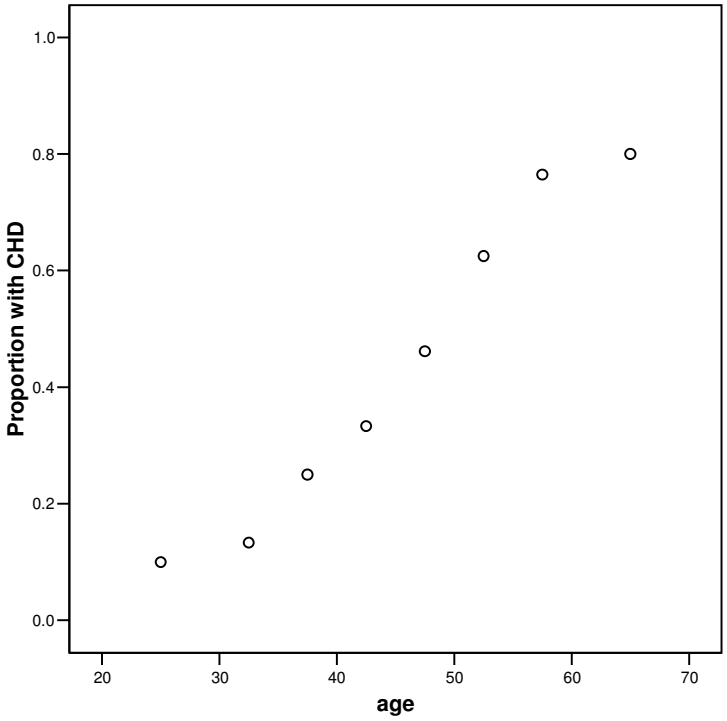


Figure 2.2: Plot of the Proportion of Subjects with CHD in Each Age Group



So what is the problem with applying the standard linear regression model?

It is possible to fit the model (2.2) using the ordinary least squares (OLS) method that we have already come across in linear regression. Indeed, such a model might produce sensible results in some cases. However:

- **The predicted values of π obtained from fitting this model may be outside the interval [0,1]**.

  Since π is a probability, its value must lie within the interval [0,1]. However, the right-hand side (RHS) of equation (2.2) is unbounded so that, theoretically, the RHS can take on values from - ∞ to ∞. This means we could get a predicted probability of, for example, 2.13 from our fitted model, which is rather non-sensical! It turns out that if $0.25 < \pi < 0.75$ the linear probability model produces fairly sensible results though.

- **The usual regression assumption of normality of *Y* is not satisfied** - *Y* is not continuous, it only takes a value of 0 or 1.

What is the solution to this problem?

Instead of fitting a model for π, we use a **transformation** of π. We shall consider the most commonly used transformation, the log of the odds of a "success" outcome, i.e. we shall model $\log\left(\frac{\pi}{1-\pi}\right)$. First, though, what do we mean by the "odds"?

3. Probabilities and Odds

The **odds** are defined as the probability of a "success" outcome divided by the probability of a "failure" outcome

$$\text{odds} = \frac{\Pr(\text{success})}{\Pr(\text{failure})} = \frac{\Pr(\text{success})}{1-\Pr(\text{success})} = \frac{\pi}{1-\pi}. \qquad (3.1)$$

It is easy to convert from probabilities to odds and back again. Note: Since π lies between 0 and 1 the odds can take values between 0 to ∞.

Examples

1. If $\pi = \mathbf{0.8}$, then the odds are equal to 0.8/(1-0.8) = 0.8/0.2 = **4**.

2. If $\pi = \mathbf{0.5}$, then the odds are equal to 0.5/(1-0.5) = 0.5/0.5 = **1**. So, if the odds are equal to 1 the chance of "success" is the same as the chance of "failure".

3. If the odds are equal to **0.3** then solving $\pi/(1 - \pi) = 0.3$ gives $\pi = 0.3/1.3 = \mathbf{0.2308}$.

So, we can think of the odds as another scale for representing probabilities. We also note here that since division by zero is not allowed, the odds will be undefined when the probability of "failure" (i.e. $1-\pi$) is 0.

4. The Logistic Regression Model

The logistic regression model can be written in terms of the **log of the odds**, called the **logit**, as

$$\log_e\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \quad . \qquad (4.1)$$

The **logit** is just the (natural) logarithm of the odds. With this model, the range of values that the left-hand side can potentially take is now between -∞ and ∞, which is the same range as that of the right-hand side. Now we have something that looks very familiar. We have a linear model on the **logit scale**. This is the most common form of the logistic regression model.

An alternative and equivalent way of writing the logistic regression model in (4.1) is in terms of the **odds**,

$$\frac{\pi}{1-\pi} = e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k)}. \qquad (4.2)$$

With this form of the model, the range of values that the right-hand side can take is now between 0 and ∞ (since the **exponential function** is non-negative), the same as for the odds.

A third way in which you will see the model written is in terms of the underlying probability of a "success" outcome,

$$\pi = \frac{e^{(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k)}}{1+e^{(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k)}} . \qquad (4.3)$$

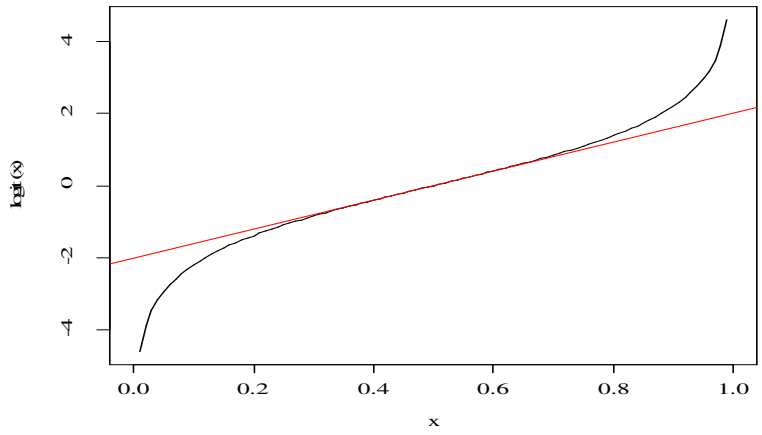
This form is just obtained by re-arranging (4.2). Notice that (4.3) can be written in a slightly different way as

$$\pi = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_kX_k)}} . \qquad (4.4)$$

We emphasise here that all three forms of the model (4.1) – (4.3) are equivalent.

Figure 4.1 shows the logit function, i.e. a plot of logit(π) against π. Note that the curve is almost linear for $0.25 < \pi < 0.75$. This is why the linear probability model produces sensible results for π within this range.

Figure 4.1: A Plot of the Logit Function: logit(π) (= $\log_e[\pi/(1-\pi)]$) vs π.



5. Performing Logistic Regression

Our aim is to quantify the relationship between the probability of a "success" outcome, π, and the explanatory variables X_1, X_2, …, X_k based on some sample data. For now, we assume that

in the population there is a relationship between π and a single continuous explanatory variable X and that this relationship is of the form

$$\text{logit}(\pi) \;=\; \log_e\left[\frac{\pi}{1-\pi}\right] \;=\; \beta_0 \;+\; \beta_1 X \;. \qquad (5.1)$$

This model can be estimated using SPSS (or practically any other general-purpose statistical software) as

$$\text{logit}(\hat{\pi}) \;=\; b_0 \;+\; b_1 X \;, \qquad (5.2)$$

where b_0 and b_1 are the estimated regression coefficients.

The estimation for logistic regression is commonly performed using the statistical method of **maximum likelihood estimation**. Explicit closed-form formulae for the estimated regression coefficients like those obtained in the case of simple linear regression (see equations (1.6) and (1.7) in the linear regression handouts) are not usually available, so numerical methods are used. A proper introduction to these methods is beyond the scope of this course, and so we'll just let SPSS do the estimation for us.

Example: Low Birth Weight Babies

Suppose that we are interested in whether or not the gestational age (GAGE) of the human foetus (number of weeks from conception to birth) is related to the birth weight. The dependent variable is birth weight (BWGHT), which is coded as 1 = normal, 0 = low. The data for 24 babies (7 of whom were classified as having low weight at birth) are shown in Table 5.1.

Table 5.1: Gestational Ages (in Weeks) of 24 Babies by Birth Weight

Normal Birth Weight (BWGHT = 1)	Low Birth Weight (BWGHT = 0)
40, 40, 37, 41, 40, 38, 40, 40 38, 40, 40, 42, 39, 40, 36, 38, 39	38, 35, 36, 37, 36, 38, 37

The model we shall fit is:

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{ GAGE} . \quad (5.3)$$

The output given by SPSS from fitting this model is shown in Figure 5.1.

Under Block 0, SPSS first produces some output which corresponds to fitting the "constant model" – this is the model $\text{logit}(\pi) = \beta_0$, i.e. the model with the predictor variable excluded.

Under Block 1, SPSS presents output for the model that includes the predictor variable. The estimated regression coefficients (see final table in the output labelled "Variables in the Equation") are

$$b_0 = \mathbf{-48.908} , \qquad b_1 = \mathbf{1.313} ,$$

and so our fitted model is

$$\text{logit}(\hat{\pi}) = \mathbf{-48.908} + \mathbf{1.313} \text{ GAGE} . \quad (5.4)$$

Two obvious questions present themselves at this stage:

1. ***How do we know if this model fits the data well?***
2. ***How do we interpret this fitted model?***

6. How Good is the Model?

We shall consider several approaches to assess the "fit" of the model. Note that in practice, reporting two or three of these is normally sufficient.

6.1 Classification Table

One way of assessing how well the model fits the observed data is to produce a **classification table**. This is a simple tool which indicates how good the model is at predicting the outcome variable. As an example, consider the fitted model (5.4).

First, we choose a "cut-off" value c (usually 0.5). For each individual in the sample we "predict" their BWGHT condition as 1 (i.e. normal) if their fitted probability of being normal birth weight is greater than c, otherwise we predict it as 0 (i.e. low). We then construct a table showing how many of the observations we have predicted correctly.

Figure 5.1: SPSS Output from Logistic Regression (Gestational Age Data)

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	24	100.0
	Missing Cases	0	.0
	Total	24	100.0
Unselected Cases		0	.0
Total		24	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
.0000	0
1.0000	1

Block 0: Beginning Block

Classification Table[a,b]

			Predicted		
			BWGHT		Percentage
	Observed		.0000	1.0000	Correct
Step 0	BWGHT	.0000	0	7	.0
		1.0000	0	17	100.0
	Overall Percentage				70.8

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.887	.449	3.904	1	.048	2.429

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	GAGE	10.427	1	.001
	Overall Statistics		10.427	1	.001

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	12.676	1	.000
	Block	12.676	1	.000
	Model	12.676	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	16.298	.410	.585

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	1.626	5	.898

Contingency Table for Hosmer and Lemeshow Test

		BWGHT = .0000		BWGHT = 1.0000		
		Observed	Expected	Observed	Expected	Total
Step 1	1	1	.951	0	.049	1
	2	2	2.518	1	.482	3
	3	2	1.753	1	1.247	3
	4	2	1.372	3	3.628	5
	5	0	.185	2	1.815	2
	6	0	.213	8	7.787	8
	7	0	.009	2	1.991	2

Classification Table[a]

			Predicted		
			BWGHT		Percentage
	Observed		.0000	1.0000	Correct
Step 1	BWGHT	.0000	5	2	71.4
		1.0000	2	15	88.2
	Overall Percentage				83.3

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	GAGE	1.313	.541	5.890	1	.015	3.716
	Constant	-48.908	20.338	5.783	1	.016	.000

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

In this example (see the penultimate table in the Figure 5.1 SPSS output under Block 1 labelled "Classification Table"), we have 24 cases altogether. Of these, 7 were observed as having low birthweight (BWGHT = 0) and 5 of these 7 we correctly predict, i.e. they have a fitted probability (calculated from our model) of less than 0.5. Similarly, 15 out of the 17 observed as having normal birthweight are correctly predicted, i.e. they have a fitted probability of greater than 0.5.

Generally, the higher the overall percentage of correct predictions (in this case 20/24 = 83%) the better the model. However, there is no formal test to decide whether a certain percentage of correct predictions is adequate. Also, it is easy to construct a situation where the logistic regression model is in fact the correct model and therefore fits, but classification is poor.

6.2 Histogram of Estimated Probabilities

The classification table described above only tells you for each group whether the predicted probability is greater or less than c (0.5). We can also produce the distribution of predicted probabilities of the "success" outcome in the form of a histogram.

Figure 6.1 gives the histogram for the birth weight data. The plotting symbol used for each case designates the group to which the case actually belongs (1 = "normal" birth weight, 0 = "low" birth weight). If the fitted model (5.4) successfully distinguishes the two groups, the "normal" birth weight cases should be to the right of 0.5 and the "low" birth weight cases should be to the left. As a general rule, the more the two groups cluster at their respective ends of the plot, the better.

If there were many cases around the centre of the graph then this means that for these cases the model is predicting a probability of around 0.5 for a "success" outcome. That is, for these cases, there is only about a 50-50 chance that the data are correctly predicted.

In this example we note there is one "normal" birth weight case having a very small predicted probability of being "normal" birth weight (around 0.17).

6.3 The Likelihood Ratio Test

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. For example,

Figure 6.1: Histogram of Estimated Probabilities (Gestational Age Data)

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

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

Here, Model 1 is said to be **nested** within Model 2 – all the explanatory variables in Model 1 (X_1) 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 $\beta_2 \neq 0$.

The **likelihood ratio test** to test this hypothesis is based on something called the likelihood function. For each model it fits, SPSS calculates the statistic which is -2*log-likelihood (which we shall refer to as -2LL). This statistic is also called the **scaled deviance** and it measures the degree of discrepancy between the observed values and predicted values from the model. In that sense, you can think of it as being analogous to the error sum of squares in linear regression – it is an indicator of how much unexplained information there is after the model has been fitted.

The value of this statistic for the "smaller" model (Model 1) will be larger than the value for the "larger" model (Model 2). The difference in the value of the -2LL statistic between Model 1 and Model 2 is called the **likelihood ratio (LR) test statistic** and this is what we use to test our null hypothesis.

For the example of gestational age and birth weight, the following table gives the values for –2LL:

Model	-2 Log Likelihood
Model 1 (constant model)	28.975
Model 2 (with GAGE)	16.298

The value of -2LL when just the constant is included is 28.975. The value for the model that includes GAGE is 16.298. SPSS gives this latter value in the "Model Summary" table (see Block 1 output in Figure 5.1). The LR test statistic is **28.975 – 16.298 = 12.676**, and SPSS gives this value in the "Omnibus Tests of Model Coefficients" table (see Figure 5.1).

The statistical theory underlying the use of this test statistic is beyond the scope of this course. We just note that if the null hypothesis (that the coefficient of GAGE is 0) is true, the LR test statistic should have a **chi-squared** distribution with one degree of freedom, as long as the sample size is "large".

So, the procedure is to observe the value of the LR test statistic and compare with the table value from a chi-squared distribution with one degree of freedom. If the LR test statistic is too large relative to the table value, then this will imply that the null hypothesis should be rejected, i.e. the simpler model does not fit the data as well as the fuller model. At the 5% level of significance, the cut-off value from the table is **3.84**, which indicates that in this example we should **reject** the null hypothesis in favour of the alternative, i.e. GAGE is significant. [Again, SPSS just provides a p-value (in the "sig" column) – if this is less than 0.05 then we reject the null hypothesis at the 5% level.] That is, the model that includes gestational age is better at predicting birth weight than the model with just the constant term.

Note: When two nested models, Model 1 and Model 2, are compared, where Model 2 has all the covariates from Model 1 and an additional p explanatory variables, the significance of the

LR test statistic is determined by comparing with the table value from a chi-squared distribution with p degrees of freedom.

6.4 The Wald Test

As already mentioned in Section 5, estimation of the coefficients (i.e. the β's) in logistic regression is performed in SPSS using the method of maximum likelihood estimation (MLE). The standard errors are also computed by SPSS, and their estimation also relies on MLE theory. Although this theory is beyond the scope of this course, there are some important results of the estimation that we need to know about and use in our interpretation of the SPSS output.

Consider our model in the example of gestational age and birth weight,

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{ GAGE} . \qquad (6.1)$$

Another way of checking if the variable GAGE should be in the model or not is to calculate the ratio of the estimate to its standard error (this is a bit like the t statistic in linear regression)

$$b_1 / s_{b_1} . \qquad (6.2)$$

If the null hypothesis that $\beta_1 = 0$ is true, then this statistic has an approximate *standard normal* distribution. So, we can compare this to values in the normal tables for a given level of significance.

Equivalently, we can calculate the **Wald statistic**, which is the square of this ratio, i.e.

$$\left(\frac{b_1}{s_{b_1}} \right)^2 .$$

If the null hypothesis that $\beta_1 = 0$ is true, then this statistic has a **chi-squared** distribution with one degree of freedom. This is what SPSS calculates and displays in the "Variables in the Equation" table (see Figure 5.1), along with an associated p-value.

In the example of gestational age and birth weight, the value of the Wald test statistic for the coefficient corresponding to the variable GAGE is **5.890** (= $[1.313/0.541]^2$). The p-value is **0.015**, indicating that the coefficient is significant at the 5% level (but not at the 1% level).

Wald's test or the likelihood ratio Test – do they lead to the same conclusion?

The simple answer is not always! In most cases, both tests would lead you to the same decision. However, in some cases the Wald test produces a test statistic that is non-significant when the likelihood ratio test indicates that the variable should be kept in the model. This is because sometimes the estimated standard errors are "too large" (this happens when the absolute value of the coefficient becomes large) so that the ratio (and thus the Wald statistic) becomes too small. The likelihood ratio test is the more robust of the two and is generally to be preferred.

6.5 Measures of the Proportion of Variation Explained

In linear regression, one measure of the usefulness of the model was the statistic R^2 which gave the proportion of variation in the outcome variable being explained by the model. Several statistics have been proposed in the case of logistic regression that are roughly equivalent in interpretation to the R^2 in linear regression.

One simple statistic is based on comparing the log-likelihood for the "constant" model (LL_0) and the model of interest (LL_m). For the gestational ages example, this would be calculated as

$$R^2 = \frac{LL_0 - LL_m}{LL_0} = \frac{-14.4875 - (-8.149)}{-14.4875} = \mathbf{0.438}.$$

SPSS gives two variations on this, **Cox and Snell's R^2** and **Nagelkerke's** $\overline{R}^2$ (or adjusted R^2). Cox and Snell's R^2 has the disadvantage that for discrete models (such as logistic regression) it may not achieve the maximum value of one, even when the model predicts all the outcomes perfectly. Nagelkerke's $\overline{R}^2$ is an improvement over Cox and Snell's R^2 that can attain a value of one when the model predicts the data perfectly.

SPSS gives the values for these two statistics in the "Model Summary" table (see Figure 5.1). For the birth weight example, the Cox and Snell R^2 is **0.410** and Nagelkerke's $\overline{R}^2$ is **0.585**. The interpretation is that the model (with gestational age as the single explanatory variable) explains about 59% of the variation in the data. However, there is no formal test that can tell us if 59% is sufficient or not.

6.6 The Hosmer-Lemeshow Goodness-of-Fit Test

A commonly used test of the overall fit of a model to the observed data is the **Hosmer and Lemeshow test**. The idea is to form groups of cases and construct a "goodness-of-fit" statistic by comparing the observed and predicted number of events in each group. In the birth weight example, the cases are divided into a number of approximately equal groups based on values of the predicted probability of being "normal" birth weight (i.e. the event occurring). The differences between the observed number and expected number (calculated by summing predicted probabilities based on the model) in each group are then assessed using a chi-square test.

The SPSS output for the Hosmer and Lemeshow test applied to the birth weight data is shown in Figure 5.1. In this example, seven groups are created. For example, the fourth group consists of the 5 cases for which GAGE = 38. Of this group, 3 were "normal" birthweight and 2 were "low" birth weight. Summing the predicted probabilities of being "normal" birth weight for these 5 cases, the predicted number of "normal" birth weight cases is 3.628 and the predicted number of "low" birth weight cases is 1.372. The Hosmer and Lemeshow goodness-of-fit statistic is then calculated as

$$\sum_{cells} \frac{(O-E)^2}{E} \quad ,$$

where O and E are the observed and expected numbers in a cell. The logic is that the closer the expected numbers are to the observed, then the smaller the value of this statistic. So, small values will indicate that the model is a good fit - large values of this statistic indicate the model is not a good fit to the data.

The value of this test statistic is **1.626** which is compared to the cut-off value from the chi-square distribution with 5 (number of groups – 2) degrees of freedom. The p-value (given by SPSS) is **0.898**, so we do not reject the null hypothesis that there is no difference between the observed and predicted values, i.e. the model appears to fit the data reasonably well.

This example illustrates the use of the Hosmer and Lemeshow test. However, to use this test sensibly you need a fairly large sample size so that the expected numbers in most groups exceeds 5 and none of the groups have expected values less than 1.

7. Interpreting the Model

Remember from Section 4 that the logistic regression model can be written on three different scales (logit, odds, or probability). It can therefore be interpreted on these different scales. Consider the example of gestational age and birth weight:

- The interpretation of the coefficient for GAGE in the model

$$\text{logit}(\hat{\pi}) = \mathbf{-48.908 + 1.313}\,\text{GAGE}\ , \qquad (7.1)$$

  is that **a unit change in GAGE increases the log odds (logit) of normal birth weight by 1.313, on average**, i.e. a one-week increase in gestational age increases the log odds of normal birth weight by **1.313**, on average.

- The model (7.1) can also be written as the **odds model** by taking the exponent of both sides, i.e.

$$\text{odds} = \frac{\hat{\pi}}{1-\hat{\pi}}$$

$$= e^{-48.908+1.313\text{GAGE}} = e^{-48.908}e^{1.313\text{GAGE}}\ . \qquad (7.2)$$

  Thus, a one-week increase in gestational age changes the odds of normal birth weight multiplicatively by a factor equal to $e^{1.313}$, i.e. by **3.716**. This factor is called an **odds ratio** (more about these later), and is computed for us by SPSS and displayed in the final column (labelled Exp(B)) of the "Variables in the Equation" table. Equivalently, we may say that a unit increase in GAGE **increases** the odds of normal birth weight by [3.716 –1] x 100%, i.e. **272%**.

- Finally, the model (7.1) can also be written on the **probability scale**, i.e.

$$\hat{\pi} = \frac{e^{-48.908+1.313\text{GAGE}}}{1+e^{-48.908+1.313\text{GAGE}}}\ . \qquad (7.3)$$

  Thus, we can predict the probability of normal birth weight for any given gestational age.

For the general fitted model equation

$$\text{logit}(\hat{\pi}) = b_0 + b_1 X \ , \qquad (7.4)$$

the value of X when $\hat{\pi} = 0.5$ is called the **median effective level**, i.e. the outcome of interest has a 50% chance of occurring. This happens when the odds = 1, i.e. when the logit (log odds) = 0, and this occurs when $X = - b_0/b_1$. For the birth weight data, this is when gestational age = 48.908/1.313 = **37.25** weeks. At this age, the baby has a 50% chance of being of normal birth weight.

What about interpretation of the constant term?

Some statisticians interpret the exponent of the constant term as the **baseline odds**. In our example this is the value of the odds when GAGE = 0. This interpretation is not really applicable here and I would think of the constant as simply a nuisance parameter to be kept in the model so that the odds are of the correct scale.

Note: When there are two or more explanatory variables, say X_1 and X_2, the interpretation of the coefficient β_1 as the change in the log odds when X_1 changes by one unit is correct only if X_1 and X_2 are unrelated (i.e. when X_1 changes, X_2 is unaffected).

8. Confidence Intervals

We can construct **confidence intervals** for the β's based on the estimated coefficients and the standard errors. In the example of gestational age and birth weight an approximate 95% confidence interval for the coefficient of GAGE (β_1) is given by

$$\mathbf{1.313} \pm 1.96 * \mathbf{0.541} \ ,$$

where 0.541 is the standard error of the estimated coefficient and 1.96 is the value from the standard normal distribution corresponding to the 95% level. So, a 95% confidence interval for β_1 is **(0.27, 2.37)**, i.e. the change in the log-odds of a normal birth weight baby per one-week increase in gestational age is estimated as **1.313** but could be as low as **0.27** or as high as **2.37** with 95% confidence.

By taking the exponents of the lower and upper limits of this confidence interval we can obtain a **confidence interval for the odds ratio** $\mathbf{e^{\beta_1}}$, which is

$$(e^{0.27}, e^{2.37}) = (\mathbf{1.31}, \mathbf{10.70}) .$$

This gives the "effect" on the odds scale. So the effect of a one week increase in gestational age on the odds of a normal birth weight baby lies between **1.31** and **10.70** (with 95% confidence), i.e. an increase between about 31% to 970%! There is an option in SPSS to compute this confidence interval.

9. Models with Categorical Explanatory Variables

Suppose that we are interested in whether or not a woman goes back to work within six months after having a baby. What are some of the factors that might influence her decision? The level of the husband's income? Her occupation? The state of health of the baby?

For simplicity, let us assume that we only have data on the husband's income (HINC) and her occupation (OCCUP) which are coded as follows:

HINC	= 1	Low level of income
	= 2	Medium
	= 3	High
OCCUP	= 1	Casual/unskilled manual
	= 2	Skilled manual/clerical
	= 3	Professional/managerial

The dependent variable, WORK, is coded 1 if the woman returns to work within six months of having a baby, and coded 0 if she does not return.

9.1 Fitting the Model

To fit this model, we need to tell SPSS that the two explanatory variables are not continuous, but are **categorical**. Once this is indicated, we also need to tell SPSS how to calculate the estimates corresponding to these categorical variables. We can either

1. Compare each category of a variable with one particular category (called the **reference category**), or

2. Compare how much better (or worse) each category is from the average effect.

These procedures are called **contrasts**, and the first method is called "**Indicator**" while the second is called "**Deviation**". SPSS creates a set of dummy variables for each categorical variable according to the type of contrast that you specify.

We will use the indicator method, choosing the **last** category of each variable as the reference category (this is the SPSS default). Remember that in linear regression we had to explicitly set up dummy variables to represent each level (apart from one) of a categorical variable. Here, we are doing exactly the same kind of thing except that SPSS will now automatically do this for us! The reference category is basically assigned an estimate of zero by SPSS and the estimates for the other categories indicate a higher (positive estimate) or lower (negative estimate) effect compared to the reference category.

Data for 41 women were used in the analysis producing the estimated logistic regression model given in Table 9.1.

Table 9.1: SPSS Output – "Variables in the Equation" Table (Back to Work Data)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	HINC			6.824	2	.033	
	HINC(1)	2.939	1.350	4.740	1	.029	18.890
	HINC(2)	2.223	1.073	4.290	1	.038	9.232
	OCCUP			7.417	2	.025	
	OCCUP(1)	-3.523	1.331	7.010	1	.008	.030
	OCCUP(2)	-1.845	1.340	1.896	1	.169	.158
	Constant	1.625	1.096	2.198	1	.138	5.077

a. Variable(s) entered on step 1: HINC, OCCUP.

So, our fitted model is

$$\text{logit}(\hat{\pi}) = \mathbf{1.625} + \mathbf{2.939}*\text{HINC}(1) + \mathbf{2.223}*\text{HINC}(2) - \mathbf{3.523}*\text{OCCUP}(1) - \mathbf{1.845}*\text{OCCUP}(2), \quad (9.1)$$

where HINC(1) is the dummy variable set up to represent the "low" level of husband's income, HINC(2) is the dummy representing the "medium" level of husband's income, etc.

To make sure there is no confusion, SPSS presents a table called "Categorical Variables Coding" which tells you exactly how the dummy variables have been set up and which category is being used as the reference (see Table 9.2).

Table 9.2: SPSS Output – "Categorical Variables Codings" Table (Back to Work Data)

Categorical Variables Codings

		Frequency	Parameter (1)	Parameter (2)
OCCUP	1.00 casual/unskilled	14	1.000	.000
	2.00 skilled/clerical	12	.000	1.000
	3.00 professional/manager	15	.000	.000
HINC	1.00 Low	9	1.000	.000
	2.00 Medium	15	.000	1.000
	3.00 High	17	.000	.000

Note: If you had chosen the **first** category as the reference category, you would have obtained different estimates, but the conclusions are essentially the same.

9.2 Interpretation of Estimates: Odds Ratios

A nice feature of logistic regression models with categorical explanatory variables is that the exponents of the parameter estimates are the **odds ratios**.

Aside

The concept of odds ratios is straightforward. Let π_1 be the probability of going back to work for a woman whose husband has low income, and let π_2 be the probability of going back to work for a woman whose husband has medium income.

Then $\text{Odds}_1 = \dfrac{\pi_1}{1-\pi_1}$ = odds of going back to work for a woman whose husband has low income, and $\text{Odds}_2 = \dfrac{\pi_2}{1-\pi_2}$ = odds of returning to work for a woman whose husband has medium income. The **ratio of the odds**, i.e. $\dfrac{\text{Odds}_1}{\text{Odds}_2} = \dfrac{\pi_1}{1-\pi_1} \times \dfrac{1-\pi_2}{\pi_2}$ conveys information of how likely women whose husbands have low income are to return to work, compared with women whose husbands have medium income.

To see this, consider as an example the fitted model for women in professional or managerial jobs whose husbands have a **low** level of income. Here, HINC(1) = 1, HINC(2) = 0, OCCUP(1) = 0 and OCCUP(2) = 0, and

$$\text{logit}(\hat{\pi}_{low}) \;=\; \mathbf{1.625} \;+\; \mathbf{2.939}\,.$$

The fitted model for women in professional or managerial jobs whose husbands have a **high** level of income (HINC(1) = 0, HINC(2) = 0, OCCUP(1) = 0 and OCCUP(2) = 0) is

$$\text{logit}(\hat{\pi}_{high}) \;=\; \mathbf{1.625}\,.$$

Taking the difference, and remembering that the logit is the log of the odds, we have

$$\text{logit}(\hat{\pi}_{low}) - \text{logit}(\hat{\pi}_{high}) \;=\; \log(\text{odds}_{low}) - \log(\text{odds}_{high})\,,$$

$$= \;\log\left(\frac{\text{odds}_{low}}{\text{odds}_{high}}\right) \;=\; \mathbf{2.939}\,.$$

Now, taking the exponent of both sides we get

$$\frac{\text{odds}_{low}}{\text{odds}_{high}} \;=\; \exp(2.939) \;=\; \mathbf{18.890}\,.$$

In other words, the exponent of the coefficient of HINC(1) has given an odds ratio of **18.890**. This means that the odds of returning to work for women whose husbands have a low level of income are **18.9 times** the odds of returning to work for women whose husbands have a high level of income (the reference category). Notice that SPSS gives the odds ratio in the final column of the "Variables in the Equation" table.

Similarly, the odds of returning to work for women whose husbands have a medium level of income are about **9 times** the odds for those women whose husbands have a high level of income. This suggests that a higher level of the husband's income is associated with less likelihood of a woman returning to work within six months of giving birth.

(The reference group here is HINC(3), i.e. 'high level of income', which you can think of as having estimate 0, so that its odds ratio is $e^0 = 1$.)

For the respondent's occupation, we can see that women employed in categories 1 and 2 are less likely to return to work within six months, compared to professional women (the reference category). The odds of returning to work for women in casual or unskilled jobs are [0.03 – 1] x 100% = -97%, i.e. **97%** lower than the odds of returning to work for women in professional or managerial jobs. This is irrespective of their husbands' levels of incomes.

9.3 Estimated Probabilities

Using the probability form of the logistic regression model

$$\hat{\pi} = \frac{e^{(b_0 + b_1 X_1 + \ldots)}}{1 + e^{(b_0 + b_1 X_1 + \ldots)}} . \qquad (9.2)$$

we can work out the probabilities of returning to work for the different groups of women. These are given in Table 9.3.

Table 9.3: Estimated Probabilities of Returning to Work after Giving Birth (n = 41)

Respondent's Occupation	Husband's Level of Income		
	Low	Medium	High
Casual/unskilled	0.74	0.58	0.13
Skilled/clerical	0.94	0.88	0.44
Professional/managerial	0.99	0.98	0.84

For example, for women in casual/unskilled jobs whose husbands have a low level of income

$$\hat{\pi} = \frac{e^{1.625+2.939-3.523}}{1 + e^{1.625+2.939-3.523}} = 0.74.$$

Note that in SPSS there is an option to calculate these probabilities.

9.4 Odds Ratios versus Relative Risks

If the odds of returning to work for women in group HINC(1) are about **19** times the odds for HINC(3), does this mean that the probability of returning to work for women in group HINC(1) is 19 times the probability for women in group HINC(3)? **The answer is NO**.

Although odds have a one-to-one relationship with probabilities, an odds ratio of, say, 2 does not mean that the probability of the numerator category is twice that of the denominator category. The **relative risk** of an outcome is the ratio of the probabilities of the outcome for two groups. For example, among casually employed women, the relative risk of returning to

work if the husband's income is "low", compared with "high" husband's income, is 0.74/0.13 = **5.69**. This is different from the odds ratio of nearly 19 which we found above.

When the probability of the outcome of interest is small (< 0.2), the odds ratios are roughly equal to the relative risks.

10. Models with Interaction Terms.

The model that we fitted in Section 9 is an example of an **additive model on a logit scale** since there are no interaction terms. This means that the "effect" of one explanatory variable on the odds is the same for each category of the other explanatory variable. Such models are the easiest in terms of interpretation.

The presence of interactions poses more of a challenge in interpreting the logistic regression model. If two explanatory variables, X_1 and X_2, are involved in a significant interaction, we can no longer talk of the "effect" of X_1 without fixing the level of X_2 - the effect of one variable depends on the level of the other.

10.1 Fitting the Model with Interactions

Let us suppose that we have data for 2,000 women who gave birth 12 months ago, and we have information on their age at last birthday (AGE), occupation (OCCUP), and the husband's level of income (HINC). The variables OCCUP and HINC are coded as in Section 9, and AGE is coded 1 if the woman's age at last birthday is under 30 and is coded 2 otherwise. The dependent variable is again WORK, as defined in Section 9.

Suppose that our data suggest that the model which includes an interaction between HINC and OCCUP, which we write in "shorthand" as

$$\text{AGE + HINC + OCCUP + HINC.OCCUP ,} \qquad (10.1)$$

is a better model than the additive model

$$\text{AGE + HINC + OCCUP .} \qquad (10.2)$$

That is, we have a significant interaction between HINC and OCCUP.

We can easily fit the model that includes the interaction term in SPSS without having to explicitly compute interaction variables – SPSS will do this automatically for us. Suppose the estimates obtained from fitting this model (with the **last** categories as the reference) are those shown in Table 10.1.

Table 10.1: Estimates from the Interaction Model (Back to Work Data)

Variable	Estimate
HINC(1)	1.30
HINC(2)	1.49
OCCUP(1)	-2.21
OCCUP(2)	-1.83
AGE(1)	-1.34
HINC(1).OCCUP(1)	-0.48
HINC(1).OCCUP(2)	1.41
HINC(2).OCCUP(1)	0.32
HINC(2).OCCUP(2)	1.53
Constant	2.31

Note the variables (all dummy variables) that have been set up to represent the interaction between HINC and OCCUP. For example, the HINC(1).OCCUP(2) term is a dummy variable created by multiplying the HINC(1) dummy by the OCCUP(2) dummy. So, the HINC(1).OCCUP(2) term is equal to 1 if both HINC(1) = 1 and OCCUP(2) = 1 (i.e. a woman in casual/unskilled employment and whose husband has a low level of income), and 0 otherwise.

With this model, we can interpret the variable AGE in the usual way - the odds of returning to work for women aged under 30 are exp(-1.34) = **0.26** times the odds of returning to work for older women, i.e. **74% lower**.

Since HINC and OCCUP are involved in a significant interaction, **we can no longer interpret HINC independently of OCCUP**. We can now only talk of the "effect" of HINC on the odds of returning to work for a fixed level of OCCUP. The best way to interpret this model is through the use of a **Multiple Classification Analysis** (**MCA**) table or an **interaction plot**.

10.2 Calculating Probabilities for a MCA Table or an Interaction Plot

An MCA table is a table of probabilities classified by one or two categorical explanatory variables. This involves calculating the probabilities of the success outcome for various categories of the explanatory variable(s) of interest, while holding all other explanatory variables in the model constant at their **mean value** (or some other suitable constant, e.g. zero).

Step 1:

How do we find these mean values?

For continuous explanatory variables, you should calculate the mean using the usual formula. For categorical variables, the means are just the observed proportions at each level. For example, suppose in our data there were 800 women aged less than 30 years and 1,200 aged 30+. The mean for the AGE(1) dummy variable is then 800/2000 = **0.4**, the observed proportion in category 1. (Note that the observed proportion in category 2 is 1200/2000 = **0.6**. However, this is the reference category and so there is no dummy variable explicitly set up to represent this category.)

Step 2

To interpret the interaction between HINC and OCCUP, we set the AGE(1) variable to its mean value and calculate the estimated probabilities for all combinations of HINC and OCCUP. For example, the probability of returning to work within six months of giving birth for a woman in casual employment (OCCUP = 1) whose husband has low income (HINC = 1) is:

$$\hat{\pi} = \frac{e^{2.31+1.3-2.21-0.48-1.34\times 0.4}}{1+e^{2.31+1.3-2.21-0.48-1.34\times 0.4}} = 0.5948.$$

Note that the estimate for AGE(1) (i.e. –1.34) is multiplied by the mean value of AGE(1), i.e. the observed proportion of women in this category.

Similarly, we can calculate the probability of returning to work for professional women (OCCUP = 3) whose husbands have high income (HINC = 3) as

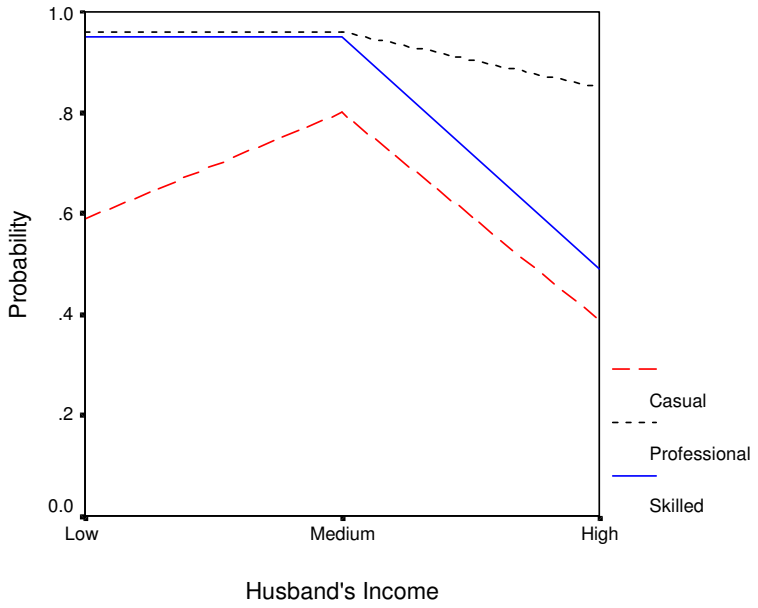
$$\hat{\pi} = \frac{e^{2.31-1.34\times 0.4}}{1+e^{2.31-1.34\times 0.4}} = 0.855.$$

Estimated probabilities for every combination of HINC and OCCUP are given in Table 10.2. These probabilities can also be presented as an interaction plot as shown in Figure 10.1.

Table 10.2: Estimated Probabilities of Returning to Work after Giving Birth

Respondent's Occupation	Husband's Level of Income		
	Low	Medium	High
Casual/unskilled	0.59	0.80	0.39
Skilled/clerical	0.95	0.95	0.49
Professional/managerial	0.96	0.96	0.85

Figure 10.1: Estimated Probabilities of Returning to Work after Giving Birth, by Husband's Income and Occupation



How do we interpret the plot?

The plot shows that overall, irrespective of husband's income level, women in professional and managerial occupations are the most likely to return to work, followed by women in skilled manual and clerical employment.

For professional women, the probabilities of returning to work are very high and about the same regardless of whether the husband's income is "low" or "medium". However, when the husband's income is "high", professional women have a slightly lower probability of returning to work.

For skilled manual, clerical, and casual employees, the probabilities of returning to work are much lower when the husband has "high" income, compared with the case when the husband has "medium" or "low" income.

Amongst women in casual/unskilled employment, the highest probability of returning to work is for women whose husbands have "medium" income.

11. Strategy for Model Selection

As in linear regression, SPSS offers a range of procedures for model selection. The following stepwise methods are available:

- Forward: conditional
- Forward: LR
- Forward: Wald
- Backward: conditional
- Backward: LR
- Backward: Wald

We will not discuss these methods in detail. As a general guide, choose:

- **Forward: LR** when you have a large number of explanatory variables.
  As in linear regression, the "forward" methods will start with no predictor variables in the model and then enter variables one at a time, at each step adding the predictor with the largest **score statistic** whose significance value is less than 0.05. At each step, SPSS will check for significance of variables already in the model to see if any should be removed. Removal is based on the likelihood ratio test. See Field (2000, Section

5.3.2, p174-192) for a nice example of interpreting the SPSS output from the "Forward: LR" method.

- **Backward: LR** when you have a small number of variables and can manage to have them all in the model.
  SPSS starts with the full model, removing variables based on the likelihood ratio test.

As in the case of linear regression, many writers would argue against the use of one of these automatic procedures since it takes any decision out of the researcher's hands and the decision to include or exclude a predictor variable is based purely on mathematical criteria. There may be existing theoretical literature which indicates the importance of certain predictors, in which case these variables should be included in the model. One situation in which the use of a stepwise procedure may be useful is when there is no previous research available to tell you which variables to expect to be reliable predictors. See Field (2000, Section 4.2.3, p119-121).

Another important consideration is the number of predictors in your model. Don't include too many – one rule of thumb is that you should have at least 15-20 cases per predictor variable.

Note that, unlike in linear regression, all dummy variables used to represent the same categorical variable are entered (or removed) from the model together in one go.

Check for intuitively meaningful interactions. However, you should avoid having too many interaction terms in your model, especially if the contribution to explanatory power of the model is small.

12. Diagnostic Methods

As in the case of linear regression, we stress the importance of examining the residuals after an analysis. Field (2000) uses the nice analogy that running a logistic regression without checking how well the model fits the data is like buying a new pair of trousers without trying them on!

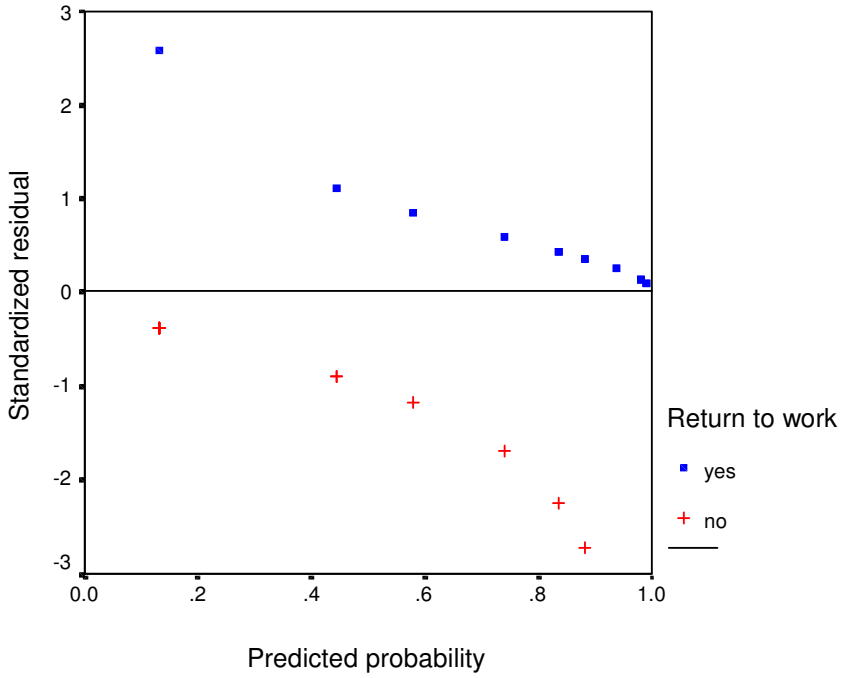
For "individual-level" binary data, the **standardized residuals** (or Pearson residuals) are given by the formula

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}, \qquad (12.1)$$

where y_i is the observed value of the response variable and $\hat{\pi}_i$ is the fitted probability for the i^{th} observation in the data. The mean of these residuals is 0 and the variance is (approximately) 1. However, they are not normally distributed – notice that y_i can equal only 0 or 1, and r_i can assume only two values for any particular combination of the predictor variables - and so the normal P-P plot and histogram that we constructed in the case of linear regression are not strictly appropriate.

The standardized residuals for "individual-level" data should still be obtained and checked for any outliers (i.e. look for residuals larger than 2 in magnitude), and also for a visual check of poor fit. Although large values of the residuals may indicate failure of the model to fit at those points, one must be cautious about regarding these large residuals as extreme. These cases should be inspected closely to try to find a good reason why they were "unusual" – don't just delete them so that the model fits better!

Figure 12.1: Plot of Standardized Residuals against Predicted Probabilities of Returning to Work after Giving Birth



For the "Back to work" data and the fitted model (9.1), Figure 12.1 plots the standardized residuals against the predicted values (probabilities). [In SPSS, you need to click the **Save** button in the logistic regression dialogue box, and then check **Probabilities** (under Predicted Values) and check **Standardized** (under Residuals). This "saves" the standardized residuals and predicted probabilities as new variables. Then use **Scatterplot**.]. The plot distinguishes between WORK = 1 or 0 to see which observations have been incorrectly predicted.

13. Logistic Regression for Grouped Data

Often we do not have the "individual-level" data available, but we may have access to tabulated (grouped) data presented in reports. Logistic regression can also be performed on **grouped data** where the aim is to model **proportions**. In this case, the data are assumed to have a *binomial distribution* where we observe the number y of **success** outcomes out of n "trials". The modelling and interpretation are the same as logistic regression for **binary data**.

13.1 An Example

To evaluate the effects of the drug AZT in slowing the development of AIDS symptoms, 338 veterans were assigned to one of two groups (see Agresti (1996), page 119): those who received AZT immediately (Use of AZT = YES), and those who didn't but waited until their T cells showed severe immune weakness (Use of AZT = NO). The race of the veteran (white or black) was also recorded.

We are interested in modelling the development of AIDS symptoms for white and black veterans, according to whether or not they used AZT. The data are shown in Table 13.1.

Table 13.1: AIDS Symptoms Data, Classified by Race and Use of AZT

Race	Use of AZT	Symptoms		
		Yes	No	Total
White	YES	14	93	**107**
	NO	32	81	**113**
Black	YES	11	52	**63**
	NO	12	43	**55**

We are interested in modelling π, the proportion of "success" outcomes for each combination of RACE and AZT. Here, a "success" outcome is the development of AIDS symptoms.

We can fit a "main effects" model (i.e. one without interactions) of the form

$$\text{logit}(\pi) = \beta_0 + \beta_1 \text{AZT} + \beta_2 \text{RACE} \ , \qquad (13.1)$$

where AZT is a dummy variable representing AZT use (1 = yes, 0 = no), and RACE is a dummy variable representing race of the veteran (1 = white, 0 = black).

13.2 Entering the Data in SPSS

In SPSS, we need to input the information for **each combination of values of the explanatory variables AZT and RACE** in a separate row in the Data Editor. In addition to the values of AZT and RACE, we need a column for the **response frequency** (we'll call it FREQ) and a column for the **number of trials** (we'll call it TOTAL).

The response frequency in this case is the number of veterans developing AIDS symptoms. For example, for the cell "white veterans who used AZT immediately" the response frequency is **14** out of a total "number of trials" of **107** veterans.

So, in SPSS the data would be entered in 4 rows as follows:

RACE	AZT	FREQ	TOTAL
1	1	14	107
1	0	32	113
0	1	11	63
0	0	12	55

13.3 Fitting the Model in SPSS

There is no drop-down menu for logistic regression for grouped data in SPSS. You need to use the **Probit Analysis** module (select **Analyze | Regression | Probit…**) and then choose the **Logit** option from the **Model** panel.

This approach of using the probit module is rather limited. It allows only one **factor** (i.e. categorical explanatory variable) and several **covariates** (continuous explanatory variables). If you have several categorical variables, you can "trick" SPSS by creating dummy variables to represent them and then specify these as "covariates". Thus, in our example, by coding AZT use and RACE as 0 or 1, we can declare these as covariates.

13.4 Interpreting the Output

The SPSS output from fitting the model (13.1) is given in Figure 13.1.

- The estimate of β_1 is **-0.719** with a standard error of **0.279**.
- The estimated odds ratio of developing AIDS symptoms for those who received AZT immediately, compared to those who waited until their T cells showed weakness, is exp(-0.719) = **0.49**. In other words, the odds of developing AIDS symptoms are about half as great for those who took AZT immediately. This is irrespective of their race.
- The estimate of β_2 is **0.055** with a standard error of **0.289**. This estimate is not significant (Wald's test), so we do not interpret it.

13.5 Goodness-of-Fit Test

For grouped data, we can think of having a residual for each "cell", i.e. for each combination of values of the explanatory variables. The **Pearson** residuals are defined as

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}, \qquad (13.2)$$

where y_i is the observed response frequency for the i^{th} cell (i.e. the number showing AIDS symptoms), $\hat{\pi}_i$ is the predicted probability (of showing AIDS symptoms), and n_i is the total number of subjects in the cell. Note: $n_i\hat{\pi}_i$ is the expected number of responses for the i^{th} cell.

A test of the null hypothesis that "the model fits well" is then based on the **chi-squared goodness-of-fit test statistic**, which is calculated from the Pearson residuals as

$$X^2 = \sum_i e_i^2 = \sum_i \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}. \qquad (13.3)$$

Figure 13.1: SPSS Output from Probit (AIDS Symptoms Data)

```
* * * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * *

DATA  Information

 4 unweighted cases accepted.

MODEL Information

      ONLY Logistic Model is requested.

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -  * * * * * *
* * *  P R O B I T    A N A L Y S I S  * * * * * *

 Parameter estimates converged after 13 iterations.
 Optimal solution found.

 Parameter Estimates (LOGIT model:
(LOG(p/(1-p))) = Intercept + BX):

           Regression Coeff.  Standard Error     Coeff./S.E.

   RACE               .05546          .28861          .19217
   AZT               -.71947          .27898        -2.57894


                   Intercept  Standard Error  Intercept/S.E.

                    -1.07355          .26294        -4.08289

Pearson  Goodness-of-Fit Chi Square =1.391  DF = 1  P =  .238

Since Goodness-of-Fit Chi square is NOT significant, no heterogeneity
factor is used in the calculation of confidence limits.
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

 Covariance(below) and Correlation(above) Matrices of Parameter Estimates

                  RACE        AZT

 RACE            .08330     .04131
 AZT             .00333     .07783



* * * * * * * * * * * * *  P R O B I T    A N A L Y S I S  * * *

 Observed and Expected Frequencies


       Number of    Observed    Expected
  RACE   Subjects   Responses   Responses    Residual     Prob

  1.00      107.0        14.0      16.010      -2.010   .14962
  1.00      113.0        32.0      29.990       2.010   .26540
   .00       63.0        11.0       8.990       2.010   .14270
   .00       55.0        12.0      14.010      -2.010   .25473

**************************************************************
```

SPSS gives this statistic as **1.391**. For this example, the test statistic is based on 1 degree of freedom (number of cells – number of parameters estimated, i.e. 4 – 3 = 1). The p-value associated with this statistic is **0.238** and so we therefore **accept H_0**. We conclude that the model fits very well. This test is valid as long as most of the fitted counts (expected responses) are ≥ 5.

Since in fact the race of the veteran was not important, we could check to see if the simpler model with only AZT fits the data well. The chi-squared statistic for this model is **1.41** (p-value > 0.49), and this suggests that the model with only AZT fits the data very well.

13.6 Residual Checks

With logistic regression for grouped data, large-sample results apply so that the residuals defined in (13.2) should approximately be normally distributed, and so are suitable for model-checking. However, SPSS does not provide any residual checks for this analysis. The Probit module does list the residuals $y_i - n_i\hat{\pi}_i$ (see Figure 13.1), so that the Pearson residuals in (13.2) can be calculated. These can then be checked for "large" values to assess where the model is a poor fit.