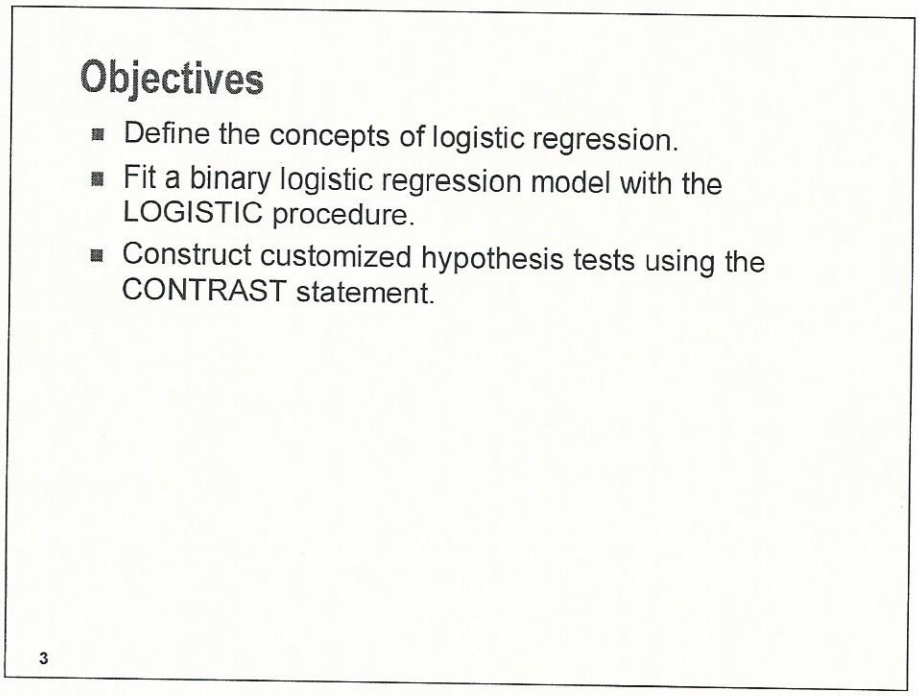
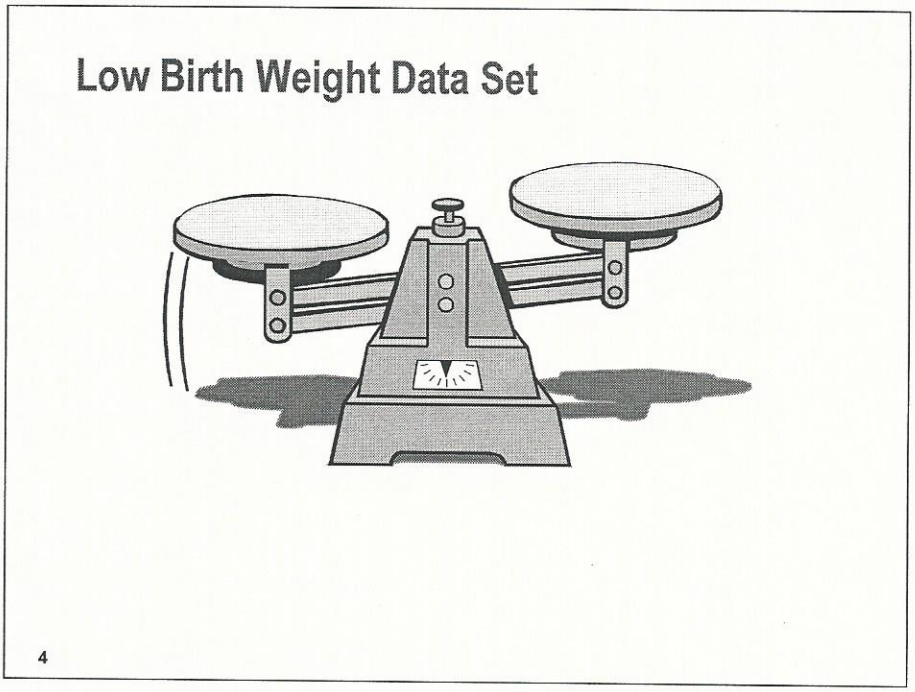
1.1 Introduction to Logistic Regression





Example: Babies with low birth weights (defined to be less than 2500 grams) are a concern because of their potential medical problems. Health researchers want to identify possible contributing factors to low birth weight and recommend strategies to reduce the number of low birth weight babies.

The variables in the data set are

low	low birth weight (1=yes, 0=no)
mother_age	age of mother in years
mother_wt	mother's weight at last menstrual period
socio	socioeconomic status (1=low, 2=medium, 3=high)
alcohol	drinking status during pregnancy (1=yes, 0=no)
prev_pretrm	history of preterm labor (0=none, 1=one or more)
hist_hyp	history of hypertension (1=yes, 0=no)
uterine_irr	presence of uterine irritability (1=yes, 0=no)
phy_visit	physician visits during the first trimester (1=yes, 0=no).

The data is stored in a SAS data set named **sasuser.birth**.

The data was modified from an example in Hosmer and Lemeshow (2000).

Linear Probability Model

$$p_i = \beta_0 + \beta_1 X_1$$

Problems

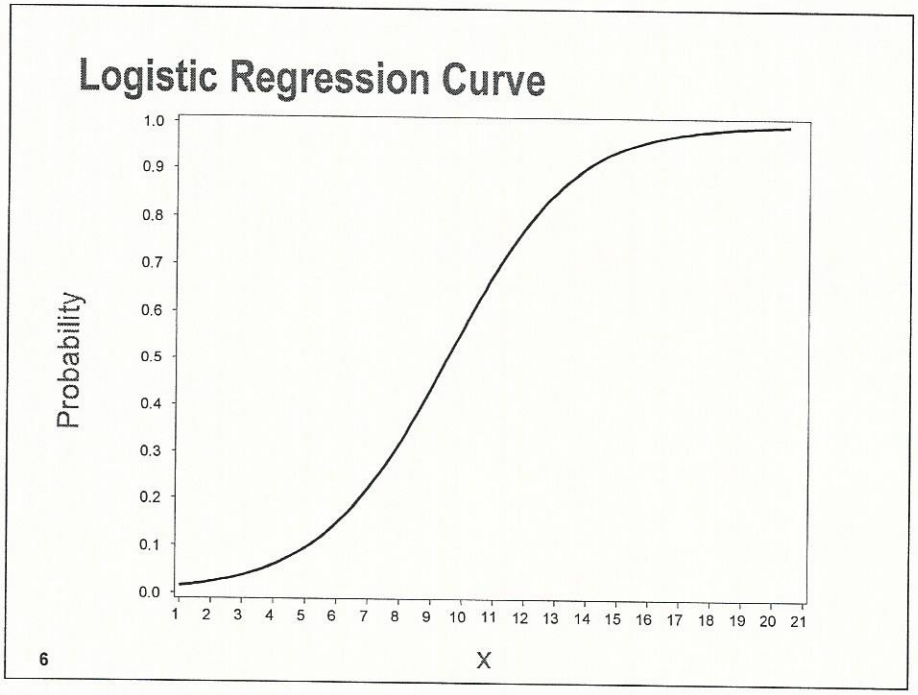
- Probabilities are bounded, but linear functions can take on any value.
- The relationship between probabilities and X is usually nonlinear.

5

One way to identify possible contributing factors to low birth weight is to build a linear probability model where the outcome is the probability of having a low birth weight baby. Such a model implies that the probability of low birth weight is a linear function of the predictor variables. The regression coefficients would therefore have a straightforward interpretation in this model. For example, you can estimate the change in the probability of low birth weight, given a one-unit change in **`alcohol`**.

Unfortunately, the linear probability model has some serious shortcomings. The predicted values from a linear model can assume, theoretically, any value. However, probabilities are by definition bounded between 0 and 1. Thus, the model can only be valid over a finite range of predictor variable values. A more appropriate model would somehow constrain the predicted probabilities to be between 0 and 1.

Another shortcoming is that the observed relationship between the probability of an outcome and the predictor variables is usually nonlinear rather than linear. For example, a one-unit change in the predictor variable may have less impact when the probability is near 0 or 1 than when the probability is near .50. In fact, the relationship often resembles an S-shaped curve rather than a linear function (Hosmer and Lemeshow 2000).



The nonlinear relationship between the probability of the outcome and the predictor variables is due to the constrained scale of the outcome. Furthermore, the relationship is fairly linear in the middle of the range of the probabilities (.20 to .80) and fairly nonlinear at the end of the range (0 to .20 and .80 to 1).

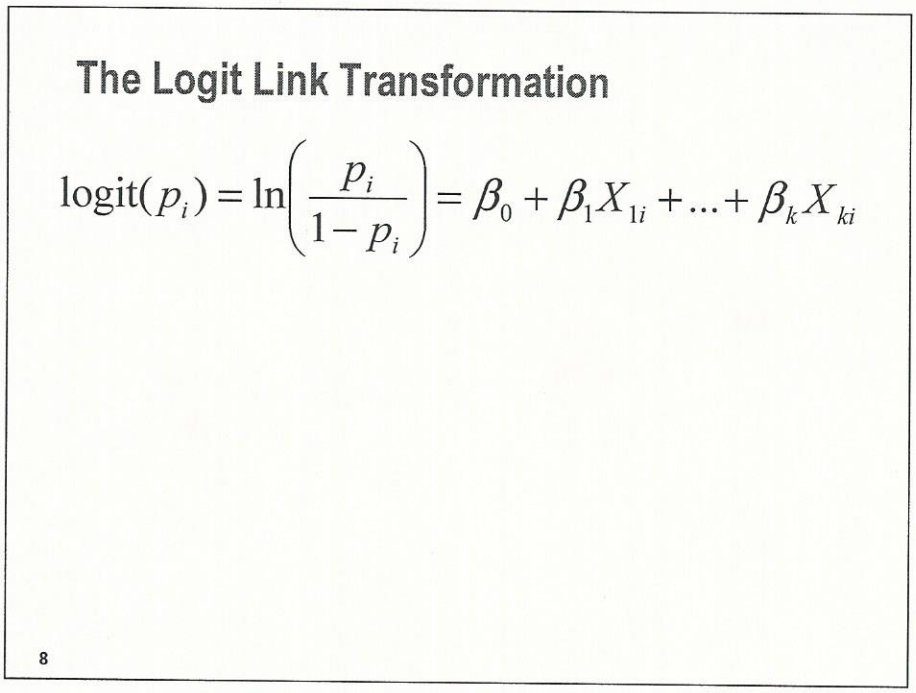
The parameter estimate of this curve determines the rate of increase or decrease of the estimated curve. When the parameter estimate is greater than 0, the probability of the outcome increases as the predictor variable values increase. When the parameter estimate is less than 0, the probability decreases as the predictor variable values increase. As the absolute value of the parameter estimate increases, the curve has a steeper rate of change. When the parameter estimate is equal to 0, the curve resembles a straight line.

Logistic Regression Model

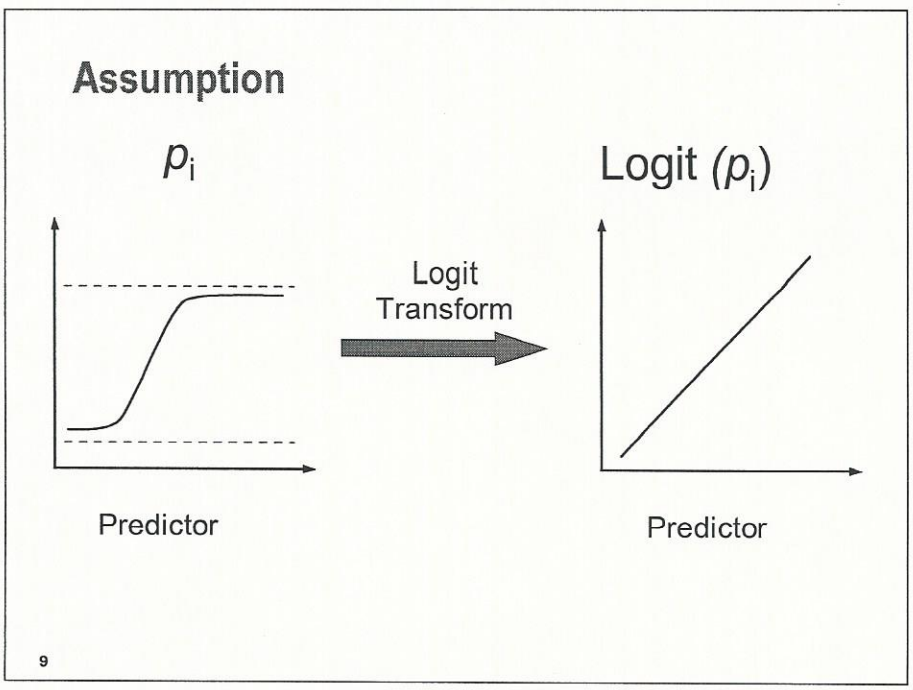
$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki})}}$$

7

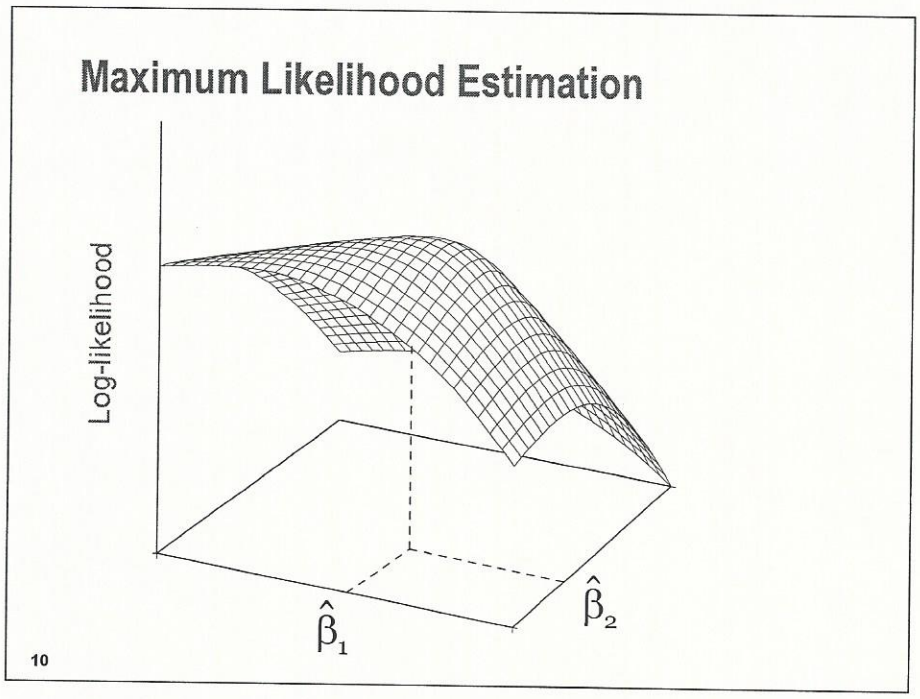
The equation for the logistic regression model that refers directly to the probability of the outcome is shown above. This equation has the desired property that the predicted probabilities will always be between 0 and 1. This model is nonlinear because the parameter estimates do not enter the model equation linearly. Furthermore, the model permits the rate of change of the probabilities to vary as the predictor variable values vary.



The *logit transformation* is the log of the odds, which is the ratio of the probability of the outcome to the probability of no outcome. To create a linear model, the logit transformation is applied to the probability. Unlike a probability, the logit is unbounded because transforming the probability to odds removes the upper bound, whereas taking the natural logarithm of the odds removes the lower bound. The model (also called the logistic regression model) is now linear because the logit is linear in its parameters. Furthermore, the model gives estimated probabilities that are between 0 and 1.



A key assumption in the logistic regression model is that the logits are linearly related to each predictor variable. For binary predictor variables, this is not a problem because a straight line connects two points. However, for ordinal or continuous predictor variables, this assumption should be examined with the use of logit plots.



Least squares estimators for the model parameters are not used in logistic regression because the variance of the outcome is not constant across the values of the predictor variable. Because of the nonconstant variance, least squares parameter estimates are not efficient (other estimation methods yield smaller standard errors) and the standard error estimates are not consistent (the estimated standard errors could be biased). Therefore, the method of maximum likelihood is used to produce estimators that are consistent, asymptotically efficient, and asymptotically normal. This method finds the parameter estimates that are most likely to occur, given the data, by maximizing the likelihood function, which expresses the probability of the observed data as a function of the unknown parameters.

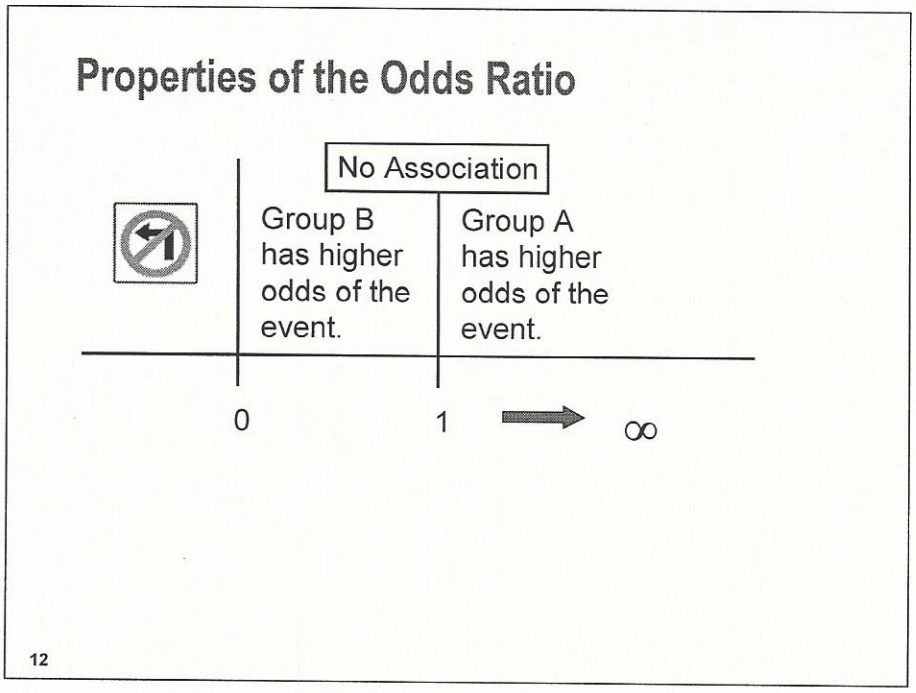
Odds Ratio

$$\text{Odds}_A = \frac{p_A}{1-p_A} \qquad \text{Odds}_B = \frac{p_B}{1-p_B}$$

$$\text{Odds Ratio} = \frac{\text{odds}_A}{\text{odds}_B}$$

11

A useful feature of the logistic regression model is that the parameter estimates can be converted to odds ratios. The odds of an outcome is the ratio of the expected number of times that the outcome will occur to the expected number of times the outcome will not occur. In other words, the odds are simply the ratio of the probability of the outcome to the probability of no outcome. The odds ratio compares the odds of the outcome in one group to the odds of the outcome in another group. For example, the odds ratio for a binary predictor variable (**X**) would compare the odds of the outcome when **X**=1 to the odds of the outcome when **X**=0.



The odds ratio shows the strength of the association between the predictor variable and the response variable. If the odds ratio is 1, then there is no association between the predictor variable and the response. If the odds ratio is greater than 1, then group A has higher odds of having the event. If the odds ratio is less than 1, then group B has higher odds of having the event. For example, an odds ratio of 3 indicates that the odds of getting the event in group A is 3 times that in group B.

Odds Ratio from a Logistic Regression Model

Estimated logistic regression model:

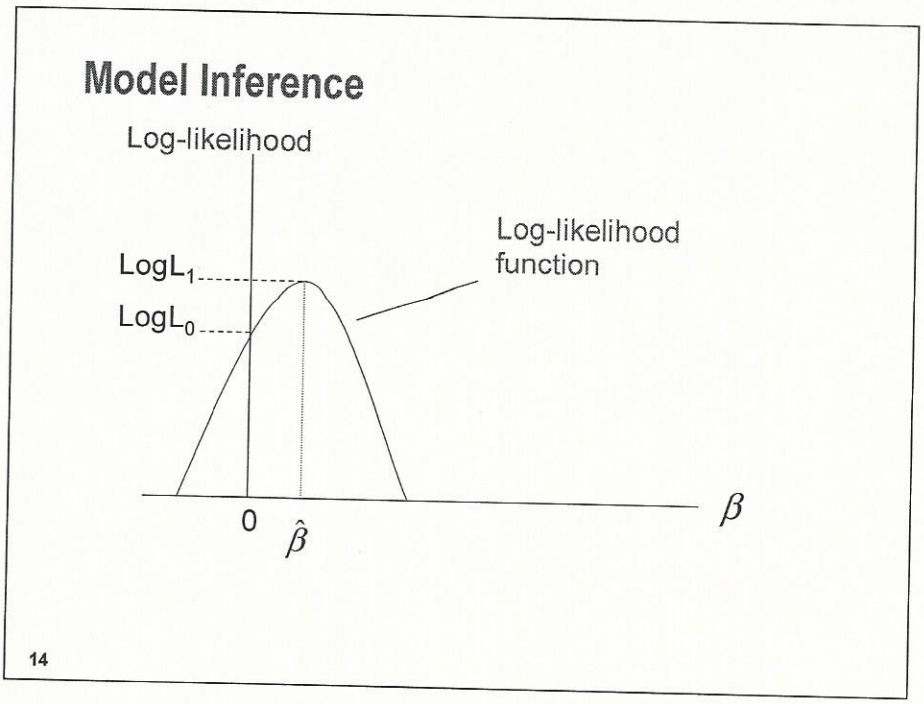
logit(p) = –.7567 + .4373*(**gender**)

Estimated odds ratio (Females to Males):

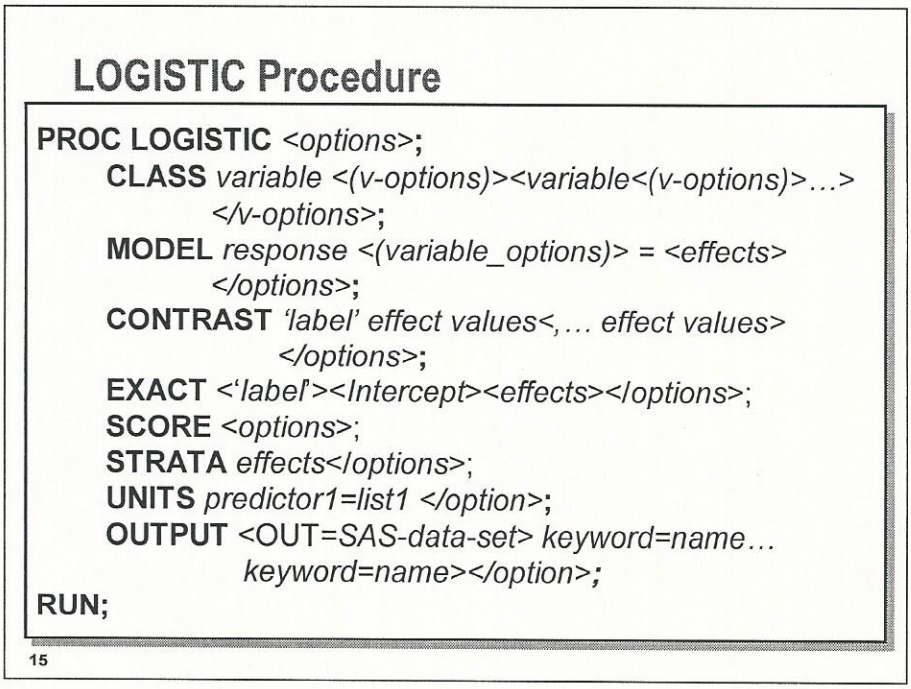
odds ratio = (e-.7567+.4373)/(e-.7567) = 1.55

13

In logistic regression, odds ratios are simple functions of the parameters. For example, suppose you want to examine the relationship between **gender** and **outcome**. The odds ratio for **gender** compares the predicted odds of females to have the outcome to the predicted odds of males. If **gender** is coded 1 for females and 0 for males, the odds ratio is simply the exponentiation of the parameter estimate for **gender**. An odds ratio of 1.55 means that females are 1.55 times more likely to have the outcome than males.



To test the null hypothesis that all regression coefficients of the model are 0, the likelihood ratio test statistic is computed. This statistic compares the log-likelihood values at the fitted parameter estimates ($LogL_1$) to the log-likelihood values when the parameter estimates are 0 ($LogL_0$). So that it follows a chi-square distribution, the statistic is computed by the formula $-2(LogL_0-LogL_1)$. In the above diagram, the likelihood ratio statistic is twice the vertical distance between the values of the log-likelihood function at $LogL_1$ and at $LogL_0$.



The LOGISTIC procedure fits logistic regression models for binary, ordinal, or nominal response data. Enhancements to PROC LOGISTIC in SAS®9 include the STRATA statement, which enables you to perform a conditional logistic regression on binary response data, and the SCORE statement, which enables you to score a data set using a previously fitted model.

PROC LOGISTIC has options that control how to select effects (either variables or interactions) in and out of the model. When there are no interaction terms, a CLASS variable can enter or leave a model in a single step. When there are interaction terms, the selection process also depends on whether you want to preserve model hierarchy (which is explained later in the course). For example, you can specify whether model hierarchy is to be preserved, how model hierarchy is applied, and whether a single effect or multiple effects can be moved in a single step.

PROC LOGISTIC provides a CONTRAST statement for specifying customized hypothesis tests concerning the model parameters. The CONTRAST statement can also be used to obtain odds ratio estimates for various levels of the CLASS variables and confidence intervals around odds ratios for variables that are involved in an interaction.

In the MODEL statement, the response variable can be specified in two ways.

- The *events/trials* syntax represents a ratio of variables where the **event** variable indicates the number of observations with the response of interest for a particular combination of predictor variable values and the **trial** variable indicates the number of observations in the same combination of predictor variable values. This response is useful when you have a summarized data set where each observation represents a unique combination of predictor variable values.
- The **response** variable can also be specified as a positive or negative response for each observation. In other words, each observation represents a single case.

Selected PROC LOGISTIC statement options:

NOPRINT suppresses all displayed output. This option temporarily disables the Output Delivery System (ODS).

NAMELEN= specifies the length of effect names in tables and output data sets to be *n* characters, where *n* is a value between 20 and 200. Because the default length is 20 characters, you may have to increase the length for some interaction terms.

Selected LOGISTIC procedure statements:

CLASS specifies the classification variables to be used in the analysis. The CLASS statement must precede the MODEL statement.

MODEL specifies the response variable (which can be binary, ordinal, or nominal) and the predictor variables (which can be character or numeric). The MODEL statement is required, and only one is allowed with each invocation of PROC LOGISTIC.

CONTRAST provides a mechanism for obtaining customized hypothesis tests. There is no limit to the number of CONTRAST statements that you can specify, but they must appear after the MODEL statement.

EXACT performs exact tests of the parameters for the specified effects and optionally estimates the parameters and outputs the exact conditional distributions. You can specify several EXACT statements, but they must follow the MODEL statement.

SCORE creates a data set that contains all the data in the DATA= data set together with posterior probabilities and, optionally, prediction confidence intervals. You can specify several SCORE statements.

STRATA names the variables that define strata or matched sets to use in a stratified conditional logistic regression of binary response data. The STRATA variables can be either character or numeric, and the formatted values of the STRATA variables determine the levels. You can use formats to group values into levels.

UNITS enables you to obtain an odds ratio estimate for a specified change in a predictor variable. The unit of change can be a number, standard deviation (SD) or a number times the standard deviation (2*SD).

OUTPUT creates an output data set containing all the variables from the input data set and the requested statistics.

Effect Coding

			Design Variables	
Class	Value	Label	1	2
socio	1	low	1	0
	2	medium	0	1
	3	high	-1	-1

16

In the CLASS statement in PROC LOGISTIC, you can specify the coding scheme for the design variables created from the CLASS variable. For effect coding (also called deviation from the mean coding), the number of design variables created is the number of levels of the CLASS variable minus 1. For example, because the variable **socio** has three levels, only two design variables are created. For the last level of the CLASS variable, all the design variables have a value of –1. Parameter estimates of the CLASS main effects using this coding scheme estimate the difference between the effect of each level and the average effect over all levels. Effect coding is the **default** in PROC LOGISTIC.

Reference Cell Coding

			Design Variables	
Class	Value	Label	1	2
socio	1	low	1	0
	2	medium	0	1
	3	high	0	0

17

For reference cell coding, parameter estimates of the CLASS main effects estimate the difference between the effect of each level and the last level. For example, the effect for the low level would estimate the difference between low and high. You can choose the reference level with the REF= option.

Testing Individual Contrasts

1. Write out the model equation.

   logit (p) = $\beta_0 + \beta_1$*`socio`

2. Write out the model equation for each level of the class variable.

   logit (low income) = $\beta_0 + \beta_1 + 0\beta_2$
   logit (medium income) = $\beta_0 + 0\beta_1 + \beta_2$
   logit (high income) = $\beta_0 + 0\beta_1 + 0\beta_2$

3. State the null hypothesis.

   H_0: low income = medium income

23

In this course, contrasts are used to compute the odds ratio that compares one level of a CLASS variable to another level. However, in order to construct a contrast, you need to be able to define the coefficients for the test of the hypothesis. For example, to compute the odds ratio that compares low to medium for the variable `socio` in a CONTRAST statement, you should first write the model equation for each of the levels.

Testing Individual Contrasts

4. Rewrite the null hypothesis using the model equations.

   H_0: $\beta_0 + \beta_1 = \beta_0 + \beta_2$

5. Rewrite the null hypothesis so that 0 is on the right side.

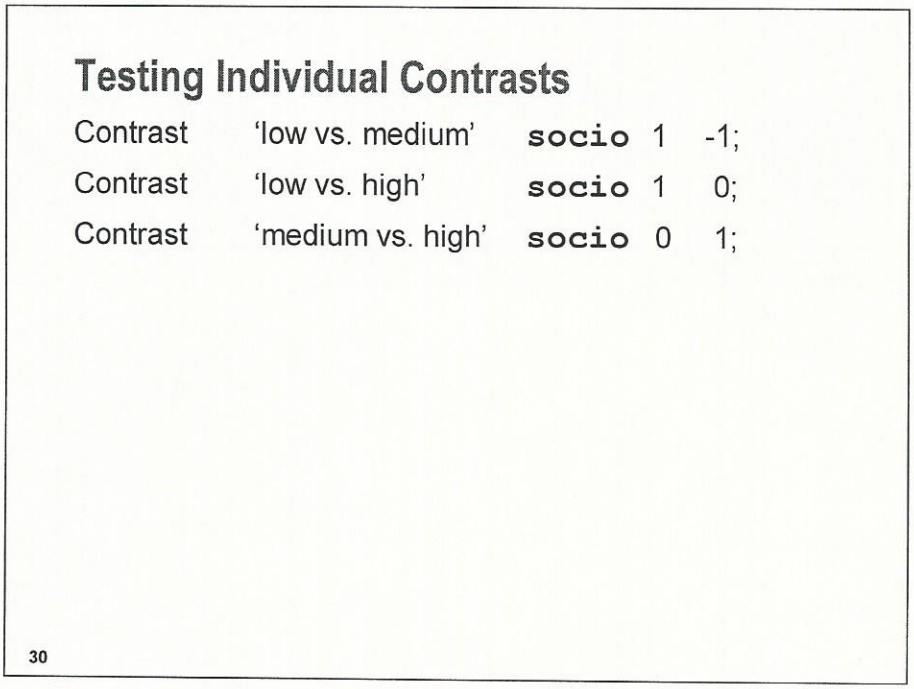
   H_0: $\beta_1 - \beta_2 = 0$

6. Identify the coefficients for the effects.

   H_0: $1\beta_1 - 1\beta_2 = 0$

29

To obtain the coefficients, take the null hypothesis of interest and rewrite the equation so that 0 is on the right side. Thus, the coefficients for the contrast comparing low versus medium are 1 –1 if you use reference cell coding.



The CONTRAST statement is made up of the following components:

label identifies the contrast in the output. A label is required for every contrast specified, and it must be enclosed in quotes.

effect identifies an effect that appears in the MODEL statement. You do not need to include all effects that are included in the MODEL statement.

values identifies the coefficients associated with the effect. To correctly specify your contrast, it is crucial to know the ordering of the parameters within each effect and the variable levels associated with each parameter.

If an effect is not specified in the CONTRAST statement, all of its coefficients are set to 0. If too many values are specified for an effect, the extra ones are ignored. If too few values are specified, the remaining ones are set to 0.