Chapter 8

Logistic Regression Model

Aim: Upon completing the chapter, the learner should be able to use a logistic regression model to estimate the magnitude of the association between exposure and disease, controlling for potential confounders.

8.1 Model Definition

The *logistic regression model* is a statistical model that can explain the behavior of a dichotomous variable or a binomial proportion through different predictor variables. In epidemiology these predictors may include variables of exposure, potential confounding variables, and other types of controlling variables. A simple binary logistic regression model (Hosmer and Lemeshow, 2000; Collett, 2002) is defined using the following expression:

$$\Pr\left(Y_i = 1\right) = p_i = \frac{e^{-(\beta_0 + \beta_1 X)}}{1 + e^{-(\beta_0 + \beta_1 X)}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

where:

- p_i indicates the probability of having the diagnosis of interest in the ith subject, that is, the probability of the ith subject being a case
- Y indicates a dichotomous variable, coded as $Y = 1$ for a case and $Y = 0$ for a control
- X indicates a predictor variable
- β_0 indicates the coefficient that is not affected by any predictor
- β_1 indicates the coefficient that affects the predictor X

By using the logarithmic transformation of $\left[p_i / (1 - p_i) \right]$, we obtain the *logit* function, as can be seen in the following:

$$g(x) = \text{logit}(p_i) = \ln\left[\frac{p_i}{1 - p_i} \right] = \beta_0 + \beta_1 x$$

where:

β_0 indicates the value of the logit(p_i) when $X = 0$

β_1 indicates changes in the logit(p_i) per unit of change in X

8.2 Parameter Estimation

The logistic regression model is adjusted by estimating the unknown parameters β_0 and β_1. One of the procedures for estimating the unknown parameters is known as the *maximum-likelihood estimation* (MLE), which is based on the likelihood function. This function is given by the joint probability of observing the sample data and is demonstrated in the following:

$$L = \Pr(Y_1, Y_2, \ldots, Y_n)$$

where n is the number of observations or the sample size. The likelihood function provides support for a particular value of the parameter β_i, given an observed data. If the observed data provide more support for one value of the parameter than for another value, then the likelihood is higher for the former parameter value (Marschener, 2015).

Under the assumption that the observed data are independent, the likelihood function can be expressed as follows:

$$L = \Pr(Y_1) * \Pr(Y_2) *, \ldots, * \Pr(Y_n)$$

Defining $\Pr(Y_i)$ depends on the probability distribution of the random variable, Y_i. In the case of binomial distribution, when Y is the binomial proportion in K groups, the likelihood function is defined as

$$L(\beta) = \prod_{i=1}^{K} C_{yi}^{ni} * p_i^{y_i} * (1 - p_i)^{n_i - y_i}$$

where:

C_{yi}^{ni} is the total number of combinations with y cases given n subjects in the ith group

p_i is the probability of the disease in the ith group

If Y is a dichotomous variable ($Y = 0, 1$), the likelihood function is defined as follows:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i} * (1 - p_i)^{1 - y_i}$$

where:

p_i is the probability of having the disease in the ith subject
n is the total number of subjects

In the case of the logistic regression model, we use the coefficients βs that produce the highest value of the likelihood function. The βs obtained in this maximization process are identified as the maximum-likelihood estimates (Hardin and Hilbe, 2001).

8.3 Programming the Logistic Regression Model

There are several commands in Stata that can be used to estimate the parameters of the logistic regression model, including *logit, logistic, binreg*, and *glm*. For example, perhaps the user is interested in exploring the statistical relationship between smoking habits and oral cavity cancer with the following data (Fu et al., 2013):

	Cancer		
Smoker	*No (0)*	*Yes (1)*	*Total*
No (0)	511	333	844
Yes (1)	346	340	686
Total	857	673	1530

Note: Codes of the categories of each variable are shown in parentheses.

The database in Stata is the following:

```
+----------------------------+
| smoker   cancer   subjects |
|----------------------------|
|      0        0        511 |
|      1        0        346 |
|      0        1        333 |
|      1        1        340 |
+----------------------------+
```

The estimate of the simple logistic regression model parameters for the above grouped data can be accomplished with different commands. The difference

between them is the default output provided and the method used to maximize the likelihood function for parameters' estimation. Some of these commands and their outputs are shown below.

8.3.1 Using glm

```
glm cancer smoker [fw=subjects], fam(bin)
```

Output

```
Generalized linear models                         No. of obs      =      1530
Optimization     : ML                             Residual df     =      1528
                                                  Scale parameter =         1
Deviance         =  2083.154244                   (1/df) Deviance = 1.363321
Pearson          =  1529.999986                   (1/df) Pearson  = 1.001309

Variance function : V(u) = u*(1-u)                [Bernoulli]
Link function     : g(u) = ln(u/(1-u))            [Logit]

                                                  AIC             =    1.364153
Log likelihood   = -1041.577122                   BIC             =   -9121.705
------------------------------------------------------------------------------
             |                 OIM
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Smoker |   0.4107339   0.1038812     3.95   0.000     0.2071305    0.6143372
       _cons |  -0.428227    0.0704269    -6.08   0.000    -0.5662612   -0.2901928
------------------------------------------------------------------------------
```

The command *fam(bin)* is used to ensure that the probability distribution of the dependent variable (cancer) will follow a binomial distribution.

8.3.2 Using logit

```
logit cancer smoker [fw=subjects]
```

Output

```
Logistic regression                               Number of obs   =       1530
                                                  LR chi2(1)      =      15.69
                                                  Prob > chi2     =     0.0001
Log likelihood = -1041.5771                       Pseudo R2       =     0.0075

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   0.4107339   0.1038812     3.95   0.000     0.2071306    0.6143373
       _cons |  -0.4282271   0.0704269    -6.08   0.000    -0.5662613   -0.2901929
------------------------------------------------------------------------------
```

8.3.3 Using logistic

```
logistic cancer smoker [fw=subjects], coef
```

Output

```
Logistic regression                     Number of obs   =       1530
                                        LR chi2(1)      =      15.69
                                        Prob > chi2     =     0.0001
Log likelihood = -1041.5771             Pseudo R2       =     0.0075

------------------------------------------------------------------------------
  cancer |   Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  smoker |  0.4107339   0.1038812     3.95   0.000     0.2071306    0.6143373
   _cons | -0.4282271   0.0704269    -6.08   0.000    -0.5662613   -0.2901929
------------------------------------------------------------------------------
```

To display the estimates of the beta parameters, the **coef** option is used.

8.3.4 Using binreg

```
binreg cancer smoker [fw=subjects], coef ml
```

Output

```
Generalized linear models                  No. of obs      =       530
Optimization     : ML                      Residual df     =      1528
                                           Scale parameter =         1
Deviance         = 2083.154244             (1/df) Deviance = 1.363321
Pearson          = 1529.999986             (1/df) Pearson  = 1.001309

Variance function: V(u) = u*(1-u)          [Bernoulli]
Link function    : g(u) = ln(u/(1-u))      [Logit]

                                           AIC             = 1.364153
Log likelihood   = -1041.577122            BIC             = -9121.705

------------------------------------------------------------------------------
             |                 OIM
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   0.4107339   0.1038812     3.95   0.000     0.2071305    0.6143372
       _cons |    -0.428227   0.0704269    -6.08   0.000    -0.5662612   -0.2901928
------------------------------------------------------------------------------
```

The **coef** option is used to display the estimates of the beta parameters. The option *ml* is for obtaining the maximum-likelihood estimates.

Therefore, the fitted model for all the commands of the simple logistic regression can be determined with the following equation:

$$\text{logit}(p) = -0.43 + 0.41 * \text{smoker}$$

8.4 Alternative Database

A logistic regression model can also be used when the data are summarized as a binomial proportion (number of cases with a characteristic of interest over the number of observations). For example, let us assume the previous example with the following format:

Smoker	*Cancer*	*Total*
No (0)	333	844
Yes (1)	340	686

In parentheses are the codes for the smoker categories.

In this case, the database is entered in Stata as follows:

```
. list
    +-------------------------+
    | smoker    cases   total |
    |-------------------------|
    |      0      333     844 |
    |      1      340     686 |
    +-------------------------+
```

The syntax using the *glm* command, under this database structure, is the following:

```
glm cases smoker, fam(bin total)
```

Output

```
Generalized linear models                         No. of obs      =         2
Optimization     : ML                             Residual df     =         0
                                                  Scale parameter =         1
Deviance         =  7.90479e-14                   (1/df) Deviance =         .
Pearson          =  1.60265e-29                   (1/df) Pearson  =         .

Variance function: V(u) = u*(1-u/total)           [Binomial]
Link function    : g(u) = ln(u/(total-u))         [Logit]

                                                  AIC             =  9.063989
Log likelihood   = -7.063989303                   BIC             =  7.90e-14

------------------------------------------------------------------------------
             |                 OIM
       cases |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   0.4107339   0.1038812     3.95   0.000     0.2071306   0.6143373
       _cons |  -0.4282271   0.0704269    -6.08   0.000    -0.5662613  -0.2901929
------------------------------------------------------------------------------
```

The **fam(bin)** option is modified to **fam(bin total)** because the dependent variable denotes the number of cases and not the presence or absence of disease (dichotomous scenario).

Similar results can be obtained with the *binreg* command if the following specifications are used:

```
binreg cases smoker, or n(total) ml
```

The *or* and *ml* options are added to estimate the *odds ratio* (OR) using the maximum likelihood method.

8.5 Estimating the Odds Ratio

One of the important applications of the logistic regression model is estimating the strength or magnitude of association between the disease and exposure under study. This model can be expressed as follows:

$$\text{odds}_i = \frac{p_i}{1-p_i} = \frac{\left[1/\left(1+e^{-(\beta_0+\beta_1 X)}\right)\right]}{\left\{1-\left[1/\left(1+e^{-(\beta_0+\beta_1 X)}\right)\right]\right\}} = e^{\beta_0+\beta_1 X}$$

Using this expression, the user can obtain the OR. For example, if we assume that X takes 0 for unexposed subjects and 1 for exposed subjects, the resulting OR will be

$$\text{OR}_{\text{exp } vs. \text{unexp}} = \frac{\text{odds}_{\text{exposed}}}{\text{odds}_{\text{unexposed}}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

In this case, the OR is the exponential of the regression coefficient associated with the exposure. The syntax in Stata to estimate the OR of the previous example (using the *glm* command) is as follows:

```
glm cases smoker, fam(bin total) ef nolog noheader
```

Output

```
------------------------------------------------------------------------------
             |                 OIM
       cases | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      smoker |   1.507924   .1566449     3.95   0.000     1.230143    1.848431
       _cons |   .6516634   .0458946    -6.08   0.000     .5676437    .7481193
------------------------------------------------------------------------------
```

The ***ef*** option is added to obtain the estimated OR. The terms ***nolog*** and ***noheader*** are used to display only the parameters of the model.

The result indicates that the odds of having oral cavity cancer among smokers is 1.51 (95% CI: 1.23, 1.85) times the odds of having oral cavity cancer among nonsmokers. This OR is known as the crude OR, because the model includes only the exposure variable.

8.6 Significance Tests

8.6.1 Likelihood Ratio Test

Hypothesis testing can be performed for the logistic regression model using the statistic known as *Deviance* (*D*), which is a measure of discrepancy between observed and fitted data. This measure is defined based on the relative comparison of two likelihood functions, as follows (McCullagh and Nelder, 1999):

$$D = -2*\ln\left(\frac{\text{likelihood function} - \text{current model}}{\text{likelihood function} - \text{best model}}\right)$$

where:

likelihood function–current model is calculated with the estimate of the parameter p of the binomial distribution using the current logistic model

likelihood function–best model is calculated with the binomial proportion using the observed data

For logistic regression, the equation for D is the following:

$$D = -2\sum_{i=1}^{n} y_i \ln\left(\frac{\widehat{p_1}}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\widehat{p_i}}{1-y_i}\right)$$

This comparison is known as the *likelihood ratio test*. The syntaxes in Stata to perform this test (with the previous database) to assess the effect of the predictor *smoker* are as follows:

```
. quietly: glm cases smoker, fam(bin total)
. estimates store model1
. quietly: glm cases, fam(bin total)
. lrtest model1 .
Likelihood-ratio test                    LR chi2(1)  =     15.69
(Assumption: . nested in model1)         Prob > chi2 =    0.0001
```

The results show that removing the predictor *smoker* from the model has a significant effect (*P*-value = .0001). Therefore, it is suggested that it not be removed from the model.

8.6.2 Wald Test

The statistical assessment of specific parameters in the logistic regression model can be performed with the likelihood ratio test. Another option is the Wald test for assessing individual parameters (H_0: $\beta j = 0$) using the following statistic:

$$Z_0 = \frac{\hat{\beta}_j}{\text{SE}\left(\hat{\beta}_j\right)}$$

where $\mathrm{SE}\left(\hat{\beta}_j\right)$ is the asymptotic (i.e., large-sample) standard error of $\hat{\beta}_j$. The test statistic Z_0 follows an asymptotic standard-normal distribution, $N(0,1)$, under the null hypothesis.

An equivalent process is to calculate the square of Z_0 and use the chi-squared distribution (χ^2) to assess the null hypothesis, H_0: $\beta_j = 0$. The use of χ^2 is recommended for *two-sided alternatives* (H_a: $\beta_i \neq 0$). For *one-sided alternatives* (H_a: $\beta_i < 0$, H_a:$\beta_i > 0$), the normal distribution is recommended.

The output of the *glm* command for the logistic model shows the Wald test for each predictor. Another Stata command that can be used to perform the Wald test is ***test***. For example, using the previous database to assess the effect of the predictor *smoker,* the syntaxes are as follows:

```
glm cases smoker, fam(bin total)
test smoker
```

Output

```
. quietly: glm cases smoker, fam(bin total)
. test smoker
( 1)  [cases]smoker = 0
          chi2(  1) =   15.63
        Prob > chi2 =    0.0001
```

The likelihood ratio test and the Wald test showed the significant effect of the predictor *smoker* in the logistic regression model (*P*-value = .0001); however, the test statistics differ (15.69 vs. 15.63).

8.7 Extension of the Logistic Regression Model

The logistic regression model can be extended to include as predictors of the potential confounders and the interaction terms formed by the product of the exposure and confounders. When more than one predictor is included, the model is called a *multivariable logistic regression model* and is expressed as follows:

$$\Pr\left(Y_i = 1\right) = p_i = \frac{1}{1 + e^{-\left(\beta_0 + \beta_E * E_i + \Sigma \beta_i C_i + \sum \gamma_{(j)i} * \left(E * C\right)_{(j)i}\right)}}$$

where:

p_i indicates the probability of the ith subject's having the disease of interest
E indicates the exposure
C_i indicates the ith potential confounding variable
γ_j indicates the jth coefficient or the interaction terms associated with the product of the exposure and the potential confounding variables *(E*C)*

These interaction terms are useful to estimate the magnitude of the association in different strata.

By way of illustration, let us continue to use the previous example, in which the predictor *sex* was included as a potential confounding variable, with the following data distribution:

Smoker	*Sex*	*Cases of Cancer*	*Total*
No (0)	Female (0)	218	477
	Male (1)	115	367
Yes (1)	Female (0)	20	42
	Male (1)	370	694

Note: In parentheses are the codes for the categories of the variables *smoker* and *sex*.

To analyze these data, the following database is created in the Stata data editor:

```
     +-------------------------------+
     | smoker   sex   cases   total  |
     |-------------------------------|
  1. |      0     0     218     477  |
  2. |      0     1     115     367  |
  3. |      1     0      20      42  |
  4. |      1     1     370     694  |
     +-------------------------------+
```

In Stata, the *glm* command can be used in conjunction with the previous database to fit a logistic regression model with interaction terms. The syntax for doing so is as follows:

```
xi: glm cases i.smoker*i.sex, fam (bin total) nolog noheader
```

Output

```
. xi: glm cases i.smoker*i.sex, fam (bin total) nolog noheader
------------------------------------------------------------------------------
             |                 OIM
       cases |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Ismoker_1 |   .0770228   .3223394     0.24   0.811    -.5547508    .7087965
      _Isex_1 |  -.612164   .1452999    -4.21   0.000    -.8969466   -.3273814
_IsmoXsex_1_1 |   .8402336   .3497938     2.40   0.016     .1546503    1.525817
        _cons |  -.172333   .0919139    -1.87   0.061    -.3524809    .0078149
------------------------------------------------------------------------------
```

The multivariate logistic regression model is determined with the following equation:

$$\text{logit}(p) = -0.17 + 0.08 * \_Ismoker\_1 - 0.6 * \_Isex\_1 + 0.84 * \_IsmoXsex\_1\_1$$

where:

_Ismoker_1 is a dummy variable with value 1 if the subject smokes, otherwise is 0
_Isex_1 is a dummy variable with value 1 for males, otherwise is 0
_IsmoXsex_1_1 is a dummy variable with value 1 if the subject smokes and is a male, otherwise is 0

Start the command with ***xi:*** in the *glm* command to indicate that some of the predictors are defined as categorical. This instruction, when placed prior to the *glm* command, enables us to define the model with interaction terms. The instruction ***i.smoker*i.sex*** indicates that the logistic regression model will use as predictors *smoker, sex,* and the interaction term formed by the product of these predictors. This form is useful when there are more than two categories in the predictor variables that are defined as being categorical.

In the previous output table, the Wald test shows that there is evidence that the interaction term ***_IsmoXsex_1_1*** affects the logit(p) estimate (P-value = .016). An alternative procedure for making a statistical assessment of the interaction term is the likelihood ratio test (*lrtest*), which is recommended when the user is interested in assessing simultaneously several interaction terms. The following commands sequence perform the *lrtest* with the previous database:

```
. quietly xi: glm cases i.smoker*i.sex, fam(bin total)
. estimates store model1
. quietly: glm cases i.smoker, fam(bin total)
lrtest model1 .
Likelihood-ratio test                    LR chi2(2)  =     18.62
(Assumption: . nested in model1)         Prob > chi2 =    0.0001
```

The results indicate that the interaction term composed of *smoker* and *sex* is statistically significant (P-value = .0001), which is similar to what was found using the Wald test. Therefore, the variable sex modifies the relationship between smoking status and cancer. As a consequence, it is recommended to estimate sex-specific OR using the *lincom* command as follows:

```
quietly xi: glm cases i.smoker*i.sex, fam(bin total)
*In females
. lincom _Ismoker_1 , or
(1)  [cases]_Ismoker_1 = 0
------------------------------------------------------------------------------
  cases | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------+---------------------------------------------------------------------
    (1) |   1.080067   .3481481     0.24   0.811     .5742153    2.031545
------------------------------------------------------------------------------
```

In females the result indicates that the odds of having oral cavity cancer among smokers is 1.08 (95% CI: 0.57, 2.03) times the odds of having oral cavity cancer among nonsmokers. However, this excess was not statistically significant (*P*-value > .1).

```
*In males

. lincom _Ismoker_1 + _IsmoXsex_1_1, or

( 1)  [cases]_Ismoker_1 + [cases]_IsmoXsex_1_1 = 0

------------------------------------------------------------------------------
       cases | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.502415   .3399329     6.75   0.000     1.917479     3.26579
------------------------------------------------------------------------------
```

In males the result indicates that the odds of having oral cavity cancer among smokers is 2.5 (95% CI: 1.92, 3.27) times the odds of having oral cavity cancer among nonsmokers. This excess was statistically significant (*P*-value < .05).

8.8 Adjusted OR and the Confounding Effect

The logistic regression model without interaction terms allows us to estimate the OR of the exposure of interest, while simultaneously adjusting for potential confounders. Let us assume the following expression of the logistic regression model without interaction terms:

$$\text{Odds} = \frac{p}{1-p} = e^{\beta_0 + \beta_E * E_i + \Sigma \beta_i C_i}$$

If we assume that E is a dichotomous exposure of interest with two categories (0 indicates the absence of exposure and 1 indicates the presence of exposure) and that C_i is a potential confounding variable, then the adjusted OR is obtained following the steps described below:

1. Calculate the odds when the exposure is absent ($E = 0$):

$$\text{Odds}_0 = \frac{p_0}{1-p_0} = e^{\beta_0 + \sum \beta_i * C_i'}$$

   where:
   C' is used to distinguish the value of the potential confounders
2. Calculate the odds when the exposure is present ($E_1 = 1$):

$$\text{Odds}_1 = \frac{p_1}{1-p_1} = e^{\beta_0 + \beta_E + \sum \beta_i * C_i}$$

3. Calculate the ratio of the odds obtained in steps (1) and (2):

$$\mathrm{OR} = \frac{\mathrm{Odds}_1}{\mathrm{Odds}_0} = e^{\beta_E + \beta_2\left(C_1 - C_1^*\right) + \cdots + \beta_p\left(C_p - C_p'\right)}$$

When $\left(C_i - C_i'\right) = 0$, that is, when we assume that the values of the potential confounding variables are equal in exposed and nonexposed subjects, we can obtain the adjusted *odds ratio* ($\mathrm{OR}_{\mathrm{adjusted}}$), as follows:

$$\mathrm{OR}_{\mathrm{adjusted}} = e^{\beta_E}$$

The syntax in Stata for obtaining the adjusted odds ratio using the previous data is as follows:

```
xi: glm cases i.smoker i.sex, fam (bin total) ef nolog noheader
```

Output

```
------------------------------------------------------------------------------
             |                 OIM
       cases | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Ismoker_1 |   2.211517   .2757527     6.37   0.000     1.732026    2.823749
     _Isex_1 |   .6247632   .0825355    -3.56   0.000     .4822243    .8094034
       _cons |   .7947862   .0708192    -2.58   0.010     .6674276    .9464473
------------------------------------------------------------------------------
```

The results indicate that the odds of having oral cavity cancer in smokers is 2.21 (95% CI: 1.73, 2.82) times the odds of having oral cavity cancer in nonsmokers, after adjusting for sex. The difference between the point estimate of the adjusted OR ($\widehat{\mathrm{OR}}_{\mathrm{adjusted}} = 2.21$) and the point estimate of the crude OR ($\widehat{\mathrm{OR}}_{\mathrm{crude}} = 1.51$) indicates that the magnitude of association given by the crude OR is underestimated. Therefore, the variable *sex* confounds the relationship between the smoking habit and oral cavity cancer.

8.9 Effect Modification

When the magnitude of the association between the exposure and the disease is explored in different strata, the user can identify effect modification. If the ORs change between strata, using a subjective assessment, then it is expected that the interaction terms in the model will be statistically significant. For example, using the previous example, if we stratify by sex, the point estimates of the ORs are very different ($\widehat{\mathrm{OR}}^{\mathrm{sex}=1}_{\mathrm{smk+vs\ smk-}} = 2.50$ vs. $\widehat{\mathrm{OR}}^{\mathrm{sex}=0}_{\mathrm{smk+vs\ smk-}} = 1.08$), as can be seen in the following results:

```
. xi: glm cases i.smoker if sex==1, fam (bin total) ef nolog noheader
------------------------------------------------------------------------------
             |                 OIM
       cases | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Ismoker_1 |    2.502415   0.3399329     6.75   0.000     1.917479    3.26579
       _cons |    0.4563492  0.0513548    -6.97   0.000     0.3660228   0.5689662
------------------------------------------------------------------------------

. xi: glm cases i.smoker if sex==0, fam (bin total) ef nolog noheader

------------------------------------------------------------------------------
             |                 OIM
       cases | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Ismoker_1 |    1.080067   0.3481481     0.24   0.811     0.5742153   2.031545
       _cons |    0.8416988  0.0773638    -1.87   0.061     0.702942    1.007845
------------------------------------------------------------------------------
```

This result indicates that the predictor *sex* has a modifying effect on the relationship between the smoking habit and oral cavity cancer, as was expected (because of the significant results in the *likelihood ratio test*).

8.10 Prevalence Ratio

Several epidemiological studies use the prevalence ratio (PR) to assess the magnitude of the association between exposure and disease. The main reason is that the OR can augment this magnitude, particularly when the prevalence of the outcome among exposure and nonexposure groups is large. To estimate the PR using the logistic regression model, the user needs to use the command ***link(log)***. For example, to estimate the PR by sex, the syntax is as follows:

```
glm cases i.smoker if sex==1, fam(bin total) ef link(log)
```

Output

```
------------------------------------------------------------------------------
             |                 OIM
       cases | Risk Ratio   Std. Err.      z    P>|z|     95% Conf. Interval]
-------------+----------------------------------------------------------------
1.smoker     |  1.701416    0.1446968     6.25   0.000     1.440191    2.010022
       _cons |  0.3133515   0.0242131   -15.02   0.000     0.2693136   0.3645904
------------------------------------------------------------------------------
glm cases i.smoker if sex==0, fam(bin total) ef link(log)
```

Output

```
------------------------------------------------------------------------------
             |                 OIM
       cases | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1.smoker     |  1.04194     0.1764579     0.24   0.808     0.7476307   1.452105
       _cons |  0.4570231   0.0228087   -15.69   0.000     0.4144356   0.5039868
------------------------------------------------------------------------------
```

Among males we can see that there is a substantial difference between the ORs and the PRs; the estimated OR is 2.50 and the estimated PR is 1.70.

Another command that can be used to obtain the PR by sex in a logistic regression model is *binreg*, as follows:

```
binreg cases smoker if sex==1, n(total) rr nolog
```

Output

```
------------------------------------------------------------------------------
        |                 EIM
  cases | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------+---------------------------------------------------------------------
 smoker |  1.701416   0.1446966     6.25   0.00     1.440191    2.010022
  _cons |  0.3133515  0.0242131   -15.02   0.000    0.2693137   0.3645904
------------------------------------------------------------------------------
```

```
binreg cases smoker if sex==0, n(total) rr nolog
```

Output

```
------------------------------------------------------------------------------
        |                 EIM
  cases | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------+---------------------------------------------------------------------
 smoker |    1.04194   0.1764578     0.24  0.808   0.7476309   1.452105
  _cons |  0.4570231   0.0228087   -15.69  0.000   0.4144356   0.5039868
------------------------------------------------------------------------------
```

The observed results (using binreg and glm), by sex, show that the estimates of the PRs are the same; only slight differences are observed in the standard errors, and these are due to the default methods used to estimate the variance; the *glm* command uses the *maximum-likelihood method,* and *binreg* uses *Fisher's scoring method* (Hardin and Hilbe, 2001; Collett, 2002).

8.11 Nominal and Ordinal Outcomes

The logistic regression model can be extended to handle polytomous outcome (i.e., more than two nominal or ordinal categories). An example of a nominal outcome in an epidemiological case-control study might be a situation in which we have cases (diseased) and two types of controls, which, for example, could be subjects with another disease (control I) and healthy subjects (control II). An example of an ordinal outcome in an epidemiological case-control study might be a situation in which cases (diseased) are categorized by their level of disease severity (e.g., high, moderate, low). In Stata, the programming for these situations can be done with the following commands: *mlogit, ologit*, and *gologit2.*

The logistic regression model in the case of a nominal outcome with more than two categories is called a *multinomial logistic regression model* or *polytomous logistic regression model.* When the outcome is ordinal, the model is called an *ordinal logistic regression model* (Kleinbaum and Klein, 2002). The mathematical expression of both models is based on the ratio of two probabilities defined according to the codes used

in the outcome *Y*. For example, assuming *Y* has *k* categories (0, 1, 2, …, *k*), then the most simple expression of the multinomial regression model is the following:

$$\ln\left(\frac{\Pr[Y=k|E]}{\Pr[Y=0\,|\,E]}\right)=\beta_0+\beta_E*E_{ii}$$

The exponential of the estimated exposure coefficient ($\hat{\beta}_E$) will provide the estimated OR between the category with code k and the category with code 0, as follows:

$$\widehat{\mathrm{OR}}_{E+vs.E-}^{(k\,vs.0)}=e^{\hat{\beta}_E}$$

The interpretation of this OR is as follows: in reference to category 0, the probability of being in category *k* among the exposure group is $e^{\hat{\beta}_E}$ times this probability among the nonexposure group's being in category *k*. For the following example, let us assume that we are working with a case-control study to assess the relationship between hepatitis C and receiving a blood transfusion (before 1992), using two types of controls (subjects with hepatitis B and healthy subjects), and using, as well, the following data:

	Hepatitis		
Blood Transfusion	*C (3)*	*B (2)*	*Healthy (1)*
Yes (1)	19	11	14
No (0)	85	63	220

Note: Codes are in parentheses.

The database in Stata for this study should be as seen below:

```
     +-------------------------+
     | hep   trans   subjects  |
     |-------------------------|
  1. |   1       1         14  |
  2. |   1       0        220  |
  3. |   2       1         11  |
  4. |   2       0         63  |
  5. |   3       1         19  |
     |-------------------------|
  6. |   3       0         85  |
     +-------------------------+
```

The syntax in Stata to run the multinomial regression model is as follows:

```
mlogit hep trans [fw=subjects], rrr
```

Output

```
Multinomial logistic regression                 Number of obs =     412
                                                LR chi2(2)    =   12.86
                                                Prob > chi2   =  0.0016
Log likelihood = -396.16997                     Pseudo R2     =  0.0160
```

```
------------------------------------------------------------------------------
         hep |       RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |  (base outcome)
-------------+----------------------------------------------------------------
2            |
       trans |   2.743767   1.17296     2.36   0.018     1.187022    6.342139
       _cons |   .2863636   .0409194   -8.75   0.000     .2164148    .3789211
-------------+----------------------------------------------------------------
3            |
       trans |   3.512603   1.316033    3.35   0.001     1.685457    7.320497
       _cons |   .3863636    .049343   -7.45   0.000     .3008072     .4962543
------------------------------------------------------------------------------
```

The above table shows that in reference to the healthy subjects: the likelihood of hepatitis C among the subjects who had had blood transfusion experience before 1992 is 3.51 (95% CI: 1.69, 7.32) times the likelihood of hepatitis C among the subjects who had not had blood transfusion experience before 1992. This excess was statistically significant (*P*-value = .001).

To use as the reference the subjects with hepatitis B instead of the healthy subjects, you need to use the option *baseoutcome,* as is done in the following:

```
mlogit hep trans [fw=subjects], rrr baseoutcome(2)
```

Output

```
------------------------------------------------------------------------------
         hep |        RRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
       trans |   0.3644624  0.1558076  -2.36   0.018     0.1576755   0.8424443
       _cons |    3.492063  0.4989922   8.75   0.000      2.639072    4.620756
-------------+----------------------------------------------------------------
2            |  (base outcome)
-------------+----------------------------------------------------------------
3            |
       trans |   1.280212    0.529671   0.60   0.550     0.5689947    2.880417
       _cons |   1.349206   0.2243001   1.80   0.072     0.9740238    1.868905
------------------------------------------------------------------------------
```

The table above indicates that in reference to the subjects with hepatitis B, the likelihood of hepatitis C among the subjects who had had blood transfusion experience before 1992 is 1.28 (95% CI: 0.57, 2.88) times the likelihood of hepatitis C among the subjects who had not had blood transfusion experience before 1992. However, this excess was not statistically significant (*P*-value > .1).

For the ordinal logistic regression model, there are different expressions, the use of each depending on the manner in which the categories are compared. When these categories are grouped and the ORs do not depend on the grouping procedure,

it is said that the *proportional odds assumption* is met. The most common expression of this model, under this assumption, is as follows:

$$\ln\left(\frac{\Pr[Y \le k|E]}{\Pr[Y > k \mid E]}\right) = \beta_0 - \beta_E * E_i$$

This model combines into two groups the categories of the outcome, as follows: those subjects with categories that are less than or equal to *k* and those with categories that are greater than *k*. The negative sign in the coefficient of the exposure occurs because of the way Stata programmed this model; therefore, caution has to be taken to interpret the output and the way the codes of the outcome categories are defined. The exponential of the estimated exposure coefficient, $\hat{\beta}_E$, will provide the estimated OR between categories with code $>k$ and categories with code $\le k$, due to the following relationship:

$$e^{-\beta_E} = \frac{1}{\widehat{\mathrm{OR}}_{E+\mathrm{vs}.E-}^{(\le k \mathrm{\ vs.} > k)}}, \text{ then,}$$

$$\widehat{\mathrm{OR}}_{E+\mathrm{vs}.E-}^{(> k \mathrm{\ vs.} \le k)} = e^{\hat{\beta}_E}$$

The interpretation of this OR is as follows: the likelihood of being in a category greater than *k* among the members of the exposure group is $e^{\hat{\beta}_E}$ times the likelihood of being in a category greater than *k* among the nonexposure group. To improve the interpretation, use high values of the outcome codes for those subjects with worst outcome. For example, assuming a case-control study to assess the relationship between glycohemoglobin and age, let us suppose that glycohemoglobin is categorized into three groups—using tertiles as the cutoff points—as follows:

Q1: ≤ 5.4 (best)
Q2: >5.4 and ≤ 5.9
Q3: >5.9 (worst)

In addition, let's assume that age was categorized into two groups (above and at or below the mean value of the study sample). Using the available data, then, the following table results:

	Glycohemoglobin Group			
Age (Years)	*Q1 (1)*	*Q2 (2)*	*Q3 (3)*	*Total*
≤45 (0)	14	7	3	24
>45 (1)	9	16	14	39
Total	23	23	17	63

Note: Codes are in parentheses.

The structure of the database in Stata is as seen below:

```
     +---------------------------+
     | glycon3    age   subjects |
     |---------------------------|
  1. |       1      0         14 |
  2. |       1      1          9 |
  3. |       2      0          7 |
  4. |       2      1         16 |
  5. |       3      0          3 |
     |---------------------------|
  6. |       3      1         14 |
     +---------------------------+
```

The ordinal logistic model can be run with the assumption that the OR depends on the cutoff point of the outcome. Therefore, for every cutoff point in the outcome, one OR is estimated. If we assume that the *proportional odds* assumption is fulfilled, then we would expect all ORs to be equal. The syntax in Stata to run the ordinal logistic model without the proportional odds assumption is as follows:

```
gologit2 glycon3 age [fw=subjects], or
```

Output

```
Generalized Ordered Logit Estimates      Number of obs =      63
                                         LR chi2(2)    =    8.83
                                         Prob > chi2   =  0.0121
Log likelihood = -64.204942              Pseudo R2     =  0.0643
------------------------------------------------------------------------------
glycon3 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
--------+---------------------------------------------------------------------
1       |
    age |   4.666667   2.622787    2.74   0.006     1.550992    14.04119
  _cons |   .7142857   .2957424   -0.81   0.416     .3172788    1.608062
--------+---------------------------------------------------------------------
2       |
    age |       3.92   2.750659    1.95   0.052     .9908299    15.50862
  _cons |   .1428571   .0881733   -3.15   0.002     .0426117    .4789332
------------------------------------------------------------------------------
```

The *gologit2* command has to be downloaded from Internet.

The numbers in the first column of the above table indicate the way the reference categories are defined. For example, the number 1 indicates that Q1 is the reference category, $\widehat{OR}_{>45 vs. \leq 45}^{(>5.4 vs. \leq 5.4)}$. The number 2 indicates that Q1 and Q2 are the reference categories, $\widehat{OR}_{>45 vs. \leq 45}^{(>5.9 vs. \leq 5.9)}$.

The interpretation of these ORs is as follows:

1. The likelihood of having a glycohemoglobin higher than 5.4 among subjects older than 45 years is 4.67 (95% CI: 1.55, 14.04) times the likelihood of having a glycohemoglobin higher than 5.4 among subjects 45 years old or younger.
2. The likelihood of having a glycohemoglobin higher than 5.9 among subjects older than 45 years is 3.92 (95% CI: 0.99, 15.51) times the likelihood of having a glycohemoglobin higher than 5.9 among subjects 45 years old or younger.

The syntax in Stata to run the ordinal logistic model, assessing the proportional odds assumption, is as follows:

```
gologit2 glycon3 age [fw= subjects], or autofit lrf
```

Output

```
------------------------------------------------------------------------------
Testing parallel lines assumption using the .05 level of
significance...

Step 1: Constraints for parallel lines imposed for age
  (P Value = 0.8004)
Step 2: All explanatory variables meet the pl assumption

Wald test of parallel lines assumption for the final model:

( 1)  [1]age - [2]age = 0

          chi2(  1) = 0.06
        Prob > chi2 = 0.8004

An insignificant test statistic indicates that the final model
does not violate the proportional odds/ parallel lines
assumption

If you re-estimate this exact same model with gologit2,
instead of autofit you can save time by using the parameter

pl(age)
------------------------------------------------------------------------------
Generalized Ordered Logit Estimates   Number of obs    =      63
                                      LR chi2(1)       =    8.77
                                      Prob > chi2      = 0.0031
Log likelihood = -64.236022           Pseudo R2        = 0.0639

( 1)  [1]age - [2]age = 0
------------------------------------------------------------------------------
 glycon3 | Odds Ratio  Std. Err.   z    P>|z|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
1        |
     age |  4.427933   2.303347  2.86 0.004 1.597417   12.27394
   _cons |  .7293552   .2970438 -0.77 0.438 .3283002   1.620343
```

```
----------+-----------------------------------------------------------------
2         |
      age | 4.427933   2.303347   2.86 0.004 1.597417    12.27394
    _cons | .1290829   .0626706  -4.22 0.000 .0498431     .334297
------------------------------------------------------------------------------
```

The first assessment in *gologit2* with the *autofit lrf* command is used to determine if there is statistical evidence, based on the likelihood ratio test, that the proportional odds assumption has been fulfilled. If this assumption has been fulfilled, the same OR is estimated for all combinations of the outcome. In this example, the output indicates that the model does not violate the proportional odds assumption (*P*-value = .8004). As a consequence we interpret only one OR, as follows:

Using as the reference category the participants with the low levels of glycohemoglobin, the likelihood of having high levels of glycohemoglobin among subjects older than 45 years is 4.43 (95% CI: 1.60, 12.27) *times* the likelihood of having high levels of glycohemoglobin among subjects 45 years old or younger.

8.12 Overdispersion

When the logistic regression model is run with grouped data (binomial proportion), the relationship between the *deviance* and the degrees of freedom can be useful in determining the model's goodness of fit (McCullagh and Nelder, 1999; Hardin and Hilbe, 2001). *Overdispersion* occurs when data exhibit more variation than expected. *Underdispersion* occurs when data exhibit less variation than expected. Because *deviance* is a random variable with chi-squared distribution, if the model is adequate to explain the binomial proportion, then it is expected that the observed deviance would be close to the degrees of freedom of the model *(equidispersion).* For example, if we run the model with only the predictor *smoker* in the example of cancer explained by smoker and sex, we discover that the deviance is 18.62, with 2 degrees of freedom; therefore, overdispersion is observed. To assess the departure between the deviance and the degrees of freedom, we can use the *P*-value to determine the statistical significance of this difference. The syntax to perform this in Stata is as follows:

```
dis chi2tail(2,18.62)
.00009051
```

The results show that there is a significant difference between the deviance and its degrees of freedom (*P*-value = 0.00009051). Therefore, the logistic regression model using only the predictor *smoker* is not adequate. Either including more predictors or exploring other models would be another option to consider at this point.

8.13 Sample Size and Statistical Power

To determine the total minimum sample size for assessing the adjusted OR in a case-control study with enough statistical power (i.e., $1 - \beta = 0.8$), a minimum significance level (i.e., $\alpha = 0.05$), and using an unconditional multivariable logistic

regression model with one exposure and different covariates, the following expression (Hosmer and Lemeshow, 2000) can be used:

$$n = \frac{(1+2P_0)}{1-\rho^2} * \frac{\left(Z_{1-\alpha}\sqrt{(1/1-\pi)+(1/\pi)} + Z_{1-\beta}\sqrt{(1/1-\pi)+(1/\pi e^{\beta_E})}\right)^2}{P_0 * \beta_E^2}$$

where:

$Z_{1-\alpha}$ and Z_{β} denote the upper α and β percentage points, respectively, of the standard normal distribution

π denotes the fraction of subjects in the study who are not exposed

P_0 denotes the probability of being a case among those who are not exposed

ρ^2 denotes the squared correlation between the observed and fitted values of the exposure (dichotomous variable) using a logistic regression model, as follows: $\text{logit}(\text{pr}[E=1]) = \beta_0 + \sum \beta_0 * X_i$

This can be used to estimate pseudo R^2, as follows:

$$\text{pseudo}\, R^2 = 1 - \frac{L_p}{L_0}$$

where:

L_0 and L_p denote the log likelihoods for models containing only the intercept and the model containing the intercept plus the p-covariates, respectively

β_E denotes the coefficient of the exposure in the multivariate logistic regression model for the outcome of interest under the alternative hypothesis. If we assume that OR = 2, then β_E is approximately 0.69314 {ln(2) = 0.69314}

Based on the data presented previously, the purpose of which was to assess the magnitude of the association between cancer and the smoking habit adjusted by sex ($\widehat{OR}_{adj} = 2.21$), the parameters needed to obtain the minimum sample size are as follows:

$$\beta_E = \ln(2.21) = 0.7929, \Pi = 0.4658, P_0 = 0.3945, \rho^2 = 0.5679,$$

$$Z_{0.95} = 1.96, Z_{0.80} = 1.28$$

Therefore,

$$n = \frac{(1+2*.3945)}{1-0.5679} * \frac{\left(1.96*\sqrt{(1/1-0.4658)+(1/0.4658)} + 1.28\sqrt{(1/1-0.4658)+(1/0.4658e^{.7929})}\right)^2}{0.3945*0.7929^2}$$

$$= 618.64$$

It is desirable for the result to be divisible by 2, given that a total sample size of about 619, or 310, per group would be the minimum required. Unfortunately, Stata does not provide the option in its power and sample size calculation tool for this formula. Therefore, a *do-file* has to be programmed with the following sequence of commands (and assuming the data of the previous example):

```
gen a=(1+2*.3945)/(1-.5679)
gen b=1.96*sqrt((1/(1-.4658))+(1/.4658))
gen c=1.28*sqrt( (1/(1-.4658)) + (1/(.4658*exp(.7929))))
gen d=.3945*.7929^2
gen n=a*((b+c)^2)/d
```

Before running these commands, go to *edit* and create a dataset with one variable, such as *id,* and one space row. The other option is to work interactively with Stata by invoking the *mata* command, as follows:

```
. mata
-------------------- mata (type end to exit) ------------------
: a=(1+2*.3945)/(1-.5679)
: b=1.96*sqrt((1/(1-.4658))+(1/.4658))
: c=1.28*sqrt( (1/(1-.4658)) + (1/(.4658*exp(.7929))))
: d=.3945*.7929^2
: n=a*((b+c)^2)/d
: n
  618.6378426
: end
-------------------------------------------------------------------
```

Mata is a matrix programming language that can be used interactively or as an extension for do-files and ado-files. To extend the information about *mata* command, we recommend checking out the book by Baum (*An Introduction to Stata Programming,* 2009).