

This chapter illustrates how to use LogXact for logistic regression on stratified binary data with covariates. Suppose there are N strata, with binary responses in each of them. Let the i th stratum have m_i responses and $n_i - m_i$ non-responses. For all $1 \leq i \leq N$, and $1 \leq j \leq n_i$, let $Y_{ij} = 1$ if the j th individual in i th stratum responded; 0 otherwise. Define $\pi_{ij} = \Pr(Y_{ij} = 1 \mid \mathbf{x}_{ij})$ where \mathbf{x}_{ij} is a p -dimensional vector of covariates for the j th individual in the i th stratum. The logistic regression model for π_{ij} is of the form

$$\log \left(\frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \gamma_i + \mathbf{x}_{ij}'\beta, \quad (8.1)$$

where γ_i is a stratum specific scalar parameter and β is a $(p \times 1)$ vector of parameters common across all N strata.

We are usually interested in inferences about β , and regard the γ_i 's as nuisance parameters. LogXact provides both asymptotic and exact inference for the β parameters. The asymptotic inference is based on maximizing the partially conditional likelihood function (see chapter 1). In this approach, one conditions on the sufficient statistics for the γ_i parameters and thereby eliminates them from the likelihood function. This approach is documented in Section A.5 of Appendix A. The exact approach is based on generating the exact permutation distribution of the sufficient statistics for the parameters of interest, conditioning on the observed values of the sufficient statistics for all the remaining parameters. This approach is documented in Section A.6 of Appendix A.

8.1 Purpose of this Chapter

In this chapter we will illustrate logistic regression for stratified data through the following four data sets:

SCHIZO.cyd — Birth Complications and Schizophrenia This is a simple example of stratified data involving only one covariate. Its primary objective is to clarify the difference between unconditional maximum likelihood inference, conditional maximum likelihood inference, and conditional exact inference. The key lies in understanding how the `Stratification` command works. This command, unlike the other statistical commands of LogXact, has not been encountered so far.

BIOMET.cyd — Cross-over Clinical Trial of Analgesics This example builds on the experience gained in the previous example by analyzing a data set with more than one covariate.

TRIESTE.cyd — Radiation and Lung Cancer A common feature of the previous two data sets is the fact that the binary outcomes were correlated. The conditional methods of this chapter provide a way to handle some correlated binary data for small data sets. However, these methods are not restricted to correlated outcomes. Their primary application has been to the analysis of matched sets with uncorrelated binary outcomes. They were popularized for this application by Breslow and Day (1980). This data set is derived from a classic case-control study of involving 18 matched sets

8 Logistic Regression for Stratified Data

with one case and several controls in each matched set. All observations are independent. The example highlights the fact that there can be striking differences between the unconditional and conditional approaches when there are a large number of matched sets with only a few observations in each matched set.

GOLD.cyd — Female Assistant Professors In the three previous data sets each stratum had either one response and several non-responses, or several responses and one non-response. LogXact is not restricted to these special situations. It can handle the general case of m_i responses and $n_i - m_i$ non-responses in stratum i . This example deals with the general case.

8.2 Summary of LogXact Results

LogXact writes the results of running a test or estimate in the `Results` window. If you were to choose from the menu:

```
Regression
Binary Response ▾
Logistic Model ...
```

and then run a test or estimate, LogXact will generate the results in the `Results` window. The `Results` window contains all the results from the current model fit. See Chapter 5 for details on the `Results` workbook.

8.3 Running a Statistical Test

1. Selecting the `Test` and `Asymptotic` options in the `Binary Regression: Logistic Model` dialog box produces the likelihood ratio, Wald, and Scores tests, based on maximizing the likelihood function, conditional on the observed values of the sufficient statistics for the stratum-specific constants. These tests are documented in Section A.5 of Appendix A.
2. Selecting the `Test` and `Exact` options produces either the exact conditional Probability test, the exact conditional Scores test, or the exact conditional Likelihood ratio test, depending on the `Type of Test` you selected in the `Options` dialog box after clicking on the `Options` button. These tests are documented in Section A.6 of Appendix A. Both tests are based on the joint permutation distribution of the sufficient statistics for the parameters being tested, conditional on the observed values of the sufficient statistics for all the remaining parameters. Mid-p-values are also produced from this distribution.
3. Selecting the `Test` and `Monte Carlo` options produces the Monte Carlo estimate of either the exact conditional Score test or the exact conditional Likelihood Ratio test.
4. Selecting the `Test` and `MCMC` options produces the Markov Chain Monte Carlo estimate of the Likelihood Ratio test.
5. Selecting the `Estimate` and `Asymptotic` options produces one line of asymptotic Results for each variable in the model. The following items appear on that line of

Results:

- the name of the variable;
 - a line-label denoting the type of inference — asymptotic;
 - the point estimate, standard error, confidence interval, and p-value of the β coefficient. These estimates are obtained by maximizing a conditional likelihood function. The likelihood function being maximized is only partially conditional. We condition on the observed sufficient statistics for the stratum specific constants only. Other nuisance parameters are estimated rather than being conditioned out of the likelihood function. Refer to Section A.5 of Appendix A for details.
6. Selecting the `Estimate` and `Exact` options produces one line of exact Results for each variable in the model. The following items appear on that line of Results:
- the name of the variable;
 - a line-label denoting the type of inference — exact;
 - the point estimate of the β coefficient. Where possible, this estimate is obtained by maximizing the conditional likelihood function (CMLE) formed by conditioning on the observed values of the sufficient statistics corresponding to all the nuisance parameters. Notice that this conditional likelihood function is in general not the same as the conditional likelihood function being maximized to obtain the point estimate of β on the asymptotic line of the Results. The latter likelihood is conditioned on the sufficient statistics for the stratum specific constants only, whereas this likelihood is conditioned on the sufficient statistics for all the parameters (other than the one being tested).
Sometimes this maximization is not possible, because the sufficient statistic of the β being estimated lies at one extreme of its range. In that case the median unbiased point estimate (MUE) is reported. See Section A.6 of Appendix A for details. When the MUE is reported a label with an arrow `<<` pointing towards the point estimate appear on the screen;
 - the exact confidence interval and exact p-value of β . These estimates are derived from the exact permutational distribution of the sufficient statistic for β , conditional on all the remaining sufficient statistics. See the documentation in Section A.6 of Appendix A for details.
7. When the asymptotic estimates fail to exist, due to non-convergence of the maximum likelihood algorithm, LogXact places “?” characters in appropriate places on the Results window.
8. When the memory requirements for doing an exact computation are exceeded, LogXact places a `NO MEMORY` sign at the appropriate place in the Results window.
9. If the number of cases in a stratum is very large, floating point values that are out of the machine’s range may have to be computed. This may result in a run-time error condition.

8 *Logistic Regression for Stratified Data*

10. The label NA, meaning “Not Applicable”, appears in the Results window whenever a particular computation is inappropriate. The value of SE (Beta) is not available when the estimates are reported on the odds ratio scale rather than the log odds ratio scale; i.e., when the Output option in the Options Global dialog box has been set to Odds. The standard errors for the β 's cannot be converted directly into standard errors for the odds ratios.
11. Selecting the Estimate and Monte Carlo options produces Monte Carlo estimate output same as the Estimate and Exact options with an additional result of standard error of the P-value.

8.4 *Birth Complications and Schizophrenia* *Monte Carlo Results*

We thank Dr. Armando Garsd for providing this example. A case-control study (Garsd, 1988) was designed to determine the role of birth complications in schizophrenics. The sample consisted of 7 families with several siblings per family. An individual within a family was classified either as normal or schizophrenic. A “birth-complications index” was available for each individual, ranging in value from 0 (uncomplicated birth) to 15 (severely complicated birth). The data are displayed below:

Family ID	Birth-Complications X	Number of Siblings		
		Normal	Y	Total
1	15	0	1	1
1	7	1	0	1
1	6	1	0	1
1	5	1	0	1
1	3	2	0	2
1	2	3	0	3
1	0	1	0	1
2	2	0	1	1
2	0	1	0	1
3	9	0	1	1
3	2	1	0	1
3	1	1	0	1
4	2	0	1	1
4	0	4	0	4
5	6	1	0	1
5	3	0	1	1
5	0	0	1	1
6	3	1	0	1
6	0	2	1	3
7	6	0	1	1
7	2	1	0	1

Is there a positive correlation between the chance of schizophrenia and the birth-complications index? The data do indeed suggest some such tendency. But the numbers are small, and the magnitude of the effect appears to vary across families. This is an ideal situation for exact logistic regression on matched sets. Treating each family as a separate matched set, one can model π_{ij} , the probability of schizophrenia for the j th sibling in the i th family in terms of the birth-complications index, x_{ij} :

$$\log \left(\frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \gamma_i + \beta x_{ij} .$$

We eliminate nuisance parameter γ_i , corresponding to the family effect, by conditioning on the total number of schizophrenics within each family. We then estimate β by the conditional methods discussed in Sections A.5 and A.6 of Appendix A.

Choose from the menu,

8 Logistic Regression for Stratified Data

File
Open

Select the `SCHIZO.cyd` file in the subdirectory `Data` in your LogXact installation directory.

The data will now appear:

	strat	y	x	
1	1	1	15	
2	1	0	0	
3	1	0	2	
4	1	0	2	
5	1	0	2	
6	1	0	6	
7	1	0	7	
8	1	0	3	
9	1	0	3	
10	1	0	5	
11	2	1	2	
12	2	0	0	
13	3	0	1	
14	3	0	2	
15	3	1	9	

After you have examined the data in the editor, choose from the main menu:

Regression
Binary Response >
Logistic Model ...

In the ensuing dialog box, select `y` as the Response variable, `strat` as the Stratum variable, and `x` as the Model Term. Then, since we wish to estimate the parameters, click on the `Estimate` and the `Exact` buttons before clicking `OK`.

At first, the `Work in Progress` box appears, showing how the estimation is progressing.

The `Work in Progress` box is described in Chapter 9.

In a few seconds, the following results appears in the `Results` window.

Binary Regression

Basic Information

Data file Schizo.cyd
 Model y(Response = 1)=x
 Link type Logit
 Weight variable <Not Specified>
 Stratum variable strat
 Informative strata 7
 Analysis type Estimate :: Exact
 Number of terms in model 1
 Number of term(s) dropped 0
 Number of observations in analysis 29
 Number of records rejected 0
 Number of groups 21

Summary Statistics

Statistics	Value	DF	P-Value
Likelihood Ratio	5.1997	1	0.0226

Parameter Estimates

Model Term	Point Estimate			Confidence Interval and P-Value for Beta			
	Type	Beta	SE(Beta)	Type	95 %CI		2*1-sided P-Value
					Lower	Upper	
x	MLE	0.3251	0.1679	Asymptotic	-0.0040	0.6542	0.0528
	CMLE	0.3251	0.1679	Exact	0.0223	0.7409	0.0333

The `Estimate Results` display the name of the data file, the model, weight variable, the stratum variable, the number of informative strata (here, the number of families), the number of terms in the model, the total number of observations in the data set, observations rejected and the total number of groups or the distinct covariate combinations in the dataset. The next section, under the headings `Summary Statistics`, displays the deviance and its degrees of freedom, and Likelihood ratio statistic and its degrees of freedom.

The Likelihood ratio statistic has a chi-squared distribution under the null hypothesis and may be used to test the over all significance of the model. For the present example, Likelihood ratio statistic is 5.1997 with $df = 1$ and the P- value is 0.0226.

In the last section i.e. in the `Parameter Estimates` , you can see the conditional asymptotic and exact point estimate, confidence interval, and the p-values for the regression coefficient corresponding to birth complications index. The asymptotic and exact P-values are 0.0528 and 0.0333. The formulae for computing these results are given in Sections A.5 and A.6 of Appendix A.

8 Logistic Regression for Stratified Data

Now rerun the model by choosing from the menu:

```
Regression
Binary Response >
Logistic Model ...
```

followed by the Test and Exact choices in the dialog box.

The screenshot shows the SPSS dialog box for Binary Regression, with the 'Exact Distribution' tab selected. The dialog displays the following information:

Analysis type: Test :: Exact
Test type: Score
Number of terms in model: 1
Number of term(s) dropped: 0
Number of observations in analysis: 29
Number of records rejected: 0
Number of groups: 21

Summary Statistics

Statistics	Value	DF	P-Value
Likelihood Ratio	5.1997	1	0.0226

Parameter Estimates

Model Term	Point Estimate			Confidence Interval and P-Value for Beta			
	Type	Beta	SE(Beta)	Type	95 %CI		2*1-sided P-Value
					Lower	Upper	
x	MLE	0.3251	0.1679	Asymptotic	-0.0040	0.6542	0.0528

Hypothesis Testing

Tests <x=0>

Type of Test	Statistics	DF	P-Value	P-Mid
Score	6.3280	1	0.0119	NA
Likelihood Ratio	5.1997	1	0.0226	NA
Wald	3.7496	1	0.0528	NA
Exact Score	6.3280	NA	0.0167	0.0147

The following details concerning the above results might be helpful to know more about how LogXact works.

- The asymptotic p-value is based on the Wald test, and so it matches the Wald test p-value.
 - The exact p-value for the Estimate is obtained by doubling the tail area of the permutation distribution of the sufficient statistic for the coefficient being tested. Thus it will not in general equal the exact p-value for the Test. The latter p-value is defined by equation (A.39) in Appendix A.
- To see the distinction between the two exact p-values, it will be helpful to examine the permutation distribution of the sufficient statistic for birth complications index conditional on the remaining sufficient statistics. To save this distribution to a file,

choose from the menu:

- Regression
- Binary Response ▷
- Logistic Model ...

and click on the `Exact` distribution check box. Now rerun the exact test procedure. To see the permutation distribution go to the `Exact Distribution` tab in the active workbook.

You now see a partial view of this conditional distribution as shown below:

		A	B	C	D	E	F	G	H
1	Sufficient Statistic(s): x(37)								
2	Sr.No	x	Count	Probability	Statistics				
3	1	6	12	0.00186667	5.08052				
4	2	7	12	0.00186667	4.41061				
5	3	8	51	0.00708333	3.78804				
6	4	9	91	0.0126389	3.21281				
7	5	10	100	0.0138889	2.68493				
8	6	11	170	0.0236111	2.20439				
9	7	12	230	0.0319444	1.77119				
10	8	13	257	0.0356944	1.38534				
11	9	14	336	0.0466667	1.04683				
12	10	15	379	0.0526389	0.755666				
13	11	16	423	0.05875	0.511846				
14	12	17	413	0.0573611	0.315369				
15	13	18	452	0.0627778	0.166236				
16	14	19	455	0.0631944	0.0644468				
17	15	20	415	0.0576389	0.0100014				
18	16	21	425	0.0590278	0.00289981				
19	17	22	401	0.0556944	0.0431421				
20	18	23	355	0.0493056	0.130728				
21	19	24	328	0.0455556	0.265668				
22	20	25	300	0.0416667	0.447932				
23	21	26	282	0.0391667	0.677549				
24	22	27	222	0.0308333	0.95451				
25	23	28	215	0.0298611	1.27882				
26	24	29	189	0.02625	1.65046				
27	25	30	124	0.0172222	2.06946				
28	26	31	122	0.0184722	2.52570				

The observed conditional score is 6.328. (The conditional score is defined precisely by equation (A.38) in Appendix A.) The exact p-value based on the conditional scores test is then the sum of probabilities of all the values of X whose corresponding scores equal or exceed 6.328. This p-value is reported in the `Results` window as 0.0167. Next, the observed value of the sufficient statistic is 37. Twice the tail area of the permutation distribution of X to the right of $X = 37$ is 0.0333. This p-value is reported in the `Results` window.

- Notice that there is no estimate for the constant term in the `Results` window. There were seven stratum-specific constants in the model, corresponding to the seven families. However, the conditional logistic regression approach is to eliminate these

8 Logistic Regression for Stratified Data

constants from the likelihood function by conditioning on their sufficient statistics, i.e., the number of schizophrenia cases in each family. (Note: Although for this data set there was only one case and n_i controls per family, the software can handle the general case of m_i cases and n_i controls for the i th matched set.) In contrast, unconditional logistic regression would create six dummy variables for the seven levels of the stratum variable and estimate all the parameters of the model:

$$\log \frac{\pi_j}{1 - \pi_j} = \gamma + \beta x_j + \sum_{l=1}^6 \lambda_l u_{lj},$$

where $u_{lj} = 1$ if the j th observation belonged to family l , and 0 otherwise. We can create such a model and estimate its parameters. Choose:

```
Regression
Binary Response >
Logistic Model ...
```

Remove STRAT as the Stratum variable, click on the **Toggle Factor** option for STRAT, select STRAT as a model term. To obtain the estimates, click on the **Estimate and Exact** options.

The **Work in Progress** box appears, showing how the estimation is progressing. Then, the **Results** window is displayed as follows:

Parameter Estimates							
Model Term	Point Estimate			Confidence Interval and P-Value for Beta			
	Type	Beta	SE(Beta)	Type	95 %CI		2*1-sided P-Value
					Lower	Upper	
%Const	MLE	-2.0468	1.8536	Asymptotic	-5.6798	1.5862	0.2695
x	MLE	0.5117	0.2325	Asymptotic	0.0560	0.9674	0.0277
	CMLE	0.3251	0.1679	Exact	0.0223	0.7409	0.0333
strat_1	MLE	-4.1119	2.7467	Asymptotic	-9.4953	1.2714	0.1344
	MUE	-1.4769	NA	Exact	-INF	2.1867	0.3718
strat_2	MLE	1.5351	2.2784	Asymptotic	-2.9305	6.0008	0.5005
	CMLE	0.3466	1.8877	Exact	-4.9524	5.6456	1.0000
strat_3	MLE	-1.0521	2.4159	Asymptotic	-5.7872	3.6830	0.6632
	CMLE	-0.7843	1.5244	Exact	-5.4251	3.8565	1.0000
strat_4	MLE	0.4063	2.1202	Asymptotic	-3.7493	4.5619	0.8480
	CMLE	-0.5816	1.6190	Exact	-5.4238	4.2607	1.0000
strat_5	MLE	1.4578	2.1833	Asymptotic	-2.8214	5.7370	0.5043
	CMLE	0.9962	1.6688	Exact	-3.9418	5.9342	1.0000
strat_6	MLE	0.4651	2.1225	Asymptotic	-3.6949	4.6251	0.8266
	CMLE	0.0771	1.7993	Exact	-5.0862	5.2403	1.0000

Now rerun the exact test for the variable X. Choose:

```
Regression
Binary Response >
Logistic Model ...
```

In the dialog box, click on the `Test` option. Toggle the `Selected for Testing Yes/No` button to choose `X` for testing. You can examine all the estimates of the stratum-specific constants in the results. The following points are worth noting about the Results:

1. The asymptotic estimates of the beta coefficient, the confidence interval, and the p-value have all changed. These were obtained by maximizing the unconditional likelihood function (A.16) instead of the conditional likelihood function (A.28) in Appendix A.
2. The exact point estimate, confidence interval, and p-value for the coefficient of `X` is unchanged relative to what was obtained by the stratified inference performed earlier. (You can confirm this by scrolling up to the previous model in the `Results` window.) As explained in Section A.6 of Appendix A, the exact inference is always conditional, regardless of whether the stratified or unstratified option is invoked by LogXact. In both cases the statistics on the `Exact` line of the variable `X` are derived from the permutation distribution (A.43) discussed in Appendix A.
3. The exact conditional scores p-value for the coefficient of `X`, displayed in the `Results` window, is the same for the stratified and unstratified models. Again, the exact inference is always conditional. For both the stratified and unstratified cases, the exact p-value is given by equation (A.39) in Appendix A.

Monte Carlo Results

To illustrate the Monte Carlo method in logistic regression, choose from the menu

```
Regression
Binary Response ▸
Logistic Model ...
```

In the ensuing dialog box, choose `Y` as the response variable and `X` as the model term. Choose `strat` as the Stratum variable. Click on the `Estimate` and `Monte Carlo` options. Click .

After the computations are completed the results appear as shown below:

8 Logistic Regression for Stratified Data

Analysis type				Estimate :: Monte Carlo			
Number of terms in model				1			
Number of term(s) dropped				0			
Number of observations in analysis				29			
Number of records rejected				0			
Number of groups				21			

Summary Statistics			
Statistics	Value	DF	P-Value
Likelihood Ratio	5.1997	1	0.0226

Model Term	Point Estimate			Confidence Interval and P-Value for Beta			
	Type	Beta	SE(Beta)	Type	99 %CI		P-Value
					Lower	Upper	
x	MLE	0.3251	0.1679	Asymptotic	-0.0040	0.6542	0.0528
	CMLE	0.3254	0.1673	Monte Carlo	0.0235	0.7423	0.0318
(Seed = 1096318156, Samples = 10000)							

The above Monte Carlo results were obtained using the computer Clock time as the Random number seed. Hence these Monte Carlo values may be different from the values you will get in your computer. It is instructive to compare the Monte Carlo results with the results obtained using the Exact option.

8.5 Cross-Over Clinical Trial of Analgesic Efficacy Monte Carlo Results

The data below are taken from a three-treatment, three-period cross-over clinical trial. The three drugs are A=New Drug, B=Aspirin, C=Placebo; the primary end-point was analgesic efficacy, here dichotomized as 0 for relief and 1 for no-relief. See Snapinn and Small (*Biometrics*, 42, 583-592, 1986) for details.

Patient	Drug Sequence	Response		
		P1	P2	P3
1	ABC	0	1	1
7	ABC	0	1	1
2	BCA	0	1	1
8	BCA	0	0	0
3	CAB	1	0	0
9	CAB	1	0	1
4	CBA	1	0	1
10	CBA	1	0	0
5	ACB	0	0	0
11	ACB	0	1	0
6	BAC	1	0	0
12	BAC	0	0	1

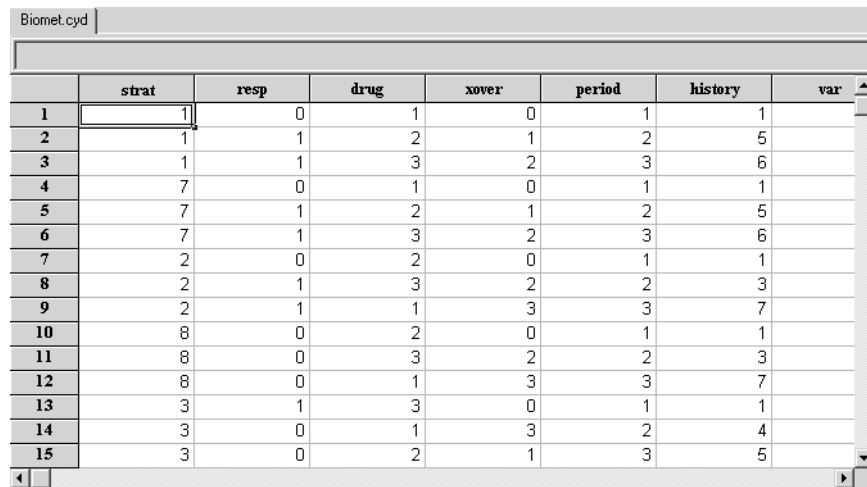
The question to be addressed is whether the three treatments are different. We answer this question by including treatment as the primary covariate in a stratified logistic regression model. In this model, treatment is included as an unordered categorical covariate at three levels and, hence, with two degrees of freedom. We regard each patient as a stratum. Within each such stratum there are three observed responses, one at each of the three time periods P1, P2, and P3. It is assumed that observations from the same patient are conditionally independent given the stratum effects. See Jones and Kenward (*Statistics in Medicine*, 6, 555-564, 1987), Kenward and Jones (*Statistics in Medicine*, 10, 1607-1619, 1991). A carry-over term is included in the model with four levels and three degrees of freedom. The first level corresponds to the first period, in which there is no carry-over. Hence one degree of freedom of the carry-over term is associated with the comparison of the first period and the average of the second and third periods. In other words the three degrees of freedom for this term include the two for carry-over effect and one of the two degrees of freedom for the period effects. The remaining one degree of freedom for the period effects is aliased with the other effects because of the small size of the data set. Thus we have omitted separate terms for the period effect from the model.

The data set, recoded in a form suitable for analysis by LogXact, is available in a file named BIOMET.CYD. To open this data set, choose from the menu:

```
File
Open
```

Select the BIOMET.CYD file in the Data subdirectory of the LogXact installation directory. The following data appears:

8 Logistic Regression for Stratified Data



	strat	resp	drug	xover	period	history	var
1	1	0	1	0	1	1	1
2	1	1	2	1	2	5	
3	1	1	3	2	3	6	
4	7	0	1	0	1	1	
5	7	1	2	1	2	5	
6	7	1	3	2	3	6	
7	2	0	2	0	1	1	
8	2	1	3	2	2	3	
9	2	1	1	3	3	7	
10	8	0	2	0	1	1	
11	8	0	3	2	2	3	
12	8	0	1	3	3	7	
13	3	1	3	0	1	1	
14	3	0	1	3	2	4	
15	3	0	2	1	3	5	

Notice that each stratum has three responses. The three drugs are coded 1, 2, and 3. The three periods are also coded 1, 2, and 3. The carry-over variable is coded 0 if the observation occurs in the first period, 1 if the subject has just crossed over from treatment A, 2 if from treatment B, and 3 if from treatment C. The “History” variable may be ignored.

To obtain the estimates, choose from the menu:

```
Regression
Binary Response ▾
Logistic Model ...
```

In the dialog box, specify `STRAT` as the Stratum variable and `RESP` as the Response variable. Choose `DRUG` and `XOVER` as the Model Terms. Since there is no reason to expect the `DRUG` or `XOVER` variables to have any natural ordering, specify that they are factor variables by clicking the `Factor` option. Then, estimate the regression coefficients by clicking on the `Estimate` and `Exact` options. Click on `OK`.

The `Work in Progress` box appears, showing how the estimation is progressing. Then, the following `Results` window is displayed.

Biomet.cyd Binary Regression

Data file: Biomet.cyd
 Model: resp(Response = 1)=drug+xover(Factor: drug xover)
 Link type: Logit
 Weight variable: <Not Specified>
 Stratum variable: strat
 Informative strata: 10
 Analysis type: Estimate :: Exact
 Number of terms in model: 2
 Number of term(s) dropped: 0
 Number of observations in analysis: 30
 Number of records rejected: 0
 Number of groups: 30

Summary Statistics

Statistics	Value	DF	P-Value
Likelihood Ratio	9.8946	5	0.0783

Parameter Estimates

Model Term	Point Estimate			Confidence Interval and P-Value for Beta			
	Type	Beta	SE(Beta)	Type	95 %CI		2*1-sided P-Value
					Lower	Upper	
drug_1	MLE	-3.0154	2.0985	Asymptotic	-7.1283	1.0975	0.1507
	MUE	-1.0860	NA	Exact	-INF	0.4962	0.1905
drug_2	MLE	-2.0277	1.3772	Asymptotic	-4.7269	0.6715	0.1409
	CMLE	-1.5522	1.2195	Exact	-5.7489	1.0461	0.3913
xover_0	MLE	0.3002	1.5626	Asymptotic	-2.7623	3.3628	0.8476
	CMLE	0.8959	1.6227	Exact	-3.9536	5.7454	1.0000
xover_1	MLE	0.5159	1.9895	Asymptotic	-3.3835	4.4153	0.7954
	CMLE	0.2027	1.6227	Exact	-4.6468	5.0523	1.0000
xover_2	MLE	1.5758	1.7593	Asymptotic	-1.8723	5.0240	0.3704
	CMLE	0.6931	1.3612	Exact	-2.6847	5.0554	1.0000

You may examine the results more carefully in the output. To perform a 2-df test that there is no drug effect, choose from the menu:

```
Regression
Binary Response >
Logistic Model ...
```

In the Binary Logistic Regression dialog box, click on the Test and Exact options. Declare DRUG as a factor variable by clicking on Toggle Factor On/Off . Select strat as a stratum variable. Select DRUG<fa> in the Model Terms section for testing, using the Selected for Testing Yes/No toggle button. Click on OK.

The Work in Progress box appears, showing how the estimation is progressing. Then, the following results are displayed.

8 Logistic Regression for Stratified Data

Parameter Estimates							
Model Term	Point Estimate			Confidence Interval and P-Value for Beta			
	Type	Beta	SE(Beta)	Type	95 %CI		2*1-sided P-Value
					Lower	Upper	
drug_1	MLE	-2.7207	1.2496	Asymptotic	-5.1698	-0.2716	0.0295
drug_2	MLE	-2.0001	1.1546	Asymptotic	-4.2630	0.2628	0.0832
xover	MLE	-0.0798	0.4279	Asymptotic	-0.9184	0.7589	0.8521

Hypothesis Testing				
Tests <drug_1=drug_2=0>				
Type of Test	Statistics	DF	P-Value	P-Mid
Score	6.8113	2	0.0332	NA
Likelihood Ratio	7.7442	2	0.0208	NA
Wald	4.8024	2	0.0906	NA
Exact Score	6.3579	NA	0.0395	0.0374

There are wide variations among the three asymptotic tests, a clear indication that the asymptotic methods might not be valid. The exact p-value for DRUG is between the Wald and Likelihood Ratio p-values.

In order to compute the exact P-value in hypothesis testing, LogXact uses conditional Score, Probability and Likelihood ratio statistics. For Score test there are two options, Score with asymptotic variance and Score with exact variance, available in LogXact. By factory default, LogXact uses Score with exact variance to compute the exact P-value when Test and Exact options are chosen. In the case of Score with asymptotic variance option LogXact uses asymptotic variance to compute the Score statistic, therefore, in this case the test statistic is same as asymptotic Score statistic. In case of Score with exact variance LogXact uses conditional exact variance to compute the Score statistic. Therefore, the test statistic with variance might be different from that of the asymptotic variance. You may examine this by doing the following:

Choose from the menu:

```
Regression
Binary Response ▾
Logistic Model ...
```

In the ensuing dialog box, choose RESP as the response variable, STRAT as the stratum variable, DRUG and XOVER as the model terms and then select Test and Exact options. Next, select DRUG in the Model Terms section for testing, using the Selected for Testing Yes/No toggle button and then click OK.

After the computations are completed the results appear as shown below:

The screenshot shows the following data:

Parameter Estimates

Model Term	Point Estimate			Confidence Interval and P-Value for Beta			
	Type	Beta	SE(Beta)	Type	95 %CI		2*1-sided P-Value
					Lower	Upper	
drug	MLE	1.2969	0.5826	Asymptotic	0.1550	2.4388	0.0260
nover	MLE	-0.1026	0.4189	Asymptotic	-0.9235	0.7183	0.8065

Hypothesis Testing
Tests <drug=0>

Type of Test	Statistics	DF	P-Value	P-Mid
Score	6.4759	1	0.0109	NA
Likelihood Ratio	7.1292	1	0.0076	NA
Wald	4.9552	1	0.0260	NA
Exact Score	6.0525	NA	0.0150	0.0107

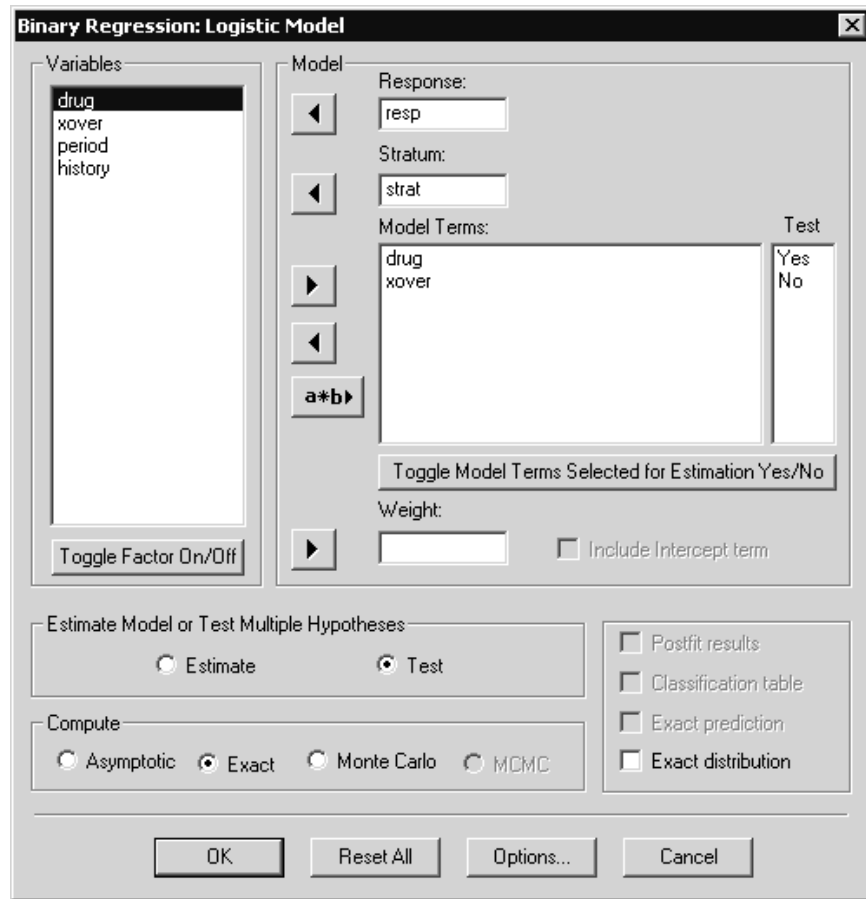
In the above result, asymptotic Score statistic is 6.4759 and the corresponding P- value is 0.0109. Also the exact Score statistic with exact variance is 6.0525 and the corresponding exact P-value is 0.0150.

Again, choose from the menu:

```
Regression
Binary Response ▸
Logistic Model ...
```

The following dialog box appears:

8 Logistic Regression for Stratified Data



Click on the Options button and then select Score test with asymptotic variance and then click OK. Next make sure DRUG is selected for testing and then click OK.

LogXact shows the following screen:

Type of Test	Statistics	DF	P-Value	P-Mid
Score	6.4759	1	0.0109	NA
Likelihood Ratio	7.1292	1	0.0076	NA
Wald	4.9552	1	0.0260	NA
Exact Score_asy	6.4759	NA	0.0150	0.0022

In the above result the asymptotic Score statistic and the exact Score statistic are equal (6.4759) and the exact P-value is 0.0150.

Monte Carlo Results

In order to illustrate the Monte Carlo method in logistic regression, choose from the menu

```
Regression
Binary Response >
Logistic Model ...
```

In the ensuing dialog box, choose RESP as the response variable and DRUG and XOVER as the model terms. Click on toggle factor on/off to make DRUG and XOVER as factored variables. Choose the Estimate and Monte Carlo options. Press OK.

After the computations are completed the results appear as shown below:

8 Logistic Regression for Stratified Data

Parameter Estimates									
Model Term	Point Estimate			Confidence Interval and P-Value for Beta					
	Type	Beta	SE(Beta)	Type	99 %CI		P-Value		SE
					Lower	Upper	2*1-sided		
drug_1	MLE	-3.0154	2.0985	Asymptotic	-7.1283	1.0975		0.1507	
	MJE	-1.0617	NA	Monte Carlo	-1.0617	0.4880		0.1870	NA
					(Seed = 1096319371, Samples = 10000)				
drug_2	MLE	-2.0277	1.3772	Asymptotic	-4.7269	0.6715		0.1409	
	CMLE	-1.5650	1.2211	Monte Carlo	-5.7773	1.0788		0.4010	0.0080
					(Seed = 1096319377, Samples = 10000)				
xover_0	MLE	0.3002	1.5626	Asymptotic	-2.7623	3.3628		0.8476	
	CMLE	0.9001	1.6163	Monte Carlo	-3.9438	5.7700		1	0.0098
					(Seed = 1096319383, Samples = 10000)				
xover_1	MLE	0.5159	1.9895	Asymptotic	-3.3835	4.4153		0.7954	
	CMLE	0.1959	1.6061	Monte Carlo	-4.6733	5.0117		1	0.0083
					(Seed = 1096319390, Samples = 10000)				
xover_2	MLE	1.5758	1.7593	Asymptotic	-1.8723	5.0240		0.3704	
	CMLE	0.7153	1.3509	Monte Carlo	-2.7451	5.1134		1	0.0099
					(Seed = 1096319398, Samples = 10000)				

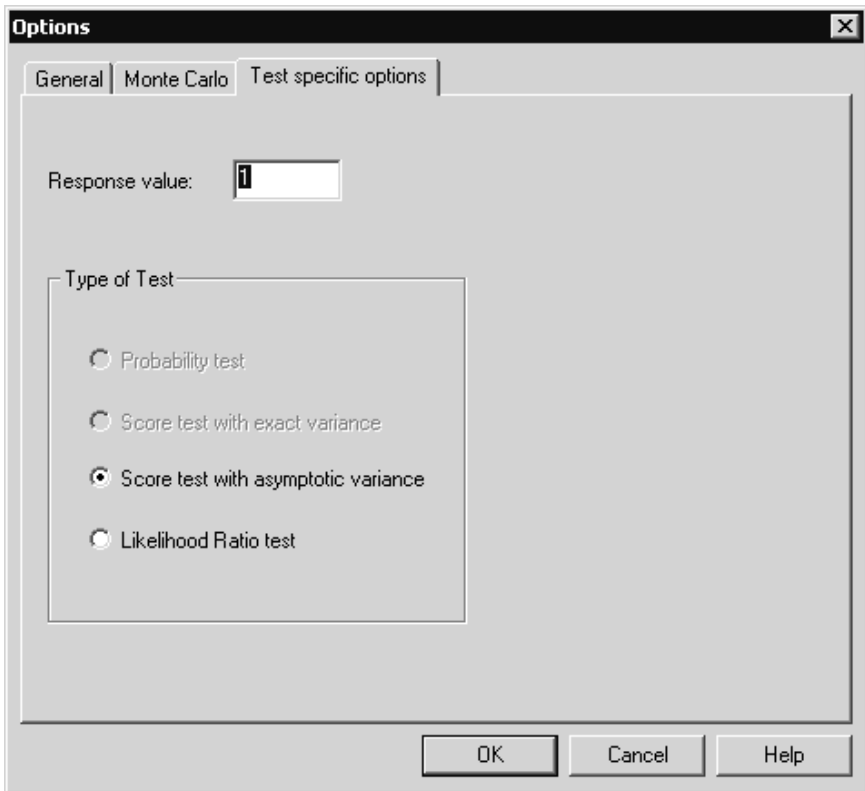
The above Monte Carlo results were obtained using the computer Clock time as the Random Number Seed. Hence these Monte Carlo values may be different from the values you will get in your computer.

To perform a test that there is no drug effect, choose from the menu:

```
Regression
Binary Response >
Logistic Model ...
```

In the Binary Regression: Logistic Model dialog box, click on Test and Monte Carlo options. Select STRAT as a stratum variable and DRUG and XOVER as the Model Terms. Select DRUG in the Model Terms section for testing, using the Selected for Testing Yes/No toggle button. Now, click on Options button.

You get the following dialog box:



Select Score test with asymptotic variance and then click OK. Next, Press OK.

8 Logistic Regression for Stratified Data

LogXact displays the following screen:

Parameter Estimates							
Model Term	Point Estimate			Confidence Interval and P-Value for Beta			
	Type	Beta	SE(Beta)	Type	95 %CI		2*1-sided P-Value
					Lower	Upper	
drug	MLE	1.2969	0.5826	Asymptotic	0.1550	2.4388	0.0260
nover	MLE	-0.1026	0.4189	Asymptotic	-0.9235	0.7183	0.8065

Hypothesis Testing				
Tests <drug=0>				
Type of Test	Statistics	DF	P-Value	Std. Error
Score	6.4759	1	0.0109	NA
Likelihood Ratio	7.1292	1	0.0076	NA
Wald	4.9552	1	0.0260	NA
Exact Score_asy:Monte	6.4759	NA	0.0153	0.0012

(10000 Monte Carlo samples with starting seed = 16018)

In the above result, the exact Score statistic is same as the asymptotic Score statistic, as both of them are using asymptotic variance to compute the test statistic. The Monte Carlo P-value is 0.0154.

8.6 Radiation and Lung Cancer

Monte Carlo Results

We thank Dr. Alexander Walker for the use of this data set. Stanta and Walker (1986) reported on a case-control study to assess the risk of developing lung cancer in women who had previously suffered from breast cancer. The covariates were radiation therapy for breast cancer and smoking history. The data consisted of 18 matched sets containing a total of 18 cases of lung cancer and 52 controls. The matching was based on age and date of diagnosis of the breast cancer. The data are displayed next:

Stratum	Case Indicator	Smoking and Radiation History			
		None	Radiation Alone	Smoking Alone	Both
1	Case	1	0	0	0
	Control	0	3	0	0
2	Case	0	0	0	1
	Control	0	1	0	0
3	Case	0	1	0	0
	Control	1	2	0	0
4	Case	0	0	0	1
	Control	0	3	0	0
5	Case	1	0	0	0
	Control	0	3	0	0
6	Case	0	0	0	1
	Control	0	2	0	1
7	Case	0	1	0	0
	Control	0	3	0	0
8	Case	0	0	0	1
	Control	0	3	0	0
9	Case	0	1	0	0
	Control	1	2	0	0
10	Case	0	1	0	0
	Control	1	1	1	0
11	Case	0	0	1	0
	Control	1	1	1	0
12	Case	0	1	0	0
	Control	0	3	0	0
13	Case	0	0	1	0
	Control	3	0	0	0
14	Case	0	0	0	1
	Control	0	2	0	1
15	Case	0	0	0	1
	Control	1	2	0	0
16	Case	0	1	0	0
	Control	1	2	0	0
17	Case	0	0	0	1
	Control	0	2	0	1
18	Case	0	0	0	1
	Control	0	3	0	0

8 Logistic Regression for Stratified Data

This data set is available in the file named TRIESTE.CYD. Bring the data by choosing from the menu:

```
File
Open
```

and selecting the TRIESTE.CYD file.

Fit the two-variable logistic regression model

```
CASE = RADIO + SMOKE
```

to the above data, but adjust for the stratum effect by declaring STRAT to be a stratum variable. To do all this, choose from the menu:

```
Regression
Binary Response ▸
Logistic Model ...
```

In the ensuing dialog box:

```
Select CASE as the Response variable
Select RADIO and SMOKE as Model Terms
Select STRAT as the Stratum variable
Click on the Estimate and Exact options
```

In this approach the 18 stratum-specific constants are not estimated. Instead they are eliminated from the conditional likelihood function. Next, perform an unconditional logistic regression analysis. This time you will actually estimate the stratum specific constants along with the coefficients of RADIO and SMOKE from the unconditional likelihood function. To do this, you must add STRAT to the model and declare it to be a factor variable. To accomplish this, choose from the main menu:

```
Regression
Binary Response ▸
Logistic Model ...
```

In the Binary Logistic Regression dialog box:

```
Select CASE as the Response variable
Select RADIO, SMOKE, and STRAT as the Model
Click on the Factor option for STRAT
Click on the Estimate and Asymptotic options
```

You could choose the Exact option instead of Asymptotic. But that would consume a large amount of memory. In any case it is unnecessary, since it would yield the very same

exact estimates as were obtained when STRAT was excluded from the model but used as a stratification variable. See Appendix A, Section A.6 for an explanation.

The purpose of this exercise is to compare the β coefficient for the SMOKE variable obtained from the conditional and the unconditional analyses. The results are tabulated below:

Inference Type	Point Estimate for SMOKE	95% C.I. for SMOKE
Unconditional MLE	4.8	(1.99 to 7.63)
Conditional MLE	3.0	(0.95 to 5.11)
Conditional Exact	2.97	(0.98 to 6.75)

Notice that there are large differences in the point and interval estimates of β from the conditional and unconditional analyses. These differences would be even more pronounced if expressed on the odds ratio scale. On the other hand the point and interval estimates obtained by the conditional asymptotic and conditional exact methods are fairly close. Based on the arguments put forward by Breslow and Day (1980), we would conclude that the conditional estimates are more appropriate than the unconditional ones for these data.

Monte Carlo Results

In order to illustrate the application of Monte Carlo method in Logistic Regression procedure let us proceed as follows:

Open the file TRIESTE.cyd in the data editor. Choose from the menu

```
Regression  
Binary Response ▸  
Logistic Model ...
```

In the ensuing dialog box, choose case as the response variable, strat as the stratum variable, radio and smoke as the model terms.

Choose Estimate and Monte Carlo options in the bottom left hand side of the dialog box. Press OK.

After the computations are completed the results appear in the screen as shown below:

8 Logistic Regression for Stratified Data

Statistics		Value	DF	P-Value
Likelihood Ratio		15.4433	2	0.0004

Model Term	Point Estimate			Confidence Interval and P-Value for Beta				
	Type	Beta	SE(Beta)	Type	99 %CI		P-Value	
					Lower	Upper	2*1-sided	SE
radio	MLE	0.1778	1.0022	Asymptotic	-1.7864	2.1420	0.8592	
	CMLE	0.1804	0.9898	Monte Carlo	-1.9612	2.9745	1.0000	0.0096
				(Seed = 1096319851, Samples = 10000)				
smoke	MLE	3.0326	1.0632	Asymptotic	0.9488	5.1164	0.0043	
	CMLE	2.9626	1.0736	Monte Carlo	0.9999	6.7532	0.0006	0.0003
				(Seed = 1096319854, Samples = 10000)				

The above Monte Carlo results were obtained using the computer Clock time as the Random Number Seed. Hence these Monte Carlo values may be different from the values you obtain on your computer.

8.7 General m_i to $n_i - m_i$ Matching

All the examples considered so far in this chapter dealt with the special case of only one response or one non-response in each stratum. LogXact is not restricted to this special case. It can handle the general case of m_i responses and $n_i - m_i$ non-responses in the i th stratum (or matched set) without any difficulty. Work out the example in Section 16.11 of Chapter 16. This is an interesting example of possible sex discrimination among college teachers. The data are stratified by time period to account for a possible cohort effect. Within each stratum there are an arbitrary number of responders (faculty members who were hired at the Assistant Professor level) and non-responders (faculty members who were hired at the Lecturer level). Covariates include gender, age, and college degree.