The Stata Journal (2003)
3, Number 3, pp. 213–225

Odds ratios and logistic regression: further examples of their use and interpretation

Susan M. Hailpern, MS, MPH Paul F. Visintainer, PhD

School of Public Health
New York Medical College
Valhalla, NY

Abstract. Logistic regression is perhaps the most widely used method for adjustment of confounding in epidemiologic studies. Its popularity is understandable. The method can simultaneously adjust for confounders measured on different scales; it provides estimates that are clinically interpretable; and its estimates are valid in a variety of study designs with few underlying assumptions. To those of us in practice settings, several aspects of applying and interpreting the model, however, can be confusing and counterintuitive. We attempt to clarify some of these points through several examples. We apply the method to a study of risk factors associated with periventricular leucomalacia and intraventricular hemorrhage in neonates. We relate the logit model to Cornfield's 2×2 table and discuss its application to both cohort and case–control study design. Interpretations of odds ratios, relative risk, and β_0 from the logit model are presented.

Keywords: st0041, cc, cci, cs, csi, logistic, logit, relative risk, case–control study, odds ratio, cohort study

1 Background

Popular methods used to analyze binary response data include the probit model, discriminant analysis, and logistic regression. Probit regression is based on the probability integral transformation. A major drawback of the probit model is that it lacks natural interpretation of regression parameters. Discriminant analysis is computationally simpler than the probit model. It assumes that predictor variables are normally distributed and that variables jointly assume a multivariate normal distribution. Because many variables in regression analysis are dichotomous or discrete, discriminant analysis assumptions are often violated. Furthermore, because discriminant analysis examines the distribution of X in terms of Y, it is dependent on Bayes theorem to extract the variable of primary interest. In contrast, the logistic regression model makes no assumption about the variable distribution. It is a direct probability model because it is stated in terms of $\Pr\{Y = 1|X\}$. Another advantage of the logit model is its ability to provide valid estimates, regardless of study design (Harrell 2001).

2 The logit model reflects the 2×2 table

The odds ratio (OR) is a popular measure of the strength of association between exposure and disease. In a cohort study, the odds ratio is expressed as the ratio of the number of cases to the number of noncases in the exposed and unexposed groups. The odds ratio and its familiar computation are attributed to Cornfield (1951), which is calculated as the ratio of the products of the pairs of diagonal elements in the 2×2 table:

$$OR = \frac{A \times D}{B \times C}$$

For illustration, data from Canterino et al. (1999) are used. This prospective cohort study investigated factors for periventricular leucomalacia and intraventricular hemorrhage in preterm neonates. Using the development of severe lesions as the disease outcome and the administration of antenatal steroids as exposure, an odds ratio is calculated using the 2×2 table.

Table 1

	Severe Lesions +	Severe Lesions −
Steroids +	26 (A)	318 (B)
Steroids −	134 (C)	584 (D)

$$OR = \frac{(26 \times 584)}{(318 \times 134)} = 0.356$$

The interpretation of the odds ratio is that the odds for the development of severe lesions in infants exposed to antenatal steroids are 64% lower than those of infants not exposed to antenatal steroids. Point estimates for the odds ratio and confidence interval are available from Stata's `cc` or `cs` command. In Stata 8, the default confidence intervals are exact. However, for purposes of comparison with logistic regression, we use the `woolf` option, which estimates the confidence interval using a Wald statistic. (The Wald statistic is a quadratic approximation of the log-likelihood curve and is most accurate in the region of the most common sample value. It is an approximation of, but less accurate than, the score statistic, which is also a quadratic approximation of the log likelihood and is most accurate in the region of the null value. Despite the fact that it is an approximation, the Wald statistic provides a simple method for estimating binomial distributions and, therefore, is widely used. Further details are found in Clayton and Hills (1993) and Rothman and Greenland (1998).)

```
. csi 26 134 318 584, or woolf
```

	Exposed	Unexposed	Total		
Cases	26	134	160		
Noncases	318	584	902		
Total	344	718	1062		
Risk	.0755814	.1866295	.1506591		
	Point estimate		[95% Conf.	Interval]	
Risk difference	-.1110481		-.1509528	-.0711434	
Risk ratio	.4049809		.2715012	.6040841	
Prev. frac. ex.	.5950191		.3959159	.7284988	
Prev. frac. pop	.1927369				
Odds ratio	.3563315		.2291072	.5542043	(Woolf)

chi2(1) = 22.41 Pr>chi2 = 0.0000

The logistic model quantifies the effect of a predictor in terms of a log-odds ratio using maximum likelihood estimation (MLE). Although computationally different, the logistic regression model produces results that are nearly identical to the 2×2 table. Notice that in the logistic model, the MLE estimation of the standard error yields a confidence interval that is quite close to the Wald confidence interval in Stata's `cc` command (or `cs` command):

```
. logistic severe ster, nolog
Logistic regression                               Number of obs   =       1062
                                                  LR chi2(1)      =      24.84
                                                  Prob > chi2     =     0.0000
Log likelihood = -437.71032                       Pseudo R2       =     0.0276
```

severe	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ster	.3563316	.0802959	-4.58	0.000	.22911	.5541973

Woolf's approach computes the standard error for the log of the odds ratio as (Collett 1991)

$$SE\ln(OR) = \sqrt{\left(\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right)}$$

The 95% confidence interval for the $\ln(OR)$ is calculated as

$$95\%\text{CI for the } \ln(OR) = \ln(OR) \pm 1.96 \times \{SE\ln(OR)\}$$

For our example, the computations are

$$\ln(OR) = \ln(.356) = -1.032$$

$$SE\ \ln(OR) = \sqrt{\frac{1}{26} + \frac{1}{318} + \frac{1}{134} + \frac{1}{584}} = 0.2253$$

$$95\%\text{CI for the } \ln(OR) = -1.032 \pm 1.96 \times .2253 = (-1.474, -.590)$$

Taking the antilog, we get the 95% confidence interval for the odds ratio:

$$95\%\text{CI for } OR = (e^{-1.474}, e^{-.590}) = (.229, .554)$$

As the investigation expands to include other covariates, three popular approaches are available in Stata to produce an adjusted odds ratio. In our example, we control for the presence of respiratory distress syndrome (RDS). The `cs` command uses a large-sample estimate of the variance (Robins, Breslow, and Greenland 1986). The `mhodds` command estimates the variance using a score variance (Clayton and Hills 1993). The variance estimate in the logistic command is based on the MLE. Given the large overall sample and individual cell sizes, all three methods produce nearly identical results. Confidence intervals are identical to two decimal places. We would, however, expect discrepancies to arise for small samples. Rothman and Greenland (1998) indicate that the Mantel–Haenszel estimates are valid for sparse data, while the MLE estimator may be biased.

Table 2

Stata command	OR	95% CI
`cs severe ster, by(rds) or`	.3357	.2127, .5298
`mhodds severe ster, by(rds)`	.3357	.2112, .5337
`logistic severe ster rds`	.3375	.2142, .5316

3 Interpretation of β_0 and other coefficients in the logit model

In epidemiology, study design determines the population parameters that may be estimated and available for interpretation. For example, relative risk, odds ratio, and incidence may be estimated from cohort studies, while of the three, only the odds ratio is available from case–control studies. Most practitioners also are aware that the intercept, β_0, in a logit model is not interpretable when the model is derived from a case–control study. Many, however, fail to recognize that these two facts are related.

3.1 Cohort studies

In a cohort study, the odds of disease given exposure, $O_{D|E+}$, is computed as (A/B). The odds of disease in the absence of exposure, $O_{D|E-}$, is (C/D). In a logistic model, the ratio of the odds (AD/BC) is expressed as the difference in the log odds. That is, the logit is the difference between the levels of the exposure odds:

$$\ln(OR) = \ln\left(\frac{AD}{BC}\right) = \ln(O_{D|E+}) - \ln(O_{D|E-})$$

where $\ln(O_{D|E+}) = \ln(\frac{A}{B})$, the log odds of disease given exposure, and $\ln(O_{D|E-}) = \ln(\frac{C}{D})$, the log odds of disease given no exposure. The log odds in the absence of exposure, $\ln(O_{D|E-})$, may be considered the "baseline" odds, or the log odds in the reference group. If $\ln(O_{D|E-})$ is represented as β_0 in the model, then $[\beta_0+\beta_1]$ represents $\ln(O_{D|E+})$. With these designations, we can express β_1:

$$\text{logit} = \ln(OR_{D|E+}) - \ln(OR_{D|E-}) = [\beta_0 + \beta_1] - [\beta_0] = \beta_1$$

In the cohort study, β_0 is interpreted as the log odds of disease in the absence of exposure, and β_1 reflects the increase in the log odds attributed to exposure beyond baseline, with e^{β_1} being the odds ratio. The logistic parameters can be used to estimate the incidence in the exposed and unexposed groups. To summarize the relations,

Table 3

Group	Odds	Risk (Incidence)
Exposed	$\frac{A}{B} = e^{(\beta_0+\beta_1)}$	$\frac{A}{A+B} = \frac{e^{(\beta_0+\beta_1)}}{1+e^{(\beta_0+\beta_1)}}$
Unexposed	$\frac{C}{D} = e^{\beta_0}$	$\frac{C}{C+D} = \frac{e^{\beta_0}}{1+e^{\beta_0}}$

Using our example, the baseline incidence is $134/718 = 0.187$, and the baseline odds, $O_{D|E-}$, are $134/584 = 0.230$. Because this is a prospective cohort study, we are able to interpret .187 and .230, respectively, as incidence and odds of lesions in the absence to exposure to steroids in this clinic sample. The $\ln(O_{D|E-})$ is ln(.230), or -1.47, and corresponds to β_0 in the logit model. The incidence of lesions given exposure is $26/344 = 0.076$, and odds of lesions given exposure are $26/318 = 0.082$. The $\ln(OR_{D|E+})$ is ln(.082) or -2.50. Thus, the difference between the two exposure log odds is $-2.50 - (-1.47) = -1.03$, which is β_1 in our logistic equation.

(*Continued on next page*)

```
. logit severe ster, nolog
Logit estimates                                   Number of obs   =       1062
                                                  LR chi2(1)      =      24.84
                                                  Prob > chi2     =     0.0000
Log likelihood = -437.71032                       Pseudo R2       =     0.0276
```

severe	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ster	-1.031894	.2253405	-4.58	0.000	-1.473553	-.5902345
_cons	-1.472061	.0957863	-15.37	0.000	-1.659799	-1.284323

The example above demonstrates what β_0 represents in a simple logistic model with a single dichotomous independent variable. In this example, when β_1 is zero, β_0 reflects a legitimate reference group. In practical analyses of cohort studies, however, models are neither simple nor are the variables always dichotomous. Most logistic models are developed using a mixture of dichotomous, ordinal, and continuous variables. In order to interpret β_0 in these settings, there must be a valid reference group when all factors in the model are set to zero. As such, situations as these are rare.

To show the error that can arise in β_0 when a continuous variable is included in a model, consider an example using maternal age, ranging between 15 years and 44 years, and the presence of severe lesions.

Table 4

```
. tab agecat severe, row
```

Key
frequency *row percentage*

agecat	SEVERE 0	1	Total
Q1: 22	269 84.86	48 15.14	317 100.00
Q2: 28	197 81.07	46 18.93	243 100.00
Q3: 31	225 88.24	30 11.76	255 100.00
Q4: 36	211 85.43	36 14.57	247 100.00
Total	902 84.93	160 15.07	1,062 100.00

Table 4 shows maternal age and the distribution of severe lesions classified into quartiles using the `xtile` command. In quartile 1, the risk of severe lesions is 15.14% and the odds of severe lesions are .1784.

We configured age differently in each of three models. In the first model, age was entered as a continuous variable. In the second model, we entered age as a single variable with four levels, using the medians of the quartile ranges. In the last model, age quartiles were entered separately with three indicator variables, with the lowest age category, Q1, used as the reference group.

Table 5

Model	Coefficients		e^{β_0}	$\frac{e^{\beta_0}}{1+e^{\beta_0}}$
1) Age as a continuous variable:				
	β_0	-1.5750	.207	.172
`logit severe age`	β_1(AGE)	$-.0054$		
2) Age is in quartiles (see Table 4):				
	β_0	-1.4465	.235	.191
`logit severe agecat`	β_1(AGE)	$-.0099$		
3) Age quartiles entered with indicators, with Q1 as reference group:				
	β_0(AGE)	-1.7235	.179	.151
`logit severe age28 age31 age36`	β_1(AGE)	.2689		
	β_2(AGE)	$-.2914$		
	β_3(AGE)	$-.0448$		

Of the models considered, only in Model 3 does β_0 represent the baseline odds, and the baseline risk of Q1 accurately (.179 and .151, respectively) when the age variables are set to zero. Thus, when interval-scaled variables are used in logistic models, we avoid interpreting β_0. Rather, appropriate estimates of odds and risk can be obtained using the `lincom` command. For example, following Model 1, if the odds and risk of lesions for a mother aged 22 (the median value of Q1) are desired, we can run

```
. lincom age*22 + _cons
 ( 1)  22 age + _cons = 0
```

severe	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	-1.693157	.1262063	-13.42	0.000	-1.940516	-1.445797

Using the coefficient -1.693, our estimate of the odds is .184, and the risk is .155, which represents the odds and risk of severe lesions for neonates of mothers aged 22.

3.2 Case–control studies

Inferences about β_0, however, cannot be made with a case–control study design. In a cohort design as above, the investigator selects subjects based on exposure status (i.e., the exposure marginals are fixed). The proportion of subjects with disease (the

incidence) and the odds of disease in the exposed and unexposed groups are independent of the size of the group. That is, $A/(A+B)$ and $C/(C+D)$ are valid estimates of incidence, and β_0 is a valid estimate of the baseline odds, $\ln(C/D)$. In a case–control study, however, the investigator fixes the number of cases and controls (the disease marginals are fixed) and, as such, artificially determines the proportion of disease in each exposure group. The estimates of $A/(A+B)$ and $C/(C+D)$, and consequently of β_0, are no longer valid estimates because they are influenced by the size of the samples drawn. Only the odds ratio remains independent of sample size. This is easily demonstrated with the following data:

Table 6

	Study A		Study B	
	Disease +	Disease −	Disease +	Disease −
Exposure +	50	20	50	100
Exposure −	10	20	10	100
Total	60	40	60	200

Suppose two case–control studies are conducted to estimate the risk of disease associated with exposure. Both studies enroll the same number of cases. However, in Study A, 40 controls are enrolled, whereas in Study B, 200 controls are enrolled. In both studies, though, the proportion exposed in the case and control groups is identical. Entering these data as cell counts, we get the following results:

Study A

```
. logit case exp [freq=sampleA], nolog
Logit estimates                                   Number of obs   =        100
                                                  LR chi2(1)      =      12.65
                                                  Prob > chi2     =     0.0004
Log likelihood = -60.974296                       Pseudo R2       =     0.0940

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   1.609438   .4690416     3.43   0.001     .6901333    2.528743
       _cons |  -.6931472   .3872983    -1.79   0.074    -1.452238    .0659436
------------------------------------------------------------------------------
```

Study B

```
. logit case exp [freq=sampleB], nolog
Logit estimates                                   Number of obs   =        260
                                                  LR chi2(1)      =      22.93
                                                  Prob > chi2     =     0.0000
Log likelihood =  -128.9871                       Pseudo R2       =     0.0816

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   1.609438   .3741657     4.30   0.000     .8760866    2.342789
       _cons |  -2.302585   .3316625    -6.94   0.000    -2.952632   -1.652539
------------------------------------------------------------------------------
```

The estimates of risk associated with exposure are the same for each study ($\beta_1 = 1.609$); however, β_0 differs in each study. Of course, the confidence intervals for β_1 are narrower in Study B because the sample size is larger.

As Gould (2000) demonstrates, one of the attractive properties of logistic regression is the constancy of the odds ratio. In the multiple logistic regression equation below, STER and RDS are dichotomous variables, and WT is a continuous measure of birth weight in 10-gram intervals.

```
. logistic severe ster rds1 wt, nolog
Logistic regression                               Number of obs   =       1060
                                                  LR chi2(3)      =      96.84
                                                  Prob > chi2     =     0.0000
Log likelihood = -399.65275                       Pseudo R2       =     0.1081

------------------------------------------------------------------------------
      severe | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ster |   .3379093   .0789938    -4.64   0.000     .2137041    .5343026
        rds1 |   2.833678   .7356233     4.01   0.000     1.703642    4.713272
          wt |   .9871476   .0027758    -4.60   0.000     .9817221    .9926031
------------------------------------------------------------------------------
```

The interpretation of weight in the model is that for each 10-gram increase in birth weight, the odds of severe lesions decrease by 1.29% $(1 - .9871)$, adjusting for steroid use and the presence of respiratory distress syndrome; that is, regardless of whether steroids were used or whether RDS was present or absent. To show the constancy of the odds ratio, we ran `predict` after the logistic model above to compute the predicted probability, p. We then generated the predicted odds from the predicted probability, as $p/(1-p)$. Finally, we randomly selected a pair of observations, which were separated by 10 grams in birth weight, from each covariate pattern of STER and RDS.

Table 7

STER	RDS	WT	p	Odds $[p/(1-p)]$	Change in odds	Change in p
0	0	890	.12584	.14395		
0	0	880	.12727	.14583	.9871	.9888
0	1	1010	.28444	.39750		
0	1	1000	.28707	.40268	.9871	.9908
1	0	1550	.02029	.02071		
1	0	1540	.02055	.02098	.9871	.9874
1	1	1350	.07963	.08652		
1	1	1340	.08059	.08765	.9871	.9881

Notice that, as expected, the ratio of the odds within each pair is constant and is identical to the model odds ratio. On the other hand, the ratio of probabilities is not constant.

We need to point out that, in the above model, we *assumed* that the increase in the odds is constant over birth weight. By entering weight as a continuous variable, we forced the model to produce a constant increase in the odds. However, this assumption of linearity of the logit may be quite untenable (Hosmer and Lemeshow 2000). Moreover, the level of significance for weight in the model provides no evidence as to whether the logit is linear in weight. Rather, graphical and statistical procedures are used to determine whether the assumption of a linear logit for a continuous variable is tenable. In Stata, useful graphical procedures are `lintrend` (see Garrett (1996)) and `lowess`. Helpful statistical procedures include `boxtid` (see Royston and Ambler (1999)) and `linktest`.

4 The logit model is applicable for both cohort and case–control studies

Prentice and Pyke (1979) have shown that the logit model can be applied to case–control designs. Mathematical computation of $\ln(OR)$ is the same irrespective of whether disease or exposure is the dependent variable (cohort or case–control study design); see Appendix A. Technically, the interpretation of the two sets of parameters and estimated coefficients will differ. Estimated parameters in the prospective model relate predictor variables to occurrence of disease, while estimated parameters in a retrospective model relate predictor variables to occurrence of exposure. In practice, however, epidemiologists do not draw this distinction (Schlesselman 1982).

An example of the "reversibility" of the logistic model is shown below. In the first model, we estimated a 62% reduction in the risk of severe lesions due to steroid use ($OR = .38$), controlling for clinical chorioamnionitis (CCA).

```
. logistic severe ster cca, nolog
Logistic regression                               Number of obs   =       1062
                                                  LR chi2(2)      =      44.80
                                                  Prob > chi2     =     0.0000
Log likelihood = -427.72883                       Pseudo R2       =     0.0498

------------------------------------------------------------------------------
      severe | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        ster |   .3803373   .0864227    -4.25   0.000     .2436426    .5937239
         cca |   2.818374   .6243629     4.68   0.000       1.8257    4.350788
------------------------------------------------------------------------------
```

In the second model, we exchanged the outcome variable (SEVERE) with the independent variable (STER). This model determines the risk of steroid use due to the presence of severe lesions and clinical chorioamnionitis. The results show that severe lesions are associated with a 62% reduction in the use of steroids, controlling for clinical chorioamnionitis.

```
. logistic ster severe cca, nolog
Logistic regression                               Number of obs   =       1062
                                                  LR chi2(2)      =      30.27
                                                  Prob > chi2     =     0.0000
Log likelihood = -653.69828                       Pseudo R2       =     0.0226

------------------------------------------------------------------------------
         ster | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      severe |   .3803373   .0864218    -4.25   0.000     .2436438     .593721
         cca |   .5842499   .1396664    -2.25   0.025     .3656931    .9334274
------------------------------------------------------------------------------
```

In other words, the relationship between severe lesions and steroids is preserved, regardless of how it is modeled. (The beta coefficient for CCA differs from the previous model because it is being associated with a different outcome.) Although identical estimates may not be necessarily attained, especially in models with continuous covariates, Breslow and Power (1978) indicate that the parameter estimates will usually become more similar with increasing degrees of adjustment (see Schlesselman (1982, 267–269) for a discussion).

5 Summary

Odds ratios and logistic regression are powerful tools for researchers. The popularity of these tools results from their versatility and relative ease of interpretation. The goal of this paper is to provide additional examples on the use and interpretation of logistic regression and odds ratios in epidemiologic and clinical research. Excellent literature exists on the mathematical basis of logistic regression, as well as its application and interpretation. The examples and interpretations presented herein highlight some of the characteristics of logistic regression that may further aid in understanding the approach.

6 Appendix A: The mathematical relationship of the logistic model between cohort and case–control designs

The material in this appendix is based on formulas from *Case–Control Studies* (Schlesselman 1982, 234–236).

The logit model expressed in terms of p_x/q_x in a cohort study is written as

$$\ln\frac{p_x}{q_x} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

The probability of disease ($d = 1$) given exposure x in a case–control study, where π_1 and π_2 represent sampling fractions for cases and control respectively, is written as

$$p'_x = \frac{\pi_1 p_x}{(\pi_1 p_x + \pi_2 q_x)}$$

where $q'_x = 1 - p'_x$ and odds of disease given the exposure are expressed as

$$\frac{p'_x}{q'_x} = \frac{\pi_1 p_x}{\pi_2 q_x}$$

Relating the log odds of disease in a case–control study to the log odds of disease in a cohort study we get

$$\ln\frac{p'_x}{q'_x} = \ln\frac{\pi_1}{\pi_2} + \ln\frac{p_x}{q_x}$$

Using substitution and where $\beta'_0 = \ln\frac{\pi_1}{\pi_2} + \beta_0$, we have

$$\ln\frac{p'_x}{p'_x} = \ln\frac{\pi_1}{\pi_2} + \ln\frac{p_x}{q_x} = \beta'_0 + \beta_1 x_1 + \cdots \beta_n x_n$$

Therefore, the only difference in a case–control study and a cohort study using the logistic model is in the β_0. All other parameters are unaffected by study design (Schlesselman 1982).

7 References

Breslow, N. and W. Power. 1978. Are there two logistic regressions for retrospective studies? *Biometrics* 34: 100–105.

Canterino, J., U. Verma, P. Visintainer, R. Figueroa, S. Klein, and N. Tejani. 1999. Maternal magnesium sulfate and the development of neonatal periventricular leucomalacia and intraventricular hemorrhage. *Obstetrics and Gynecology* 93: 396–402.

Clayton, D. and M. Hills. 1993. *Statistical Models in Epidemiology*. Oxford: Oxford University Press.

Collett, D. 1991. *Modelling Binary Data*. New York: Chapman & Hall.

Cornfield, J. 1951. A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute* 11: 1269.

Garrett, J. M. 1996. sg50: Graphical assessment of linear trend. *Stata Technical Bulletin* 30: 9–15. In *Stata Technical Bulletin Reprints*, vol. 5, 152–160. College Station, TX: Stata Press.

Gould, W. 2000. sg124: Interpreting logistic regression in all its forms. *Stata Technical Bulletin* 53: 19–29. In *Stata Technical Bulletin Reprints*, vol. 9, 257–270. College Station, TX: Stata Press.

Harrell, F. 2001. *Regression Modeling Strategies*. New York: Springer.

Hosmer, D. W., Jr. and S. Lemeshow. 2000. *Applied Logistic Regression*. 2d ed. New York: John Wiley & Sons.

Prentice, R. and R. Pyke. 1979. Logistic disease incidence models and case–control studies. *Biometrika* 66: 403–411.

Robins, J. M., N. Breslow, and S. Greenland. 1986. Estimators of the Mantel–Haenszel variance consistent in both sparse data and large-strata limiting models. *Biometrics* 42: 311–323.

Rothman, K. and S. Greenland. 1998. *Modern Epidemiology*. 2d ed. Philadelphia: Lippincott–Raven.

Royston, P. and G. Ambler. 1999. sg112.1: Nonlinear regression models involving power or exponential functions of covariates: update. *Stata Technical Bulletin* 50: 26. In *Stata Technical Bulletin Reprints*, vol. 9, 180. College Station, TX: Stata Press.

Schlesselman, J. 1982. *Case–Control Studies*. Oxford: Oxford University Press.

About the Authors

Susan Hailpern is a research associate at the Albert Einstein School of Medicine at Yeshiva University. She is also a lecturer in biostatistics and epidemiology at the School of Public Health at New York Medical College.

Paul Visintainer is Professor and Program Director for Health Quantitative Sciences in the School of Public Health at New York Medical College. He is interested in perinatal epidemiology, autism, and cerebral palsy.