CHAPTER 6

Application of Logistic Regression with Different Sampling Models

6.1 INTRODUCTION

Up to this point we have assumed that our data have come from a simple random sample. Considerable progress has been made in recent years to extend the use of the logistic regression model to other types of sampling. In this chapter we begin with a review of the classic cohort study. Next we consider the case-control study and the stratified case-control study. We conclude with a section that deals with fitting models when data come from a complex sample survey. The goals are to briefly describe some of the mathematics involved in fitting the model, to indicate how the model can be fit using available software and to discuss the interpretation of the estimated parameters. References to the literature for more detailed treatment of these topics are provided.

Throughout this chapter we assume that the outcome variable is dichotomous, coded as 0 or 1, and that its conditional probability given a vector of covariates is the logistic regression model. In addition, we assume that the number of covariate patterns is equal to the sample size. Modifications to allow for replication at covariate patterns are a notational detail, not a conceptual problem.

6.2 COHORT STUDIES

Several variations of the cohort (or prospective) study are in common use. In the simplest design, a simple random sample of subjects is chosen and the values of the covariates are determined. These subjects are then followed for a fixed period of time and the outcome variable is measured. This type of sample is identical to what is often referred to as the *regression sampling model*, in which we assume that

Applied Logistic Regression, Third Edition.
David W. Hosmer, Jr., Stanley Lemeshow, and Rodney X. Sturdivant.

the values of the covariates are fixed and measured without error and the outcome is measured conditionally on the observed values of the covariates. Under these assumptions and independence of the observations, the likelihood function for a sample of size n is simply

$$l_1(\boldsymbol{\beta}) = \prod_{i=1}^{n} \Pr(Y_i = y_i | \mathbf{x}_i). \tag{6.1}$$

When the observed values of y and the logistic regression model are substituted into the expression for the conditional probability, $l_1(\boldsymbol{\beta})$ simplifies to the likelihood function in equation (1.3).

A modification of this situation is a randomized trial where subjects are first chosen via a simple random sample and then allocated independently and with known probabilities into "treatment" groups. Subjects are followed over time and the outcome variable is measured for each subject. If the responses are such that a normal errors model is appropriate we would be naturally led to consider a normal theory analysis of covariance model which would contain appropriate design variables for treatment, relevant covariates, and any interactions between treatment and covariates deemed necessary. The extension of the likelihood function in equation (6.1) to incorporate treatment and covariate information when the outcome is dichotomous is obtained by including these variables in the logistic regression model.

Another modification is for the design to incorporate a stratification variable such as location or clinic. In this situation the likelihood function is the product of the stratum-specific likelihood functions, each of which is similar in form to $l_1(\boldsymbol{\beta})$. We would perhaps add terms to the model to account for stratum-specific responses. These might include a design variable for stratum and interactions between this design variable and other covariates.

In each of these designs we use the likelihood function $l_1(\boldsymbol{\beta})$ as a basis for determining the maximum likelihood estimates of the unknown parameters in the vector $\boldsymbol{\beta}$. Tests and confidence intervals for the parameters follow from well-developed theory for maximum likelihood estimation [see Cox and Hinkley (1974)]. The estimated parameters may be used in the logistic regression model to estimate the conditional probability of response for each subject. The fact that the estimated logistic probability provides a model-based estimate of the probability of response permits the development of methods for assessment of goodness of fit such as those discussed in Chapter 5. Chambless and Boyle (1985) extend $l_1(\boldsymbol{\beta})$ to the setting where the data come from a stratified simple random sample.

In some prospective studies the outcome variable of interest is the time to the occurrence of some event. In these studies the time to event nowadays is most often modeled using the proportional hazards model or another regression model [see Hosmer et al. (2008)]. In these situations a method of analysis that is sometimes used is to ignore the actual failure time and model the occurrence or nonoccurrence of the event via logistic regression. This method of analysis was popular before easily used software became available in the major software packages to model time-to-event data. However, now such software is just as available and just as

easy to use, and as such we see no need to use logistic regression analysis to model time to event data.

6.3 CASE-CONTROL STUDIES

One of the major reasons the logistic regression model has seen such wide use, especially in epidemiologic research, is the ease of obtaining adjusted odds ratios from the estimated slope coefficients when sampling is performed conditional on the outcome variables, as in a case-control study. Breslow (1996) has written an excellent review paper. Besides tracing the development of the case-control study he describes the statistical issues and controversies surrounding some famous studies such as the first Surgeon General's report on smoking and health [Surgeon General (1964)]. He presents some of the newer innovative applications involving nesting and matching as well as some of the current limitations of this study design. We encourage any reader not familiar with this powerful and frequently employed study design to read this paper. We only consider the use of logistic regression in the simplest case-control designs in this section. More advanced applications may be found in Breslow (1996) and cited references.

As noted by Breslow (1996), Cornfield (1951) is generally given credit for first observing that the odds ratio is invariant under study design (cohort or case-control). However, it was not until the work of Farewell (1979) and Prentice and Pyke (1979) that the mathematical details justifying the common practice of analyzing case-control data as if they were cohort data were worked out.

In contrast to cohort studies, the binary outcome variable in a case-control study is fixed by stratification. The dependent variables in this setting are one or more primary covariates, exposure variables in $\mathbf{x}$. In this type of study design, samples of fixed size are chosen from the two strata defined by the outcome variable. The values of the primary exposure variables and the relevant covariates are then measured for each subject selected. The covariates are assumed to include all relevant exposure, confounding, and interaction terms. The likelihood function is the product of the stratum-specific likelihood functions and depends on the probability that the subject was selected for the sample, and the probability distribution of the covariates.

It is not difficult algebraically to manipulate the case-control likelihood function to obtain a logistic regression model in which the dependent variable is the outcome variable of interest to the investigator. The key steps in this development are two applications of Bayes' theorem. As the likelihood function is based on subjects selected, we need to define a variable that records the selection status for each subject in the population. Let the variable s denote the selection ($s = 1$) or nonselection ($s = 0$) of a subject. The full likelihood for a sample of size n_1 cases ($y = 1$) and n_0 controls ($y = 0$) is

$$\prod_{i=1}^{n_1} \Pr(\mathbf{x}_i | y_i = 1, s_i = 1) \prod_{i=1}^{n_0} \Pr(\mathbf{x}_i | y_i = 0, s_i = 1). \tag{6.2}$$

For an individual term in the likelihood function shown in equation (6.2) the first application of Bayes' theorem yields

$$\Pr(\mathbf{x}|y, s = 1) = \frac{\Pr(y|\mathbf{x}, s = 1)\ \Pr(\mathbf{x}|s = 1)}{\Pr(y|s = 1)}. \tag{6.3}$$

The second application of Bayes' theorem is to the first term in the numerator of equation (6.3). This yields, when $y = 1$,

$$\Pr(y = 1|\mathbf{x}, s = 1)$$
$$= \frac{\Pr(y = 1|\mathbf{x})\ \Pr(s = 1|\mathbf{x}, y = 1)}{\Pr(y = 0|\mathbf{x})\ \Pr(s = 1|\mathbf{x}, y = 0) + \Pr(y = 1|\mathbf{x})\ \Pr(s = 1|\mathbf{x}, y = 1)}. \tag{6.4}$$

Assume that the selection of cases and controls is independent of the covariates with respective probabilities τ_1 and τ_0; then

$$\tau_1 = \Pr(s = 1|y = 1, \mathbf{x}) = \Pr(s = 1|y = 1),$$

and

$$\tau_0 = \Pr(s = 1|y = 0, \mathbf{x}) = \Pr(s = 1|y = 0).$$

Substitution of τ_1, τ_0 and the logistic regression model, $\pi(\mathbf{x})$, for $\Pr(y = 1|\mathbf{x})$, into equation (6.4) yields

$$\Pr(y = 1|\mathbf{x}, s = 1) = \frac{\tau_1 \pi(\mathbf{x})}{\tau_0[1 - \pi(\mathbf{x})] + \tau_1 \pi(\mathbf{x})}. \tag{6.5}$$

If we divide the numerator and denominator of the expression on the right-hand side of equation (6.5) by $\tau_0[1 - \pi(\mathbf{x})]$, the result is a logistic regression model with intercept term $\beta_0^* = \ln(\tau_1/\tau_0) + \beta_0$. To simplify the notation, let $\pi^*(\mathbf{x})$ denote the right-hand side of equation (6.5). As we assume that sampling is carried out independent of covariate values, $\Pr(\mathbf{x}|s = 1) = \Pr(\mathbf{x})$, where $\Pr(\mathbf{x})$ denotes the probability distribution of the covariates. The general term in the likelihood shown in equation (6.3) then becomes, for $y = 1$,

$$\Pr(\mathbf{x}|y = 1, s = 1) = \frac{\pi^*(\mathbf{x})\Pr(\mathbf{x})}{\Pr(y = 1|s = 1)}. \tag{6.6}$$

A similar term for $y = 0$ is obtained by replacing $\pi^*(\mathbf{x})$ by $[1 - \pi^*(\mathbf{x})]$ in the numerator and $\Pr(y = 1|s = 1)$ by $\Pr(y = 0|s = 1)$ in the denominator of equation (6.6). If we let

$$l^*(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi^*(\mathbf{x}_i)^{y_i}[1 - \pi^*(\mathbf{x}_i)]^{1-y_i},$$

the likelihood function shown in equation (6.2) becomes

$$l^*(\beta) \prod_{i=1}^{n} \left[\frac{\Pr(\mathbf{x}_i)}{\Pr(y_i|s_i = 1)} \right]. \tag{6.7}$$

The first term in equation (6.7), $l^*(\boldsymbol{\beta})$, is the likelihood obtained when we pretend the case-control data were collected in a cohort study, with the outcome of interest modeled as the dependent variable. If we assume that the probability distribution of $\mathbf{x}$, $\Pr(\mathbf{x})$, contains no information about the coefficients in the logistic regression model, then maximization of the full likelihood with respect to the parameters in the logistic model, $\pi^*(\mathbf{x})$, is only subject to the restriction that $\Pr(y_i = 1|s_i = 1) = n_1/n$ and $\Pr(y_i = 0|s_i = 1) = n_0/n$. The likelihood equation obtained by differentiating with respect to the parameter β_0^* assures that this condition is satisfied. Thus, maximization of the full likelihood with respect to the parameters in $\pi^*(\mathbf{x})$ need only consider that portion of the likelihood which looks like a cohort study. The implication of this is that *analysis of data from case-control studies via logistic regression may proceed in the same way and using the same computer programs as cohort studies*. Nevertheless, inferences about the intercept parameter β_0 are not possible without knowledge of the sampling fractions within cases and controls, τ_0 and τ_1.

The assumption that the marginal distribution of $\mathbf{x}$ contains no information about the parameters in the logistic regression model requires additional discussion, as it is not true in one historically important situation, the normal theory discriminant function model. This model was discussed briefly in Chapters 1 and 2. When the assumptions for the normal discriminant function model hold, the maximum likelihood estimators of the coefficients for the logistic regression model obtained from conditional likelihoods such as those in equations (6.2) and (6.7) are less efficient than the discriminant function estimator shown in equation (2.11) [see Efron (1975)]. However, the assumptions for the normal theory discriminant function model are rarely, if ever, attained in practice. Application of the normal discriminant function when its assumptions do not hold may result in substantial bias, especially when some of the covariates are dichotomous variables. As a general rule, estimation should be based on equations (6.2) and (6.7), unless there is considerable evidence in favor of the normal theory discriminant function model.

Prentice and Pyke (1979) have shown that the maximum likelihood estimators obtained by pretending that the case-control data resulted from a cohort sample have the usual properties associated with maximum likelihood estimators. Specifically, they are asymptotically normally distributed, with covariance matrix obtained from the inverse of the information matrix. Thus, percentiles from the $N(0, 1)$ distribution may be used in conjunction with estimated standard errors produced from standard logistic regression software to form Wald statistics and confidence interval estimates. The theory of likelihood ratio tests may be employed to compare models via the difference in the deviance of the two models, assuming of course that the models are nested. Scott and Wild (1991) have shown that inferences based on this approach are sensitive to incorrect specifications of the logit function. They show that failure to include necessary higher order terms in the logit produces a model with estimated standard errors that are too small. These results are special cases of more general results obtained by White (1982).

Modification of the likelihood function to incorporate additional levels of stratification beyond case-control status follows in the same manner as described for

cohort data (i.e., inclusion of relevant design variables and interaction terms). Thus, model building and inferences from fitted models for case-control data may proceed using the methods developed for cohort data, as described in Chapters 4 and 5. However, this approach is not valid for matched or highly stratified data. Appropriate methods for the analysis of the latter are presented in detail in Chapter 7.

Fears and Brown (1986) proposed a method for the analysis of stratified case-control data that arise from a two-stage sample. Breslow and Cain (1988) and Scott and Wild (1991) provide further discussion and refinement of the method. This approach requires that we know the sampling rates for the first stage and the total number of subjects in each stratum. This information is used to define the relative sampling rates for cases and controls within each stratum. The ratio of these is included in the model in the form of an additional known constant added to the stratum-specific logit. Specifically, suppose we let n_j be the total number of subjects with $y = j$ observed out of a possible N_j and let the kth stratum-specific quantities be n_{jk} and N_{jk}, $j = 0, 1$, and $k = 1, 2, \ldots, K$. The relative stratum-specific sampling rates are $w_{1k} = (n_{1k}/N_{1k})/(n_1/N_1)$ and $w_{0k} = (n_{0k}/N_{0k})/(n_0/N_0)$. The Fears and Brown model uses stratum-specific logits of

$$g_k(\mathbf{x}) = \ln\left(\frac{w_{1k}}{w_{0k}}\right) + \beta_0 + \boldsymbol{\beta}'\mathbf{x},$$

$k = 1, 2, \ldots, K$. This model may be handled with standard logistic regression software by defining a new variable, typically referred to as an offset, which takes on the value $\ln(w_{1k}/w_{0k})$ and forcing it into the model with a coefficient equal to 1.0.

Breslow and Cain (1988) show that the estimator proposed by Brown and Fears is asymptotically normally distributed and derive an estimator of the covariance matrix. Breslow and Zhao (1988) and Scott and Wild (1991) point out that the estimated standard errors produced when standard logistic regression software is used to implement the Brown and Fears method overestimate the true standard errors. They provide expressions for a covariance matrix that yields consistent estimates of the variances and covariances of the estimated regression coefficients. The matrix is complicated to compute, as it requires a special purpose program or a high degree of skill in using a package allowing matrix calculations such as SAS, STATA, or R [R Development Core Team (2010)]. For these reasons we do not present the variance estimator in detail. We note that Breslow and Zhao use a slightly different offset, $\ln[(n_{1k}/N_{1k})/(n_{0k}/N_{0k})]$, which yields the same estimates of the regression coefficients but a different intercept.

Before leaving our discussion of logistic regression in the case-control setting, we briefly consider the application of the chi-square goodness of fit tests for the logistic regression model presented in Section 5.2. The essential feature of these tests is that for a particular covariate pattern, the number of subjects with the response of interest among m sampled is distributed binomially with parameters m and response probability given by the hypothesized logistic regression model. Recall that for cohort data, the likelihood function was parameterized directly in terms of the logistic probability. For case-control data, the function $\pi^*(\mathbf{x})$ is the probability $P(y = 1|\mathbf{x}, s = 1)$. For a particular covariate pattern, conditioning on

the number of subjects m observed to have a given covariate pattern is equivalent to conditioning on the event, $(\mathbf{x}, s = 1)$. Thus, for case-control studies in which the logistic regression model assumption is correct, the conditional distribution of the number of subjects responding among the m observed to have a particular covariate pattern is binomial with parameters m and $\pi^*(\mathbf{x})$. Hence, the results developed in Chapter 5 based on m-asymptotics also apply. Nagelkerke et al. (2005) propose a test based on the effect on the estimated coefficients of weighting the observed outcomes. This test is focused specifically on model misspecification in the covariates and thus has, in simulations, higher power than the decile of risk test discussed in Chapter 5 when this is the source of lack of fit. The test is modestly complicated to calculate and as yet has not found its way into software packages. As such, we do not consider it further.

It is often the case that data from case-control studies do not arise from simple random samples within each stratum. For example, the design may call for the inclusion of all subjects with $y = 1$ and a sample of subjects with $y = 0$. For these designs there is an obvious dependency among the observations. If this dependency is not too great, or if we appeal to a super-population model [see Prentice (1986)], then employing a theory that ignores it should not bias the results significantly.

6.4 FITTING LOGISTIC REGRESSION MODELS TO DATA FROM COMPLEX SAMPLE SURVEYS

Some of the more recent improvements in logistic regression statistical software include routines to perform analyses with data obtained from complex sample surveys. These routines may be found in STATA, SAS, SUDAAN [Shah et al. (2002)], and other less well-known special-purpose packages. Our goal in this section is to provide a brief introduction to these methods and to illustrate them with an example data set. The reader who needs more detail is encouraged to see Korn and Graubard (1990), Roberts et al. (1987), Skinner et al. (1989), and Thomas and Rao (1987).

The essential idea, as discussed in Roberts et al. (1987), is to set up a function that approximates the likelihood function in the finite sampled population with a likelihood function formed from the observed sample and known sampling weights. Suppose we assume that the population may be broken into $k = 1, 2, \ldots, K$ strata, $j = 1, 2, \ldots, M_k$ primary sampling units in each stratum and $i = 1, 2, \ldots, N_{kj}$ elements in the kj^{th} primary sampling unit. Suppose our observed data consist of n_{kj} elements from m_k primary sampling units from stratum k. Denote the total number of observations as $n = \sum_{k=1}^{K} \sum_{j=1}^{m_k} n_{kj}$. Denote the known sampling weight for the kji^{th} observation as w_{kji}, the vector of covariates as $\mathbf{x}_{kji}$ and the dichotomous outcome as y_{kji}. The approximate log-likelihood function is

$$\sum_{k=1}^{K} \sum_{j=1}^{m_k} \sum_{i=1}^{n_{kj}} [w_{kji} \times y_{kji}] \times \ln[\pi(\mathbf{x}_{kji})] + [w_{kji} \times (1 - y_{kji})] \times \ln[1 - \pi(\mathbf{x}_{kji})]. \tag{6.8}$$

Differentiating this equation with respect to the unknown regression coefficients yields the vector of $p + 1$ score equations

$$\mathbf{X}'\mathbf{W}(\mathbf{y} - \boldsymbol{\pi}) = \mathbf{0}, \tag{6.9}$$

where $\mathbf{X}$ is the $n \times (p + 1)$ matrix of covariate values, $\mathbf{W}$ is an $n \times n$ diagonal matrix containing the weights, $\mathbf{y}$ is the $n \times 1$ vector of observed outcomes, and $\boldsymbol{\pi} = [\pi(\mathbf{x}_{111}), \ldots, \pi(\mathbf{x}_{K m_K n_{Kj}})]'$ is the $n \times 1$ vector of logistic probabilities. In theory, any logistic regression package that allows weights could be used to obtain the solutions to equation (6.9). The problem comes in obtaining the correct estimator of the covariance matrix of the estimator of the coefficients. Naive use of a standard logistic regression package with weight matrix $\mathbf{W}$ would yield estimates on the matrix $(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}$ where $\mathbf{D} = \mathbf{W}\mathbf{V}$ is an $n \times n$ diagonal matrix with general element $w_{kji} \times \hat{\pi}(\mathbf{x}_{kji})[1 - \hat{\pi}(\mathbf{x}_{kji})]$. The correct estimator is

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{S}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}, \tag{6.10}$$

where $\mathbf{S}$ is a pooled within-stratum estimator of the covariance matrix of the left-hand side of equation (6.9). Denote a general element in the vector in equation (6.9) as $\mathbf{z}'_{kji} = \mathbf{x}'_{kji} w_{kji}(y_{kji} - \pi(\mathbf{x}_{kji}))$, the sum over the n_{kj} sampled units in the jth primary sampling unit in the kth stratum as $\mathbf{z}_{kj} = \sum_{i=1}^{n_{kj}} \mathbf{z}_{kji}$ and their stratum-specific mean as $\bar{\mathbf{z}}_k = 1/m_k \sum_{j=1}^{m_k} \mathbf{z}_{kj}$. The within-stratum estimator for the kth stratum is

$$\mathbf{S}_k = \frac{m_k}{m_k - 1} \sum_{j=1}^{m_k} (\mathbf{z}_{kj} - \bar{\mathbf{z}}_k)(\mathbf{z}_{kj} - \bar{\mathbf{z}}_k)'.$$

The pooled estimator is $\mathbf{S} = \sum_{k=1}^{K} (1 - f_k)\mathbf{S}_k$. The quantity $(1 - f_k)$ is called the *finite population correction factor*, where $f_k = m_k / M_k$ is the ratio of the number of observed primary sampling units to the total number of primary sampling units in stratum k. In settings where M_k is unknown it is common practice to assume it is large enough that f_k is quite small and the correction factor is equal to 1.

The likelihood function in equation (6.8) is only an approximation to the true likelihood. Thus, inferences about model parameters should be based on univariable and multivariable Wald statistics rather than likelihood ratio tests. Wald tests are formed by comparing an estimated coefficient to an estimate of its standard error, or variance, computed from specific elements of equation (6.10) in the same manner as described in Chapter 2. However, simulations in Korn and Graubard (1990) as well as Thomas and Rao (1987) show that when data come from a complex sample survey from a finite population, use of a modified Wald statistic and the F distribution, described below, yield tests with better adherence to the stated alpha level. STATA and SUDAAN report results from these modified Wald tests. The problem is that none of the simulations referred to actually examines logistic regression models fit using continuous and categorical covariates with estimates obtained from equation (6.9) and variances from equation (6.10). Korn and Graubard appear to use a linear regression with normal errors model and refer to theoretical results

in Anderson (1984) that depend on rather stringent assumptions of multivariate normality. Thomas and Rao examine models with a dichotomous or polychotomous outcome and a few categorical covariates. Another problem, in our opinion, is the fact that software packages, for example STATA, use the t distribution to assess significance of Wald statistics for individual coefficients. Given the paucity of appropriate simulations and theory we are not convinced that there is sufficient evidence to support the use of the modified Wald statistic with the F distribution with logistic regression models. One possible justification is that the use of the modified Wald statistic with the F distribution is conservative in that significance levels using this approach are, in general, larger than those obtained from treating the Wald statistics as being multivariate normal for sufficiently large samples (as is assumed in previous chapters). We present results based on both tests in the example.

The relationship between the Wald test and the modified Wald test is as follows. Let W denote the Wald statistic for testing that all p slope coefficients in a fitted model are equal to 0, that is

$$W = \hat{\boldsymbol{\beta}}'[\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})_{p\times p}]^{-1}\hat{\boldsymbol{\beta}}, \tag{6.11}$$

where $\hat{\boldsymbol{\beta}}$ denotes the vector of p slope coefficients and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})_{p\times p}$ is the $p \times p$ submatrix obtained from the full $(p + 1) \times (p + 1)$ matrix in equation (6.10). That is, one leaves out the row and column for the constant term. The p-value is computed using a chi-square distribution with p degrees of freedom as $\Pr[\chi^2(p) \geq W]$.

The adjusted Wald statistic is

$$F = \frac{(s - p + 1)}{sp}W, \tag{6.12}$$

where $s = \left(\sum_{k=1}^{K} m_k\right) - K$ is the total number of sampled primary sampling units minus the number of strata. The p-value is computed using an F distribution with p and $(s - p + 1)$ degrees of freedom as $\Pr[F(p, s - p + 1) \geq F]$.

For purposes of illustration we use selected variables (see Table 1.11) from the 2009–2010 cycle of the National Health and Nutrition Examination Study [NHANES III Reference Manuals and Reports (2012)]. We describe the data in Section 1.6.7. It should be noted that the NHANES, like just about any other large survey, suffers from the fact that complete data are not available for every subject. This problem is exacerbated in complex sample surveys because every subject carries along a unique statistical weight based on the number of individuals in the population he or she represents. Hence, if that subject is missing a measurement on just one of the variables involved in a multivariable problem, then that subject will be eliminated from the analysis and the sum of the statistical weights of the subjects remaining will not equal the size of the population for which inference is to be made.

Survey statisticians have studied this problem extensively. Solutions to it range from redistributing the statistical weights of the dropped subjects among the subjects remaining, to imputing every missing value so that the weights will be preserved. Another, perhaps simplistic, approach is simply to run the analyses

with the subjects having complete data and assume that the relationships would not change had all subjects been used. Because it is our intention in this book to demonstrate the use of logistic regression analysis with complex survey data rather than to obtain precise population parameter estimates, we will follow this simple approach. (NHANES actually advocates this approach if the number of missing observations is small, less than 10%.)

For purposes of illustrating fitting logistic models to sample survey data in Section 6.4 we chose selected variables, see Table 1.11, from the 2009–2010 cycle of the NHANES III Reference Manuals and Reports (2012) and made some modifications to the data. This is a stratified multistage probability sample of the civilian noninstitutionalized population of the United States.

As an example we fit a logistic regression model to data from the 2009–2010 cycle of the National Health and Nutrition Examination Study [NHANES III Reference Manuals and Reports (2012)] described in Section 1.6.7. The model, shown in Table 6.1, contains age in decades (AGE10), diastolic blood pressure (DBP), gender (GENDER), walk or bike to work (WLKBIK), participates in vigorous recreational activities (VIGRECEXR), moderate work activity (MODWRK), and participates in moderate recreational activities (MODRECEXR). The 5858 subjects used in the analysis represent 204,203,191 individuals between 16 and 80 years of age living in the United States in 2009–2010.

We assessed the overall significance of the model via the multivariable Wald test and adjusted Wald test for the significance of the seven regression coefficients in the model. For the model in Table 6.1 the value of the Wald test in equation (6.11) is

$$W = \hat{\boldsymbol{\beta}}'[\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})_{7\times 7}]^{-1}\hat{\boldsymbol{\beta}} = 179.0189,$$

where $\hat{\boldsymbol{\beta}}$ is the vector of the seven estimated slope coefficients and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})_{7\times 7}$ is the 7×7 sub-matrix computed using equation (6.10). The significance level of the test is $\Pr[\chi^2(7) \geq 179.0189] < 0.001$. The value of s for the adjusted Wald

Table 6.1 Estimated Coefficients, Standard Errors, *z*-Scores, Two-Tailed *p*-Values, and 95% Confidence Intervals for a Logistic Regression Model for the Modified NHANES Study with Dependent Variable OBESE, *n* = 5858

	Coeff.	Std. Err.	*t*	*p*	95% CI	
AGE10[a]	0.001	0.0258	0.05	0.962	−0.054,	0.056
DBP	0.019	0.0046	4.09	0.001	0.009,	0.029
GENDER	0.467	0.1240	3.76	0.002	0.202,	0.731
WLKBIK	0.489	0.0920	5.32	<0.001	0.293,	0.685
VIGRECEXR	0.801	0.1100	7.29	<0.001	0.567,	1.036
MODWRK	−0.027	0.0923	−0.29	0.773	−0.224,	0.170
MODECEXR	0.330	0.1721	1.92	0.074	−0.037,	0.697
Constant	−4.610	0.4237	−10.88	<0.001	−5.513,	−3.707

[a] $\mathrm{AGE10} = \frac{\mathrm{AGE}}{10}$.

test is $30 - 15 = 15$ and the adjusted Wald test from equation (6.12) is

$$F = \frac{(15 - 7 + 1)}{15 \times 7} \times 179.0189 = 15.3444,$$

and $p = \Pr[F(7, 9) \geq 15.3444] < 0.001$. Both tests indicate that at least one of the coefficients may be different from 0.

The results in Table 6.1 indicate, on the basis of the individual p-values for the Wald statistics, that age, moderate work activity and moderate recreation may not be significant at the 5% level. As age ranges from 16 to 80 and there is evidence that obesity is most prevalent in middle age we suspect that the logit may be nonlinear in age. Hence, for subject matter reasons we do not consider age for exclusion from the model at this time. As we noted, the function in equation (6.8) is not a true likelihood function. Thus, we cannot use the partial likelihood ratio test to compare a smaller model to a larger model. In this case we must test for the significance of the coefficients of excluded covariates using a multivariable Wald test based on the estimated coefficients and estimated covariance matrix from the 8×8 larger model.

Application of the Wald test to assess the significance of the coefficients for MODWRK and MODRECEXR from the model in Table 6.1 uses the vector of estimated coefficients

$$\hat{\boldsymbol{\beta}}' = (-0.027088, 0.329987),$$

and the 2×2 sub-matrix of estimated variances and covariances obtained from the full matrix (not shown) computed using equation (6.10)

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})_{2\times 2} = \begin{bmatrix} 0.00852197 & -0.00873523 \\ -0.00873523 & 0.02962245 \end{bmatrix}.$$

The Wald test statistic is

$$W = \hat{\boldsymbol{\beta}}'[\widehat{\text{Var}}(\hat{\boldsymbol{\beta}})_{2\times 2}]^{-1}\hat{\boldsymbol{\beta}} = 4.5052,$$

with a p-value obtained as $P[\chi^2(2) \geq 4.5052] = 0.1051$. The adjusted Wald test is

$$F = \frac{(15 - 2 + 1)}{15 \times 2} \times 4.5052 = 2.1024,$$

and $p = \Pr[F(2, 14) \geq 2.1024] = 0.1591$. We note that the p-value for the adjusted Wald test is slightly larger than that of the Wald test; however, neither is significant. Thus, both tests indicate that we do not have sufficient evidence to conclude that the coefficients for MODWRKL and MODRECEXR are significantly different from 0. We now fit the reduced model.

The results of fitting the model deleting MODEXR and MODREXEXR are shown in Table 6.2. The first thing we do is to compare the magnitude of the coefficients in Table 6.2 to those in Table 6.1 to check for confounding due to the excluded covariates. As can be seen there is virtually no difference in the two sets of coefficients suggesting that neither covariate removed is a confounder of the relationship between any of the remaining covariates and obesity (BMI > 35).

Table 6.2 Estimated Coefficients, Standard Errors, *z*-Scores, Two-Tailed *p*-Values, and 95% Confidence Intervals for a Logistic Regression Model for the Modified NHANES Study with Dependent Variable OBESE, *n* = 5859

	Coeff.	Std. Err.	*t*	*p*	95% CI
AGE10[a]	0.001	0.0254	0.03	0.980	−0.054, 0.055
DBP	0.019	0.0045	4.09	0.001	0.009, 0.028
GENDER	0.458	0.1221	3.75	0.002	0.198, 0.718
WLKBIK	0.477	0.0898	5.32	<0.001	0.286, 0.669
VIGRECEXR	0.894	0.1040	8.59	<0.001	0.672, 1.115
Constant	−4.474	0.4471	−10.01	<0.001	−5.427, −3.521

[a] $\text{AGE10} = \frac{\text{AGE}}{10}$.

Following the guidelines we established in previous chapters, at this point in the analysis we would:

- Determine whether the continuous covariates in the model are linear in the logit.
- Determine whether there are any significant interactions among the independent variables in the model.
- Assess model calibration and discrimination through goodness of fit tests and area under the ROC curve.
- Examine the case-wise diagnostic statistics to identify poorly fit and influential covariate patterns.

Unfortunately, most of these procedures are not easily performed when modeling data from complex sample surveys. However, there is much that can be done to approximate the correct analysis by using a weighted ordinary logistic regression.

We can check for nonlinearity in the logit by using fractional polynomials with weights equal to the sampling weights within the ordinary logistic regression program. If a significant nonlinear transformation is found then we can fit the model accounting for the sample weights and with the correct standard error estimates to see if the coefficients remain significant. In any case, any nonlinear transformation must make clinical sense. We applied a weighted fractional polynomial analysis to age and diastolic blood pressure. We found that the (3, 3) transformation for age was significantly better than the linear and one term transformation, (3), using the closed test procedure. There was no evidence for nonlinearity in the logit for diastolic blood pressure. The fit of the model using the $m = 2$ fractional polynomial transformation for age is shown in Table 6.3.

We leave as an exercise demonstrating, using methods illustrated in Chapter 4, that the shape of the logit in the two-term fractional polynomial in age rises gradually from age 16 to its maximum at age 55 and then descends to its minimum at age 80. We also include in this exercise a demonstration that this transformation is better statistically and makes more clinical sense than a model quadratic in age (i.e., one with age and age^2).

Table 6.3 Estimated Coefficients, Standard Errors, z-Scores, Two-Tailed p-Values, and 95% Confidence Intervals for a Logistic Regression Model for the Modified NHANES Study with Dependent Variable OBESE, $n = 5859$

	Coeff.	Std. Err.	t	p	95% CI
AGEFP1[a]	0.019	0.0061	3.12	0.007	0.006, 0.032
AGEFP2[a]	−0.009	0.0029	−3.21	0.006	−0.016, −0.003
DBP	0.014	0.0051	2.66	0.018	0.003, 0.025
GENDER	0.457	0.1224	3.73	0.002	0.196, 0.718
WLKBIK	0.480	0.0928	5.17	<0.001	0.282, 0.677
VIGRECEXR	0.878	0.1014	8.66	<0.001	0.662, 1.094
Constant	−4.419	0.4526	−9.76	<0.001	−5.384, −3.454

[a] $\text{AGE10} = \frac{\text{AGE}}{10}$, $\text{AGEFP1} = (\text{AGE10})^3$, $\text{AGEFP2} = (\text{AGE10})^3 \times \ln(\text{AGE10})$.

It was decided that the only interactions that made clinical sense were those involving gender. None of these were found to be significant at the 5% level when added to the model in Table 6.3. Thus, our preliminary final model is the one shown in Table 6.3.

We noted earlier that one is able to obtain the correct value of the estimator of the coefficients by using a weighted ordinary logistic regression program. Some programs (e.g., STATA) can perform the decile of risk test following this weighted fit. The problem is that it does not test for fit of the model in the correct way. When it uses weights, the ordinary logistic regression program assumes that the value of the weights corresponds to actual observations on subjects, rather than what they really are: statistical weights. Hence the test statistic has an enormously large value and the values of the observed and expected frequencies in the 2×10 table have no relationship to the actual sample values.

Archer et al. (2007) describe an extension of the decile of risk test to sample survey data that correctly tests for model fit. Its implementation in STATA's survey commands is described in Archer and Lemeshow (2006). The test is calculated as follows:

1. Ten groups are formed from the sorted estimated probabilities from the fitted model in such a way that the sum of the sample weights in each group is approximately 10% of the total sum of the sample weights. Thus, the 10 groups are not deciles of risk in the sense used in Chapter 5.
2. Using the sample weights, calculate the weighted mean of the model's residuals, $(y - \hat{\pi})$, within each of the 10 groups. Denote these means as $\widehat{M}_k, k = 1, 2, \ldots, 10$. If the model does not fit then we expect the weighted means of the residuals to be different from 0.
3. A linearized estimator of the covariance matrix derived by Archer (2001), $\hat{\mathbf{V}}(\hat{\mathbf{M}})$, of the $\widehat{M}'s$ is then used to calculate the Wald test of the hypothesis that the means are equal to 0:

$$\widehat{W}_{\widehat{M}} = \widehat{\mathbf{M}}'[\hat{\mathbf{V}}(\widehat{\mathbf{M}})]^{-1}\widehat{\mathbf{M}}.$$

4. The Wald statistic is modified to form an F-corrected test statistic

$$F_{\widehat{M}} = \frac{(s - 10 + 2)}{s \times 10} \hat{W}_{\widehat{M}},$$

and the associated p-value is calculated as

$$p = \Pr[F(10 - 1, s - 10 + 2,) > F_{\widehat{M}}].$$

We note that this test can be used with any number of groups. It is described here with 10 groups because that is the default number in STATA. The number of groups used must be less than $s + 2$. One disadvantage of this test is that when the test rejects fit, we do not have a 2×10 table of observed and estimated expected frequencies to assist us in finding areas where the model does not fit.

Evaluating the test for the fitted model in Table 6.3 yields $F_{\widehat{M}} = 1.6984$ and $p = 0.2487$ (i.e., $\Pr[F(9, 7) > 1.6984] = 0.2487$). Hence the test supports model fit.

Roberts et al. (1987) extend the diagnostics discussed in Chapter 5 to the survey sampling setting. However, the diagnostic statistics have not, as yet, been implemented into any of the commonly available packages. The computations required to obtain the measures of leverage and the contribution to fit are not trivial and require considerable skill in programming matrix calculations. In addition, the version of Cook's distance is not an easily computed function of leverage and contribution to fit.

A "better than doing nothing at all" diagnostics evaluation can be based on fitting the model using an ordinary logistic regression program and obtaining the diagnostic statistics described in Chapter 5. An improvement on the values of the diagnostic statistics can be obtained from the ordinary logistic regression model using, as an initial guess, the values of the coefficients from Table 6.3 and setting the number of iterations to 0. This forces the fit to yield the coefficients in Table 6.3. Options to set the initial guess and control the number of iterations are available in most logistic regression packages. Diagnostic statistics are then calculated, saved and plotted as described in Chapter 5. We leave the details of this as an exercise. The reader might want to take a quick look at Table 6.4 where we show that the differences between the two possible sets of coefficients that one could use to calculate the diagnostics statistics differ by 10% or more for five of the seven values. We did evaluate the diagnostic statistics and found that a few observations are poorly fit but their deletion did not produce important changes in the coefficients. Thus we use the model in Table 6.3 as our final model.

Statistical analyses of survey data that take the survey design (stratification and clustering) and statistical weights into consideration are generally called *design-based*. When such features are ignored and the data are handled as if they arose from a simple random sample, the resulting statistical analyses are termed *model-based*. One approach that analysts have used when dealing with survey data is to estimate parameters using design-based methods but to use model-based methods to perform other functions. For example, in this analysis, determination of linearity of

Table 6.4 Coefficients and 95% Confidence Intervals for Covariates in Table 6.3 Using "Design-Based" versus "Model-Based" Analysis

	"Design-Based" Analysis		"Model-Based" Analysis		
Variable	Coeff.	95% CI	Coeff.	95% CI	Pct. Diff.[a]
AGEFP1	0.019	0.006, 0.032	0.021	0.013, 0.029	11.4
AGEFP2	−0.009	−0.016, −0.003	−0.010	−0.015, −0.006	11.7
DBP	0.014	0.003, 0.025	0.012	0.006, 0.019	−9.0
GENDER	0.457	0.196, 0.718	0.519	0.367, 0.671	13.7
WLKBIK	0.480	0.282, 0.677	0.412	0.233, 0.591	−14.1
VIGRECEXR	0.878	0.662, 1.094	0.665	0.440, 0.890	−24.3
Constant	−4.419	−5.384, −3.454	−4.066	−4.583, −3.549	8.0

[a] $\text{Pct.Diff.} = 100 \times \frac{(\hat{\beta}_{\text{Model}} - \hat{\beta}_{\text{Design}})}{\hat{\beta}_{\text{Design}}}$.

the logit for the continuous covariates in the model, assessment of model calibration and examination of diagnostic statistics could be carried out by treating the data as if they resulted from a simple random sample. Any discoveries made in those analyses would then be implemented in the final design-based analysis. For example, we used fractional polynomial analysis to find that the logit was not linear in age. This knowledge, obtained from the model-based analysis may then be implemented into the more appropriate design-based analysis to obtain the slope coefficients and estimated odds ratios.

It should also be noted that for *linear estimates* such as means, totals and proportions, design-based standard errors are typically much larger than model-based standard errors. In fact, for linear estimates, the design effect (defined as the ratio of the variance under design-based analysis to the variance under simple random sampling) is typically much larger than 1. This measure reflects the inflation in variance that occurs due to homogeneity within clusters and can be expressed as $1 + (n - 1)\rho_y$, where ρ_y is the intracluster correlation coefficient (ICC) and n is the average number of units in the sampled cluster. These ICCs can range from small negative values (when the data within clusters are highly heterogeneous) to unity (when the data in clusters are highly correlated). Only when the data are highly heterogeneous within clusters will the design effect be less than 1. However, as described by Neuhaus and Segal (1993), design effects for *regression coefficients* can be expressed as $1 + (n - 1)\rho_x\rho_y$. Note that in this expression the ICC for the independent variable is multiplied by the ICC for the dependent variable. Both of these quantities are, by definition, less than 1. As a result, the design effect will be smaller than what would be observed for means, totals, or proportions. We also note that because ρ_x and ρ_y are not necessarily in the same direction, the product of the intracluster correlation coefficients could be negative and the resulting design effect could be smaller than 1.

The estimated coefficients and their 95% confidence intervals under both design-based and model-based scenarios and the percentage difference in the two sets of coefficients are presented in Table 6.4. In this example, both modeling approaches

produce coefficients of the same order of magnitude but they do differ by anywhere from 8% to 24%.

In summary, we fit logistic regression models to data obtained from complex sample surveys via an approximate likelihood that incorporates the known sampling weights. We assess the overall model significance as well as tests of subsets of coefficients using multivariable *F*-adjusted Wald tests. However, the interpretation of odds ratios from a fitted model is the same as for models fit to less complicated sampling plans. We note that work needs to be done to make available the case-wise diagnostics obtained from complex sample surveys to the typical user of logistic regression software.

EXERCISES

1. Fit the model in Table 6.4 using a model quadratic in AGE10. Graph the logit functions for the (3, 3) model and (1, 2) model using the method shown in Chapter 5. Which model do you prefer and why? Estimate the odds ratio, with 95% confidence intervals, using the estimated coefficients from the model you prefer.

2. Using all of the covariates in Table 6.1 build, using purposeful selection, a model assessing risk factors for obesity, BMI $>$ 35.

3. Assess the fit and evaluate the diagnostics for the model developed in Problem 2.

4. Estimate the odds ratios and confidence intervals for obesity using your final model from Problem 2 and interpret them in context.