The SRS-based options assume simple random sampling with replacement. Under the weighted SRS option, it is assumed that the domain proportions are consistently estimated using the appropriate element weights, and a binomial covariance matrix is assumed for these proportions. Under the unweighted SRS option, simple random sampling with replacement is assumed, and the data set is assumed to be self-weighting. Thus, all the complexities of the sampling design are ignored.

Because the two versions of the SRS-based option are not valid for complex surveys involving clustering, they will be used as reference options for the design-based option and in the construction of appropriate generalized design-effect matrices. The weighted SRS option is used when assessing the magnitude of the clustering effects on results from multivariate analyses, and the unweighted SRS option can be used as a reference option for the design-based option when examining the effects of all the complexities of the sampling design on analysis results, including the effect of weighting procedures.

The analysis options with respect to sampling design are summarized below:

Option	Allowing weights	Allowing stratification	Allowing clustering
Design-based	Yes	Yes	Yes
Weighted SRS	Yes	No	No
Unweighted SRS	No	No	No

It should be noticed that in multivariate survey analysis, as in the analysis of two-way tables, the design-based approach to inference also constitutes inference on the parameters of the corresponding superpopulation model, provided that the finite population is large (see Rao and Thomas 1988).

8.3 ANALYSIS OF CATEGORICAL DATA

The GWLS method of generalized weighted least squares estimation provides a simple technique for the analysis of categorical data with ANOVA-type logit and linear models on domain proportions. Allowing all the complexities of a sampling design including stratification, clustering and weighting, the design-based option provides a generally valid GWLS analysis. Analysis under the weighted or unweighted SRS options assuming simple random sampling serves as a reference when studying the effects of clustering and weighting on results.

The GWLS method is computationally simple because it is noniterative for both logit and linear models on proportions. The alternative PML and GEE methods of pseudolikelihood and generalized estimating equations for logit models are, as iterative methods, computationally more demanding. For logit regression with

continuous predictors, which are not categorized, the PML and GEE methods can be used but the GWLS method is inappropriate. The application area of the GWLS method is thus more limited than that of PML and GEE methods.

In surveys with large samples, closely related results are usually attained by any of the methods. But in fitting ANOVA-type models there can be many multi-class predictors included in the model and, therefore, the number of domains can be large, and a large element-level sample size is required to obtain a reasonably large number of observations falling in each domain. This is especially important for the GWLS method, which is mainly used in large-scale surveys where the sample sizes can be in thousands of persons, as is the case in the OHC and MFH Surveys. For proper behaviour of GWLS, PML and GEE methods, a large number of sample clusters is beneficial. Recall that this property holds for the OHC Survey.

We consider the GWLS method for a binary response variable and a set of categorical predictors. The data can thus be arranged into a multidimensional table, such as Table 8.1, where the u domains are formed by cross-classifying the categorical predictors and the proportions p_j of the binary response are estimated in each domain. The consistent estimates $\hat{p}_j$, used under the design-based and weighted SRS options, are weighted ratio-type estimators of the form $\hat{p}_j = \hat{n}_{j1}/\hat{n}_j$, where $\hat{n}_{j1}$ is the weighted sample sum of the binary response in domain j, and $\hat{n}_j$ are weighted domain sample sizes. The unweighted proportion estimates $\hat{p}_j^U$, used under the unweighted SRS option, are obtained using the unweighted counterparts n_{j1} and n_j.

When applying the GWLS method for logit and linear modelling under an analysis option, the starting point is the calculation of the corresponding proportion estimate vector and its covariance-matrix estimate. By using these estimates, the model coefficients are estimated, together with a covariance matrix of the estimated coefficients, and using these, fitted proportions and their covariance-matrix estimates are obtained. Further, the Wald test of goodness of fit of the model, and desired Wald tests of linear hypotheses on the model coefficients, are executed. Finally, residual analysis is carried out to more closely examine the fit of the selected model.

Design-based GWLS Estimation

Under the design-based option, a consistent *GWLS estimator* $\hat{\mathbf{b}}_{des}$, denoted $\hat{\mathbf{b}}$ for short in this section, of the $s \times 1$ model coefficient vector $\mathbf{b}$ for a model $F(\mathbf{p}) = \mathbf{Xb}$ is given by

$$\hat{\mathbf{b}} = (\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}F(\hat{\mathbf{p}}), \tag{8.5}$$

where $\hat{\mathbf{V}}_{des}$ is a consistent estimator of the covariance matrix of the consistent domain proportion estimator vector $\hat{\mathbf{p}}$, and $\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H}$ is a covariance-matrix estimator of the function vector $F(\hat{\mathbf{p}})$. An estimate $\hat{\mathbf{V}}_{des}$ is obtained using, for example, the linearization method as described in Chapter 5. The GWLS estimating

equations (8.5) are thus based on the consistently estimated functions $F(\hat{p}_j)$ and their design-based covariance-matrix estimate. The equations also indicate that no iterations are needed to obtain the estimates $\hat{b}_k$. A justification for the label 'GWLS' is that element weights are used in obtaining the proportion vector estimate and its covariance-matrix estimate, which are supplied to the GLS estimating equations.

The GWLS estimator $\hat{\mathbf{b}}$ from (8.5) applies for both logit and linear models on domain proportions. But the matrix $\mathbf{H}$ in the covariance-matrix estimator of the function vector differs. In the logit model, the diagonal $u \times u$ matrix $\mathbf{H}$ of partial derivatives of the functions $F(\hat{p}_j)$ has diagonal elements of the form $h_j = 1/(\hat{p}_j(1 - \hat{p}_j))$. And in the linear model, the matrix $\mathbf{H}$ is an identity matrix with ones on the main diagonal and zeros elsewhere.

Under a partial parametrization of a logit ANOVA model (see Section 8.2), where the columns of the model matrix $\mathbf{X}$ corresponding to the classes of the predictors are binary variables, a log odds ratio interpretation can be given to the estimates $\hat{b}_k$. Thus, an estimate $\exp(\hat{b}_k)$ is the *odds ratio* for the corresponding class with respect to the reference class adjusted for the effects of the other terms in the model. This interpretation of the estimated model coefficients is common in epidemiology and also in social sciences.

A covariance-matrix estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ of the estimated model coefficients $\hat{b}_k$ from (8.5) is used in obtaining Wald test statistics for the coefficients. This $s \times s$ covariance matrix is given by

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}}) = (\mathbf{X}'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}\mathbf{X})^{-1}. \tag{8.6}$$

With proper choice of $\mathbf{H}$, this estimator applies again for both logit and linear models. Diagonal elements of $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ provide the design-based variance estimates $\hat{v}_{des}(\hat{b}_k)$ of the estimated coefficients $\hat{b}_k$ to be used in obtaining the corresponding standard-error estimates $\text{s.e}_{des}(\hat{b}_k) = \hat{v}_{des}^{1/2}(\hat{b}_k)$. Under a logit model, using these standard-error estimates, for example, an approximative 95% confidence interval for an odds ratio $\exp(b_k)$ can be calculated as follows:

$$\exp(\hat{b}_k \pm 1.96 \times \text{s.e}_{des}(\hat{b}_k)). \tag{8.7}$$

Two additional covariance-matrix estimators are useful in practice. These are the $u \times u$ covariance-matrix estimator $\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})$ of the vector $\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{b}}$ of the fitted logits and the covariance-matrix estimator $\hat{\mathbf{V}}_{des}(\hat{\mathbf{f}})$ of the vector $\hat{\mathbf{f}} = F^{-1}(\mathbf{X}\hat{\mathbf{b}})$ of the fitted proportions. These are

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}}) = \mathbf{X}\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})\mathbf{X}' \tag{8.8}$$

and

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{f}}) = \hat{\mathbf{H}}^{-1}\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})\hat{\mathbf{H}}^{-1}. \tag{8.9}$$

For a linear model, these covariance matrices obviously coincide, because the fitted functions are equal to the fitted proportions. For a logit model, the diagonal matrix $\hat{\mathbf{H}}$ has diagonal elements of the form $\hat{h}_j = 1/(\hat{f}_j(1-\hat{f}_j))$, and the terms $\hat{f}_j = f_j(\hat{\mathbf{b}})$ are elements of the vector $\hat{\mathbf{f}}$ of fitted proportions calculated using the equation

$$\hat{\mathbf{f}} = \mathbf{f}(\hat{\mathbf{b}}) = \exp(\mathbf{X}\hat{\mathbf{b}})/(1+\exp(\mathbf{X}\hat{\mathbf{b}})). \tag{8.10}$$

The diagonal elements of the covariance-matrix estimates (8.8) and (8.9) are needed to obtain the design-based standard errors of the fitted functions and of the fitted proportions.

Goodness of Fit and Related Tests

Examining goodness of fit of the model is an essential part of a logit and linear modelling procedure on domain proportions. Various goodness-of-fit statistics can be obtained by first partitioning the total variation *(total chi-square)* in the table into the variation due to the model *(model chi-square)* and into the residual variation *(residual chi-square)*. Hence, we have

$$\textit{total chi-square} = \textit{model chi-square} + \textit{residual chi-square}$$

similar to the partition of the total sum of squares for usual linear regression and ANOVA. A design-based Wald test statistic X^2_{des} measuring the residual variation is commonly used as an indicator of goodness of fit of the model. This statistic is given by

$$X^2_{des} = (F(\hat{\mathbf{p}}) - \mathbf{X}\hat{\mathbf{b}})'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}(F(\hat{\mathbf{p}}) - \mathbf{X}\hat{\mathbf{b}}), \tag{8.11}$$

which is asymptotically chi-squared with $u-s$ degrees of freedom under the design-based option. A small value of this statistic, relative to the residual degrees of freedom, indicates good fit of the model, and obviously, the fit is perfect for a saturated model. A Wald statistic denoted by $X^2_{des}(\textit{overall})$, measuring the variation due to the overall model, is used to test the hypothesis that all the model coefficients are zero. It is given by

$$X^2_{des}(\textit{overall}) = F(\hat{\mathbf{p}})'(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H})^{-1}F(\hat{\mathbf{p}}) - X^2_{des}, \tag{8.12}$$

where the first quadratic form measures the total variation and the second is the residual chi-square (8.11) for the model under consideration. This statistic is asymptotically chi-squared with s degrees of freedom. Also, a Wald statistic denoted by $X^2_{des}(\textit{gof})$ can be constructed for the hypothesis that all the model parameters, except the intercept, are zero. This statistic is defined as the difference of the observed values of the residual chi-square statistic (8.11) for the model where only the intercept is included and for the model including all the terms of the

current model, and therefore, it is asymptotically chi-squared with $s-1$ degrees of freedom. The statistic $X^2_{des}(overall)$ is sometimes called a test for the overall model, and $X^2_{des}(gof)$ a test of goodness of fit. Note that all these test statistics apply for both logit and linear models on domain proportions.

Linear hypotheses H_0: $\mathbf{Cb} = \mathbf{0}$ on the model coefficient vector $\mathbf{b}$ can be tested using the Wald statistic

$$X^2_{des}(\mathbf{b}) = (\mathbf{C}\hat{\mathbf{b}})'(\mathbf{C}\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})\mathbf{C}')^{-1}(\mathbf{C}\hat{\mathbf{b}}), \tag{8.13}$$

where $\mathbf{C}$ is the desired $c \times s$ ($c \leq s$) matrix of contrasts. The statistic is asymptotically chi-squared with c degrees of freedom under the design-based option. This statistic is used, for example, in the testing of hypotheses $H_0 : b_k = 0$ on single parameters of the model using the Wald statistics

$$X^2_{des}(b_k) = \hat{b}^2_k / \hat{v}_{des}(\hat{b}_k), \qquad k = 1, \ldots, s,$$

which are asymptotically chi-squared with one degree of freedom. Note that for the corresponding *t-test statistic* the equation $t^2_{des}(b_k) = X^2_{des}(b_k)$ holds.

Another asymptotically valid testing procedure for linear hypotheses on model parameters is based on a second-order Rao–Scott adjustment to a binomial-based Wald test statistic using the Satterthwaite method. This technique is similar to that used in Chapter 7 on the Pearson and Neyman test statistics. We first calculate the GWLS estimate $\hat{\mathbf{b}} = \hat{\mathbf{b}}_{bin}$ by using in (8.5) the binomial covariance-matrix estimate $\hat{\mathbf{V}}_{bin}$ of $\hat{\mathbf{p}}$ in place of $\hat{\mathbf{V}}_{des}$, and construct the corresponding Wald test statistic $X^2_{bin}(\mathbf{b})$:

$$X^2_{bin}(\mathbf{b}) = (\mathbf{C}\hat{\mathbf{b}})'(\mathbf{C}\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})\mathbf{C}')^{-1}(\mathbf{C}\hat{\mathbf{b}}),$$

where $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})$ is the covariance-matrix estimate of the binomial GWLS estimates obtained by using the estimate $\hat{\mathbf{V}}_{bin}$ in place of $\hat{\mathbf{V}}_{des}$ in (8.6). The second-order corrected Wald statistic is given by

$$X^2_{bin}(\mathbf{b}; \hat{\delta}_\bullet, \hat{a}^2) = \frac{X^2_{bin}(\mathbf{b})}{\hat{\delta}_\bullet(1+\hat{a}^2)}, \tag{8.14}$$

where the first-order and second-order adjustment factors $\hat{\delta}_\bullet$ and $(1+\hat{a}^2)$ are calculated from the $c \times c$ generalized design-effects matrix estimate

$$\hat{\mathbf{D}} = (\mathbf{C}\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})\mathbf{C}')^{-1}(\mathbf{C}\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})\mathbf{C}') \tag{8.15}$$

so that

$$\hat{\delta}_\bullet = \mathrm{tr}(\hat{\mathbf{D}})/c$$

is the mean of the eigenvalues $\hat{\delta}_k$ of the generalized design-effects matrix estimate, and

$$(1 + \hat{a}^2) = \sum_{k=1}^{c} \hat{\delta}_k^2 / (c\hat{\delta}_\bullet^2),$$

where the sum of squared eigenvalues is calculated by the formula

$$\sum_{k=1}^{c} \hat{\delta}_k^2 = \mathrm{tr}(\hat{\mathbf{D}}^2).$$

The second-order adjusted statistic $X^2_{bin}(\mathbf{b}; \hat{\delta}_\bullet, \hat{a}^2)$ is asymptotically chi-squared under the design-based option with Satterthwaite adjusted degrees of freedom $\mathrm{df}_S = c/(1 + \hat{a}^2)$. If $c = 1$, as in tests on separate parameters of a model, we have $(1 + \hat{a}^2) = 1$ because the generalized design-effects matrix reduces to a scalar and the adjustment reduces to a first-order adjustment. The test statistics are available in software products for the analysis of complex surveys.

Unstable Situations

Because the Wald statistics X^2_{des}, $X^2_{des}(overall)$ and $X^2_{des}(gof)$ of goodness of fit, and the statistic $X^2_{des}(\mathbf{b})$ of linear hypotheses on model parameters, are asymptotically chi-squared under the design-based option, they can be expected to work reasonably well if the number m of sample clusters is large relative to the number u of domains. But the test statistics can become overly liberal relative to the nominal significance levels if the covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ appears unstable. This can happen if the degrees of freedom $f = m - H$ are small for an estimate $\hat{\mathbf{V}}_{des}$, relative to the residual or model degrees of freedom.

There are certain F-corrected Wald test statistics available to protect against the effects of instability similar to those used in Chapter 7 for hypotheses of homogeneity and independence. For the goodness-of-fit test statistic (8.11), these degrees-of-freedom corrections are

$$F_{1.des} = \frac{f - (u - s) + 1}{f(u - s)} X^2_{des}, \tag{8.16}$$

referred to the F-distribution with $(u - s)$ and $(f - (u - s) + 1)$ degrees of freedom, and

$$F_{2.des} = X^2_{des}/(u - s), \tag{8.17}$$

referred in turn to the F-distribution with $(u - s)$ and f degrees of freedom. These F-corrections can also be derived for the Wald statistics $X^2_{des}(overall)$ and $X^2_{des}(gof)$, using the corresponding degrees of freedom s or $(s - 1)$ in place of $(u - s)$.

Similar F-corrections can be derived for the Wald test statistics of linear hypotheses on model parameters. For the statistic (8.13), these are

$$F_{1.des}(\mathbf{b}) = \frac{f - c + 1}{fc} X^2_{des}(\mathbf{b}) \tag{8.18}$$

and

$$F_{2.des}(\mathbf{b}) = X^2_{des}(\mathbf{b})/c, \tag{8.19}$$

referred to the F-distributions with c and $(f - c + 1)$, and c and f degrees of freedom, respectively.

Second-order Rao–Scott adjustments can be expected to be robust to instability problems. However, for the second-order corrected statistic (8.14), an F-correction can be derived. It is given by

$$F_{bin}(\mathbf{b}; \hat{\delta}_{\bullet}, \hat{a}^2) = (1 + \hat{a}^2) X^2_{bin}(\mathbf{b}; \hat{\delta}_{\bullet}, \hat{a}^2)/c = X^2_{bin}(\mathbf{b})/(c\hat{\delta}_{\bullet}), \tag{8.20}$$

which is referred to the F-distribution with df_S and f degrees of freedom.

The impact of these F-corrections on p-values of the tests is small if f is large. However, if f is relatively small, and especially if f and the residual degrees of freedom are close, the corrections can be effective. Under serious instability, the statistics $F_{1.des}$, and $F_{1.des}(\mathbf{b})$ or $F_{bin}(\mathbf{b}; \hat{\delta}_{\bullet}, \hat{a}^2)$, are preferable. These corrections have been implemented as testing options in software products for the analysis of complex surveys.

Residual Analysis

It is desirable to examine more closely the fit of the selected model by calculating the raw and standardized residuals. These can be used in detecting possible outlying domain proportions. The raw residuals are simple differences $(\hat{p}_j - \hat{f}_j)$ of the fitted proportions $\hat{f}_j$ from the corresponding observed proportions $\hat{p}_j$. Under the design-based option, the standardized residuals are calculated by first obtaining a covariance-matrix estimate $\hat{\mathbf{V}}_{res}$ of the raw residuals given by

$$\hat{\mathbf{V}}_{res} = \mathbf{H}^{-1}(\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H} - \hat{\mathbf{V}}_{des}(\hat{\mathbf{F}}))\mathbf{H}^{-1}, \tag{8.21}$$

where $\mathbf{H}\hat{\mathbf{V}}_{des}\mathbf{H}$ and $\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})$ are the design-based covariance-matrix estimates of the vector $F(\hat{\mathbf{p}})$ of the observed functions and the vector $\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{b}}$ of the fitted functions, respectively, and the matrix $\mathbf{H}$ depends on which model type, logit or linear, is fitted. Using (8.21), the standardized residuals are calculated as

$$\hat{e}_j = (\hat{p}_j - \hat{f}_j)/\sqrt{\hat{v}_j}, \quad j = 1, \ldots, u, \tag{8.22}$$

where $\hat{v}_j$ are the diagonal elements of the residual covariance matrix $\hat{\mathbf{V}}_{res}$. A large standardized residual indicates that the corresponding domain is poorly accounted for by the model. Because the standardized residuals are approximate standard normal variates, they can be referred to critical values from the $N(0,1)$ distribution.

Design Effect Estimation

A principal property of the GWLS method is its flexibility, not only for various model formulations but also for alternative sampling designs. The design-based GWLS method appeared valid under the design-based option involving a complex multi-stage design with clustering and stratification. But the GWLS method can also be used for simpler designs with the choice of an appropriate proportion estimator and its covariance-matrix estimator reflecting the complexities of the sampling design.

Under the weighted SRS option, the consistent proportion estimate $\hat{\mathbf{p}}$ and its binomial covariance-matrix estimate $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}})$ are used in equations (8.5) and (8.6) to obtain the corresponding GWLS estimate $\hat{\mathbf{b}}$ of model coefficients and the covariance-matrix estimate $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})$. The same holds for the unweighted SRS option, where the unweighted counterparts $\hat{\mathbf{p}}^U$ and $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{p}}^U)$ are used. The GWLS estimating equations indicate that the estimates $\hat{b}_k$ obtained under the SRS-based options would not numerically coincide with those from the design-based option.

The SRS-based options are restrictive in the sense that the effect of clustering on standard-error estimates of estimated model coefficients cannot be accounted for. This effect is indicated in design-effect estimates of model coefficient estimates. The design-effect estimates are calculated by using the diagonal elements of the covariance-matrix estimates $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ and $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}}^*)$ of the model coefficients. Hence, we have

$$\hat{d}(\hat{b}_k) = \hat{v}_{des}(\hat{b}_k)/\hat{v}_{bin}(\hat{b}_k^*), \quad k = 1, \ldots, s, \tag{8.23}$$

where $\hat{b}_k^*$ denotes the estimated model coefficients obtained under the weighted or unweighted SRS option. Under the unweighted SRS option, these design-effect estimates indicate the contribution of all the sampling complexities, and under the weighted SRS option, the contribution of clustering is indicated. It is often instructive to calculate the design-effect estimates under both SRS options, because then the contribution of the weighting to design effects can be examined.

Criteria for Choosing a Model Formulation

Which one of the model formulations for proportions, logit or linear, should be chosen? In certain sciences, one type is more standard than the other, but

taking an explicit position in favour of either of the types generally is not possible. It appears that there are gains with the logit formulation, such as possibilities for interpretation with odds ratios, and in certain cases with standard independence concepts. Moreover, being a member of the broad category of so-called exponential family models, a logit model for binomial proportions involves convenient statistical properties that are not shared with linear models for binomial proportions. Although these properties do not necessarily apply to logit models in complex surveys, attention has also been directed to the use of logit models for this kind of survey.

The linear model formulation on proportions, on the other hand, provides a simple modelling approach that is especially convenient for those familiar with linear ANOVA on continuous measurements. Being additive on a linear scale, the coefficients of a linear model describe differences of the proportions themselves, not their logits. In practice, however, logit and linear GWLS estimation results on model coefficients do not markedly differ if proportions are in the range 0.2–0.8, say. In the following example, we compare the logit and linear model formulations in a typical health sciences analysis.

Example 8.1

Logit and linear ANOVA with the GWLS method. Let us apply the GWLS method for logit and linear modelling on domain proportions in the simple OHC Survey setting displayed in Table 8.1. Our aim is to model the variation of domain proportions of the binary response variable PSYCH, measuring overall psychic strain, across the $u = 8$ domains formed by sex and age of respondent, and the variable PHYS describing the respondent's physical working conditions. Table 8.2 provides a more complete description of the analysis situation. The original domain sample sizes $\hat{n}_j$ and the number m_j of sample clusters covered by each domain are included in addition to the domain proportions $\hat{p}_j$, standard errors s.e_j and design effects $\hat{d}_j$. Note that the domain proportions vary around the value 0.5.

The design-based option provides valid GWLS logit and linear modelling in this analysis. The sampling design involves clustering effects, as indicated by design-effect estimates of proportions being on average greater than one. The average design-effect estimate is 1.28. Further, the domains constitute cross-classes, which is indicated by the fact that each domain covers a reasonably large number of sample clusters. More apparently, this property can be seen from the design-based covariance-matrix estimate $\hat{\mathbf{V}}_{des}$ of domain proportions displayed in Figure 8.1. It can be noted that there exist nonzero covariance terms in the off-diagonal part of the covariance-matrix estimate. The estimate also seems relatively stable, because covariance estimates are much smaller than the corresponding variance estimates. The condition number of $\hat{\mathbf{V}}_{des}$ is 12.1, which also indicates stability. The corresponding binomial covariance-matrix estimate $\hat{\mathbf{V}}_{bin}$ is displayed for comparison.

Table 8.2 Proportion $\hat{p}_j$ of persons in the upper psychic strain group, with standard error estimates s.e_j and design-effect estimates $\hat{d}_j$ of the proportions, and domain sample sizes $\hat{n}_j$ and the number of sample clusters m_j (the OHC Survey).

Domain j	SEX	AGE	PHYS	$\hat{p}_j$	s.e_j	$\hat{d}_j$	$\hat{n}_j$	m_j
1	Males	−44	0	0.419	0.0128	1.16	1734	230
2			1	0.472	0.0145	1.33	1578	198
3		45−	0	0.461	0.0178	0.88	690	186
4			1	0.520	0.0247	1.18	483	138
5	Females	−44	0	0.541	0.0125	1.23	1966	240
6			1	0.620	0.0270	1.38	447	152
7		45−	0	0.532	0.0236	1.65	740	185
8			1	0.700	0.0391	1.48	203	101
All				0.500	0.0073	1.69	7841	250

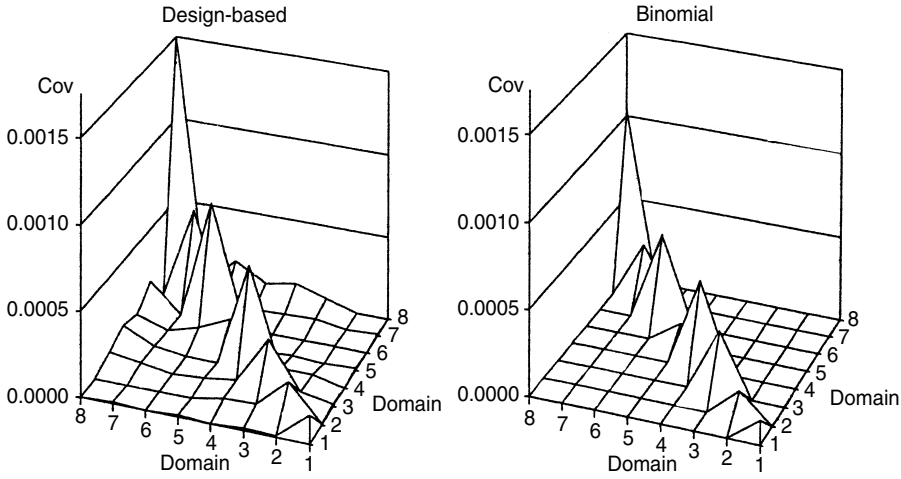

Figure 8.1 Design-based and binomial covariance-matrix estimates $\hat{\mathbf{V}}_{des}$ and $\hat{\mathbf{V}}_{bin}$ of domain proportion estimates $\hat{p}_j$.

We consider the model-building process under the design-based option, and use the unweighted SRS option as a reference. There are three predictors, and together with their main effects, an intercept, and four interaction terms, a total of eight model terms appear in the saturated logit and linear ANOVA models, which can be written in the form

$$\text{F(P)} = \text{INTERCEPT} + \text{SEX} + \text{AGE} + \text{PHYS} + \text{SEX} * \text{AGE} + \text{SEX} * \text{PHYS} + \text{AGE} * \text{PHYS} + \text{SEX} * \text{AGE} * \text{PHYS},$$

where the function is $F(P) = \log(P/(1-P))$ for the logit model and $F(P) = P$ for the linear model, and P stands for proportions of the upper PSYCH group.

In the model-building process, we first fit the saturated logit and linear models and test the significance of the interaction term of all the three predictors. If it appears nonsignificant, we remove the term, and study the two-variable interactions, in turn, for further reduction of the model. Model building is completed when a reasonably well-fitting reduced model is attained. This stepwise process is an example of the so-called *backward elimination* common in fitting of log-linear and logit ANOVA models.

Let us consider more closely the results on logit model fitting. Under the design-based option, the main effects model appeared reasonably well-fitting and could not be further reduced. Results for the model reduction are given in Table 8.3. There, the values of X^2_{des} for a difference Wald statistic are obtained, for example, in the comparison of the saturated model 5 and the model 4. The difference statistic is calculated as $X^2_{des}(overall; 5) - X^2_{des}(overall; 4) = 78.84 - 76.90 = 1.94$, and compared to the chi-squared distribution with one degree of freedom attains a nonsignificant p-value 0.1635, and thus, the interaction term can be removed from the model 5. The observed value of the Wald statistic of goodness of fit of the main effects model (Model 1) is $X^2_{des} = 78.84 - 72.39 = 6.45$, which with 4 degrees of freedom attains a p-value 0.1681, indicating reasonably good fit.

Substantial reduction of the saturated logit model was possible, and the model-building procedure produced quite a simple structure including the main effects terms only. So, the suspected interaction of SEX and PHYS appeared nonsignificant. We return to this conclusion later when fitting logit models under the SRS-based analysis options.

Table 8.3 Observed values of the Wald statistics X^2_{des} (*overall*) for overall models, and the differences statistics X^2_{des} when compared with reduced logit ANOVA models, under the design-based analysis option.

		Overall		Model		Difference	
Model	df	X^2_{des}	p-value	comparison	df	X^2_{des}	p-value
5	8	78.84	0.0000	—	1	—	—
4	7	76.90	0.0000	5–4	1	1.94	0.1635
3	6	76.09	0.0000	4–3	1	0.81	0.3693
2	5	74.78	0.0000	3–2	1	1.31	0.2533
1	4	72.39	0.0000	2–1	1	2.39	0.1218

Model 5: SEX + AGE + PHYS + SEX*AGE + SEX*PHYS + AGE*PHYS + SEX*AGE*PHYS
Model 4: SEX + AGE + PHYS + SEX*AGE + SEX*PHYS + AGE*PHYS
Model 3: SEX + AGE + PHYS + SEX*PHYS + AGE*PHYS
Model 2: SEX + AGE + PHYS + SEX*PHYS
Model 1: SEX + AGE + PHYS

In the partial parametrization used here, for each predictor the model coefficient for the first class is set to zero. The first class of the last domain is the reference domain—here domain 7 in Table 8.2. There are four coefficients b_k to be estimated in the main effects models. GWLS estimates $\hat{b}_k$ are actually obtained under the following model matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}.$$

The fitted models can be written with $\hat{b}_k$ and the model matrix as

$$F(\hat{f}_j) = \hat{b}_1 + \hat{b}_2(\text{SEX})_j + \hat{b}_3(\text{AGE})_j + \hat{b}_4(\text{PHYS})_j, \quad j = 1, \ldots, 8,$$

where $F(\hat{f}_j) = \log(\hat{f}_j/(1-\hat{f}_j))$ for the logit model, and $F(\hat{f}_j) = \hat{f}_j$ for the linear model, and the indicator variable values for SEX, AGE and PHYS are in the second, third and fourth columns of the model matrix **X**.

Let us consider more closely the estimation and test results for the main effects logit model. The estimation results for the model coefficients are displayed in Table 8.4.

Table 8.4 Estimates from design-based logit ANOVA on overall psychic strain (model fitting by the GWLS method).

Model term	Beta coefficient	Design effect	Standard error	*t*-test	*p*-value	Odds ratio	95% confidence interval for OR	
							Lower	Upper
Intercept	−0.3282	1.32	0.0635	−7.02	0.0000	0.72	0.66	0.79
Sex								
Males*	0	n.a.	0	n.a.	n.a.	1	1	1
Females	0.4663	1.44	0.0579	8.06	0.0000	1.59	1.42	1.79
Age								
−44*	0	n.a.	0	n.a.	n.a.	1	1	1
45–	0.1385	1.23	0.0570	2.43	0.0159	1.15	1.03	1.28
Physical health hazards								
No*	0	n.a.	0	n.a.	n.a.	1	1	1
Yes	0.2568	1.30	0.0574	4.48	0.0000	1.29	1.16	1.45

* Reference class; parameter value set to zero.
n.a. not available.

In the table, a positive value of the estimated coefficients $\hat{b}_2$ and $\hat{b}_3$ for females and for the older group is obtained as expected, and the corresponding t-tests attain significant p-values. The sex−age adjusted estimate $\hat{b}_4$ for the PHYS class of more hazardous work is positive, involving a clearly significant t-test. It should be noticed that the absolute value of the t-test statistic used here corresponds to the square root of the F-corrected Wald statistic (8.19). The design-effect estimates $\hat{d}(\hat{b}_k)$ of the estimated model coefficients are larger than one owing to the clustering effect. Thus, binomial standard-error estimates of the model coefficients would be smaller than the corresponding design-based estimates.

Using the estimate $\hat{b}_4 = 0.2568$ for the interesting parameter of the PHYS class of more hazardous work, the corresponding sex−age adjusted odds ratio estimate with its 95% confidence interval can be obtained by (8.7). The odds ratio (OR) estimate is $\exp(\hat{b}_4) = 1.29$, and its 95% confidence interval is calculated as

$$\exp(0.2568 \pm 1.96 \times 0.0574) = (1.16, 1.45).$$

The sex−age adjusted odds of experiencing a higher level of psychic strain is thus 1.3 times higher for persons under more hazardous working conditions than for those in the group of less hazardous work. This result is consistent with the t-test results, because the 95% confidence interval does not include the value one, which is the odds ratio for the reference group.

We next turn to the test results on the model terms in the final main effects ANOVA model (Table 8.5). There is a set of observed values from different Wald test statistics and their F-corrections. Let us consider more closely the tests for the model terms. The first test statistic corresponds to the original design-based Wald statistic (8.13), and the second statistic is the F-corrected statistic (8.18). The third statistic is the Satterthwaite corrected binomial statistic (8.14), and finally, the fourth statistic is the F-corrected statistic (8.20). The design-based Wald statistic $X^2_{des}(\mathbf{b})$ and the second-order corrected binomial statistic $X^2_{bin}(\mathbf{b}; \hat{\delta}_{\bullet}, \hat{a}^2)$ provide similar results. The design-based Wald statistic thus works adequately in this

Table 8.5 Observed values and p-values of test statistics for model terms in the final logit ANOVA model on overall psychic strain (model fitting by the GWLS method).

Contrast	Df	(1) Design-based Wald test	p-value	(2) F-correction to (1)	p-value	(3) Rao−Scott 2nd order adjustment to binomial Wald test	p-value	(4) F-correction to (3)	p-value
SEX	1	64.92	0.0000	64.92	0.0000	64.92	0.0000	64.92	0.0000
AGE	1	5.90	0.0151	5.90	0.0159	5.90	0.0153	5.90	0.0159
PHYS	1	20.04	0.0000	20.04	0.0000	20.04	0.0000	20.04	0.0000

(1) Equation (8.13), (2) Equation (8.18), (3) Equation (8.14), (4) Equation (8.20)

case, which is primarily due to the stability of the covariance-matrix estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$. Because there is a large number of degrees of freedom $f = 245$ for an estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$, the F-corrected tests do not contribute substantially to the p-values of the original tests.

Although there is no controversy about the results from the alternative test statistics in this analysis situation, there can be situations where the choice of an adequate statistic is crucial. This is especially so if the number m of sample clusters is small and the number of domains u is close to m. Then, some of the F-corrected statistics can be chosen to protect against the effects of instability.

For a more detailed examination of the model fit, let us now calculate the fitted proportions and the raw and standardized residuals for a residual analysis. These are displayed in Table 8.6.

The observed and fitted proportions are close, except in the last three domains where the largest raw residuals can be obtained. The standardized residuals in the last two groups exceed the 5% critical value 1.96 from the $N(0,1)$ distribution; so the model fit is somewhat questionable for these domains. It should be noticed that the fitted proportions and the residuals are independent of the parametrization of the model.

It would be useful to consider briefly the logit analysis under the other analysis options as a reference to the results from the design-based option. In this, we are especially interested in the importance of the term SEX*PHYS, describing the interaction of SEX and PHYS, which appeared nonsignificant under the design-based option. The results from the Wald tests are in Table 8.7.

The interaction of SEX and PHYS appears significant when ignoring the clustering effect by using the unweighted SRS option. A more complex model is thus obtained than under the design-based option. These results suggest further warnings on ignoring the clustering effect even if it is not very serious as indicated in the medium-sized domain design-effect estimates.

Table 8.6 Observed and fitted PSYCH proportions $\hat{p}_j$ and $\hat{f}_j$ with their standard errors, and raw and standardized residuals $(\hat{p}_j - \hat{f}_j)$ and $\hat{e}_j$ for the logit ANOVA Model 1 under the design-based option.

Domain	SEX	AGE	PHYS	$\hat{p}_j$	s.e $(\hat{p}_j)$	$\hat{f}_j$	s.e $(\hat{f}_j)$	$(\hat{p}_j - \hat{f}_j)$	$\hat{e}_j$
1	Males	−44	0	0.419	0.0128	0.419	0.0114	0.0000	0.0000
2			1	0.472	0.0145	0.482	0.0122	−0.0100	−1.270
3		45−	0	0.461	0.0178	0.453	0.0142	0.0082	0.771
4			1	0.520	0.0247	0.517	0.0167	0.0029	0.160
5	Females	−44	0	0.541	0.0125	0.534	0.0115	0.0062	1.306
6			1	0.620	0.0270	0.597	0.0160	0.0222	2.012
7		45−	0	0.532	0.0236	0.569	0.0156	−0.0363	−2.073
8			1	0.700	0.0391	0.630	0.0199	0.0692	1.993

Table 8.7 Wald tests $X^2(\mathbf{b})$ for the significance of the interaction term SEX*PHYS in Model 2 under the design-based and unweighted SRS analysis options.

		Design-based		Unweighted SRS	
Term	df	X^2_{des}	p-value	X^2_{bin}	p-value
SEX*PHYS	1	2.39	0.1218	3.97	0.0463

Let us turn to the corresponding design-based analysis with a linear model for the proportions of Table 8.2. In this situation, logit and linear formulations of an ANOVA model lead to similar results because proportions do not deviate much from the value 0.5. The main effects model (Model 1) is chosen, and results on model fit, residuals, and on significance of the model terms, are close to those for the logit model. But the estimates of the model coefficients differ and are subject to different interpretations. For the logit model with the partial parametrization, an estimated coefficient indicates differential effect on a logit scale of the corresponding class from the estimated intercept being the fitted logit for the reference domain. And for the linear model, an estimated coefficient indicates differential effect on a linear scale of the corresponding class from the estimated intercept, which is now the fitted proportion for the reference domain.

The linear model formulation thus involves a more straightforward interpretation of the estimates of the model coefficients. Under Model 1, these estimates are as follows:

$$\begin{aligned}
\hat{b}_1 &= 0.5705 \quad \text{(Intercept)} \\
\hat{b}_2 &= -0.1172 \quad \text{(Differential effect of SEX} = \text{Males)} \\
\hat{b}_3 &= -0.0355 \quad \text{(Differential effect of AGE} = -44) \\
\hat{b}_4 &= 0.0650 \quad \text{(Differential effect of PHYS} = 1).
\end{aligned}$$

The fitted proportion for falling into the upper psychic strain group is thus 0.57 for females in the older age group whose working conditions are less hazardous, and for males in the same age group, $0.57 - 0.12 = 0.45$. The highest fitted proportion, $0.57 + 0.07 = 0.64$, is for the older age group females doing more hazardous work. Also, the fitted proportions are close to those obtained with the corresponding logit ANOVA model.

8.4 LOGISTIC AND LINEAR REGRESSION

The PML method of pseudolikelihood is often used on complex survey data for logit analysis in analysis situations similar to the GWLS method. But the applicability of the PML method is wider, covering not only models on domain proportions of

a binary or polytomous response but also the usual regression-type settings with continuous measurements as the predictors. We consider in this section first a PML analysis on domain proportions and then a more general situation of logit modelling of a binary response with a mixture of continuous measurements and categorical variables as predictors. Finally, an example is given of linear modelling for a continuous response variable in an ANCOVA setting.

In PML estimation of model coefficients and their asymptotic covariance matrix, we use a modification of the maximum likelihood (ML) method. In the ML estimation for simple random samples, we work with unweighted observations and appropriate likelihood equations can be constructed, based on standard distributional assumptions, to obtain the ML estimates of the model coefficients and the corresponding covariance-matrix estimate. Using these estimates, standard likelihood ratio (LR) and binomial-based Wald test statistics can be used for testing the model adequacy and linear hypotheses on the model coefficients.

Under more complex designs involving element weighting and clustering, an ML estimator of the model coefficients and the corresponding covariance-matrix estimator are not consistent and, moreover, the standard test statistics are not asymptotically chi-squared with appropriate degrees of freedom. For consistent estimation of model coefficients, the standard likelihood equations are modified to cover the case of weighted observations. In addition to this, a consistent covariance-matrix estimator of the PML estimators is constructed such that the clustering effects are properly accounted for. Using these consistent estimators, appropriate asymptotically chi-squared test statistics are derived.

The PML method can be conveniently introduced in a setting similar to the GWLS method, assuming again a binary response variable and a set of categorical predictors. The data set is arranged in a multidimensional table, such as Table 8.1, with u domains, and our aim is to model the variation of the domain proportion estimates $\hat{p}_j$ across the domains. The variation is modelled by a logit model of the type given in (8.1) and (8.2). A PML logit analysis for domain proportions, covering logit ANOVA, ANCOVA and regression models with categorical predictors can be carried out under any of the analysis options previously introduced by using the corresponding domain proportion estimator vector and its covariance-matrix estimate, and the steps in model-building are equivalent to those in the GWLS method. The design-based analysis option provides a generally valid PML logit analysis for complex surveys. In practice, a PML logit analysis under the design-based option requires access to specialized software for survey analysis.

Design-based and Binomial PML Methods

Under both design-based and weighted SRS options, a consistent *PML estimator* $\hat{\mathbf{b}}_{pml}$ for the vector $\mathbf{b}$ of the s model coefficients b_k in a logit model $F(\mathbf{p}) = \mathbf{Xb}$ is obtained by iteratively solving the PML estimating equations

$$\mathbf{X}'\mathbf{W}\mathbf{f}(\hat{\mathbf{b}}_{pml}) = \mathbf{X}'\mathbf{W}\hat{\mathbf{p}}, \tag{8.24}$$

where $\mathbf{W}$ is a $u \times u$ diagonal weight matrix with weights $w_j = \hat{n}_j$ on the main diagonal, and $\mathbf{f} = \exp(\mathbf{Xb})/(1 + \exp(\mathbf{Xb}))$ is the inverse function of the logit function. It is essential in (8.24) that the weighted domain sample sizes $\hat{n}_j$ and the weighted proportion estimates $\hat{p}_j$ be used, not their unweighted counterparts n_j and $\hat{p}_j^U$ as in the ML method, i.e. under the unweighted SRS option. This is for consistency of the PML estimators. The corresponding vector (8.5) of the GWLS estimates can be used as an initial value for the PML iterations. Note that under the linear formulation of the ANOVA model, the function vector $\mathbf{f}(\hat{\mathbf{b}}_{pml})$ would be linear in $\hat{b}_k$ and, thus, no iterations are needed. Henceforth, in this section we denote the vector of PML estimates of logit model coefficients by $\hat{\mathbf{b}}$ for short.

Because the vector $\hat{\mathbf{b}}$ of PML estimates is equal under the design-based and weighted SRS options, so also are the vectors $\hat{\mathbf{F}} = \mathbf{X}\hat{\mathbf{b}}$ and $\hat{\mathbf{f}} = F^{-1}(\mathbf{X}\hat{\mathbf{b}})$ of fitted logits and fitted proportions. The equality also holds for estimated odds ratios, which can be obtained as $\exp(\hat{b}_k)$ under the partial parametrization of the model. Fitted proportions $\hat{f}_j = f_j(\hat{\mathbf{b}})$ are estimated under both options by the formula

$$\hat{\mathbf{f}} = \mathbf{f}(\hat{\mathbf{b}}) = \exp(\mathbf{X}\hat{\mathbf{b}})/(1 + \exp(\mathbf{X}\hat{\mathbf{b}})). \tag{8.25}$$

Let us derive under the weighted SRS and design-based options the $s \times s$ covariance-matrix estimators of the PML estimator vector $\hat{\mathbf{b}}$ calculated by (8.24). Assuming simple random sampling, the covariance-matrix estimator is given by

$$\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}}) = (\mathbf{X}'\mathbf{W}\hat{\Delta}\mathbf{W}\mathbf{X})^{-1}, \tag{8.26}$$

where the diagonal elements of the diagonal $u \times u$ matrix $\hat{\Delta}$ are binomial-type variances $\hat{f}_j(1 - \hat{f}_j)/\hat{n}_j$. The binomial covariance-matrix estimator (8.26) is not consistent for complex sampling designs involving clustering. For these designs, we derive a more complicated consistent covariance-matrix estimator that is valid under the design-based option:

$$\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}}) = \hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}})\mathbf{X}'\mathbf{W}\hat{\mathbf{V}}_{des}\mathbf{W}\mathbf{X}\hat{\mathbf{V}}_{bin}(\hat{\mathbf{b}}). \tag{8.27}$$

This estimator is of a 'sandwich' form such that the design-based covariance-matrix estimator $\hat{\mathbf{V}}_{des}$ of the proportion vector $\hat{\mathbf{p}}$ acts as the 'filling'.

Approximate confidence intervals for odds ratio estimates $\exp(b_k)$ under the design-based and weighted SRS options can be calculated by (8.7) using the corresponding variance estimates $\hat{v}_{des}(\hat{b}_k)$ and $\hat{v}_{bin}(\hat{b}_k)$ of the PML estimates $\hat{b}_k$, as in the GWLS method. Also, the design-effect estimates $\hat{d}(\hat{b}_k)$ of the model coefficients $\hat{b}_k$ can be obtained by (8.23), again analogously to the GWLS method.

Expressions for the consistent covariance-matrix estimators $\hat{\mathbf{V}}_{des}(\hat{\mathbf{F}})$ and $\hat{\mathbf{V}}_{des}(\hat{\mathbf{f}})$ of the vector $\hat{\mathbf{F}}$ of fitted logits and the vector $\hat{\mathbf{f}}$ of fitted proportions are similar under the design-based option to those of the GWLS method, as given in equations (8.8) and (8.9). The PML analogue $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ from (8.27) and the corresponding

matrix $\hat{\mathbf{H}}$ must of course be used in the equations. And under the weighted SRS option, the covariance-matrix estimators $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{F}})$ and $\hat{\mathbf{V}}_{bin}(\hat{\mathbf{f}})$ are derived similarly by using the binomial estimator (8.26) in the equations in place of its design-based counterpart.

A residual covariance-matrix estimator is needed for conducting a proper residual analysis under the design-based option. This $u \times u$ estimator is given by

$$\hat{\mathbf{V}}_{res} = \mathbf{A}\hat{\mathbf{V}}_{des}\mathbf{A}', \tag{8.28}$$

where the matrix $\mathbf{A}$ is obtained by the formula

$$\mathbf{A} = \mathbf{I} - \hat{\Delta}\mathbf{W}\mathbf{X}(\mathbf{X}'\mathbf{W}\hat{\Delta}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$$

with $\mathbf{I}$ being a $u \times u$ identity matrix. Using this estimate, design-based standardized residuals of the form (8.22) can then be calculated.

There are thus many similarities between the PML formulae and those derived for the GWLS method. The main differences lie in the way the estimates of model coefficients and their covariance-matrix estimate are calculated. More similarities are evident in the testing procedures. All the test statistics derived for the GWLS method are also applicable to the PML method.

Under the design-based option, goodness of fit of the model can be tested with the design-based Wald statistic X^2_{des} given by (8.11). When examining the model fit more closely, PML analogues to the Wald statistics $X^2_{des}(overall)$ and $X^2_{des}(gof)$ can be used. The Wald statistics (8.13) and (8.14) for linear hypotheses on model parameters are applicable as well. Finally, in unstable situations, the F-corrected Wald and Rao–Scott statistics (8.16)–(8.20) can be used. It should be noted that the PML estimates from (8.24) and the corresponding covariance-matrix estimate (8.27) must be used in the calculation of these test statistics under the design-based option. These test statistics are available in commonly used software products for logit analysis for complex survey data.

In testing procedures for the weighted and unweighted SRS options, the corresponding binomial covariance-matrix estimates are used in the test statistics in place of those from the design-based option. As an alternative to the Wald statistics, LR test statistics can be used, which for the design-based option should be adjusted using the Rao–Scott methodology. A second-order adjustment to LR test statistics similar to (8.14) for the binomial-based Wald statistic provides asymptotically chi-squared test statistics. The residual covariance-matrix estimate (8.28) can be used in deriving an appropriate generalized design-effects matrix estimate for the adjustments.

The main application area of the PML method for complex surveys is under the design-based option, and the weighted and unweighted SRS options are used as the reference when examining the effects of weighting and intra-cluster correlation on standard-error estimates of model coefficients and on p-values of Wald test statistics.

Logistic Regression

The PML method can also be used in strictly regression-type logit analyses on a binary response variable from a complex survey, where the predictors are continuous measurements. In logistic regression, we work with an element-level data set without aggregating these data into a multidimensional table. So, the measured values of the continuous predictor variables constitute the columns in an $n \times s$ model matrix $\mathbf{X}$ for a logistic regression model. But all the other elements of the PML estimation remain unchanged, and consistent PML estimates with their consistent covariance-matrix estimate are obtained in a way similar to that described for the design-based analysis option. Moreover, a logistic ANCOVA can be performed by incorporating categorical predictors into the logistic regression model. Then, interaction terms of the continuous and categorical predictors can also be included.

A logistic regression model is usually built by entering predictors into the model using subject-matter criteria or significance measures of potential predictors. In this, t-tests $\mathrm{t}_{des}(b_k)$, or the corresponding Wald tests $X^2_{des}(b_k)$, on model coefficients can be used as previously and, under the design-based option, asymptotic properties of these test statistics remain unchanged.

Instability of an estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ from (8.27) can destroy the distributional properties of the test statistics on model coefficients in such small-sample situations where the number of sample clusters is small. Usual degrees-of-freedom, F-corrections to the Wald and t-test statistics can then be used.

The GEE methodology of generalized estimating equations can also be used for logistic modelling on complex survey data. In this method, the model coefficients are estimated using the multivariate quasilikelihood technique, and intra-cluster correlations are taken as nuisances. Using an estimated intra-cluster correlation structure, a 'robust' estimator of the covariance matrix of the model coefficients can be obtained, basically similar to the 'sandwich' form in the PML method. Thus, the GEE method can be used to account for the clustering effects. We describe only briefly the method and give an example for logistic ANCOVA in the OHC Survey.

The GEE method was originally developed for accounting for the possible correlation of observations in fitting generalized linear models in the context of longitudinal surveys (Liang and Zeger 1986). The methodology has been further described and illustrated in Liang *et al.* (1992) and Diggle *et al.* (2002).

Two alternatives of the GEE method have been presented. A preliminary GEE method with an independent correlation assumption relates to the standard PML method where observations are assumed independent within clusters for the estimation of the regression coefficients, but are allowed to be correlated for the estimation of the covariance matrix of the estimated regression coefficients. In covariance-matrix estimation, a 'sandwich' form of estimator is used. In a more advanced GEE method, assuming an exchangeable correlation structure, observations are allowed to be correlated within clusters in the estimation of both

regression coefficients and the covariance matrix of estimated regression coefficients. There, a 'working' intra-cluster correlation is estimated and incorporated in the estimation procedure of regression coefficients and the covariance matrix of estimated coefficients.

A generalized linear model can be compactly written as

$$\mathrm{E_M}(\mathrm{g}(\mathbf{y})) = \mathbf{Xb}, \tag{8.29}$$

where $\mathrm{E_M}$ refers to the expectation under the model and the function g refers to the so-called link function postulating a relationship between the expectation of the response variable vector $\mathbf{y}$ and the linear part $\mathbf{Xb}$ of the model. Special cases of link functions are identity, logistic and logarithmic functions used in linear models for continuous responses, logistic models for binary responses and log-linear models for count data, respectively.

The covariance structure of observations within clusters is modelled by

$$\mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{R}(\alpha) \mathbf{A}_i^{1/2}, \qquad i = 1, \ldots, m, \tag{8.30}$$

where $\mathbf{A}_i$ is a diagonal matrix of variances $V(y_k)$ in cluster i and $\mathbf{R}(\alpha)$ is the 'working' correlation matrix specified by the (possibly vector-valued) correlation parameter α of observations in cluster i. The parameter ϕ denotes the dispersion parameter of the corresponding member of the exponential family of distributions. Under an independent correlation assumption, all off-diagonal elements α of the 'working' correlation matrix are set to zero. Under an exchangeable correlation of pairs of observations within a cluster, the parameter α is a scalar and requires estimation. In an estimation procedure to obtain an estimate $\hat{\mathbf{b}}$, Newton–Raphson-type algorithms are usually used. The covariance-matrix estimate $\hat{\mathbf{V}}_{des}(\hat{\mathbf{b}})$ is obtained using a 'sandwich' type estimator (see equation (8.27)). Element weights can be incorporated in a GEE estimation procedure. GEE and the weighted analogue can be applied using suitable software for the analysis of complex surveys.

The GEE method has been shown to produce consistent estimates of model parameters and their covariance matrices, independently of a correct specification of the 'working' correlation structure. In the next two examples, we apply logistic ANCOVA first with the PML method and then with the GEE method assuming an exchangeable intra-cluster correlation structure. For further training on the PML and GEE methods in logistic modelling on the OHC Survey data, the reader is advised to visit the web extension of the book.

Example 8.2

Logistic ANCOVA with the PML method. Let us consider in a slightly more general setting the analysis situation of Example 8.1, where a logit ANOVA model was fitted by the GWLS method to proportions in a multidimensional table. We now

fit a logistic ANCOVA model using the PML method, by entering some of the predictors as continuous measurements in the model. The design-based analysis option is applied, providing valid PML analysis.

The binary response variable PSYCH measures high psychic strain, and we take the variables AGE, PHYS (physical working conditions) and CHRON (chronic morbidity) as continuous predictors such that AGE is measured in years and PHYS and CHRON are binary. Thus there are four predictors, of which SEX is taken as a qualitative predictor. So, the interaction of SEX with AGE, PHYS and CHRON can also be examined.

A model with SEX, AGE, PHYS and CHRON as the main effects and an interaction term of SEX and AGE was taken as the final model, because the other interactions appeared nonsignificant at the 5% level. Results of the model coefficients are displayed in Table 8.8.

The fitted logit ANCOVA model can be written using the estimated coefficients $\hat{b}_k$ and the corresponding model matrix $\mathbf{X}$ similar to the ANOVA modelling in Example 8.1:

$$F(\hat{f}_l) = \hat{b}_1 + \hat{b}_2(\text{SEX})_l + \hat{b}_3(\text{AGE})_l + \hat{b}_4(\text{PHYS})_l + \hat{b}_5(\text{CHRON})_l + \hat{b}_6(\text{SEX} * \text{AGE})_l,$$

where $l = 1, \ldots, 7841$, and $F(\hat{f}_l) = \log(\hat{f}_l/(1-\hat{f}_l))$. The values for the model terms are obtained from the corresponding columns of the 7841×6 model matrix $\mathbf{X}$. There, SEX, PHYS and CHRON are binary, and AGE has its original values (age

Table 8.8 Design-based logistic ANCOVA on overall psychic strain with the PML method.

Model term	Beta coefficient	Design effect	Standard error	*t*-test	*p*-value	Odds ratio	95% confidence interval for OR	
							Lower	Upper
Intercept	0.1964	1.56	0.1572	1.25	0.2127	1.22	0.89	1.66
Sex								
Males	−0.9926	1.43	0.2033	−4.88	0.0000	0.37	0.25	0.55
Females*	0	n.a.	0	n.a.	n.a.	1	1	1
Age	−0.0046	1.55	0.0041	−1.12	0.2624	1.00	0.99	1.00
Physical health hazards	0.2765	1.39	0.0596	4.64	0.0000	1.32	1.17	1.48
Chronic morbidity	0.5641	1.17	0.0575	9.82	0.0000	1.76	1.57	1.97
Sex, Age								
Males	0.0131	1.41	0.0051	2.56	0.0111	1.01	1.00	1.02
Females*	0	n.a.	0	n.a.	n.a.	1	1	1

* Reference class; parameter value set to zero.
n.a. not available.

in years). Note the difference in the ANCOVA model matrix when compared with that for the ANOVA model.

The *t*-tests on model coefficients indicate that the coefficients for the interesting predictors, physical working conditions and chronic morbidity are strongly associated with experiencing psychic strain. Persons in hazardous work, and chronically ill persons are more likely to suffer from psychic strain than healthy persons and persons whose working conditions are less hazardous. Note that the sex−age adjusted coefficient $\hat{b}_5$ for CHRON is larger than $\hat{b}_4$ for PHYS. Thus, in the model, chronic morbidity is more important as a predictor of psychic strain. This can also be seen in the odds ratio (OR) estimates provided in Table 8.8.

Odds ratios with their approximative 95% confidence intervals (in parenthesis) thus are

$$\text{PHYS: Odds ratio} = \exp(0.2765) = 1.32 \quad (1.17,\ 1.48),$$

$$\text{CHRON: Odds ratio} = \exp(0.5641) = 1.76 \quad (1.57,\ 1.97).$$

We may thus conclude that odds for experiencing a higher level of psychic strain, adjusted for sex, age and chronic morbidity, is about 1.3 times higher for those in more hazardous work than for those in less hazardous work. This conclusion was similar in Example 8.1, where a closely related odds ratio and confidence interval were obtained. Furthermore, the odds of experiencing much psychic strain, adjusted for sex, age and working conditions, are about 1.8 times higher for chronically ill persons than for healthier persons. Because neither of the 95% confidence intervals covers the value one, the corresponding odds ratios differ significantly (at the 5% level) from one. It should be noted that the binomial-based confidence intervals would be narrower especially for the predictor PHYS, for which the design-effect estimate is larger than for CHRON.

An analysis under the SRS options yield the same final model as the design-based analysis, but the observed values of the test statistics are somewhat larger and thus more liberal test results are attained.

Finally, let us examine more closely the fitted proportions $\hat{f}_l$ for the upper psychic strain group under the present model. The results are summarized in Figure 8.2 by plotting the proportions against the predictors included in the model. Fitted proportions increase with increasing age for males, and decrease for females. At a given age, the proportions are larger for the chronically ill and for those in more hazardous work than in the reference groups. Also, in females the fitted proportions tend to be larger than in males in all the corresponding domains, although the differences decline with increasing age.

Example 8.3

Logistic ANCOVA with the GEE method. Let us consider further the analysis situation of Example 8.2, where a logistic ANCOVA model was fitted by the PML method. We now fit a logistic ANCOVA model using the GEE method with

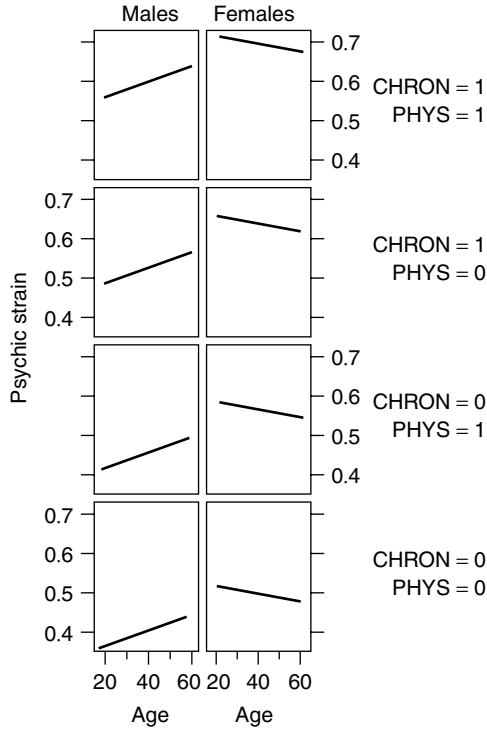

Figure 8.2 Fitted proportions of falling into the high psychic strain group for the final logistic ANCOVA model.

an assumed exchangeable correlation of pairs of observations within a cluster. Similarly as in Example 8.2, our response variable is the binary PSYCH measuring psychic strain. The variable SEX is included in the model as a categorical predictor and AGE, PHYS (physical working conditions) and CHRON (chronic morbidity) as continuous predictors such that AGE is measured in years and PHYS and CHRON are binary. We fit the same model as in Example 8.2.

Results are shown in Table 8.9. A comparison with logistic ANCOVA with the PML method in Example 8.2 indicates that the results are quite similar, and our inferential conclusions remain the same. There are, however, certain differences. First, the estimated beta coefficients have changed. Absolute values of estimates are larger than in the PML application, except for the CHRON effect. Standard-error estimates are somewhat smaller than the PML counterparts. Hence, the observed t-statistics tend to be larger involving slightly more liberal tests than in the PML case. These differences are due to the fact that in the GEE method with an exchangeable correlation structure, the correlation of observations also contributes to the estimation of the beta parameters. The 'working' intra-cluster

Table 8.9 Design-based logistic ANCOVA on overall psychic strain with the GEE method under exchangeable intra-cluster correlation structure.

Model Term	Beta coefficient	Design effect	Standard error	*t*-test	*p*-value
Intercept	0.2292	1.44	0.1524	1.50	0.1338
Sex					
Males	−1.0290	1.36	0.2000	−5.14	0.0000
Females*	0	n.a.	0	n.a.	n.a.
Age	−0.0057	1.43	0.0039	−1.45	0.1489
Physical health hazards	0.3011	1.31	0.0587	5.13	0.0000
Chronic morbidity	0.5569	1.14	0.0568	9.81	0.0000
Sex, Age					
Males	0.0144	1.33	0.0050	2.88	0.0044
Females*	0	n.a.	0	n.a.	n.a.

* Reference class; parameter value set to zero.
n.a. not available.

correlation is estimated as $\hat{\alpha} = 0.0189$. Using the expression $\text{deff} = 1 + (\overline{m} - 1)\hat{\alpha}$, where $\overline{m}$ is the average cluster size, this corresponds to an average design effect of 1.57.

Linear Modelling on Continuous Responses

We have extensively considered the modelling of binary response variables from complex surveys. The GWLS, PML and GEE methods were used, covering logit and linear modelling on categorical data and logit modelling with continuous predictors. These types of multivariate models are most frequently found in analytical surveys, for example, in social and health sciences. But in some instances it is appropriate to model a quantitative or continuous response variable, such as the number of physician visits or blood pressure. We discuss briefly the special features of multivariate analysis in such cases, and give an illustrative example of a special case of linear ANCOVA.

Linear modelling provides a convenient analysis methodology for analysis situations with a continuous response variable and a set of predictors. This situation was present in Examples 8.2 and 8.3, where the dichotomized PSYCH was analysed with a logistic ANCOVA model. There the original continuous variable on psychic strain could be taken as the response variable as well, leading to linear ANCOVA modelling. For a simple random sample, the analysis would be based on ordinary least squares (OLS) estimation with a standard program for linear modelling. For the OHC Survey data set, which is based on cluster sampling, the design-based approach with weighted least squares (WLS) estimation provides proper linear modelling.