■ 연구논문

Logistic Regression for Retrospective Studies

Mi-Young Shin
Dept. of Mathematics, Song Sim University

Abstract

We consider logistic models based on retrospective, case-control data with stratified samples and study the Weighted Exogeneous Sampling Maximum Likelihood (WESML). We develop a consistent estimator of the asymptotic covariance matrix of the WESML estimator.

1. Introduction

The logistic regression model has been applied to two distinct sampling designs: prospective sampling and retrospective sampling. In prospective sampling, a sample of subjects is chosen from the population of interest and the values of the response and explanatory variables are determined. In a retrospective case-control study, separate samples of fixed size are taken from cases ($Y=1$) and controls ($Y=0$) respectively, then the values of the explanatory variables are measured for each subject selected. For practical reasons, sampling is often stratified by factors such as age, gender or location.

Since maximum likelihood methods may be difficult to implement for complex sampling designs, there has been research on adapting standard technique for fitting prospective logistic models to complex sampling such as stratified case control sampled data. Prentice and Pyke (1979) have shown that valid estimates of the odds-ratio parameters β and their asymptotic covariance in a logistic regression model may be obtained from simple case-control data by fitting the model as if the data had been obtained in a prospective sampling. That is, the sampling scheme can be ignored and the model fitted using a standard logistic regression program as if the data were prospective. Manski and Lerman (1977) proposed the Weighted Exogeneous Maximum Likelihood (WESML) estimator which is based on the weighted log-likelihood function for hypothetical prospectively sampled data. Fears and Brown (1986) developed another procedure

for logistic regression analysis of stratified case-control data. They assumed known sampling rates as well as known total populations of each stratum. They fitted the usual prospective model to the case-control data, treating case-control status as a binary outcome variable. In order to adjust for case-control sampling, they included the logarithm of the sampling fraction in the regression equation. Breslow and Zhao (1988) showed that the estimation procedure suggested by Fears and Brown was equivalent to the conditional maximum Likelihood (CML) estimator developed by Manski and McFadden (1981).

In this paper we consider stratified case-control sampling and study the weighted exogeneous sampling maximum likelihood estimate, which was introduced for simple choice-based sampling by Manski and Lerman (1977). We develop a consistent estimator of the asymptotic covariance matrix of the WESML estimator of the logistic regression coefficient. We adapt White's consistent covariance estimation (1980) for estimating the standard error in logistic regression model under stratified case-control sampling.

2. Notation and Assumptions

Suppose we have a finite known population of N_{1g} cases and N_{0g} controls in each stratum $g=1, \ldots, G$; a $(p\times 1)$ vector of stratum explanatory variables $\underline{Z}_1$; a $(q\times 1)$ vector of discrete within-stratum explanatory variables $\underline{Z}_2$.

Suppose we have sample of size n_{1g} are selected from the cases and n_{0g} from the controls independently. n_g and N_g denote the total sample and population size of stratum g respectively.

The logistic regression model to be fitted has the following form:

$$Pr(Y=1 \mid \underline{z}_{2j}, \text{stratum}=g) = \exp(\underline{z}_{1g}'\underline{\alpha}+\underline{z}_{2j}'\underline{\beta})/\{1+\exp(\underline{z}_{1g}'\underline{\alpha}+\underline{z}_{2j}'\underline{\beta})\}$$
$$= \exp(\underline{z}_{gj}'\underline{\theta})/\{1+\exp(\underline{z}_{gj}'\underline{\theta})\}$$

, where $\underline{\theta}=(\underline{\alpha}', \underline{\beta}')'$ and $\underline{Z}_{gj}=(\underline{Z}_{1g}', \underline{Z}_{2j}')$.

Let $p_{igj}=Pr(Y=i \mid \underline{z}_{2j}, \text{stratum}=g)$, $\mu_{igj}=Pr(Y=i, \underline{z}_2=\underline{z}_{2j}, \text{stratum}=g)$ and $\mu_{ig}=Pr(Y=i, \text{stratum}=g)$. Let p^*_{igj} and μ^*_{igj} be p_{igj} and μ_{igj} evaluated at the true parameter $\underline{\theta}^*$ respectively.

The following assumptions will be maintained throughout this paper.

Assumption 1: There exist constants $\rho_{ig}>$ and $\zeta_g>0$ such that $\lim_{n_g\to\infty}\frac{n_{ig}}{n_g}=\rho_{ig}$,

and $\lim_{n\to\infty} \frac{n_g}{n} = \zeta_g$ $(i=0, 1; g=1, .., G)$.

Assumption 2: $n_g / N_g \to v_g$ and $n/N \to v$ for some positive v_g and v.

Assumption 3: $\underline{C} = \sum_g \sum_j \frac{\zeta_g}{\mu_g} (\mu^*_{0gj} + \mu^*_{1gj})\, p^*_{1gj}\, p^*_{0gj}\, \underline{z}_{gj}\, z_{gj}'$ exists and nonsingular.

3. WESML Estimator

If the data had been obtained by stratified prospective sampling with n_g observation per stratum, the cell frequencies n_{1gj} would have independent binomial distributions with parameters n_{+gj} and p_{1gj} for $g=1, \ldots, G$; $j=1, \ldots, r$. The joint log-likelihood function for the stratified prospective sampling could be written

$$\log L \propto \sum_g \sum_j n_{1gj} \underline{z}_{gj}' \underline{\theta} - \sum_g \sum_j n_{+gj} \log\{1+\exp(\underline{z}_{gj}' \underline{\theta})\}.$$

The estimate which maximizes this log-likelihood function would be consistent under prospective sampling. In general this estimate is known to be inconsistent in the stratified case-control sampling situation.

To obtain a consistent estimate we use a weighted log-likelihood function which was introduced by Manski and Lermann (1977) for simple choice-based sample. The weights which restore the expected proportions of cases and controls under stratified prospective sampling is $w_g(i) = (N_{ig}\ n_g)/(N_g\ n_{ig})$. Thus, the estimated prospective cell frequencies are $\tilde{n}_{igj} = n_{igj}\, w_g(i)$.

The WESML estimator, $\hat{\underline{\theta}}_{\text{wesml}}$ maximizes the weighted prospective log likelihood function

$$\frac{1}{n} \sum_i \sum_g \sum_j w_g(i)\, n_{igj} \log p_{igj}.$$

That is, $\hat{\underline{\theta}}_{\text{wesml}}$ is a root of the equation

$$\sum_g \sum_j \underline{z}_{gj} (\tilde{n}_{1gj} - \tilde{n}_{+gj}\, p_{1gj}) = \underline{0}. \tag{1}$$

By the generalization of Manski and Lermann (1977) results we get the following theorem.

Theorem 1. The distribution of $\sqrt{n}(\hat{\underline{\theta}}_{\text{wesml}} - \underline{\theta})$ is asymptotically normal with mean zero and covariance matrix

$$\begin{aligned}\underline{\Delta}_{\text{wesml}} &= \underline{C}^{-1}\,\underline{\Lambda}\,\underline{C}^{-1} \\ &= \underline{C}^{-1}\{\sum_i\sum_g (\mu_{ig}/\mu_g)^2\,(\zeta_g/\rho_{ig})\,\underline{a}_{ig}\,\underline{\Omega}_{ig}\,\underline{a}_{ig}'\}\,\underline{C}^{-1},\end{aligned}$$

where $\underline{a}_{1g} = (p_{0g1}\,\underline{z}_{g1}, \ldots\ldots, p_{0gr}\,\underline{z}_{gr})$, $\underline{v}_{1g} = (\mu_{1g1}/\mu_{1g}, \ldots\ldots, \mu_{1gr}/\mu_{1g})$,
$\underline{\Omega}_{1g} = \text{diag}(\mu_{1g1}/\mu_{1g}, \ldots, \mu_{1gr}/\mu_{1g}) - \underline{v}_{1g}'\,\underline{v}_{1g}$.
$\underline{a}_{0g}$, $\underline{v}_{0g}$ and $\underline{\Omega}_{0g}$ are analogously defined.

Proof. The proof is given in Appendix.

4. The Consistent Covariance Estimate

A significant advantage of the WESML estimator is its computational simplicity. Existing logistic regression software is easily modified to yield the WESML estimate and its asymptotic variance matrix. The problem with WESML is that the standard errors of the regression coefficients printed by logistic regression software would be incorrect. If it were possible to replace $\underline{\Delta}_{\text{wesml}}$ with a consistent estimator, the usual asymptotic tests could be performed. It is natural to estimate $\underline{C}$ by its consistent estimator

$$\underline{C}_n = \frac{1}{n}\sum_g\sum_j \tilde{n}_{+gj}\,\hat{p}_{1gj}\,\hat{p}_{0gj}\,\underline{z}_{gj}\,\underline{z}_{gj}',$$

where $\hat{p}_{igj}$ is fitted probability of being $Y = i$.
The difficulty arises in estimating $\underline{\Lambda}$. In the proof of theorem 1, we derived the asymptotic variance $\underline{\Lambda}$ from following equation

$$\frac{1}{\sqrt{n}}\sum_k\sum_j(\tilde{n}_{1gj} - \tilde{n}_{+gj}\,p_{1gj})\,\underline{z}_{gj} = \frac{1}{\sqrt{n}}\sum_i\sum_g\sum_j w_g(i)\,n_{igj}\,(i - p_{1gj})\,\underline{z}_{gj}.$$

This equation can be written in disaggregated form as

$$\frac{1}{\sqrt{n}}\sum_i\sum_g\sum_{k=1}^{n_{ig}} w_g(i)\,(i - p_{1gk})\,\underline{z}_{igk} \qquad (2)$$

where $\underline{z}_{igk}$ denotes the $(p+q)\times 1$ vector of design variables for the kth of the n_{ig} subjects with $Y=i$ in stratum g.
For notational simplicity let us define

$$\hat{\underline{z}}_{igk} = w_g(i)(i-\hat{p}_{1gk})\underline{z}_{igk}, \text{ and } \bar{\hat{\underline{z}}}_{ig} = \sum_{k=1}^{n_{ig}} \frac{z_{igk}}{n_{ig}}.$$

Then equation (2) can be written as $\frac{1}{\sqrt{n}} \sum_i \sum_g \sum_{k=1}^{n_{ig}} \hat{z}_{igk}$.
The variance of equation (2) would naturally be estimated by

$$E_n = \frac{1}{n} \sum_i \sum_g \sum_{k=1}^{n_{ig}} (\hat{\underline{z}}_{igk} - \bar{\hat{\underline{z}}}_{ig})(\hat{\underline{z}}_{igk} - \bar{\hat{\underline{z}}}_{ig})'.$$

This leads us to consider the following consistent covariance matrix estimate.

Theorem 2. A consistent estimate of the asymptotic covariance matrix $\underline{\Delta}_{\text{wesml}}$ under stratified case-control sampling is $\hat{\underline{\Delta}}_{\text{wesml}} = \underline{C}_n^{-1} \underline{E}_n \underline{C}_n^{-1}$.

Proof. The proof is given in Appendix.

5. Simulation Study and Conclusions

A small scale simulation study was carried out in order to evaluate the accuracy of the proposed consistent covariance matrix estimator in theorem 2. We use the same setup as in Breslow and Zhao's simulation (1988).

The range of the stratum variable z_1 is (1, 2, 3) and the explanatory variable z_2 is binary. The true parameter value is set at $(\alpha_0, \alpha_1, \beta) = (-4, 1, 1)$ and the ratio of population frequencies N_{1g}/N_{0g} is set to (0.1, 0.2, 0.5) for $g=1, 2, 3$. Balanced case and control samples of size $n_{ig}=25$, 50 and 100 are drawn from each of the three strata and binomial observations of the number with $z_2=1$ out of n_{ig} are generated for each strata using $p_{ig}=(z_2=1 \mid Y=i, \text{stratum}=g)$ which is calculated later. Each experiment is replicated 1,000 times.
The model to be fitted is

$$p(Y=1 \mid z_1, \text{stratum}=g) = \{1+\exp(-\alpha_0 - \alpha_1 z_1 - \beta z_2)\}^{-1}.$$

Let N_{igj} be the numbe of population with $Y=i$ and $Z_2=j$ in the gth stratum. Then $p_{ig}=N_{ig1}/(N_{ig0}+N_{ig1})$ and $q_{ig}=1-p_{ig}$.

Using the relations $N_{1g0}=N_{0g0}\exp(\alpha_0+\alpha_1 Z_{1g})$ and $N_{1g1}=N_{0g1}\exp(\alpha_0+\alpha_1 Z_{1g}+\beta)$, we obtain

$$N_{1g}/N_{0g} = \exp(\alpha_0+\alpha_1 Z_{1g})(q_{og}+e^{\beta}p_{0g}) \quad (3$$

From equation (3) we can find p_{0g} and p_{1g}.

〈Table 1〉, 〈Table 2〉 and 〈Table 3〉 show the the average values of the WESML estimates over the 1,000 replicates($\hat{\underline{\theta}}_{\text{wesml}}$), the empirical standard deviation of the WESML estimates (Empirical s.d.), and the average of the estimated standard deviation using $\hat{\underline{\Delta}}_{\text{wesml}}$ over the 1,000 replicates (average of estimated s.d.).

Even with moderate sample sizes n_{ig} ($i=0, 1$; $g=1, 2, 3$) the empirical average values of the WESML estimates are close to the true parameter (-4, 1, 1) and the estimators approach the true parameter as n_{ig} increase. In 〈Table 1〉 (total sample size=150), the average values of the WESML estimate are (-4.042, 1.014, 1.025). In 〈Table 3〉 (total sample size=600), we have the average values of the WESML estimates (-4.013, 1.004, 1.010). WESML estimates are different from the true parameter by at most two digits in the third decimal place. The standard deviation estimates for WESML are very consistent with the empirical standard deviation.

〈 Table 1 〉 Simulation results with n_{ig} = 25

parameter	$\hat{\underline{\theta}}_{\text{wesml}}$	Empirical s.d.	average of estimated s.d.
α_0	-4.042	0.3944	0.3823
α_1	1.014	0.1099	0.1035
β	1.025	0.4024	0.3931

〈 Table 2 〉 Simulation results with n_{ig} = 50

parameter	$\hat{\underline{\theta}}_{\text{wesml}}$	Empirical s.d.	average of estimated s.d.
α_0	-4.023	0.2655	0.2676
α_1	1.008	0.0747	0.0724
β	1.011	0.2704	0.2770

〈 Table 3 〉 Simulation results with $n_{ig} = 100$

parameter	$\hat{\theta}_{wesml}$	Empirical s.d.	average of estimated s.d.
α_0	−4.013	018634	0.1883
α_1	1.004	0.0502	0.0505
β	1.010	0.1919	0.1951

The 〈Table 4〉 shows the percents of the WESML estimates within 1, 2, and 3 standard deviation of the average. Comparing the percents for WESML estimates with percents for the normal distribution we conclude that the WESML estimates well approximated by a normal distribution for moderate sample sizes.

〈 Table 4 〉 Normal counts percents

	normal	α_0	α_1	β
ave ± 1s.d.	68.26%	71.4%	72.6%	70.4%
ave ± 2s.d.	95.45%	96.2%	96.2%	96.3%
ave ± 3s.d.	99.73%	99.6%	99.5%	99.8%

References

[1] Breslow, N. E. and Zhao, L. P. (1988), "Logistic regression for stratified case control studies," *Biometrics*, Vol. 44, pp. 891−899.

[2] Fears, T. R. and Brown, C. C. (1986), "Logistic regression methods for retrospective case control studies using complex sampling procedures," *Biometrics*, Vol. 42, pp. 955−960.

[3] Manski, C. F. and Lerman, S. (1977), "The estimation of choice probabilities from choice based samples," *Econometrica*, Vol. 45, pp. 1977−1989.

[4] Manski, C. F. and McFadden, D. (1981), *Structural analysis of discrete data with econometric applications* (Cambridge, Massachusetts: The MIT Press), pp. 1−49.

[5] Prentice, R. L. and Pyke, R. (1980), "Logistic disease incidence models and case control studies," *Biometrika*, Vol. 66, pp. 403−411.

[6] White, H. (1980), "A heteroskedasticity consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, Vol. 48, pp. 817−838.

APPENDIX

Lemma 1. Under case-control sampling, $\sum_j \frac{p_{0gj}\, n_{1gj}}{\sqrt{n_{1g}}}\, \underline{z}_{gj}$ is asymptotically normal with mean $\sqrt{n_{1g}}\ \underline{a}_{1g}\ \underline{v}_{1g}'$ and covariance matrix $\underline{a}_{1g}\ \underline{\Omega}_{1g}\ \underline{a}_{1g}'$.

Proof: Under case-control sampling, $(n_{1g1}, \ldots, n_{1gr})$ has a multinomial distribution with parameters n_{1g} and $\underline{v}_{1g}$ for $g=1, \ldots, G$. The multivariate Lindeverg-Levy Central Limit Theorem implies that $\frac{1}{\sqrt{n_{1g}}}(n_{1g1}, \ldots, n_{1gr})$ is asymptotically normal with mean $\sqrt{n_{1g}}\ \underline{v}_{1g}$ and covariance matrix $\underline{\Omega}_{1g}$. Q.E.D.

Similarly $\sum_j \frac{p_{1gj}\, n_{0gj}}{\sqrt{n_{0g}}}\, \underline{z}_{gj}$ is asymptotically normal with mean $\sqrt{n_{0g}}\ \underline{a}_{0g}\ \underline{v}_{0g}'$ and covariance matrix $\underline{a}_{0g}\ \underline{\Omega}_{0g}\ \underline{a}_{0g}'$.

Lemma 2. $\frac{1}{\sqrt{n}}\sum_g\sum_j (\tilde{n}_{1gj} - \tilde{n}_{+gj}\, p_{1gj})\, \underline{z}_{gj}$ converges in distribution to the normal distribution with mean zero and covariance matrix $\underline{\Lambda}$.

Proof: The proof can be done by combining Assumption 1 and Lemma 1 with Slutsky's theorem. Q.E.D.

The proof of theorem 1.

A first-order Taylor expansion of $\hat{p}_{1gj}$ about the true value $\underline{\theta}^*$ gives

$$\hat{p}_{1gj} = p_{1gj} + \underline{z}_{gj}'\,(\hat{\underline{\theta}}_{\text{wesml}} - \underline{\theta}^*)\, p^{\circ}_{1gj}\, p^{\circ}_{0gj}, \tag{4}$$

where p°_{1gj} is p_{1gj} evaluated at $\underline{\theta}^{\circ}$ which is a point between $\underline{\theta}^*$ and $\hat{\underline{\theta}}_{\text{wesml}}$. Substituting $\hat{p}_{1gj}$ from (4) into equation (1), we obtain

$$\sum_g\sum_j \underline{z}_{gj}\,\{(\tilde{n}_{1gj} - \tilde{n}_{+gj}\, p_{1gj}) - \tilde{n}_{+gj}\, \underline{z}_{gj}'\, p^{\circ}_{1gj}\, p^{\circ}_{0gj}\,(\hat{\underline{\theta}}_{\text{wesml}} - \underline{\theta}^*)\} = \underline{0}$$

Hence

$$\{\sum_g\sum_j \frac{1}{n}\,\tilde{n}_{+gj}\ \underline{z}_{gj}\ \underline{z}_{gj}'\ p^{\circ}_{1gj}\, p^{\circ}_{0gj}\}\sqrt{n}\,(\hat{\underline{\theta}}_{\text{wesml}} - \underline{\theta}^*)$$
$$= \sum_g\sum_j \frac{1}{n}(\tilde{n}_{1gj} - \tilde{n}_{+gj}\, p_{1gj})\, \underline{z}_{gj}. \tag{5}$$

Since the right side of equation (5) converges in distribution to $N(\underline{0}, \underline{\Lambda})$ by Lemma 2, we need to show that $\sum_g \sum_j \frac{1}{n} \{\tilde{n}_{+gj}\, \underline{z}_{gj}\, \underline{z}_{gj}'\, \hat{p}^{\circ}_{1gj}\, \hat{p}^{\circ}_{0gj}\}$ converges in probability to $\underline{C}$. Now

$$\sum_g \sum_j \frac{1}{n} \{\tilde{n}_{+gj}\, \underline{z}_{gj}\, \underline{z}_{gj}'\, \hat{p}^{\circ}_{1gj}\, \hat{p}^{\circ}_{0gj}\}$$
$$= \sum_g \sum_j \frac{1}{n} \{ \frac{N_{1g}}{N_g} \frac{n_g}{n_{1g}} n_{1gj} + \frac{N_{0g}}{N_g} \frac{n_g}{n_{0g}} n_{0gj} \} \underline{z}_{gj}\, \underline{z}_{gj}'\, \hat{p}^{\circ}_{1gj}\, \hat{p}^{\circ}_{0gj}\}. \qquad (6)$$

By the WLLN n_{igj}/n_{ig} converges in probability to μ^*_{igj}/μ^*_{igj}. Since $\hat{\underline{\theta}}_{\text{wesml}}$ converges in probability to $\underline{\theta}^*$, $\underline{\theta}^{\circ}$ which is a point between $\hat{\underline{\theta}}_{\text{wesml}}$ and $\underline{\theta}^*$ also converges in probability to $\underline{\theta}^*$. And $\hat{p}^{\circ}_{1gj}$ converges in probability to p^*_{1gj} by the continuity property of p_{1gj}. Hence equation (6) converges in probability to $\underline{C}$.

Q.E.D.

The proof of theorem 2.

By the same argument we used in theorem 1 to prove (6) converges in probability to $\underline{C}$ we can show $\underline{C}_n$ converges in probability to $\underline{C}$. To prove $\underline{E}_n$ converges in probability to $\underline{\Lambda}$ we rewrite $\underline{E}_n$ in aggregated form again.

$$\underline{E}_n = \frac{1}{n} \sum_i \sum_g \sum_{k=1}^{n_{ig}} \hat{\underline{z}}_{igk}\, \hat{\underline{z}}_{igk}' - \frac{1}{n} \sum_i \sum_g \frac{1}{n_{ig}} \Big(\sum_{k=1}^{n_{ig}} \hat{\underline{z}}_{igk} \Big) \Big(\sum_{k=1}^{n_{ig}} \hat{\underline{z}}_{igk} \Big)'$$
$$= \frac{1}{n} \sum_i \sum_g \sum_j n_{igj}\, w_g^2(i) (i - \hat{p}_{1gj})^2\, \underline{z}_{gj}\, \underline{z}_{gj}'$$
$$- \frac{1}{n} \sum_i \sum_g \frac{1}{n_{ig}} \Big(\sum_j n_{igj}\, w_g(i) (i - \hat{p}_{1gj})\, \underline{z}_{gj} \Big) \Big(\sum_j n_{igj}\, w_g(i) (i - \hat{p}_{1gj})\, \underline{z}_{gj} \Big)'$$
$$= \sum_i \sum_g \Big(\frac{N_{ig}}{N_g} \Big)^2 \frac{n_g}{n_{ig}} \frac{n_g}{n} \Big\{ \sum_j \frac{n_{igj}}{n_{ig}} (i - \hat{p}_{1gj})^2\, \underline{z}_{gj}\, \underline{z}_{gj}'$$
$$- \Big(\sum_j \frac{n_{igj}}{n_{ig}} (i - \hat{p}_{1gj})\, \underline{z}_{gj} \Big) \Big(\sum_j \frac{n_{igj}}{n_{ig}} (i - \hat{p}_{1gj})\, \underline{z}_{gj} \Big)' \Big\}.$$

The proof is done by Assumption 1, Lemma 1 and the fact $\hat{p}_{igj}$ converges in probability to p_{igj}.

Q.E.D.