Binary logistic regression with stratified survey data

Nicklas Pettersson [1]

[1] Stockholm University, Sweden

e-mail: nicklas.pettersson@stat.su.se

Abstract

Standard inference techniques are only valid if the design is ignorable. Two approaches that take the design into account are compared using binary logistic regression. The modelbased approach includes relevant design variables as independents and the designbased approach use design weights. The approaches are exemplified using a cross-sectional stratified mail survey, where associations between urinary incontinence and health related variables among women are studied. Stratification variable age is strongly associated with the dependent variable while stratification variable geography is primarily administrative. The final models contain the same variables irrespective of which approach is used, and the design seem to be only slightly informative. The modelbased approach gives more efficient estimates since geography is not included in the model. But since point estimates differ slightly there is a tradeoff between bias and efficiency, so a designbased approach which includes geography in the design weights might also be advocated.

1 Introduction

From a frequentist view a survey can be used to make likelihood-based inference about a parameter. When for example a logistic regression is fitted, the parameters are assumed to be defined with respect to a stipulated model. This model is usually a superpopulation model which assume that samples are drawn from a population that is generated from a larger superpopulation, which is governed by a parameter θ. It can also be a finite population model which assumes that samples are drawn from a finite population, and inference concern the finite population parameter θ_U. Indexation with U indicate a finite population consisting of $i = 1,...,N$ subjects. With a large population, θ_U can be considered approaching θ, see Skinner, Holt and Smith (1989).

It is assumed that all subjects have values on variables $Y = \left(Y^{(k)}, X^{(j)}\right)$ for $k = 1,...,p$ and $j = 1,...,q$, goverened by θ. A probability function for the population values from a superpopulation can be denoted by $f(y_U;\theta)$. Lower-case letters indicate realisations, so y_U is a realisation of variable Y_U, which itself is generated from Y.

Samples are represented with an indicator variable M_U where the realisation $m_{U,i}=1$ if subject $i=1,...,N$ is in the sample or $m_{U,i}=0$ if not. The sample mechanism is then a function $f(m_U)$ that gives every sample a selection probability. A common feature of most surveys is stratification on design variables Z_U, whose realisations can be described by a probability function $f(z_U;\psi)$, indexed by a parameter ψ.

In simple random sample (SRS) design weights are $w_i=N/n$ for all subjects $i=1,...,n$ in a sample, and with stratification $w_{hi}=N_h/n_h$ for subject $i=1,...,n_h$ in stratum $h=1,...,H$. Assuming that nonresponse is missing completely at random (MCAR) within each stratum, sample weights are calculated as $\widetilde{w}_{hi}=w_h/r_h$, where r_h is the response rate in stratum h. Missing data is left out until section 3.

1.1 Inference and accounting for design

According to Chambers and Skinner (2003) the reason to adjust for the design is that inference may be invalid if the design is non-ignorable. If inference about θ differ when the chosen model is $f(y_U;\theta)$ or $f(y_U,m_U;\theta)$, then inference is non-ignorable of $f(m_U)$ and should be based on $f(y_U,m_U;\theta)$. Otherwise it is ignorable and can be based on $f(y_U;\theta)$.

SRS assumes that y_U is independent of m_U, in other words that $f(m_U)$ is non-informative about $f(y_U;\theta)$. It then follws that $f(y_U,m_U;\theta)=f(y_U;\theta)f(m_U)$ so that $f(m_U)$ is also ignorable. If instead m_U and y_U are dependent, thus $f(m_U)$ is informative about $f(y_U;\theta)$, inference based on $f(y_U;\theta)$ is usually non-ignorable of $f(m_U)$. Since the conditions for ignorability are difficult to verify, see Sugden and Smith (1984), it is sufficient to specify a noninformative model in order for valid inference. Testing for informativeness are described by Chambers, Dorfman och Sverchkov (2003).

Assume $f(m_U)$ is informative on $f(y_U;\theta)$ due to stratification on z_U. A modelbased approach then takes the design into account through conditioning on z_U through specifying $f(y_U,m_U \mid z_U;\phi)$. If $f(y_U,m_U;\phi,\psi)$ can be written $f(y_U \mid z_U;\phi)f(m_U \mid z_U)f(z_U;\psi)$, ψ does not contain elements from ϕ and

$f(y_U \mid z_U;\phi)f(z_U;\psi)$ can be used as a model. Chambers et al (2003) are annoyed by the change of parameter from θ to ϕ, which is due to conditioning on z_U. They advocate only to condition on variables in z_U that are relevant for the model, thus preferably not variables that are only used for administrative purposes. This is refered to as a modelbased approach.

Another approach is the designbased approach, where stratification of z_U is taken into account through including the weights w_{hi} in the estimation. This creates a fictitious total population. Both approaches are usually considered complementary. It is also possible to use weights in a modelbased approach, see Pfeffermann (1993).

Whether modelbased or designbased approach is preferred depends on if the specified model is true or not. If the model is true, both approaches will give consistent estimates and inference will be valid. But a modelbased approach could be better since it gives more efficient estimates due to the stronger assumptions about parameters, and as variation among w_{hi} increases, the efficiency of estimates in the designbased approach is worsened. A finite population correction for without replacement designs has a similar effect. The effect however diminishes with increased sample size.

If the chosen model is misspecified a designbased approach will still give consistent estimates for the finite population and valid inference, while a modelbased approach may give non-consistent estimates with invalid variances and invalid inference. In this sense, the designbased approach can be said to better protect against an informative design or a misspecified model, see Lohr (1999).

The two approaches are compared using a stratified mail survey where logistic regression is used to study urinary incontinence (UI) in relation to aspects of general health, living conditions, personal habits and socioeconomics. UI is known to be associated with other diseases, and can have a negative impact on quality of life. The results have been presented previously in Pettersson and Ewerdahl (2007).

2 Logistic regression

Binary logistic regression can be used to regress a binary dependent variable $Y^{(1)}$ on X, with the purpose of studying associations, classifying or predicting. The model assumes that the logarithm of an odds for $Y^{(1)}$ is linearly dependent on X

$$g(x_i) = \ln\left[\frac{p(Y^{(1)} = 1 \mid X = x_i)}{p(Y^{(1)} = 0 \mid X = x_i)}\right] = \ln\left[\frac{p(x_i)}{1 - p(x_i)}\right] = x_i^t \beta\,. \qquad (1)$$

For each β_j, the likelihood function is set equal to zero and then derived with respect to β_j to obtain $j = 1,...,q$ score equations $\partial \log L(\beta) / \partial \beta_j = u(\beta_j) = 0$. The equations are

$$\sum_{i=1}^{n} x_i^{(j)} \left(y_i^{(1)} - \frac{\exp\left[x_i^t \beta\right]}{1 + \exp\left[x_i^t \beta\right]} \right) = 0 \,, \; j = 1,...,q \,. \tag{2}$$

Error terms are assumed to follow a shifted binomial instead of a normal distribution, so to estimate β maximum likelihood (ML) has to be used, and an algorithm such as Iteratively reweighted least-squares (IRLS) can be used to obtain $\hat{\beta}_j$, see Collett (1991). The information matrix $I(\hat{\beta})$ can be estimated as $\hat{I}(\hat{\beta}) = x^t v x$, where $v = diag\{\hat{p}(x)[1 - \hat{p}(x)]\}$, and its inverse $\hat{I}^{-1}(\hat{\beta})$ can be used as an approximation of the covariance matrix $Var(\hat{\beta})$.

2.1 Modelbased and designbased logistic regression

If Z depend on X or $Y^{(1)}$, a stratified survey design could be informative. A modelbased approach include Z in the model and equation (1) becomes

$$g(x_{hi}, z_{hi}) = \ln\left[\frac{p(x_{hi}, z_{hi})}{1 - p(x_{hi}, z_{hi})} \right] = z_h^t \alpha + x_{hi}^t \beta + z_h^t x_{hi}^t \gamma_h \,. \tag{3}$$

The interaction between z_h and x_{hi} with parameter γ_k allows for different effects among strata. With many estimated parameters in relation to n, Kleinbaum, Kupper, Muller and Nizam (1998) suggest a conditional model. However, if Z only depend on $Y^{(1)}$ and not X, Prentice and Pyke (1979) show that the only parameter affected is the intercept.

With a designbased approach w_{hi} are included in the log-likelihood equations so that an enlarged fictitious total population is obtained. Equation (2) becomes

$$\sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} x_{hi}^{(j)} \left(y_{hi}^{(1)} - \frac{\exp\left[x_{hi}^t \beta\right]}{1 + \exp\left[x_{hi}^t \beta\right]} \right) = 0 \,, \; j = 1, ...,q \,. \tag{4}$$

Since ML only is based on true observations a modified method called pseudo-maximum-likelihood (PML) can be applied according to Skinner et al (1989). The score equations are $wu(\beta_j) = 0$, and Hosmer et al (2000) give an estimator of the covariance matrix of $\hat{\beta}$ as $V\hat{a}r(\hat{\beta}) = (x^t wvx)^{-1} \hat{s} (x^t wvx)^{-1}$, where $\hat{s} = \sum_{i=1}^{n_h} (1 - n_h / N_h) \hat{s}_h$ is the pooled within-stratum covariance matrix and w a weight matrix. $V\hat{a}r(\hat{\beta})$ is approximated with Taylor series.

3 Material and subjects

A health questionnaire on issues of public health was sent to sampled subjects 18-79 years of age in Örebro county, in March 2000. An additional UI-questionnaire "*For Persons with Problems with Involuntary Urine Loss*" was also enclosed. Statistics Sweden conducted the study at the request of the county council and was responsible for selecting the sample, distributing the questionnaires (including three reminders) and coding the data. The sample was randomly selected from the Population Register of Sweden and stratified for gender $\left(Z^{(a\bullet\bullet)},\ a=1,2\right)$ four age groups (18-34; 35-49; 50-64; 65-79) $\left(Z^{(\bullet b\bullet)}, b=1,2,3,4\right)$and 16 geographical areas $\left(Z^{(\bullet\bullet c)}, c=1,...,16\right)$ with $n_h = 120$. Only women were kept in this study $\left(Z^{(1\bullet\bullet)}, N = 99\,679, n = 7\,680,\ H = 64\right)$. $Z^{(1\bullet\bullet)}$ was compared to population values, and nonresponse was assumed to be MCAR within each stratum.

Modelbased (PROC LOGISTIC) and designbased (PROC SURVEYLOGISTIC) logistic regression was carried out using SAS 9.1.3. The dependent variable $Y^{(1)}$ was defined as $\left(y_i^{(1)} = 1\right)$ for subjects answering the UI-questionnaire and claiming problems of UI or $\left(y_i^{(1)} = 0\right)$ for subjects not answering the UI-questionnaire or not claiming problems of UI. A group of clinicians identified independent variables $X^{(j)}$ from health questionnaire and public registers. Three steps were taken for each model:

1. $Y^{(1)}$ regressed on each variable in $X^{(j)}$ and $Z^{(1\bullet\bullet)}$ univariately, with threshold set at p-value< 0.25 using Pearson χ^2 test for inclusion in step 2.
2. $Y^{(1)}$ regressed on all variables in $X^{(j)}$ and $Z^{(1\bullet\bullet)}$ from step 1, with stepwise selection until p-value< 0.01 for remaining variables using likelihood-ratio χ^2 test.
3. Recoding of some variables, including categorisation of age, and plausible interaction effects were tried out before the final model was reached.

4 Results

In total 5 279 women responded, but 670 were excluded due to partial nonresponse in the health questionnaire, leaving 1 332 persons classified as suffering and 3 277 persons as not suffering from UI. Differences between $Z^{(1\bullet\bullet)}$ and population values were mainly due to the unequal selection probabilities in strata. Sample weights varied modestly (range: 6,5-75,0) with largest contribution from $Z^{(1\bullet c)}$. Both methods included the same variables in the models, see table 1. The effect of age dominated in both cases. It was first included as continuous, but in order to provide a better fit to the data it was split into a continuous part

(18-49 years of age) plus a dummy variable (65-79 years of age). This split was in accordance with the stratification. All other variables, except for BMI, had rather small estimated effects but were reasonable according to the clinical group. No interaction effects were significant.

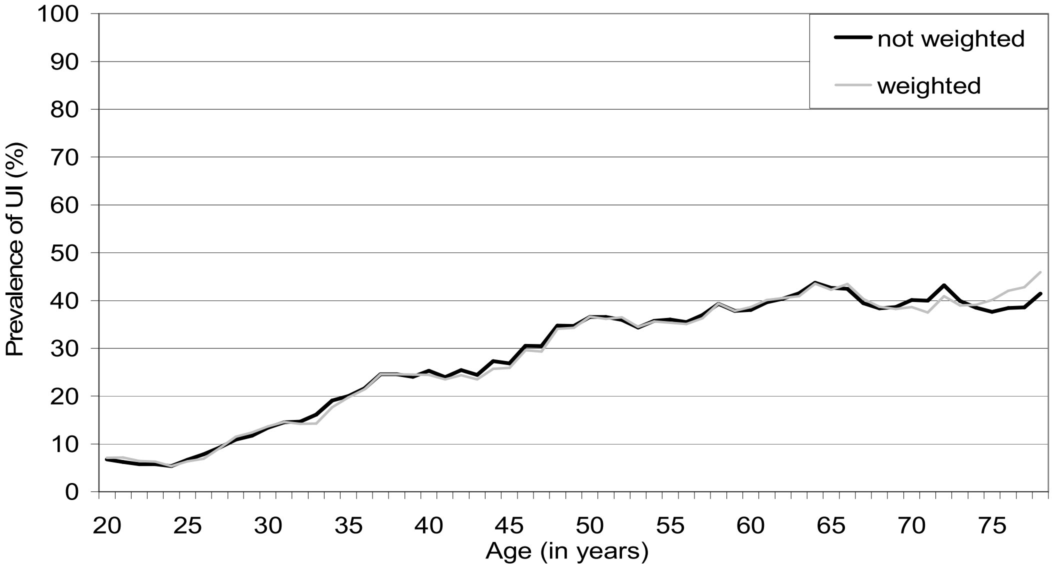
Table1.Oddsratios (OR) and confidence intervals (CI) for final model variables (n=4 609)

	Univariate OR (99% CI)		Multivariate OR (99% CI)	
	Model	Design	Model	Design
Intercept	0.41 (0.37-0.44)	0.39 (0.35-0.42)	0.01 (0.00-0.01)	0.01 (0.00-0.01)
Age continuous 18-49 years	1.08 (1.06-1.09)	1.08 (1.06-1.09)	1.07 (1.06-1.09)	1.07 (1.06-1.09)
Age - dummy for 65-79 years	1.95 (1.61-2.36)	2.17 (1.75-2.68)	1.22 (0.98-1.52)	1.30 (1.02-1.66)
Age (dummy) 18-49 years (reference)	1	1	1	1
BMI (kg/m2); Obese (>=30)	2.99 (2.35-3.80)	3.25 (2.48-4.26)	2.04 (1.58-2.63)	2.10 (1.59-2.76)
BMI (kg/m2); Overweight (25-29.99)	1.63 (1.35-1.98)	1.72 (1.39-2.13)	1.19 (0.97-1.46)	1.18 (0.95-1.47)
BMI (kg/m2); Low/normal (<25) (reference)	1	1	1	1
Bothered by musculoskeletal pain*	2.50 (1.96-3.19)	2.36 (1.80-3.11)	1.58 (1.21-2.07)	1.42 (1.06-1.92)
Somewhat bothered by musculoskeletal pain	1.48 (1.13-1.92)	1.44 (1.07-1.94)	1.29 (0.98-1.70)	1.23 (0.90-1.68)
Not bothered by musculoskeletal pain (reference)	1	1	1	1
Felt the need to seek medical care but have declined it*	1.63 (1.34-1.99)	1.69 (1.35-2.12)	1.38 (1.10-1.72)	1.45 (1.14-1.85)
Have not declined medical care (reference)	1	1	1	1
Bothered by tiredness, weakness and sleeping disorder*	2.18 (1.67-2.83)	2.18 (1.61-2.94)	1.60 (1.20-2.13)	1.64 (1.19-2.26)
Somewhat bothered by tiredness weakness and sleeping disorder	1.54 (1.17-2.03)	1.50 (1.10-2.06)	1.47 (1.10-1.97)	1.47 (1.06-2.05)
Not bothered by tiredness, weakness and sleeping disorder (reference)	1	1	1	1
Problems manage daily expenses or not able to raise 18 000 SEK*	1.30 (1.03-1.66)	1.33 (1.02-1.75)	1.37 (1.05-1.79)	1.41 (1.05-1.90)
Have not had these financial problems (reference)	1	1	1	1
Been humiliated or ridiculed*;**	1.04 (0.79-1.37)	0.99 (0.72-1.35)	1.39 (1.02-1.90)	1.37 (0.98-1.92)
Somewhat been humiliated or ridiculed	1.06 (0.88-1.27)	1.03 (0.84-1.27)	1.28 (1.05-1.58)	1.30 (1.04-1.63)
Have not had that experience (reference)	1	1	1	1

**Reference period is "during the last three months". **Univariate p-value >0,25 but included in the further analysis due to significant effect when conditioned on age.*

Both univariate and multivariate OR were very similar among the two approaches. Most OR diminished from the univariate to the multivariate regression. Since age was strongly associated with UI, it was also significant whether the approach, see figure 1. Standard errors in the final models were always smaller in the modelbased approach. Geography was not significant in any approach.

Figure1.Age (5-year-average) in relation to prevalence of UI (n=4 609).



5 Discussion

The approaches should be considered as complementary. From the results it was not obvious which approach was preferred. The same variables were significant with quite similar estimates but with less efficient estimates in the designbased model due to the sample weights.

Age was included as semi continuous in both models due to its association with UI. This should have eliminated its potential informativeness in the modelbased approach. Its contribution to the variation in design weights was also small, and therefore had a minimal effect on the estimates and their errors in the designbased approach. One could therefore argue that age should be excluded from the design weights in the designbased approach. Since geography is primarily an administrative variable and thus is not likely to be informative, one could also argue that it should be excluded from the design weights. However, the differences in estimates are mainly due to geography. This is the core tradeoff between the approaches, designbased protect against a misspecified model but at the cost of efficiency. One way to decide could be to test for informativeness. With design weights excluded the only difference between the two approaches would be the finite population correction, whose efficiency reduction was marginal due to the large sample size and relatively few strata, and the nonresponse adjustment in the designbased approach.

A potentially larger problem could be if nonresponse was non-ignorable. UI is known to be correlated with other diseases, which were either not asked for or the questions were inaccurately quoted, and it is therefore reasonable that response rate will be smaller among less-healthier. But nonresponse need not affect any other estimates than the intercept if associations are independent of the level of health. However, the estimated effects should perhaps be considered more of indicators rather than true effects. Many variables were also significant in the univariate analysis, and the changes in effects to multivariate models also show that many effects are perhaps not stable. This also emphasize that it is always important to have a good model, or at least not a very bad one.

It would be interesting to try and adjust for the nonresponse as part of the sensitivity analysis, rather than just assuming MCAR within strata. One way is to poststratify and calibrate on some of the register variables or to impute the missing values. However, an advantage of this study is that UI with accompanying problems are to a large extent a subjective symptom. It is therefore reasonable to assume that an anonymous mail survey could have smaller missclassification and underestimation than other estimation methods.

References

Chambers, R. L. Dorfman A. H. Sverchkov M.Y. (2003). Nonparametric regression with complex survey data. In *Analysis of survey data* (eds Chambers, R. L. Skinner, C. J.) Chichester: Wiley.

Chambers, R. L. Skinner, C. J. (eds) (2003). *Analysis of survey data*. Chichester: Wiley.

Collett, D. (1991). *Modelling binary data*. London: Chapman & Hall.

Hosmer, D.W. Lemeshow, S. (2000). *Applied logistic regression.* 2nd edition. New York: Wiley.

Kleinbaum, D.G. Kupper, L.L. Muller, K.E. Nizam, A. (1998). *Applied regression analysis and multivariate methods*. 3rd edition. Pacific Grove: Brooks/Cole publishing company.

Lohr, S. L. (1999). *Sampling: design and analysis*. Pacific Grove: Brooks/Cole publishing company.

Pettersson, N. Ewerdahl, D. (2007). Binary logistic regression-description and application to stratified surveydata. Bachelor Thesis (in Swedish), Department of Statistics, Stockholm University

Pfeffermann, D. (1993). The role of sampling weights when modelling survey data. International statistical review, 61(2), 317-337.

Prentice, R.L. Pyke, R. (1979). Logistic disease incidence models and case-control studies. Biometrika, 66, 403-411.

Skinner, C. J. Holt D. Smith, T.M.F. (1989). *Analysis of complex surveys*. West Sussex: Wiley.

Sugden, R. A. Smith, T.M.F. (1984) Ignorable and informative designs in survey sampling inference. Biometrika, 71, 495-506.