

Outcome-related, auxiliary variable sampling designs for longitudinal binary data:

eAppendix

Jonathan S. Schildcrout, Enrique F. Schisterman, Melinda C. Aldrich, Paul J. Rathouz

LUNG HEALTH STUDY DATA

The data from the Lung Health Study were download from the database for Genotype and Phenotypes (dbGaP; <https://www.ncbi.nlm.nih.gov/gap>) that was designed to share data from studies that have examined genotype–phenotype associations in humans. Instructions for applying for dbGaP data can be found at <https://www.ncbi.nlm.nih.gov/books/NBK5294/>, and instructions for downloading and extracting data can be found at <https://www.ncbi.nlm.nih.gov/books/NBK5291/>.

SEQUENTIAL OFFSETTED REGRESSIONS ANALYSES

Under the proposed auxiliary variable sampling designs, the sample is not representative of the target population; rather, it is representative of a pseudo–population that is enriched with values $Y_{ij} = 1$ and with values of $V_i = 1$ under exposure and auxiliary variable sampling designs. For a given design, we denote the mean of this pseudo–population with $\mu_{ij}^s = E(Y_{ij} \mid \mathbf{X}_{ij}, S_i = 1) = \Pr(Y_{ij} = 1 \mid \mathbf{X}_{ij}, S_i = 1)$, where superscript s signifies ‘sample’ to emphasize that, in general, $\mu_{ij}^s \neq \mu_{ij}^p$. We interpret the pseudo–mean μ_{ij}^s as simply the mean of Y_{ij} that would be obtained, on average, in a naive analysis of data from this design, i.e., an analysis that ignores the enhanced sampling design. As such, in most circumstances, if we conduct the proposed design and analyze the data as if it was collected from a simple random sample, we are likely to obtain biased parameter estimates.

Here, we briefly outline an analysis procedure described more completely in reference [1]. The procedure yields valid analyses under the proposed designs in the sense that estimates of $(\beta_0, \beta_1, \beta_v)$ are approximately unbiased, as are standard errors of those estimates; hypothesis tests and confidence intervals are also therefore correct. Our algorithm is based on a sequence of two offsetted logistic regression analyses. The first regression estimates the relationship of auxiliary variable Z_i to response Y_{ij} and covariate \mathbf{X}_{ij} data. The results from this model are then used as an offset in the outcome model that captures the relationship between the response and covariate data. We now detail both models.

Auxiliary Variable Model

Because the sampling procedure involves a known quantity, we may rely on a well–known result from case–control studies. Namely, the prospective, population model is preserved under case–control sampling with logistic regression analysis of binary response data, with the exception that the intercept is shifted by a

known constant, i.e., the log transformed ratio of sampling probabilities for cases versus controls. Exploiting this fact for the auxiliary sampling variable Z_i , we let $\lambda_{ij}^p(y_{ij}, \mathbf{X}_{ij}) = \Pr(Z_i = 1 \mid Y_{ij} = y_{ij}, \mathbf{X}_{ij}, V_i)$ in the population, and $\lambda_{ij}^s(y_{ij}, \mathbf{X}_{ij}) = \Pr(Z_i = 1 \mid Y_{ij} = y_{ij}, \mathbf{X}_{ij}, V_i, S_i = 1)$ in the sample. By applying Bayes' Theorem, the odds model for Z_i in the sample is given by

$$\frac{\lambda_{ij}^s(y_{ij}, \mathbf{X}_{ij})}{1 - \lambda_{ij}^s(y_{ij}, \mathbf{X}_{ij})} = \frac{\lambda_{ij}^p(y_{ij}, \mathbf{X}_{ij})}{1 - \lambda_{ij}^p(y_{ij}, \mathbf{X}_{ij})} \frac{\pi(1, V_i)}{\pi(0, V_i)}, \quad (1)$$

where $y_{ij} \in \{0, 1\}$.

We assume $\lambda_{ij}^p(y_{ij}, \mathbf{X}_{ij})$ can be approximated with a logistic regression model in Y_{ij} , $\mathbf{W}_{1,ij}$, and $Y_{ij} \cdot \mathbf{W}_{2,ij}$, where $\mathbf{W}_{1,ij}$ and $\mathbf{W}_{2,ij}$ are subsets of $(\mathbf{X}'_{ij}, V_i)'$. We may then fit an offsetted logistic regression model to the biased sample in order to estimate parameters in the sampling variable population model λ_{ij}^p . Estimates from this intermediate model are used to identify the offset in the second, outcome model discussed next. It is important to note that $\mathbf{W}_{1,ij}$ and $\mathbf{W}_{2,ij}$ must be sufficiently rich to include all important predictors of Z_i contained in $(\mathbf{X}'_{ij}, V_i)'$ for this approach to be valid.

Outcome Model

Similar to the auxiliary variable model, the population and sample odds for the primary outcome model can be related to one another by applying Bayes' theorem

$$\frac{\mu_{ij}^s}{1 - \mu_{ij}^s} = \frac{\mu_{ij}^p}{1 - \mu_{ij}^p} \cdot \frac{\rho_{ij}(1, \mathbf{X}_{ij})}{\rho_{ij}(0, \mathbf{X}_{ij})} \quad (2)$$

where $\rho_{ij}(y_{ij}, \mathbf{X}_{ij}) = \Pr(S_i = 1 \mid Y_{ij} = y_{ij}, \mathbf{X}_{ij}, V_i)$. Using the estimated sampling variable model ($\hat{\lambda}_{ij}^p$), along with the known sampling probabilities, $\pi(Z_i, V_i)$, we estimate $\rho_{ij}(y_{ij}, \mathbf{X}_{ij}, V_i)$ as

$$\hat{\rho}_{ij}(y_{ij}, \mathbf{X}_{ij}, V_i) = \pi(0, V_i) \{1 - \hat{\lambda}_{ij}^p(y_{ij}, \mathbf{X}_{ij})\} + \pi(1, V_i) \hat{\lambda}_{ij}^p(y_{ij}, \mathbf{X}_{ij}). \quad (3)$$

We may then use the log transformed ratio $\hat{\rho}_{ij}(y_{ij}, \mathbf{X}_{ij}, V_i) / \hat{\rho}_{ij}(y, \mathbf{X}_{ij}, V_i)$ as an offset in the logistic regression mode of Y_{ij} on $(\mathbf{X}'_{ij}, V_i)'$ in sampled subjects. Because the offset is estimated with uncertainty, standard errors are not straightforward, but can be obtained in steps described in [1, 2]. In certain situations, one may use non-independence working covariance weighting to gain efficiency for parameters estimated in μ_{ij}^p . We note, however, that the associated working covariance parameters α do not represent population parameters and should not be interpreted as such.

Summary of Sequential Offsetted Regressions

The following summarizes steps for parameter estimation using the sequential offsetted regressions approach for longitudinal logistic regression analyses.

1. Using $\log\{\pi(1, V_i)/\pi(0, V_i)\}$ as an offset, estimate parameters for λ_{ij}^p using offsetted logistic regression.
2. Combining the known sampling probabilities $\pi(0, V_i)$ and $\pi(1, V_i)$ with estimates $\widehat{\lambda}_{ij}^p$, use equation (3) to calculate $\widehat{\rho}_{ij}(y_{ij}, \mathbf{X}_{ij})$ for all i and j .
3. Using $\log\{\widehat{\rho}_{ij}(1, \mathbf{X}_{ij})/\widehat{\rho}_{ij}(0, \mathbf{X}_{ij})\}$ as an offset, estimate parameters for μ_{ij}^p with a logistic regression model using GEE.

References

- [1] J. S. Schildcrout and P. J. Rathouz, “Longitudinal studies of binary response data following case-control and stratified case-control sampling: design and analysis,” *Biometrics*, vol. 66, pp. 365–373, Jun 2010.
- [2] J. S. Schildcrout, S. L. Mumford, Z. Chen, P. J. Heagerty, and P. J. Rathouz, “Outcome-dependent sampling for longitudinal binary response data based on a time-varying auxiliary variable,” *Stat Med*, vol. 31, pp. 2441–2456, Sep 2012.

eFigure 1: Relative Variance across 250 replicates for the COPD-free outcome when sampling 500 subjects. We show $RV = \overline{\widehat{var}}_{RS}(\hat{\beta}) / \overline{\widehat{var}}_{AVS:SOR}(\hat{\beta})$ for a number of data features. Black diamonds denote the analyses presented in Table 3 and grey diamonds show the result of perturbing one data, analysis or design feature. Panels show the impact on relative variance of response prevalence (A), strength of the $Z \sim Y$ relationship (B), amount of response dependence (C), richness of the auxiliary variable model for Z (D), and asthma exposure and auxiliary variable sampling (E).

eFigure 2: Relative Variance across 250 replicates for the severe COPD outcome when sampling 500 subjects. We show $RV = \overline{\widehat{var}}_{RS}(\hat{\beta}) / \overline{\widehat{var}}_{AVS:SOR}(\hat{\beta})$ for a number of data features. Black diamonds denote the analyses presented in Table 3 and grey diamonds show the result of perturbing one data, analysis or design feature. Panels show the impact on relative variance of response prevalence (A), strength of the $Z \sim Y$ relationship (B), amount of response dependence (C), richness of the auxiliary variable model for Z (D), and asthma exposure and auxiliary variable sampling (E).