# A Pseudolikelihood Approach to Correcting for Ascertainment Bias in Family Studies

Daniel Rabinowitz

Department of Statistics, Columbia University, New York

## Summary

**The Cannings and Thompson approach to correcting for ascertainment bias in family studies is extended to settings with multiple ascertainment. The extension is based on maximizing a pseudolikelihood. Two approaches to computing standard errors for the maximum pseudolikelihood estimate are described. One is especially simple to compute, while the other is more generally applicable. Simulation experiments suggest that the standard-error computations can be quite accurate.**

## Introduction

In genetic epidemiology, it is common for the collection of family data to occur in stages. In the first stage, individuals or portions of pedigrees come to the attention of investigators. In the second stage, related individuals are chosen sequentially for enrollment. When a pedigree comes to the attention of the investigators through an individual, that individual is usually termed a *proband.* The events that bring a pedigree to the attention of investigators are sometimes termed the *ascertainment events.*

The factors that influence ascertainment events may include phenotypes of pedigree members. When this is the case, the pedigrees included in a study may not be representative of the population of interest. It has long been known that this discrepancy must be corrected for during data analysis, if ascertainment bias is to be avoided (see, for example, Fisher [1934] or Morton [1959]).

Methods for avoiding ascertainment bias have been a subject of considerable interest (e.g., see Elandt-Johnson 1971; Elston and Stewart 1971; Cannings and Thompson 1977; Elston and Sobel 1979; Thompson and Cannings 1979; Boehnke and Greenberg 1984; Ewens and Shute 1986; Hodge and Boehnke 1986; Hodge 1988; Shute and Ewens 1988*a,* 1988*b*). Several methods based on conditioning have been advocated. One is to condition on the phenotypic information that influenced the selection of the pedigrees (see Cannings and Thompson 1977). Another is to condition on all the phenotypic information in a pedigree that is relevant to whether the pedigree could be included in the data (see Ewens and Shute 1986). Elston and Sobel (1979) describe an alternative conditioning approach that requires both a model for the probability of becoming a proband and that the population may be divided into nonoverlapping pedigrees.

Vieland and Hodge (1995) suggest that the applicability of the first of the conditioning approaches is limited to situations in which no individual is included in pedigrees from separate ascertainment events. They also point out that the second approach may not be applicable in cases where the conditioning event is not well defined or in cases where the marginal likelihood of the conditioning event is not observed. The purpose of this note is to propose an extension of the Cannings and Thompson approach that is applicable to multiple ascertainment.

## Correcting for Ascertainment

The approach proposed here involves the construction of a pseudolikelihood. The pseudolikelihood contains a term for each ascertainment event. The contribution from each ascertainment event to the pseudolikelihood is the term that would be included in the conditional likelihood of Cannings and Thompson (1977), had the sequential enrollment choices been made blinded to the other ascertainment events and their sequentially enrolled pedigrees: even if individuals associated with one ascertainment event are related to individuals associated with another, the terms in the pseudolikelihood must depend only on the data that would have been collected had sequential enrollment been made without awareness of the relatedness.

With single ascertainment, the pseudolikelihood is identical to the conditional likelihood of Cannings and Thompson (1977). The pseudolikelihood approach presented here may therefore be thought of as an extension of the Cannings and Thompson (1977) approach. Elston

Received November 1, 1995; accepted for publication June 11, 1996.

Address for correspondence and reprints: Dr. Daniel Rabinowitz, Department of Statistics, Mathematics Building, Columbia University, New York, NY 10027.

0002-9297/96/5903-0028$02.00

(1995) suggests a pseudolikelihood approach to estimation. Implicit in the suggestion is that some form of correction for ascertainment is applied to each term in the pseudolikelihood. The approach presented here may therefore also be thought of as an implementation of Elston's (1995) suggestion. Each term in the pseudolikelihood behaves, by itself, as in the case of single ascertainment. Therefore, as with the conditional likelihood of Cannings and Thompson (1977), the gradient of the log of each term has expectation zero. This ensures that the maximum pseudolikelihood estimate is asymptotically unbiased (e.g., see Gong and Samaniego 1981; Gourieroux et al. 1984; Kalbfleisch 1986; Davidian and Carroll 1988; Heyde 1989; Barndorff-Nielsen 1991).

Just as in the case Cannings and Thompson's approach with single ascertainment, for traits that are not associated with patterns of mating or family size, a sufficient condition for the pseudolikelihood approach to produce asymptotically unbiased estimates is that the enrollment choices made during the sequential data collection are influenced by phenotypic information only through the phenotypic information that has been previously observed. Recall that the terms in the pseudolikelihood are defined to be what would have been the result of sequential sampling, had the sequential sampling associated with each ascertainment event been blinded to the data collection associated with the other ascertainment events. The enrollment choices referred to in the sufficient condition are to be interpreted as the choices defined implicitly by the construction of the pseudolikelihood. In order to apply the pseudolikelihood approach, the phenotypic information that influences the occurrence of each ascertainment event must be available to the data analyst. The requisite information may be characterized in terms of an independence condition. Associated with each ascertainment event is phenotypic information $I$, available to the analyst, with the property that *the conditional probability of the ascertainment event, given all phenotypic information, is the same as the conditional probability of the ascertainment event, given the information $I$.* The information $I$ is the information that is to be conditioned on when computing a term in the pseudolikelihood. Availability of the information $I$ is determined (just as in the case of single ascertainment) by the kinds of data that are recorded during sampling: unless these data about the ascertainment events are available to the investigator, they cannot be used to adjust for the ascertainment process.

The example discussed by Vieland and Hodge (1995) can be used to illustrate the construction of the pseudolikelihood. In that example, a universe of sibships each with three sibs was assumed, and it was assumed that the three sibs were ordered by age and that each sib could be affected or unaffected. It was assumed that every affected sib had equal probability, $\pi$, of becoming a proband, independently of all other sibs. Unaffected sibs could not become probands, so the affected status of the proband is the phenotypic information that should be conditioned on. If more than one sib were affected, then each of the affected sibs might separately become a proband. The sequential sampling protocol was to enroll all relatives immediately adjacent in age to a proband. Therefore, if only the oldest or the youngest sib were a proband, two family members would be enrolled. If the middle sib were the proband, however, all three sibs would be enrolled. Similarly, if either two or three sibs separately were probands, then all sibs would be enrolled.

The following notation will facilitate examination of the example. Let $p_0$ denote the marginal probability that no sibs are affected. Let $p_1$ denote the probability that the youngest sib is affected but the others are not. $p_1$ is also the probability that the middle sib is affected but the others are not, and it is also the probability that the oldest is affected but the others are not. Let $p_2$ be the probability that the youngest sib is not affected but that the others are. $p_2$ is also the probability that the middle is not affected but the others are, and also the probability that the oldest is not affected but the younger two are. Finally, let $p_3$ denote the probability that all three sibs are affected. Note that $p_0 + 3p_1 + 3p_2 + p_3 = 1$. Although the notation does not reflect any particular genetic model, the $p_i$ should be thought of as functions of the parameters of the relevant model. This section concludes with an examination of two different scenarios in the context of the example. Suppose that, in a family with all three sibs affected, only the youngest becomes a proband. Since the sampling protocol is to include only sibs immediately adjacent in age to the proband, the data observed for the sibship would be that the youngest and middle sib are affected. The marginal probability of the observed data would thus be $p_2 + p_3$. The conditioning event is that the youngest sib is affected. The conditioning event thus has probability $p_1 + 2p_2 + p_3$. The contribution of the sibship to the likelihood would thus be $(p_2 + p_3)/(p_1 + 2p_2 + p_3)$.

Suppose that, in a family with the youngest and middle sibs affected, both the youngest and the middle sibs become probands. In this case, each proband would contribute a term to the pseudolikelihood. Consider first the term associated with the middle sib. The sampling protocol is to include all sibs adjacent in age, so the data associated with the middle sib being a proband would be that the youngest and middle sibs are affected and that the oldest is not. The contribution to the likelihood associated with the middle sib would thus be $p_2/(p_1 + 2p_2 + p_3)$. Consider now the term associated with the youngest sib. The data associated with the youngest sib being a proband would be that the youngest and middle sibs are affected. Note that, even though it would be

**Table 1**

**Quantiles of the Empirical Distribution of the Normalized Estimate, Using the Uncorrected Standard Error**

| | | Quantile | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *N* | $\pi$ | .05th | .10th | .25th | .50th | .75th | .90th | .95th |
| 200 | .8 | −2.11 | −1.65 | −.86 | 0 | .85 | 1.60 | 2.05 |
| 400 | .6 | −2.00 | −1.55 | −.82 | 0 | .80 | 1.52 | 1.96 |
| 800 | .5 | −1.94 | −1.51 | −.79 | 0 | .78 | 1.50 | 1.93 |

known through the other proband in the family that the oldest sib is not affected, that information would not be used in computing the term for the youngest sib. This is because the sampling protocol, when the youngest sib is a proband, requires that only the middle sib be enrolled. The contribution to the likelihood for the youngest sib would thus be $(p_2 + p_3)/(p_1 + 2p_2 + p_3)$.

## Standard Errors

Elston (1995) points out that, in computing standard errors for the maximum pseudolikelihood estimate, it is not valid to proceed exactly as if the pseudolikelihood were a likelihood. As in the usual likelihood theory, an information matrix may be defined as the expected matrix of mixed partials of the log pseudolikelihood. However, unlike as in the usual likelihood theory, the inverse of the observed information matrix is not asymptotically unbiased for the variance of the parameter estimates. This is because the information matrix does not account for correlation induced by the same or related individuals being associated with more than one ascertainment event. The usual expansions applied to the log pseudolikelihood can be used to show that the covariance matrix of the maximum pseudolikelihood estimate is asymptotic to the covariance matrix of the gradient of the log pseudolikelihood, pre- and postmultiplied by the inverse of the information matrix. The covariance of the gradient is equal to the sum of the covariances of terms corresponding to each of the ascertainment events plus the sum of the cross-covariances between correlated terms. The observed information matrix is asymptotically unbiased for the sum of the covariances alone. Accurate standard error calculations must take into account cross-covariances.

These considerations suggest that one approach to computing standard errors for the estimates obtained from the pseudolikelihood is to inflate the inverse information matrix by $(1 + k/n)$, where $k$ is the number of ordered pairs of ascertainment events whose associated pedigrees share relatives, and $n$ is the number of ascertainment events. The inflation factor reflects that the variance of the score is not just the sum of $n$ ascertainment event–specific variance terms but that it also involves $k$ ascertainment event pair–specific covariance terms. An appealing aspect of the correction factor is its simplicity. The approach will be most accurate when covariances between terms in the gradient of the log pseudolikelihood are equal on average to the average of the variances of the terms in the gradient.

An alternative approach to computing standard errors is to pre- and postmultiply an estimate of the covariance of the gradient of the log likelihood by the observed information matrix. The estimate of the covariance can be computed by first forming partial sums of correlated terms in the gradient evaluated at the maximum pseudolikelihood estimate. The estimate of the covariance is then computed as the sum of the cross-products of the partial sums. That is, if $S_i$ are the terms in the gradient of the log likelihood, where $i$ indexes the ascertainment events, and if $\tau_j$, $j = 1, 2, \ldots, m$ is the partition of the

**Table 2**

**Quantiles of the Empirical Distribution of the Normalized Estimate, Using the Correction Factor**

| | | Quantile | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *N* | $\pi$ | .05th | .10th | .25th | .50th | .75th | .9[2*ex*]0th | .95th |
| 200 | .8 | −1.69 | −1.32 | −.69 | 0 | .67 | 1.26 | 1.60 |
| 400 | .6 | −1.68 | −1.30 | −.69 | 0 | .66 | 1.25 | 1.61 |
| 800 | .5 | −1.67 | −1.30 | −.68 | 0 | .67 | 1.28 | 1.64 |

**Table 3**

**Quantiles of the Empirical Distributions from the Second Set of Simulations**

| | Quantile | | | | | | |
|---|---|---|---|---|---|---|---|
| Correction | .05th | .10th | .25th | .50th | .75th | .90th | .95th |
| No correction | −1.94 | −1.50 | −.79 | 0 | .78 | 1.49 | 1.91 |
| Simple correction | −1.58 | −.65 | −.36 | 0 | .64 | 1.22 | 1.56 |
| Second correction | −1.66 | −.68 | −.39 | 0 | .67 | 1.29 | 1.64 |

indices of the ascertainment events into subsets so that the indices of ascertainment events associated with the same or or related individuals are all together in the same subset, then the estimate of the covariance is given by

$$\sum_{j=1}^{m} \left(\sum_{i \in \tau_j} S_i\right)^T \left(\sum_{i \in \tau_j} S_i\right).$$

The basis for the second approach to standard-error calculation is that with the true parameters substituted for the maximum pseudolikelihood estimates, the sum of cross-products has expectation equal to the variance of the gradient of the log likelihood.

Although the second approach is more computationally involved than the first, it is accurate in a broader class of settings. The first approach to standard-error calculation treats covariances between terms in the gradient of the log pseudolikelihood as if they were equal on average to the average of the variances of the terms in the gradient. This is appropriate in settings where pedigrees that share related individuals usually overlap completely and overlapping pedigrees do not differ systematically from the other pedigrees in the data set. In settings where these conditions do not hold, the second approach to variance calculations is advisable.

## Simulation Results

Two sets of simulation experiments were performed, to examine the accuracy of the approaches to computing standard errors. Each of the experiments in both sets consisted of 50,000 replications.

In the first set of experiments, the family structure and sampling scheme were taken to be as in the example of Vieland and Hodge (1995). The data were generated as if all matings were between an affected heterozygote and an unaffected homozygote for an autosomal dominant trait with complete penetrance. The model was parameterized as by Khoury et al. (1993, p. 237): $p_0$, $p_1$, $p_2$, and $p_3$ were thus $(1-\theta)^3$, $\theta(1-\theta)^2$, $\theta^2(1-\theta)$, and $\theta^3$, respectively, where $\theta$ represents the marginal probability that a given offspring is affected. In each experiment, in each replication, for a population of $N$ sibships, affected status was assigned independently with the probability $\theta = 1/2$ to all of the sibs. Then, proband status was assigned to the affected sibs, each with probability $\pi$. The maximum pseudolikelihood estimate, the observed information, and the correction factor were then computed. Only the first approach to standard-error calculations is considered.

The difference between the estimate and $1/2$, the true value of the parameter, was normalized by the square root of the observed information and by the corrected observed information. Quantiles of the normalized estimates for three experiments are recorded in tables 1 and 2 for $(N, \pi)$ equal to (200, 0.8), (400, 0.6), and (800, 0.5). Table 1 reports the quantiles for the statistics with the uncorrected standard error. Table 2 reports the quantiles for the statistics with the the correction factor. The empirical quantiles in the tables should be compared to the corresponding quantiles of the standard normal distribution, −1.64, −1.28, −0.67, 0.00, 0.67, 1.28, and 1.64. The simulation experiments indicate that using the observed information matrix can be significantly anticonservative and that the simple approach to standard-error calculations can be accurate.

The genetic model used in the second set of experiments was as in the first: $\theta$ was $1/2$, and the $(n, \pi)$ pairs were (200, 0.8), (400, 0.6), and (800, 0.5). All sibships had 11 sibs, however, and only the youngest and oldest sib could be probands. When the oldest sib was a proband, affected status of the oldest eight sibs was observed, and when the youngest sib was a proband, affected status of the youngest eight sibs was observed. All affected oldest and youngest sibs became probands, and no other sib could become a proband. In each iteration, there were 400 sibships that could give rise to ascertainment events.

Note that the ascertainment events lead to overlapping pedigrees whenever both a youngest and oldest sib are probands. Note also that the pedigrees do not overlap completely. It is expected therefore that the uncorrected standard error is anticonservative but that the simple correction would be conservative. These results were observed in the simulations. The first and second rows of table 3 report the quantiles from the empirical distribution from the simulations of the standardized estimates using the uncorrected standard error and the simple correction, respectively. The third row of table 3 reports the quantiles from the empirical distribution when the more computationally intensive standard-error calculation was used. The quantiles in the third row are close to the quantiles of a standard normal.

## Acknowledgments

The author is grateful to Ruth Ottman, Susan Hodge, and Veronica Vieland for stimulating conversations and encouragement.

## References

Barndorff-Nielsen O (1991) Likelihood theory. In: Hinkley DV, Reid N, Snell EJ (eds) Statistical theory and modelling: in honour of Sir David Cox. Chapman & Hall, London, pp 232–264

Boehnke M, Greenberg DA (1984) The effects of conditioning on probands to correct for multiple ascertainment. Am J Hum Genet 36:1298–1308

Cannings C, Thompson EA (1977) Ascertainment in the sequential sampling of pedigrees. Clin Genet 12:208–212

Davidian P, Carroll R (1988) A note on extended quasi-likelihood. J R Stat Soc [B] 50:74–82

Elandt-Johnson RC (1970) Segregation analysis for complex modes of inheritance. Am J Hum Genet 22:129–144

Elston RC (1995) 'Twixt cup and lip: how intractable is the ascertainment problem? Am J Hum Genet 56:15–17

Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21:523–542

Elston RC, Sobel E (1979) Sampling considerations in the gathering and analysis of pedigree data. Am J Hum Genet 31:62–69

Ewens WJ, Shute NCE (1986) A resolution of the ascertainment sampling problem. I. Theory. Theor Popul Biol 30: 523–542

Fisher RS (1934) The effects of methods of ascertainment upon the estimation of frequencies. Ann Eugenics 6:13–25

Heyde C (1989) On efficiency for quasi-likelihood and composite quasi-likelihood methods. In: Dodge Y (ed) Statistical data analysis and inference. North-Holland, New York, pp 209–213

Hodge SE (1988) Conditioning on subsets of the data: applications to ascertainment and other genetic problems. Am J Hum Genet 43:364–373

Hodge SE, Boehnke M (1986) A note on Cannings and Thompson's sequential sampling scheme for pedigrees. Am J Hum Genet 39:274–281

Gong G, Samaniego FJ (1981) Pseudo maximum likelihood estimation: theory and applications. Ann Stat 9:861–869

Gourieroux C, Monfort A, Trognon A (1984) Pseudo maximum likelihood methods: theory. Econometrica 52:681–700

Kalbfleisch JD (1986) Pseudo-likelihood. In: Kotz S (ed) Encyclopedia of statistical science. Vol 7. John Wiley & Sons, New York, pp 324–327

Khoury MJ, Beaty TH, Cohen BH (1993) Fundamentals of genetic epidemiology. Oxford University Press, New York

Morton NE (1959) Genetic tests under incomplete ascertainment. Am J Hum Genet 11:1–16

Thompson EA, Cannings C (1979) Sampling schemes and ascertainment. In: Sing CF, Skolnick M (eds) Genetic analysis of common disease: applications to predictive factors in coronary disease. Alan R Liss, New York

Shute NCE, Ewens WJ (1988*a*) A resolution of the ascertainment sampling problem. II. Generalizations and numerical results. Am Hum Genet 43:374–386

——— (1988*b*) A resolution of the ascertainment sampling problem. III. Pedigrees. Am J Hum Genet 43:387–395

Vieland JV, Hodge SE (1995) Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. Am J Hum Genet 56:33–43