Statistics
in Medicine

Research Article

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Likelihood-based analysis of outcome dependent sampling designs with longitudinal data: supplementary materials

Leila R. Zelnick[a]*, Jonathan S. Schildcrout[b], and Patrick J. Heagerty[c]

[a] *Department of Medicine, University of Washington, Seattle, WA, 98195, USA*
[b] *Department of Biostatistics, Vanderbilt School of Medicine, Nashville, TN, 37203, USA*
[c] *Department of Biostatistics, University of Washington, Seattle, WA, 98195, USA*
[*] *Correspondence to: Department of Medicine, University of Washington, Seattle, WA, 98195, USA. Email: lzelnick@uw.edu*

Appendix A: Likelihood-based estimator using subsampled subjects only and with covariates

The likelihood SO,WC (Equation 4) differs from that of likelihood SO,NC (Equation 3) by the second term in Equation 5. Without loss of generality, suppose M_i is a binary marker. Define $a_0 = AC_0(M_i = 0, \boldsymbol{T}_i; \boldsymbol{\theta})$ and $a_1 = AC_0(M_i = 1, \boldsymbol{T}_i; \boldsymbol{\theta})$. Under a complete and balanced design, $P(M_i = m_i, \boldsymbol{T}_i; \boldsymbol{\Gamma}) = P(M_i = m_i; \boldsymbol{\Gamma})$ and $\mathbb{E}_{M_i|\boldsymbol{T}_i}[AC_0(M_i, \boldsymbol{T}_i)] = \mathbb{E}_{M_i}[AC_0(M_i, \boldsymbol{T}_i)]$, and the ith person's ascertainment correction contribution to the likelihood (Equation 5 in main text) can be written as follows:

$$\begin{aligned}
\frac{AC_0(M_i, \boldsymbol{T}_i; \boldsymbol{\theta}) \cdot P(M_i = m_i, \boldsymbol{T}_i; \boldsymbol{\Gamma})}{\mathbb{E}_{M_i|\boldsymbol{T}_i}[AC_0(M_i, \boldsymbol{T}_i; \boldsymbol{\theta})]} &= \frac{AC_0(M_i, \boldsymbol{T}_i; \boldsymbol{\theta}) \cdot P(M_i = m_i; \boldsymbol{\Gamma})}{\mathbb{E}_{M_i}[AC_0(M_i, \boldsymbol{T}_i; \boldsymbol{\theta})]} \\
&= \frac{a_1^m a_0^{1-m} \cdot p^m (1-p)^{1-m}}{pa_1 + (1-p)a_0} \\
&= \frac{(a_1 p)^m \left(a_0(1-p)\right)^{1-m}}{pa_1 + (1-p)a_0} \\
&= \frac{\left(\frac{a_1}{a_0} p\right)^m (1-p)^{1-m}}{\frac{a_1}{a_0} p + 1 - p} \\
&= \frac{\frac{a_0}{a_1 p}}{\frac{a_0}{a_1 p}} \cdot \frac{\left(\frac{a_1}{a_0} p\right)^m (1-p)^{1-m}}{\frac{a_1}{a_0} p + 1 - p} \\
&= \left(\frac{1}{1 + \frac{a_0}{a_1} \cdot \frac{1-p}{p}}\right)^m \left(\frac{\frac{a_0}{a_1} \cdot \frac{1-p}{p}}{1 + \frac{a_0}{a_1} \cdot \frac{1-p}{p}}\right)^{1-m} \\
&= \xi^m (1-\xi)^{1-m},
\end{aligned}$$

where $\xi = \left(\frac{1}{1 + \frac{a_0}{a_1} \cdot \frac{1-p}{p}}\right)$. For balanced and complete designs, a_0 and a_1 do not vary across subjects, so this term's contribution to likelihood SO,WC (Equation 4) can be seen as a reparameterization of the marginal distribution of M, which contains no information about $\boldsymbol{\theta}$. Thus, adding this term to the likelihood SO, NC (Equation 3) will add no information to the resulting inference, and in fact will yield the same estimate, up to the chosen tolerance of the Newton-Raphson algorithm.

Supplementary Table 1

Supplementary Table 1. Percent bias and relative efficiency for likelihood-based ODS estimators, under low subject-to-subject heterogeneity. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator. NA for percent bias indicates that percent bias is undefined.

	Estimator								
Design	**Likelihood analysis**	β_0	β_T	β_M	$\beta_{M\times T}$	γ_1	γ_2	γ_3	γ_4
Full cohort	Standard	0[3.91]	0[3.88]	-1[3.98]	0[4.02]	0[4.14]	NA[4.39]	0[4.14]	NA[3.71]
Random sample	Standard	0[1.00]	1[1.00]	0[1.00]	-1[1.00]	-2[1.00]	NA[1.00]	-1[1.00]	NA[1.00]
Intercept	SO, NC	0[1.65]	-1[1.00]	1[1.48]	-1[1.07]	-1[1.52]	NA[1.87]	1[0.95]	NA[0.92]
	SO, WC	0[1.65]	-1[1.00]	1[1.48]	-1[1.07]	-1[1.52]	NA[1.87]	1[0.95]	NA[0.92]
	SU, NC	0[2.96]	0[1.44]	2[1.54]	0[1.14]	0[3.72]	NA[3.69]	0[2.43]	NA[3.71]
	$SU, NC + PI$	0[3.42]	0[2.56]	2[1.56]	-1[1.20]	0[3.72]	NA[3.89]	0[2.87]	NA[3.71]
	SU, WC	0[3.44]	0[2.86]	2[1.52]	0[1.18]	0[3.73]	NA[3.89]	0[2.94]	NA[3.71]
	UC	0[3.43]	0[2.86]	2[1.51]	0[1.18]	0[3.83]	NA[3.90]	0[2.94]	NA[3.70]
Slope	SO, NC	0[1.08]	0[1.80]	2[1.17]	2[1.37]	-1[0.96]	NA[1.70]	1[1.36]	NA[1.03]
	SO, WC	0[1.08]	0[1.80]	2[1.17]	2[1.37]	-1[0.96]	NA[1.70]	1[1.36]	NA[1.03]
	SU, NC	0[2.69]	0[2.24]	3[1.10]	2[1.39]	0[3.65]	NA[3.50]	0[2.57]	NA[3.71]
	$SU, NC + PI$	0[3.14]	0[3.16]	2[1.18]	1[1.46]	0[3.83]	NA[3.65]	0[3.06]	NA[3.71]
	SU, WC	0[3.16]	0[3.42]	3[1.07]	2[1.42]	0[3.82]	NA[3.71]	0[3.14]	NA[3.71]
	UC	0[3.19]	0[3.43]	3[1.15]	2[1.48]	0[3.85]	NA[3.75]	0[3.33]	NA[3.71]

Supplementary Table 2

Supplementary Table 2. Percent bias and relative efficiency for likelihood-based ODS estimators, under high subject-to-subject heterogeneity. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 4, \sigma^2_e = 4,$ and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator. NA for percent bias indicates that percent bias is undefined.

Design	**Estimator Likelihood analysis**	β_0	β_T	β_M	$\beta_{M\times T}$	γ_1	γ_2	γ_3	γ_4
Full cohort	Standard	0[3.98]	0[3.97]	0[4.09]	0[4.46]	-1[4.39]	NA[4.61]	0[4.09]	0[3.90]
Random sample	Standard	0[1.00]	1[1.00]	0[1.00]	-1[1.00]	-2[1.00]	NA[1.00]	-1[1.00]	0[1.00]
Intercept	$SO + NC$	0[1.62]	1[0.97]	0[1.72]	2[1.15]	-1[1.38]	NA[1.94]	-1[1.01]	0[1.02]
	SO, WC	0[1.62]	1[0.97]	0[1.72]	2[1.15]	-1[1.38]	NA[1.94]	-1[1.01]	0[1.02]
	SU, NC	0[2.33]	6[1.53]	0[1.61]	2[1.03]	-1[3.70]	NA[4.16]	-1[3.73]	0[3.88]
	$SU, NC + PI$	0[3.32]	-1[2.28]	-1[1.72]	2[1.11]	-1[3.80]	NA[4.28]	0[3.92]	0[3.89]
	SU, WC	0[3.39]	0[2.98]	-1[1.69]	2[1.08]	-1[3.82]	NA[4.30]	0[3.93]	0[3.89]
	UC	0[3.38]	0[2.97]	-1[1.71]	2[1.08]	-1[4.01]	NA[4.33]	0[3.92]	0[3.89]
Slope	SO, NC	0[0.96]	0[1.68]	0[1.07]	2[1.84]	-3[1.12]	NA[1.89]	-1[1.40]	0[1.06]
	SO, WC	0[0.96]	0[1.68]	0[1.07]	2[1.84]	-3[1.12]	NA[1.89]	-1[1.40]	0[1.06]
	SU, NC	0[1.55]	4[2.37]	0[0.95]	1[1.74]	-1[4.02]	NA[4.32]	0[3.39]	0[3.88]
	$SU, NC + PI$	0[2.96]	0[3.53]	0[1.07]	2[1.84]	-1[4.18]	NA[4.36]	0[3.43]	0[3.89]
	SU, WC	0[2.96]	0[3.55]	1[1.05]	2[1.84]	-1[4.18]	NA[4.39]	0[3.43]	0[3.89]
	UC	0[2.97]	0[3.56]	0[1.07]	2[1.86]	-1[4.17]	NA[4.40]	0[3.69]	0[3.89]

Supplementary Table 3

Supplementary Table 3. Percent bias and relative efficiency for time-specific predicted means and predicted difference in means, under high subject-to-subject heterogeneity. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 4, \sigma^2_e = 4$, and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator.

Estimator	$\mu_0(1)$	$\mu_1(1)$	Δ_1	$\mu_0(6)$	$\mu_1(6)$	Δ_6
SO, NC_{int}	0 [1.36]	0 [1.41]	-4 [1.37]	0 [0.99]	0 [1.16]	2 [1.13]
SO, WC_{int}	0 [1.36]	0 [1.41]	-4 [1.37]	0 [0.99]	0 [1.16]	2 [1.13]
SU, NC_{int}	0 [2.20]	0 [1.38]	-4 [1.23]	0 [1.59]	0 [1.12]	2 [1.00]
$SU, NC + PI_{int}$	0 [3.46]	0 [1.44]	-5 [1.34]	-1 [2.99]	0 [1.17]	3 [1.09]
SU, WC_{int}	0 [3.49]	0 [1.40]	-5 [1.32]	0 [3.07]	0 [1.14]	2 [1.06]
UC_{int}	0 [3.50]	0 [1.40]	-6 [1.32]	0 [3.07]	1 [1.14]	3 [1.06]
SO, NC_{slope}	0 [1.14]	0 [1.13]	-4 [1.11]	0 [1.65]	0 [1.74]	2 [1.71]
SO, WC_{slope}	0 [1.14]	0 [1.13]	-4 [1.11]	0 [1.65]	0 [1.74]	2 [1.71]
SU, NC_{slope}	0 [1.90]	0 [1.13]	-2 [1.00]	-1 [2.43]	0 [1.74]	2 [1.60]
$SU, NC + PI_{slope}$	0 [3.24]	0 [1.19]	-3 [1.09]	0 [3.57]	0 [1.77]	2 [1.70]
SU, WC_{slope}	0 [3.22]	0 [1.18]	-2 [1.08]	0 [3.58]	0 [1.78]	3 [1.70]
UC_{slope}	0 [3.23]	0 [1.21]	-3 [1.11]	0 [3.58]	0 [1.80]	2 [1.72]

Supplementary Table 4

Supplementary Table 4. Percent bias and relative efficiency for time-specific predicted means and predicted difference in means, under low subject-to-subject heterogeneity. Results shown summarize 1000 replications with $N = 1000, \boldsymbol{\beta} = (\beta_0, \beta_T, \beta_M, \beta_{M\times T}) = (10, -0.25, -0.75, 0.5), \sigma^2_{b_0} = 4, \sigma^2_{b_1} = 0.25, \sigma^2_e = 1,$ and $\rho = 0$. $N_S = 250$ subjects were subsampled on average. Percent bias defined as the 100 $\times$ the difference between estimator mean and parameter value, divided by parameter value, and relative efficiency for an estimator is defined as the ratio of variances between a random sample of $N_S = 250$ and the estimator.

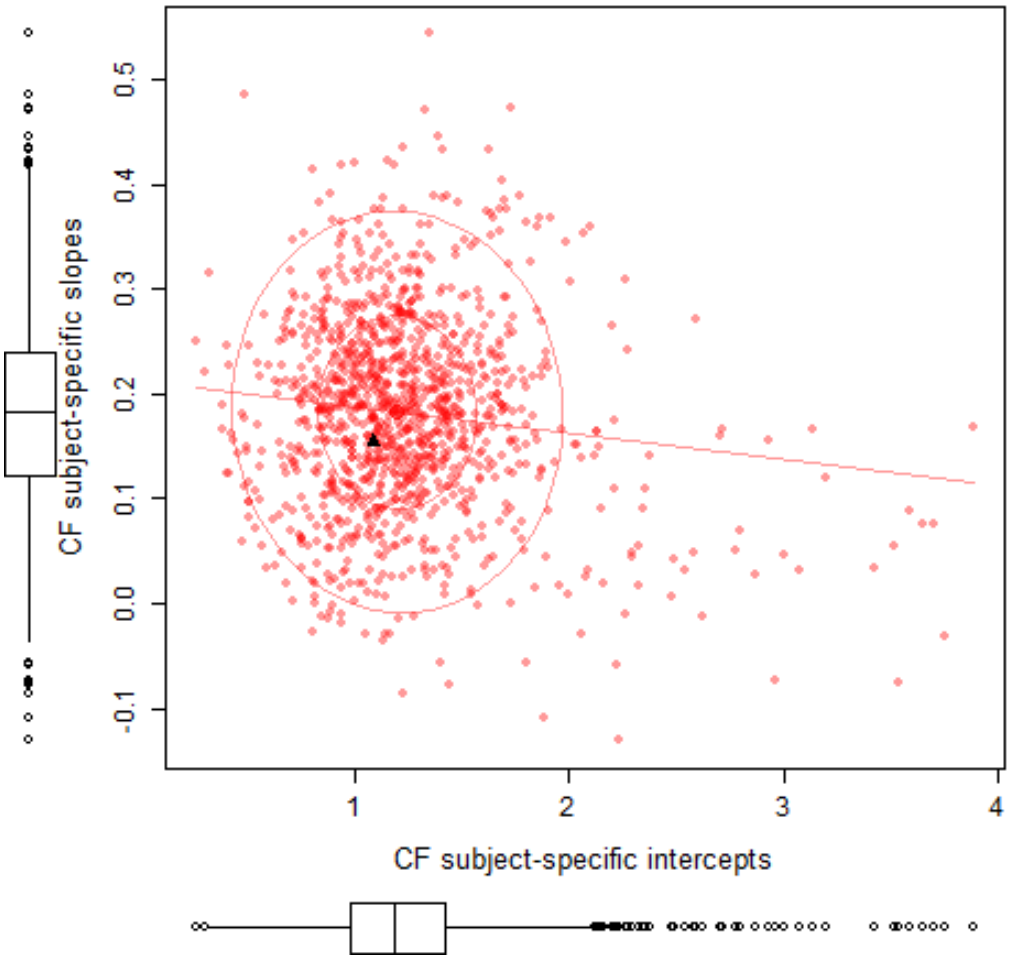
Estimator	$\mu_0(1)$	$\mu_1(1)$	Δ_1	$\mu_0(6)$	$\mu_1(6)$	Δ_6
SO, NC_{int}	0 [1.62]	0 [1.46]	5 [1.46]	0 [1.15]	0 [1.16]	-1 [1.15]
SO, WC_{int}	0 [1.62]	0 [1.46]	5 [1.46]	0 1.15]	0 [1.16]	-1 [1.15]
SU, NC_{int}	0 [3.24]	0 [1.54]	5 [1.49]	0 [2.02]	0 [1.15]	0 [1.18]
$SU, NC + PI_{int}$	0 [3.47]	0 [1.57]	6 [1.54]	0 [2.91]	0 [1.32]	-2 [1.27]
SU, WC_{int}	0 [3.48]	0 [1.52]	6 [1.50]	0 [3.13]	0 [1.29]	-1 [1.23]
UC_{int}	0 [3.48]	0 [1.51]	5 [1.49]	0 [3.13]	0 [1.29]	-1 [1.23]
SO, NC_{slope}	0 [1.10]	0 [1.17]	1 [1.21]	0 [1.53]	0 [1.31]	2 [1.40]
SO, WC_{slope}	0 [1.10]	0 [1.17]	1 [1.21]	0 [1.53]	0 [1.31]	2 [1.40]
SU, NC_{slope}	0 [2.94]	0 [1.17]	4 [1.12]	0 [2.84]	0 [1.28]	2 [1.34]
$SU, NC + PI_{slope}$	0 [3.26]	0 [1.26]	4 [1.21]	0 [3.42]	0 [1.42]	1 [1.45]
SU, WC_{slope}	0 [3.24]	0 [1.16]	6 [1.12]	0 [3.53]	0 [1.40]	2 [1.42]
UC_{slope}	0 [3.30]	0 [1.21]	4 [1.18]	0 [3.59]	0 [1.41]	1 [1.44]

Supplementary Table 5. Baseline characteristics of Cystic Fibrosis Foundation Registry cohort (N = 3,141).

Characteristic	Value
N	3,141
Age (years)	8.8 (1.1)
Male sex, N (%)	1584 (50.4)
Height (cm)	125.4 (8.4)
Weight (kg)	25.9 (6.0)
Presence of *S. aureus*, N (%)	2206 (70.2)
FEV_1 (L)	1.4 (0.4)

Values are mean (SD), except as noted.

Supplementary Figure 1



Supplementary Figure 1. Subject-specific intercepts and slopes resulting from regressing a CF patient's FEV_1 longitudinal outcome on time. The distribution of intercepts and slopes do not appear to be clearly inconsistent with bivariate normality.