Nested Case-Control Studies With Multiple Measurements: Estimating Marginal Relationships

Joshua N. Sampson* Paul S. Albert† Mark P. Purdue‡

Appendix A1

For reasons that will become clear later, we assume that X_{ij} is actually a Q-length vector (e.g. however when considering a single biomarker and no time-varying effect, Q = 1). For any matched set, the probability that the case has the disease follows equation 1. Here, i and i' index a case and a matched control respectively; and j and j' are any of the measurements with $j, j' \in \{1, ..., J\}$.

$$P_{ii'jj'} = P(Y_i(T_i) = 1 | Y_i(T_i) + Y_{i'}(T_i) = 1, X_{ij\cdot}, X_{i'j'\cdot}) = \frac{exp(\sum_q \beta_q X_{ijq})}{exp(\sum_q \beta_q X_{ijq}) + exp(\sum_q \beta_q X_{i'j'q})} \quad (1)$$

For combination k (i.e. there are $K = J \times J$ combinations), let $j(k)$ and $j'(k)$ be the

*Corresponding Author, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville MD, USA

†Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville MD, USA

‡Occupational and Environmental Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville MD, USA

corresponding measures and abbreviate $P_{ik} \equiv P_{ii'j(k)j'(k)}$, $X_{ik} \equiv X_{ij(k)}$, and $X_{i'k} \equiv X_{i'j(k)}$. Then, our pseudo-likelihood for our study is defined by

$$L = \prod_{i=1}^{N_A} \prod_{k=1}^{K} P_{ik}(\beta) = \prod_{i=1}^{N_A} \prod_{k=1}^{K} \frac{exp(\sum_q \beta_q X_{ikq})}{exp(\sum_q \beta_q X_{ikq}) + exp(\sum_q \beta_q X_{i'kq})}$$

Note, this would have been a true likelihood had all $N_A \times K$ pairs been independent of each other. The pseudo-log-likelihood can be defined by

$$\ell = \sum_{i=1}^{N_A} \sum_{k=1}^{K} (\sum_q \beta_q X_{ikq} - log(exp(\sum_p \beta_p X_{ikq}) + exp(\sum_q \beta_q X_{i'kq}))$$

The pseudo-score equations (aka estimating equations)

$$U(\beta) = \sum_{i=1}^{N_A} \sum_{k=1}^{K} (X_{ik\cdot} - \overline{X}_{ik\cdot}(\beta)) = \sum_{i}^{N_A} U_{i\cdot}(\beta)$$

where $\overline{X}_{ik\cdot}(\beta) = P_{ik}(\beta) X_{ik\cdot} + (1 - P_{ik}(\beta)) X_{i'k\cdot}$. Note, U is q-length column vector. Then, our estimate of β is the solution to $U(\beta) = 0$.

$$U(\hat{\beta}) = 0$$

We can then estimate the variance, Σ, of $\hat{\beta}$ by $\hat{\Sigma} = (D^{-1})^t V D^{-1}$.

$$V = \sum_i U_i(\hat{\beta})(U_i(\hat{\beta}))^t$$

and D is a q $\times$ q matrix where the $D[t,u]$ entry is $\partial U_t / \partial \beta_u$

$$D[t,u] = \sum_{i=1}^{N_A} \sum_{k=1}^{K} P_{ik}(\hat{\beta})(1 - P_{ik}(\hat{\beta}))(X_{iku} - X_{i'ku})(X_{ikt} - X_{i'kt})$$

Appendix A2

We will assume that $\beta(t - R_{ij})$ has a specific form, $\beta(t - R_{ij}) = \sum_{q=1}^{Q} \beta_q f_q(t - R_{ij})$, where $\{f_1, ..., f_q\}$ are a set of known basis functions. Letting $X'_{ijq} = X'_{ijq}(R_{ij}, t) = X_{ij} f_q(t - R_{ij})$

$$log(P(Y_i(t) = 1|T_i \geq t, X_{ij})) = h_0(t) + \sum_{q=1}^{Q} \beta_q f_k(t - R'_{ij}) X_{ij} = h_0(t) + \sum_{q=1}^{Q} \beta_q X'_{ijq}$$

Then, for our case-control study, we can define $X'_{ijq} = f_q(T_i - R_{ij}) X_{ij}$ and describe the conditional probability for a case/control pair of measures by

$$P_{ik} = P(Y_i(T_i) = 1|Y_i(T_i) + Y_{i'}(T_i) = 1, X'_{ij\cdot}, X'_{i'j'\cdot}) = \frac{exp(\sum_q \beta_q X'_{ijq})}{exp(\sum_q \beta_q X'_{ijq}) + exp(\sum_q \beta X'_{i'j'q})}$$

We can now estimate $\{\beta_1, ..., \beta_Q\}$ and their confidence intervals using the above steps. These estimates, in turn, allow us to estimate $\hat{\beta}(\cdot)$ by $\sum_{q=1}^{Q} \hat{\beta}_q f_q(\cdot)$ and point-wise confidence intervals for $\beta(t)$ by noting that the variance for $\beta(t)$ is $[f_1(t), ..., f_q(t)]\hat{\Sigma}[f_1(t), ..., f_Q(t)]^t$.

We may let $f_1(t) = 1$ and be independent of time. Then, we test whether the log(HR) varies over time by testing $H_0 : \beta_{-1} = [\beta_2, ..., \beta_Q]^t = 0$. Letting Σ_{M11} be the $(Q-1) \times (Q-1)$ matrix that omits the first row and column of Σ, we can define our test statistic to be $S = \beta_{-1}^t \Sigma_{M11}^{-1} \beta_{-1}$ and our p-value for testing H_0 by $p = P(\chi^2_{Q-1} < S)$.

Appendix A3

Here, we review Generalized Estimating Equations (GEE). For individual (aka cluster) i, we have J data points: $\{X_{i11}, ..., X_{i1q}, Y_i\}, \ldots, \{X_{iJ1}, ..., X_{iJq}, Y_i\}$. Note, compared to other clustered data sets, the value Y_i is the same for each data point within a cluster. However, we review the GEE theory here for the sole purpose of showing that, at no point, does the theory claim that Y_i's cannot be identical.

We can use the logistic-link and estimate the odds ratio

$$\mu_{ij}(\beta) = \frac{exp(\sum_q \beta_q X_{ijq})}{1 + exp(\sum_q \beta_q X_{ijq})}$$

We estimate β by the values that solve the GEE (i.e.$GEE(\hat{\beta}) = 0$) where

$$GEE(\beta) = \sum_i X_i(Y_i - \mu_i(\beta))$$

Here, $Y_i - \mu_i$ is the $J \times 1$ vector

$$\begin{bmatrix} Y_i - \mu_{i1}(\beta) \\ ... \\ Y_i - \mu_{iJ}(\beta) \end{bmatrix}$$

and X_i is the $J \times Q$ matrix

$$\begin{bmatrix} X_{i11} & X_{i12} & ... & X_{i1Q} \\ X_{i21} & X_{i22} & ... & X_{i2Q} \\ \vdots & \vdots & ... & \vdots \\ X_{iJ1} & X_{iJ2} & ... & X_{iJQ} \end{bmatrix}$$

The Taylor Series expansion around the true value of β, β^*, yields

$$\sum_i X_i^t(Y_i - \mu_i(\hat{\beta})) \approx \sum_i X_i^t(Y_i - \mu_i(\beta^*)) + \sum_i (X_i^t V_i(\beta^*) X_i)(\hat{\beta} - \beta^*)$$

where

V_i is the $J \times J$ variance matrix

$$V_i(\beta) = \begin{bmatrix} \mu_{i1}(\beta)(1-\mu_{i1}(\beta)) & 0 & ... & 0 \\ 0 & \mu_{i2}(\beta)(1-\mu_{i2}(\beta)) & ... & 0 \\ \vdots & \vdots & ... & \vdots \\ 0 & 0 & ... & \mu_{iJ}(\beta)(1-\mu_{iJ}(\beta)) \end{bmatrix}$$

which leads to

$$(\hat{\beta} - \beta^*) \approx (\sum_i X_i^t V_i(\beta^*) X_i)^{-1} (\sum_i X_i^t (Y_i - \mu_i(\beta^*))$$

Since, $Y_i - \mu_i(\beta^*)$ is still a random variable and assuming $V_i(\beta^*) \approx V_i(\hat{\beta})$ we can approximate the variance by

$$var((\hat{\beta} - \beta^*)) \approx (\sum_i X_i^t V_i(\hat{\beta}) X_i)^{-1} (\sum_i X_i^t (Y_i - \mu_i(\hat{\beta})(Y_i - \mu_i(\hat{\beta}))^t X_i) (\sum_i X_i^t V_i(\hat{\beta}) X_i)^{-1}$$

Appendix A4

To use clogitRV, please download and load the R package using

install.packages("devtools")

devtools::install_github("sampsonj74/clogitRV")

library(clogitRV)

And to see examples of how to run clogitRV, including how to model time-varying coefficients, please scroll to the bottom of the help page accessed by

?clogitRV