Counter matching

James A. Deddens

COUNTER-MATCHING IN NESTED CASE CONTROL STUDIES

by

JAMES A. DEDDENS
DEPT OF MATHEMATICAL SCIENCES
UNIVERSITY OF CINCINNATI
and NIOSH

Joint work with
KYLE STEENLAND, NIOSH, Emory University

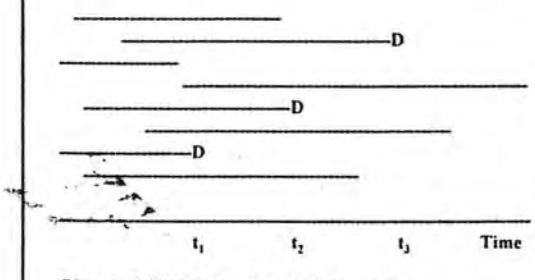
INCREASED PRECISION USING COUNTER-MATCHING IN NESTED CASECONTROL STUDIES, EPIDEMIOLOGY, 8(1997)238-242, COMMENTARY PP-227-229

MORTALITY/MORBIDITY STUDIES INVOLVING TIME UNTIL AN EVENT

- OUTCOME OF INTEREST
  (eg, DEATH FROM LUNG CANCER)
- EXPOSURE OF INTEREST
  (eg, CUMULATIVE EXPOSURE TO DUST)
- OFTEN THE EXPOSURE OF INTEREST IS EXPENSIVE TO MEASURE ON ALL SUBJECTS

RISK SET APPROACH TO COHORT DATA

D = failure

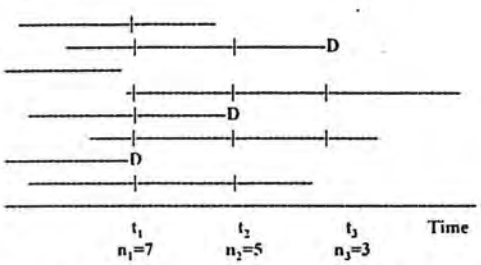


Observe failure times $0 < t_1 < t_2 < t_3 \ldots$

RISK SET APPROACH TO COHORT DATA

D = failure

| = at risk



R_j = j^{th} risk set = all members of cohort "at risk" just prior to t_j

---

HAZARD RATE: $h(t, Z_j(t)) = h_0(t)\, f(\beta, Z_j(t))$
where $h_0(t)$ is the baseline hazard function, and f is a some function of exposure and other covariates

COX REGRESSION RELATIVE RISK MODEL

$$R(\beta, Z_j(t)) = h(t) / h_0(t) = \exp(\beta\, Z_j(t))$$

ESTIMATION BASED ON PARTIAL LIKELIHOOD

$$PL = \prod_j \frac{R[\beta, Z_{i_j}(t_j)]}{\sum_{i \in R_j} R[\beta, Z_i(t_j)]}$$

where R_j is the risk set at time t_j and i_j is the individual in the risk set R_j failing at time t_j

---

NESTED CASE-CONTROL SAMPLING

(Thomas 1977, Oakes 1981)

- At time t_j select a random sample of controls from among the $n(t_j)$-1 non-failures in the risk set R_j
- Need only collect exposure and confounder data on those in the "sampled risk set". If cost of collecting data is expensive, NCCS can save considerable time and money
- Smaller data set means faster estimation of parameters in models (if large number of failures)
- Often 5 to 10 controls per case are randomly selected
- Sometimes controls are matched to cases by age, etc

NESTED CASE-CONTROL SAMPLING

- Estimation is done using the usual Partial Likelihood
- Usually one forms the risk sets, selects the controls, and uses PROC PHREG with STRATA=risk set, and defining the case's failure time to occur before the time assigned to the controls
- Efficiency is known to be (m-1)/m when the exposure response coefficient β is 0, where efficiency is defined as the ratio of the variance of the estimator of β using the full cohort to the variance of the estimator of β using Nested Case-Control Sampling

$$EFF = \sigma_{FULL}^2 / \sigma_{NCCS}^2$$

COUNTER-MATCHING IN NESTED CASE-CONTROL(CM)

(Langholz & Clayton 1994, Langholz & Borgan, etc)

GOAL: FURTHER IMPROVE EFFICIENCY OF NCCS

SUPPOSE for all individuals in the cohort one knows a covariate X(t) (surrogate) that is correlated with the true exposure Z(t)

- Suppose each risk set R_j is divided into k strata (based on X(t)) containing $n_1, n_2, n_3, \ldots, n_k$ subjects. One then randomly selects $m_1, m_2, m_3, \ldots, m_k$ subjects from each strata in such a way that if the case is in strata i one selects only m_i-1 controls from strata i.
- 1-1 counter-matching requires dividing each risk set into 2 strata and then picking 1 control from the strata not containing the case.
- Maximize the heterogeneity of exposure in the sampled risk sets

- Since counter-matching does not involve random sampling, one must compensate by introducing weights into the partial likelihood equation: For each risk set R_j let:

$$\text{weight} = W_i(t_j) = n_i / m_i$$

where n_i is the number of subjects in strata i, and m_i is the number of controls selected from strata i.

- Estimates are then based on the modified partial likelihood.

$$PL = \prod_j \frac{R[\beta \cdot Z_i(t_j)] \cdot W_i(t_j)}{\sum_{l \in R_j} R[\beta \cdot Z_l(t_j)] \cdot W_l(t_j)}$$

- Weights can be incorporated into the estimation process using the offset function in PROC PHREG

- If the surrogate X(t) is a 0/1 variable, then 1-1 counter-matching requires picking a control with X(t)=0 if the case has X(t)=1, and visa versa.
- This maximizes the number of "discordant" pairs with respect to the surrogate X(t) and possibly with respect to Z(t).
- If the surrogate is a 0/1 variable, then 3-1 counter-matching might consist picking 2 controls with X(t)=0 and 1 control with X(t)=1 if the case has X(t)=1, etc (m_1=2, m_2=2), etc

- If the surrogate X(t) is a continuous variable, then one might create 4 strata for each risk set R_j, based on cutpoints of X(t)
- For 3-1 counter-matching one would pick one control from each strata except the strata of the case.
- For 7-1 counter-matching one might pick 1 control from the case's strata, and 2 from each of the other three strata (m_1= m_2= m_3= m_4=2). Of course, there are many other possible choices of m_1, m_2, m_3, m_4 summing to 8
- The question is then "How to form the strata in each risk set?"
- The natural method would be to use "quartiles" (or percentiles)

There are several ways to pick the percentiles. For illustration purposes suppose one does a 3-1 counter-matching with 4 strata.

METHOD 1: Pick quartiles of the covariate X(t) within each risk set. If equal numbers are sampled from each quartile, then all weights are equal, and hence cancel out of the partial likelihood.

METHOD 2: Pick quartiles of the covariate X(t) for the cases. Then weights will vary, since the size of the strata will vary within each risk set and between risk sets.

In our study METHOD 2 proved to be more efficient.

Suppose X(t) and Z(t) are both 0/1 variables, with sensitivity = P(X=1|Z=1) specificity = P(X=0|Z=0), I.e., with a given degree of association between the surrogate and the true exposure..

Define the asymptotic relative efficiency (ARE) :

$$ARE = \sigma_{NCCS}^2 / \sigma_{CM}^2$$

as the ratio of the variance of the simple 1-1 nested case-control sampling to the variance of 1-1 counter-matching.

Under $H_0 : \beta=0$, Langholz and Clayton showed:

ARE	SPECIFICITY		
SENSITIVITY	1.0	0.9	0.8
0.4	0.80	0.84	0.88
0.5	1.00(1)	1.00	1.00
0.7	1.40(2)	1.32	1.24
0.8	1.60(4)	1.48	1.36
0.9	1.80(19)	1.64	1.48

- 1-1 counter-matching is "better" than simple 1-1 nested as long as the sensitivity is greater than .5
- If sensitivity is .8 and specificity is 1.0 then 1-1 counter-matching is equivalent to 4-1 nested sampling

- Langholz & Clayton also showed that if exp(β)=4 and P(Z=1)=.1 then

ARE	SPECIFICITY		
SENSITIVITY	1.0	0.9	0.8
0.4	1.79	1.38	1.14
0.5	2.17	1.65	1.33
0.7	2.89	2.17	1.69
0.8	3.23	2,41	1.86
0.9	3.54	2.64	2.02

- Langholz & Clayton also showed that relative efficiency increased with rarity of exposure.

In general the efficiency of CM versus NCCS depends on:
(Borgan & Olsen 1999)

- sensitivity and specificity
- fraction exposed
- baseline hazard
- relative risk coefficient β
- censoring distribution

Most general results have been developed for dichotomous (0/1) exposures $Z(t)$ and their dichotomous (0/1) surrogates $X(t)$.

We were interested in continuous exposures and surrogates.

CRYSTALLINE SILICA EXPOSURE AND SILICOSIS IN GOLD MINERS

- cohort of 3,300 gold miners, 1940-1965
- 170 cases of silicosis
- cumulative silica exposure estimated via repeated dust sampling
- case control study would avoid much cost of dust collection for whole cohort
- a surrogate for cumulative exposure, duration employed is known AND highly correlated with cumulative silica exposure

Steenland & Brown, Silicosis among gold miners, Amer. J. Public Health, 85(1995)1372-1377

GOLD MINERS COHORT

We wanted to see the effect of various sampling strategies: full cohort, nested case-control sampling (m=3,10,20,100), and counter-matching (m=3).

We performed 50 sample selections for each method

We performed the analysis on the full cohort, which showed a very strong dose-response, and also on a reconstructed data set with a reduced dose response.

RESULTS : Average estimates of the dose response coefficient β and its standard error:

DESIGN	ORIGINAL	REDUCED
FULL	1.56(.110)	0.844(.078)
NCCS(100)	1.56(.112)	0.843(.078)
NCCS(20)	1.54(.119)	0.845(.081)
NCCS(10)	1.54(.127)	0.840(.084)
NCCS(3)	1.46(.155)	0.876(.103)
CCM(3) M1	1.53(.157)	0.842(.093)
CCM(3) M2	1.61(.125)	0.865(.085)

- averages based on 50 sample selections

RESULTS : Estimated relative efficiency = $100*\sigma_{FULL}^2/\sigma_{CM}^2$

DESIGN	ORIGINAL	REDUCED
NCCS(100)	96.4%	98.8%
NCCS(20)	85.4%	91.9%
NCCS(10)	73.9%	85.7%
NCCS(3)	50.0%	59.1%
CCM(3) M1	49.0%	71.6%
CCM(3) M2	77.7%	83.1%

- averages based on 50 sample selections
- counter-matching m=3 equivalent to matching m=10 to 15
- NCCS and CM are more efficient for reduced dose-response effect

CONCLUSIONS:

- Method 2 (strata formed based on cases' exposure) is superior to Method 1 (strata formed based on all members of risk set)
- Counter-matching is superior to NCCS
- Counter-matching with 3 controls is about the same as NCCS with 10-15 controls, i.e., counter-matching achieves same efficiency as NCCS using fewer selected controls (LESS EXPENSIVE)
- Counter-matching is even more efficient for reduced dose response situation

GENERAL CONSIDERATIONS

- Counter-matching requires knowing some information on all subjects in the cohort, before randomly sampling
- if the information (eg, duration) is correlated with exposure then one should use it to COUNTER-MATCH
- on the other hand, if the information is a confounder, then one should use it to MATCH
- counter-matching on a confounder will result in a loss of efficiency for estimating exposure
- matching on a variable related to exposure will result in being unable to estimate the exposure effect

REFERENCES

Steenland K. & Deddens JA. Increase precision using counter-matching in nested case-control studies. Epidemiology. 8.(1997)238-242. Commentary pp-227-229

Langholz B. & Clayton D. Sampling strategies in nested case-control studies. Env. Health Pers. 102(1994)47-51

Langholz B. & Borgan O. Counter-matching: a stratified nested case-control sampling method. Biometrika. 82(1995)69-79

Langholz B. & Goldstein L. Risk set sampling in epidemiologic studies. Stat. Sci.. 11(1996)35-53

Borgan O. & Olsen E. The efficiency of simple and counter-marched nested case-control sampling. Scan J Stat. 26(1999)493-509

Andrieu N. Goldstein AM. Thomas DC. & Langholz B. Counter-matching in studies of gene-environment interaction: Efficiency and feasibility. AJE. 153(2001)265-274

16th **EPICOH**

Congress on Epidemiology in Occupational Health

and

Jack Pepys Symposium on Prevention of Occupational Asthma

and

3rd International Congress on Women's Health: Occupation, Cancer and Reproduction

2002

Course on New methods in Epidemiology

Saturday, 14 September 2002

Barcelona, Spain

IMAS

IPI