R code examples for nested case-control studies and case-cohort studies [†]

平成 28 年 11 月 5 日

野間久史 [‡]

1. ネステッドケースコントロール研究

ネステッドケースコントロール研究の最も古典的な解析方法は、時点ごとのケース・コントロールマッチングを行ったサンプルを“matched case-control data”と見なして、ロジスティック回帰の条件付き尤度に基づいて推定する方法である [1]。これは、survival パッケージの clogit などで容易に実行することができるため、ここでは割愛する。

一方、近年の研究により、この古典的な条件付き尤度に基づく方法よりも、比例ハザードモデルの重み付き部分尤度を用いたセミパラメトリック推測のほうが、推定精度（検出力）が高いことが知られており [2, 3]、実践的にはそちらのほうが推奨できるといえるだろう。本稿でも、次節のケースコホート研究の解析に対応させて、multipleNCC パッケージを使用した重み付き部分尤度に基づく推測手法について紹介する。パッケージ全般の解説に関しては、Støer and Samuelsen [4] をご参照いただきたい。

重み付き解析は、wpl 関数で実行することができる。事例コードを以下に示す。

```
---
require(multipleNCC)

attach(CVD_Accidents)
CVD_Accidents$samplestat2 <- samplestat
CVD_Accidents$samplestat2[samplestat==3] <- 2
wpl(Surv(agestart,agestop,dead24) ~ factor(smoking3gr)+bmi+factor(sex),
data=CVD_Accidents, samplestat=CVD_Accidents$samplestat2, m=1,
match.var=cbind(CVD_Accidents$sex, CVD_Accidents$bmi), match.int=c(0,0,-2,2),
weight.method="glm")

Endpoint 1 :
Call:
wpl.formula(formula = Surv(agestart, agestop, dead24) ~ factor(smoking3gr) +
    bmi     +     factor(sex),     data     =     CVD_Accidents,     samplestat     =
CVD_Accidents$samplestat2,
```

[†] 第 27 回日本疫学会学術総会 疫学セミナー「追跡データ分析の A to Z」講演資料，平成 29 年 1 月 25 日

[‡] 大学共同利用機関法人 情報・システム研究機構 統計数理研究所 データ科学研究系，リスク解析戦略研究センター　e-mail: noma(at)ism.ac.jp，URL: http://www.ism.ac.jp/~noma/

```
    m = 1, match.var = cbind(CVD_Accidents$sex, CVD_Accidents$bmi),
    match.int = c(0, 0, -2, 2), weight.method = "glm")


                       coef exp(coef) se(coef) robust se     z       p
factor(smoking3gr)2  0.2602     1.297   0.1946    0.2406  1.08 2.8e-01
factor(smoking3gr)3  1.2404     3.457   0.1658    0.2121  5.85 5.0e-09
bmi                  0.0811     1.085   0.0163    0.0245  3.32 9.1e-04
factor(sex)2        -1.2662     0.282   0.1435    0.2137 -5.92 3.1e-09

  n= 566, number of events= 296
---
```

CVD_Accidents は、multipleNCC パッケージに実装されている事例データである。1974 年から 2000 年に行われた Norway の集団検診を受けた 3933 人の 2000 年までのフォローアップのデータがまとめられている。agestart, agestop は、追跡開始時とイベント発生 or 打ち切り時のそれぞれの年齢である。dead24 は、"death from cardiovascular disease, alcohol abuse, liver disease,violence or accidents" というイベントの指示変数である。smoking3gr は、3 水準のカテゴリカル変数で「1=never smoked, 2=former smoker, 3=smoker」と定義されている。bmi は BMI、sex は性別を表す。詳細については、"?CVD_Accidents" をご参照いただきたい。samplestat は、著者らが行った仮想的なネステッドケースコントロール研究における、サンプリングの有無を表す変数で [1]「0=non-sampled subjects in the cohort, 1=sampled controls, 2=dead from cardiovascular disease, 3=dead from alcohol abuse, liver disease, violence or accidents」という定義になる。ここでは、一般的な単一のアウトカムを対象としたネステッドケースコントロール研究を想定し、イベントの種類を区別しない samplestat2 という変数を作成している（2 と 3 をまとめて、イベントを 1 種類に統一している）。

wpl 関数の Surv 以下の式は、Cox 回帰と同様、生存時間回帰のモデルを指定したものである。data はデータセットを指定する引数、samplestat はネステッドケースコントロール研究のサンプリングの有無を表す指示変数を指定する引数である。m はマッチング比を表す数で、時点ごとのコントロールのマッチング数を指定する。match.var はマッチングに用いた変数で、match.int はキャリパーマッチングのキャリパー幅を表すベクトルとなる（0 とした場合、exact なマッチング）。weight.method は、重みを求める方法を表す。4 種類のオ

[1] 実際には、ネステッドケースコントロール研究は行われていない。このコホートを対象として 1 つのネステッドケースコントロール研究を、著者らがシミュレーションし、そのときのサンプリング結果を表す変数を加えているという意味である。

プションがあるが、デザインに基づく重み（KM）もしくは、ロジスティック回帰モデルから推定される重み（glm）の 2 つが一般的である。後者のほうが精度の高い推定値が得られる。詳細については、”?wpl” でヘルプファイルをご参照いただきたい。

wpl 関数の出力の見方については、coxph などと同じである。ただし、SE の推定値には、robust se のほうを参照してほしい（se(coef)は誤った推定値。過小推定のバイアスが加わる。P 値はロバスト分散による擬似 Wald 検定によって計算されている）。

2. ケースコホート研究

ケースコホート研究も、比例ハザードモデルによる解析には、修正部分尤度を用いる必要がある。層別を行わない単純なケースコホート研究における代表的な方法としては、Prentice [5], Self-Prentice [6], Lin-Ying [7]の 3 つの方法があるが、これは、survival パッケージの cch で実行できる。

以下は、R に実装されている Wilms 腫瘍の臨床試験のデータ [8] を事例としたプログラムであるが、関数そのものは単純な構造であるため、解析の内容については、コードを見ていただければおわかりいただけるだろう。詳細については、各データセット，関数のヘルプファイルをご参照いただきたい。

```
---
require(survival)

## The complete Wilms Tumor Data
## (Breslow and Chatterjee, Applied Statistics, 1999)
## subcohort selected by simple random sampling.
##

subcoh <- nwtco$in.subcohort
selccoh <- with(nwtco, rel==1|subcoh==1)
ccoh.data <- nwtco[selccoh,]
ccoh.data$subcohort <- subcoh[selccoh]
## central-lab histology
ccoh.data$histol <- factor(ccoh.data$histol,labels=c("FH","UH"))
## tumour stage
ccoh.data$stage <- factor(ccoh.data$stage,labels=c("I","II","III","IV"))
ccoh.data$age <- ccoh.data$age/12 # Age in years

##
```

```
## Standard case-cohort analysis: simple random subcohort
##

fit.ccP <- cch(Surv(edrel, rel) ~ stage + histol + age, data =ccoh.data,
   subcoh = ~subcohort, id=~seqno, cohort.size=4028)
fit.ccP

Case-cohort analysis,x$method, Prentice
 with subcohort of 668 from cohort of 4028

Call: cch(formula = Surv(edrel, rel) ~ stage + histol + age, data = ccoh.data,
    subcoh = ~subcohort, id = ~seqno, cohort.size = 4028)

Coefficients:
              Value         SE        Z            p
stageII   0.73457084 0.16849620 4.359569 1.303187e-05
stageIII  0.59708356 0.17345094 3.442377 5.766257e-04
stageIV   1.38413197 0.20481982 6.757803 1.400990e-11
histolUH  1.49806307 0.15970515 9.380180 0.000000e+00
age       0.04326787 0.02373086 1.823274 6.826184e-02

fit.ccSP <- cch(Surv(edrel, rel) ~ stage + histol + age, data =ccoh.data,
   subcoh = ~subcohort, id=~seqno, cohort.size=4028, method="SelfPren")
summary(fit.ccSP)

Case-cohort analysis,x$method, SelfPrentice
 with subcohort of 668 from cohort of 4028

Call: cch(formula = Surv(edrel, rel) ~ stage + histol + age, data = ccoh.data,
    subcoh = ~subcohort, id = ~seqno, cohort.size = 4028, method = "SelfPren")

Coefficients:
          Coef    HR  (95%   CI)     p
stageII   0.736 2.088 1.501 2.905 0.000
stageIII  0.597 1.818 1.294 2.553 0.001
stageIV   1.392 4.021 2.692 6.008 0.000
```

```
histolUH 1.506 4.507 3.295 6.163 0.000
age      0.043 1.044 0.997 1.094 0.069
---
```

subcoh という引数で、サブコホートの指示変数を指定する。method では、デフォルトが Prentice 法である。SelfPren, LinYing のオプションを用いることも可能である。いずれの方法も、理論的な妥当性は保証されている。詳細については、”?cch” をご参照いただきたい。

層別サンプリングを行う場合には、Borgan et al. [9]の Type-I, II 推定量を用いることができる。それぞれ Self-Prentice 法、Lin-Ying 法の一般化であるが、後者のほうがいわゆる一般的な IPW 推定量である。

```
---
##
## (post-)stratified on instit
##
stratsizes <- table(nwtco$instit)
fit.BI <- cch(Surv(edrel, rel) ~ stage + histol + age, data =ccoh.data,    subcoh
= ~subcohort, id=~seqno, stratum=~instit, cohort.size = stratsizes,
method="I.Borgan")
summary(fit.BI)

Exposure-stratified case-cohort analysis, I.Borgan method.
            1   2
subcohort  952 202
cohort    3622 406
Call: cch(formula = Surv(edrel, rel) ~ stage + histol + age, data = ccoh.data,
    subcoh = ~subcohort, id = ~seqno, stratum = ~instit, cohort.size = stratsizes,
    method = "I.Borgan")

Coefficients:
          Coef    HR  (95%   CI)     p
stageII  0.737 2.090 1.501 2.909 0.000
stageIII 0.602 1.825 1.301 2.561 0.000
stageIV  1.395 4.036 2.702 6.029 0.000
histolUH 1.522 4.580 3.450 6.080 0.000
```

```
age       0.043 1.044 0.996 1.093 0.072

fit.BII <- cch(Surv(edrel, rel) ~ stage + histol + age, data =ccoh.data,
subcoh = ~subcohort, id=~seqno, stratum=~instit, cohort.size = stratsizes,
method="II.Borgan")
summary(fit.BII)

Exposure-stratified case-cohort analysis, II.Borgan method.
             1   2
subcohort  952 202
cohort    3622 406
Call: cch(formula = Surv(edrel, rel) ~ stage + histol + age, data = ccoh.data,
    subcoh = ~subcohort, id = ~seqno, stratum = ~instit, cohort.size = stratsizes,
    method = "II.Borgan")

Coefficients:
          Coef    HR  (95%   CI)     p
stageII  0.693 1.999 1.453 2.751 0.000
stageIII 0.640 1.896 1.370 2.625 0.000
stageIV  1.303 3.681 2.538 5.341 0.000
histolUH 1.498 4.473 3.456 5.789 0.000
age      0.045 1.046 1.001 1.093 0.045
---
```

stratum で層を指定し、cohort.size で層ごとの人数の分母を指定することにより、層別を考慮した解析を行うことができる。ここでは、いわゆる「事後層別」による解析を行っているが、一定の条件下でハザード比の推定精度が向上することが知られている。

生存時間アウトカムだけではなく、2 値アウトカムの回帰モデルについても、同様の IPW 法による解析で曝露効果の指標の推定値を得ることができる。詳細については、Noma and Tanaka [10] をご参照いただきたい。

Auxiliary variables の情報を用いることで、さらに推定精度（検出力）の高い解析を行うことも可能である [11]。これらのモジュールも、survey パッケージに実装されているが、やや複雑なコードを組む必要がある。ご関心のある方は、CRAN の survey のチュートリアルなどをご参照いただきたい。

文献

1. Thomas DC. Addendum to a paper by Liddell F. D. K., McDolad J. C., Thomas D. C., and Cunliffe S. V. *Journal of the Royal Statistical Society, Series A* 1977; **140**: 483-485.
2. Samuelsen SO. A pseudolikelihood approach to analysis of nested case-control data. *Biometrika* 1997; **84**: 379-394. DOI: 10.1093/biomet/84.2.379.
3. Støer NC, Samuelsen SO. Comparison of estimators in nested case-control studies with multiple outcomes. *Lifetime Data Analysis* 2012; **18**: 261-283. DOI: 10.1007/s10985-012-9214-8.
4. Støer NC, Samuelsen SO. multipleNCC: Inverse probability weighting of nested case-control data. *The R Journal* 2016; Available online: https://journal.r-project.org/archive/accepted/stoer-samuelsen.pdf.
5. Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* 1986; **73**: 1-11. DOI: 10.1093/biomet/73.1.1.
6. Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics* 1988; **16**: 64-81. DOI: 10.1214/aos/1176350691.
7. Lin DY, Ying Z. Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association* 1993; **88**: 1341-1349. DOI: 10.2307/2291275.
8. Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms Tumour Prognosis. *Journal of the Royal Statistical Society, Series C* 1999; **48**: 457-468. DOI: 10.1111/1467-9876.00165.
9. Borgan Ø, Langholz B, Samuelsen SO, Goldstein DR, Pogoda J. Exposure stratified case-cohort designs. *Lifetime Data Analysis* 2000; **6**: 39-58. DOI: 10.1023/A:1009661900674.
10. Noma H, Tanaka S. Analysis of case-cohort designs with binary outcomes: Improving the efficiency using whole cohort auxiliary information. *Statistical Methods in Medical Research* 2016. DOI: 10.1177/0962280214556175.
11. Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology* 2009; **169**: 1398-1405. DOI: 10.1093/aje/kwp055.

```r
## R code examples for IPW analyses of nested case-control studies and case-cohort
studies

## 1. Nested case-control studies

require(multipleNCC)

data(CVD_Accidents)
attach(CVD_Accidents)
CVD_Accidents$samplestat2 <- samplestat
CVD_Accidents$samplestat2[samplestat==3] <- 2

wpl(Surv(agestart,agestop,dead24) ~ factor(smoking3gr)+bmi+factor(sex),
 data=CVD_Accidents, samplestat=CVD_Accidents$samplestat2, m=1,
 match.var=cbind(CVD_Accidents$sex, CVD_Accidents$bmi),
 match.int=c(0,0,-2,2), weight.method="KM")  # IPW analysis by design weights

wpl(Surv(agestart,agestop,dead24) ~ factor(smoking3gr)+bmi+factor(sex),
 data=CVD_Accidents, samplestat=CVD_Accidents$samplestat2, m=1,
 match.var=cbind(CVD_Accidents$sex, CVD_Accidents$bmi),
 match.int=c(0,0,-2,2), weight.method="glm")  # IPW analysis by estimated weights




## 2. Case-cohort studies

require(survival)

## The complete Wilms Tumor Data
## (Breslow and Chatterjee, Applied Statistics, 1999)
## subcohort selected by simple random sampling.
##

subcoh <- nwtco$in.subcohort
selccoh <- with(nwtco, rel==1|subcoh==1)
ccoh.data <- nwtco[selccoh,]
ccoh.data$subcohort <- subcoh[selccoh]
## central-lab histology
ccoh.data$histol <- factor(ccoh.data$histol,labels=c("FH","UH"))
## tumour stage
ccoh.data$stage <- factor(ccoh.data$stage,labels=c("I","II","III","IV"))
ccoh.data$age <- ccoh.data$age/12 # Age in years

##
## Standard case-cohort analysis: simple random subcohort
##

fit.ccP <- cch(Surv(edrel, rel) ~ stage + histol + age, data =ccoh.data,
   subcoh = ~subcohort, id=~seqno, cohort.size=4028)
summary(fit.ccP)  # Prentice method

fit.ccSP <- cch(Surv(edrel, rel) ~ stage + histol + age, data =ccoh.data,
   subcoh = ~subcohort, id=~seqno, cohort.size=4028, method="SelfPren")
summary(fit.ccSP)  # Self-Prentice method

fit.ccLY <- cch(Surv(edrel, rel) ~ stage + histol + age, data =ccoh.data,
   subcoh = ~subcohort, id=~seqno, cohort.size=4028, method="LinYing")
summary(fit.ccLY)  # Lin-Ying method

##
## (post-)stratified on instit
##
```

```
stratsizes <- table(nwtco$instit)
fit.BI <- cch(Surv(edrel, rel) ~ stage + histol + age, data =ccoh.data,
   subcoh = ~subcohort, id=~seqno, stratum=~instit, cohort.size = stratsizes, method
   ="I.Borgan")
summary(fit.BI)  # Borgan's type-I estimator

fit.BII <- cch(Surv(edrel, rel) ~ stage + histol + age, data =ccoh.data,
   subcoh = ~subcohort, id=~seqno, stratum=~instit, cohort.size = stratsizes, method
   ="II.Borgan")
summary(fit.BII)  # Borgan's type-II estimator
```