Two-phase designs in epidemiology

Thomas Lumley

December 9, 2005

This document explains how to analyse case–cohort and two-phase case–control studies with the "survey" package, using examples from `http://faculty.washington.edu/norm/software.html`. Some of the examples were published by Breslow & Chatterjee (1999).

The data are relapse rates from the National Wilm's Tumor Study (NWTS). Wilm's Tumour is a rare cancer of the kidney in children. Intensive treatment cures the majority of cases, but prognosis is poor when the disease is advanced at diagnosis and for some histological subtypes. The histological characterisation of the tumour is difficult, and histological group as determined by the NWTS central pathologist predicts much better than determinations by local institution pathologists. In fact, local institution histology can be regarded statistically as a pure surrogate for the central lab histology.

In these examples we will pretend that the (binary) local institution histology determination (`instit`) is avavailable for all children in the study and that the central lab histology (`histol`) is obtained for a probability sample of specimens in a two-phase design. We treat the initial sampling of the study as simple random sampling from an infinite superpopulation. We also have data on disease stage, a four-level variable; on relapse; and on time to relapse.

Case–control designs

Breslow & Chatterjee (1999) use the NWTS data to illustrate two-phase case–control designs. The data are available at `http://faculty.washington.edu/norm/software.html` in compressed form; we first expand to one record per patient.

```
> library(survey)
> load(system.file("doc", "nwts.rda", package = "survey"))
> nwtsnb <- nwts
> nwtsnb$case <- nwts$case - nwtsb$case
> nwtsnb$control <- nwts$control - nwtsb$control
> a <- rbind(nwtsb, nwtsnb)
> a$in.ccs <- rep(c(TRUE, FALSE), each = 16)
> b <- rbind(a, a)
> b$rel <- rep(c(1, 0), each = 32)
> b$n <- ifelse(b$rel, b$case, b$control)
> index <- rep(1:64, b$n)
> nwt.exp <- b[index, c(1:3, 6, 7)]
> nwt.exp$id <- 1:4088
```

As we actually do know `histol` for all patients we can fit the logistic regression model with full sampling to compare with the two-phase analyses

```
> glm(rel ~ factor(stage) * factor(histol), family = binomial,
+     data = nwt.exp)
```

```
Call:  glm(formula = rel ~ factor(stage) * factor(histol), family = binomial,      data = nwt.exp)

Coefficients:
                    (Intercept)                     factor(stage)2
                        -2.7066                             0.7679
                  factor(stage)3                     factor(stage)4
                         0.7747                             1.0506
              factor(histol)2  factor(stage)2:factor(histol)2
                         1.3104                             0.1477
factor(stage)3:factor(histol)2  factor(stage)4:factor(histol)2
                         0.5942                             1.2619

Degrees of Freedom: 4087 Total (i.e. Null);  4080 Residual
Null Deviance:           3306
Residual Deviance: 2943          AIC: 2959
```

The second phase sample consists of all patients with unfavorable histology as determined by local institution pathologists, all cases, and a 20% sample of the remainder. Phase two is thus a stratified random sample without replacement, with strata defined by the interaction of `instit` and `rel`.

```
> dccs2 <- twophase(id = list(~id, ~id), subset = ~in.ccs, strata = list(NULL,
+     ~interaction(instit, rel)), data = nwt.exp)
> summary(svyglm(rel ~ factor(stage) * factor(histol), family = binomial,
+     design = dccs2))

Call:
svyglm(rel ~ factor(stage) * factor(histol), family = binomial,
    design = dccs2)

Survey design:
twophase(id = list(~id, ~id), subset = ~in.ccs, strata = list(NULL,
    ~interaction(instit, rel)), data = nwt.exp)

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -2.5701     0.1283 -20.024  < 2e-16 ***
factor(stage)2                    0.5482     0.1979   2.769 0.005708 **
factor(stage)3                    0.4791     0.2032   2.359 0.018514 *
factor(stage)4                    1.0037     0.2592   3.872 0.000114 ***
factor(histol)2                   1.3505     0.3107   4.347 1.50e-05 ***
factor(stage)2:factor(histol)2   0.1152     0.4410   0.261 0.793886
factor(stage)3:factor(histol)2   0.5066     0.4241   1.194 0.232570
factor(stage)4:factor(histol)2   0.9785     0.6214   1.575 0.115625
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.000876)

Number of Fisher Scoring iterations: 5
```

Disease stage at the time of surgery is also recorded. It could be used to further stratify the sampling, or, as in this example, to post-stratify. We can analyze the data either pretending that the sampling was stratified or using `calibrate` to post-stratify the design.

```
> dccs8 <- twophase(id = list(~id, ~id), subset = ~in.ccs, strata = list(NULL,
+     ~interaction(instit, stage, rel)), data = nwt.exp)
> gccs8 <- calibrate(dccs2, phase = 2, formula = ~interaction(instit,
+     stage, rel))
> summary(svyglm(rel ~ factor(stage) * factor(histol), family = binomial,
+     design = dccs8))

Call:
svyglm(rel ~ factor(stage) * factor(histol), family = binomial,
    design = dccs8)

Survey design:
twophase(id = list(~id, ~id), subset = ~in.ccs, strata = list(NULL,
    ~interaction(instit, stage, rel)), data = nwt.exp)

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -2.71604    0.10768 -25.222  < 2e-16 ***
factor(stage)2                 0.78141    0.14725   5.307 1.34e-07 ***
factor(stage)3                 0.80093    0.15248   5.253 1.79e-07 ***
factor(stage)4                 1.07293    0.17816   6.022 2.32e-09 ***
factor(histol)2                1.45836    0.31771   4.590 4.92e-06 ***
factor(stage)2:factor(histol)2 -0.04743   0.43497  -0.109    0.913
factor(stage)3:factor(histol)2  0.28064   0.41299   0.680    0.497
factor(stage)4:factor(histol)2  0.90983   0.63776   1.427    0.154
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.000876)

Number of Fisher Scoring iterations: 5

> summary(svyglm(rel ~ factor(stage) * factor(histol), family = binomial,
+     design = gccs8))

Call:
svyglm(rel ~ factor(stage) * factor(histol), family = binomial,
    design = gccs8)

Survey design:
calibrate(dccs2, phase = 2, formula = ~interaction(instit, stage,
    rel))

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   -2.71604    0.10767 -25.225  < 2e-16 ***
factor(stage)2                 0.78141    0.14724   5.307 1.34e-07 ***
```

```
factor(stage)3                          0.80093   0.15249   5.252 1.79e-07 ***
factor(stage)4                          1.07293   0.17803   6.027 2.26e-09 ***
factor(histol)2                         1.45836   0.31652   4.607 4.54e-06 ***
factor(stage)2:factor(histol)2 -0.04743   0.43419  -0.109    0.913
factor(stage)3:factor(histol)2  0.28064   0.41257   0.680    0.497
factor(stage)4:factor(histol)2  0.90983   0.62990   1.444    0.149
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1.000876)

Number of Fisher Scoring iterations: 5
```

Case–cohort designs

In the case–cohort design for survival analysis, a $P\%$ sample of a cohort is taken at recruitment for the second phase, and all participants who experience the event (cases) are later added to the phase-two sample.

Viewing the sampling design as progressing through time in this way, as originally proposed, gives a double sampling design at phase two. It is simpler to view the process *sub specie aeternitatis*, and to note that cases are sampled with probability 1, and controls with probability $P/100$. The subcohort will often be determined retrospectively rather than at recruitment, giving stratified random sampling without replacement, stratified on case status. If the subcohort is determined prospectively we can use the same analysis, post-stratifying rather than stratifying.

There have been many analyses proposed for the case–cohort design (Therneau & Li, 1999). We consider only those that can be expressed as a Horvitz–Thompson estimator for the Cox model.

First we load the data and the necessary packages. The version of the NWTS data that includes survival times is not identical to the data set used for case–control analyses above.

```
> library(survey)
> library(survival)

Loading required package: splines

> data(nwtco)
> ntwco <- subset(nwtco, !is.na(edrel))
```

Again, we fit a model that uses `histol` for all patients, to compare with the two-phase design

```
> coxph(Surv(edrel, rel) ~ factor(stage) + factor(histol) + I(age/12),
+      data = nwtco)

Call:
coxph(formula = Surv(edrel, rel) ~ factor(stage) + factor(histol) +
    I(age/12), data = nwtco)


                 coef exp(coef) se(coef)    z       p
factor(stage)2 0.6673      1.95   0.1216 5.49 4.0e-08
factor(stage)3 0.8174      2.26   0.1208 6.77 1.3e-11
```

4

```
factor(stage)4  1.1537      3.17   0.1349  8.55 0.0e+00
factor(histol)2 1.5839      4.87   0.0887 17.86 0.0e+00
I(age/12)       0.0679      1.07   0.0149  4.55 5.4e-06

Likelihood ratio test=395  on 5 df, p=0  n= 4028
```

We define a two-phase survey design using simple random superpopulation sampling for the first phase, and sampling without replacement stratified on `rel` for the second phase. The `subset` argument specifies that observations are in the phase-two sample if they are in the subcohort or are cases. As before, the data structure is rectangular, but variables measured at phase two may be `NA` for participants not included at phase two.

We compare the result to that given by `survival::cch` for Lin & Ying's (1993) approach to the case–cohort design.

```
> (dcch <- twophase(id = list(~seqno, ~seqno), strata = list(NULL,
+     ~rel), subset = ~I(in.subcohort | rel), data = nwtco))

Two-phase design:twophase(id = list(~seqno, ~seqno), strata = list(NULL, ~rel),
    subset = ~I(in.subcohort | rel), data = nwtco)
Phase 1:
Independent Sampling design (with replacement)
svydesign(id = ~seqno)
Phase 2:
Stratified Independent Sampling design
svydesign(id = ~seqno, strata = ~rel, fpc = `*phase1*`)

> svycoxph(Surv(edrel, rel) ~ factor(stage) + factor(histol) +
+     I(age/12), design = dcch)

Two-phase design:twophase(id = list(~seqno, ~seqno), strata = list(NULL, ~rel),
    subset = ~I(in.subcohort | rel), data = nwtco)
Phase 1:
Independent Sampling design (with replacement)
svydesign(id = ~seqno)
Phase 2:
Stratified Independent Sampling design
svydesign(id = ~seqno, strata = ~rel, fpc = `*phase1*`)
Call:
svycoxph(Surv(edrel, rel) ~ factor(stage) + factor(histol) +
    I(age/12), design = dcch)


                 coef exp(coef) se(coef)     z        p
factor(stage)2 0.6927      2.00    0.163  4.25 2.1e-05
factor(stage)3 0.6269      1.87    0.168  3.73 1.9e-04
factor(stage)4 1.2995      3.67    0.189  6.88 6.1e-12
factor(histol)2 1.4583     4.30    0.145 10.02 0.0e+00
I(age/12)      0.0461      1.05    0.023  2.00 4.5e-02

Likelihood ratio test=NA  on 5 df, p=NA  n= 1154
```

```
> subcoh <- nwtco$in.subcohort
> selccoh <- with(nwtco, rel == 1 | subcoh == 1)
> ccoh.data <- nwtco[selccoh, ]
> ccoh.data$subcohort <- subcoh[selccoh]
> cch(Surv(edrel, rel) ~ factor(stage) + factor(histol) + I(age/12),
+     data = ccoh.data, subcoh = ~subcohort, id = ~seqno, cohort.size = 4028,
+     method = "LinYing")

Case-cohort analysis,  LinYing  method,
 with subcohort of 668 from cohort of 4028

Call: cch(formula = Surv(edrel, rel) ~ factor(stage) + factor(histol) +
    I(age/12), data = ccoh.data, subcoh = ~subcohort, id = ~seqno,
    cohort.size = 4028, method = "LinYing")

Coefficients:
                    Value         SE          Z          p
factor(stage)2   0.69265646 0.16274542   4.256073 2.080485e-05
factor(stage)3   0.62685179 0.16816124   3.727683 1.932483e-04
factor(stage)4   1.29951229 0.18896863   6.876868 6.118217e-12
factor(histol)2  1.45829267 0.14553560  10.020178 0.000000e+00
I(age/12)        0.04608972 0.02300671   2.003317 4.514329e-02
```

Barlow (1994) proposes an analysis that ignores the finite population correction at the second phase. This simplifies the standard error estimation, as the design can be expressed as one-phase stratified superpopulation sampling. The standard errors will be somewhat conservative. More data preparation is needed for this analysis as the weights change over time.

```
> nwtco$eventrec <- rep(0, nrow(nwtco))
> nwtco.extra <- subset(nwtco, rel == 1)
> nwtco.extra$eventrec <- 1
> nwtco.expd <- rbind(subset(nwtco, in.subcohort == 1), nwtco.extra)
> nwtco.expd$stop <- with(nwtco.expd, ifelse(rel & !eventrec, edrel -
+     0.001, edrel))
> nwtco.expd$start <- with(nwtco.expd, ifelse(rel & eventrec, edrel -
+     0.001, 0))
> nwtco.expd$event <- with(nwtco.expd, ifelse(rel & eventrec, 1,
+     0))
> nwtco.expd$pwts <- ifelse(nwtco.expd$event, 1, 1/with(nwtco,
+     mean(in.subcohort | rel)))
```

The analysis corresponds to a cluster-sampled design in which individuals are sampled stratified by subcohort membership and then time periods are sampled stratified by event status. Having individual as the primary sampling unit is necessary for correct standard error calculation.

```
> (dBarlow <- svydesign(id = ~seqno + eventrec, strata = ~in.subcohort +
+     rel, data = nwtco.expd, weight = ~pwts))

Stratified 2 - level Cluster Sampling design (with replacement)
With (1154, 1239) clusters.
svydesign(id = ~seqno + eventrec, strata = ~in.subcohort + rel,
    data = nwtco.expd, weight = ~pwts)
```

```
> svycoxph(Surv(start, stop, event) ~ factor(stage) + factor(histol) +
+     I(age/12), design = dBarlow)
Stratified 2 - level Cluster Sampling design (with replacement)
With (1154, 1239) clusters.
svydesign(id = ~seqno + eventrec, strata = ~in.subcohort + rel,
    data = nwtco.expd, weight = ~pwts)
Call:
svycoxph(Surv(start, stop, event) ~ factor(stage) + factor(histol) +
    I(age/12), design = dBarlow)


                  coef exp(coef) se(coef)    z       p
factor(stage)2  0.7359      2.09   0.1699 4.33 1.5e-05
factor(stage)3  0.5976      1.82   0.1753 3.41 6.5e-04
factor(stage)4  1.3907      4.02   0.2078 6.69 2.2e-11
factor(histol)2 1.5045      4.50   0.1641 9.17 0.0e+00
I(age/12)       0.0432      1.04   0.0243 1.78 7.5e-02

Likelihood ratio test=NA  on 5 df, p=NA  n= 1239
```

In fact, as the finite population correction is not being used the second stage of the cluster sampling could be ignored. We can also produce the stratified bootstrap standard errors of Wacholder et al (1989), using a replicate weights analysis

```
> (dWacholder <- as.svrepdesign(dBarlow, type = "bootstrap", replicates = 500))

Call: as.svrepdesign(dBarlow, type = "bootstrap", replicates = 500)
Survey bootstrap with 500 replicates.

> svycoxph(Surv(start, stop, event) ~ factor(stage) + factor(histol) +
+     I(age/12), design = dWacholder)

Call: as.svrepdesign(dBarlow, type = "bootstrap", replicates = 500)
Survey bootstrap with 500 replicates.
Call:
svycoxph(Surv(start, stop, event) ~ factor(stage) + factor(histol) +
    I(age/12), design = dWacholder)


                  coef exp(coef) se(coef)    z       p
factor(stage)2  0.7359      2.09   0.1770 4.16 3.2e-05
factor(stage)3  0.5976      1.82   0.1814 3.29 9.9e-04
factor(stage)4  1.3907      4.02   0.2012 6.91 4.8e-12
factor(histol)2 1.5045      4.50   0.1714 8.78 0.0e+00
I(age/12)       0.0432      1.04   0.0253 1.70 8.9e-02

Likelihood ratio test=NA  on 5 df, p=NA  n= 1239
```

Exposure-stratified designs

Borgan et al (2000) propose designs stratified or post-stratified on phase-one variables. The examples at http://faculty.washington.edu/norm/software.html use a different subcohort sample for this stratified design, so we load the new subcohort variable

7

```
> load(system.file("doc", "nwtco-subcohort.rda", package = "survey"))
> nwtco$subcohort <- subcohort
> d_BorganII <- twophase(id = list(~seqno, ~seqno), strata = list(NULL,
+     ~interaction(instit, rel)), data = nwtco, subset = ~I(rel |
+     subcohort))
> (b2 <- svycoxph(Surv(edrel, rel) ~ factor(stage) + factor(histol) +
+     I(age/12), design = d_BorganII))

Two-phase design:twophase(id = list(~seqno, ~seqno), strata = list(NULL, ~interaction(instit,
    rel)), data = nwtco, subset = ~I(rel | subcohort))
Phase 1:
Independent Sampling design (with replacement)
svydesign(id = ~seqno)
Phase 2:
Stratified Independent Sampling design
svydesign(id = ~seqno, strata = ~interaction(instit, rel), fpc = `*phase1*`)
Call:
svycoxph(Surv(edrel, rel) ~ factor(stage) + factor(histol) +
    I(age/12), design = d_BorganII)


                  coef exp(coef) se(coef)      z       p
factor(stage)2  0.4629      1.59   0.1809  2.559 1.0e-02
factor(stage)3  0.5831      1.79   0.1785  3.267 1.1e-03
factor(stage)4  1.0597      2.89   0.2052  5.163 2.4e-07
factor(histol)2 1.5974      4.94   0.1332 11.994 0.0e+00
I(age/12)       0.0299      1.03   0.0334  0.897 3.7e-01

Likelihood ratio test=NA  on 5 df, p=NA  n= 1062
```

We can further post-stratify the design on disease stage and age with `calibrate`

```
> d_BorganIIps <- calibrate(d_BorganII, phase = 2, formula = ~age +
+     interaction(instit, rel, stage))
> svycoxph(Surv(edrel, rel) ~ factor(stage) + factor(histol) +
+     I(age/12), design = d_BorganIIps)

Two-phase design:calibrate(d_BorganII, phase = 2, formula = ~age + interaction(instit,
    rel, stage))
Phase 1:
Independent Sampling design (with replacement)
svydesign(id = ~seqno)
Phase 2:
Stratified Independent Sampling design
calibrate(phase2, formula, population, ...)
Call:
svycoxph(Surv(edrel, rel) ~ factor(stage) + factor(histol) +
    I(age/12), design = d_BorganIIps)


                  coef exp(coef) se(coef)      z       p
```

```
factor(stage)2  0.6701      1.95   0.1427  4.70 2.6e-06
factor(stage)3  0.7593      2.14   0.1421  5.34 9.2e-08
factor(stage)4  1.2705      3.56   0.1521  8.35 1.1e-16
factor(histol)2 1.5730      4.82   0.1300 12.10 0.0e+00
I(age/12)       0.0314      1.03   0.0337  0.93 3.5e-01

Likelihood ratio test=NA  on 5 df, p=NA  n= 1062
```

References

Barlow WE (1994). Robust variance estimation for the case-cohort design. *Biometrics* 50: 1064-1072

Borgan Ø, Langholz B, Samuelson SO, Goldstein L and Pogoda J (2000). Exposure stratified case-cohort designs, *Lifetime Data Analysis* 6:39-58

Breslow NW and Chatterjee N. (1999) Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics* 48:457-68.

Lin DY, and Ying Z (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association* 88: 1341-1349.

Therneau TM and Li H., Computing the Cox model for case-cohort designs. *Lifetime Data Analysis* 5:99-112, 1999

Wacholder S, Gail MH, Pee D, and Brookmeyer R (1989) Alternate variance and efficiency calculations for the case-cohort design *Biometrika*, 76, 117-123