

# **An Introduction to Statistical Analysis with SUDAAN**

M. E. Thompson  
University of Waterloo

SWORDC opening  
December 2001

Assistance from Christian Boudreau is gratefully  
acknowledged.

## OUTLINE

WHAT IS SUDAAN?

HOW DOES SUDAAN ESTIMATE SURVEY POPULATION TOTALS?

OTHER KINDS OF QUANTITIES?

WHAT ARE THE “DESIGN =” CHOICES IN SUDAAN?

FOR ESTIMATING FROM SURVEY DATA, WHAT WOULD A SUDAAN PROGRAM LOOK LIKE?

EXAMPLES:

- PROC DESCRIPT
- PROC LOGISTIC
- PROC SURVIVAL
- PROC LOGLINK

---

WHAT IS SUDAAN?

“Software for the Statistical Analysis of Correlated Data”

- SUDAAN was developed around 1980 at the Research Triangle Institute by B. V. Shah
  - statistical packages of the time assumed data “points” obtained by simple random sampling from a large (hypothetical) population
  - aim of SUDAAN was to provide SAS procedures for the treatment of data from stratified multistage sampling designs
  - specifically, aim was to produce correct standard errors for estimates of totals, means, proportions, ratios, regression coefficients
  - more procedures for analysis of survey data, or correlated data from other kinds of study, were added
  - Version 7, released in 1996, advertised “linear regression, logistic regression, multinomial logistic regression, proportional hazards modelling of time-to-event data, and descriptive data analysis”
-

## VERSION NOW AVAILABLE

- we have Release 8.0.0 (2001) for standalone PC; manual has about 850 pages
- additions include “replicate weight jackknife” variance estimation and new features within the modelling procedures
- syntax is still SAS-based but the data input procedures are not as flexible as the procedures for modern SAS
- interface displays an *Input window* (text editor for SUDAAN programs), and an *Output window* which displays the results of running the programs

SUDAAN outside the RDC is somewhat expensive (US \$770 plus renewal each year per copy) and protected by strict licensing agreement.

---

## WHAT PROCEDURES ARE AVAILABLE IN SUDAAN?

CROSSTAB provides weighted frequency and percentage distributions for one-way and multi-way tabulations

RATIO provides estimates of population ratios and standard errors for the estimates

DESCRIPT provides estimates and standard errors for population means, totals, proportions, geometric means, medians and other quantiles

LOGLINK fits log-linear regression models to count data

LOGISTIC fits logistic regression models to dichotomous outcome variables (yes/no data)

MULTILOG generalizes the modelling capabilities of LOGISTIC to include categorical outcomes with two or more categories

REGRESS fits linear regression models to continuous outcomes

SURVIVAL provides proportional hazards modelling for failure time outcomes

---

## HOW DOES SUDAAN ESTIMATE SURVEY POPULATION TOTALS?

For example, consider a stratified multi-stage sample of households, with first stage sample “without replacement”:

$$\begin{aligned} N_h &= \text{number of PSUs in stratum } h \\ n_h &= \text{number of PSUs sampled in stratum } h \\ \Pi_{hr} &= \text{inclusion probability of PSU } r \text{ of stratum } h \end{aligned}$$

For short, let “PSU  $hr$ ” denote PSU  $r$  of stratum  $h$ .

$\Pi_{hr} = n_h/N_h$  if the design at the first stage within stratum  $h$  is simple random sampling without replacement (stratified random sampling)

Within selected PSU  $hr$ , SUDAAN would assume subsequent sampling stages to be taken by simple random sampling or stratified simple random sampling, with or without replacement

---

Suppose we wanted to estimate

the total of dollars spent on food in a reference week by the households of the entire population

$\mathcal{U}$ .

This is the *population total*

$$T_y = \sum_{i \in \mathcal{U}} y_i$$

$y_i$  = the amount spent on food by household  $i$ .

Also,

$$T_y = \sum_h \sum_r T_{hr}$$

where

$T_{hr}$  = total food expenditure in PSU  $hr$ .

Two ways of expressing the estimator of the total:

(i) If  $s$  denotes the sample of households drawn,

$$\hat{T}_y = \sum_{i \in s} w_i y_i$$

where  $w_i$  is the *survey weight* for household  $i$ , chosen so that:

–  $\hat{T}_y$  is exactly or approximately unbiased for  $T_y$  under the sampling design;  $w_i \simeq 1/\pi_i$  where  $\pi_i$  is the inclusion probability for household  $i$

– roughly,  $w_i$  can be interpreted as the number of households in the population represented by household  $i$ .

(ii)

$$\hat{T}_y = \sum_h \sum_{r \in s_{Ih}} \frac{\hat{T}_{hr}}{\Pi_{hr}}$$

–  $s_{Ih}$  is the first stage sample (of PSUs) in stratum  $h$ .

–  $\hat{T}_{hr}$  is based on the subsample within  $hr$ , and is exactly or approximately unbiased for  $T_{hr}$ .

Clearly,

$$\frac{\hat{T}_{hr}}{\Pi_{hr}} = t_{hr} = \sum_{i \in s_{hr}} w_i y_i$$

where  $s_{hr}$  is the subsample of households in PSU  $hr$ .

Variance estimator (first stage without replacement):

$$v(\hat{T}_y) = \sum_h \sum_r \sum_q \frac{\Pi_{hr}\Pi_{hq} - \Pi_{hrq}}{2\Pi_{hrq}} \left( \frac{\hat{T}_{hr}}{\Pi_{hr}} - \frac{\hat{T}_{hq}}{\Pi_{hr}} \right)^2 + \sum_h \sum_r \frac{\hat{V}_{hr}}{\Pi_{hr}}$$

–  $\hat{V}_{hr}$  estimates the variance of  $\hat{T}_{hr}$  as an estimator of  $T_{hr}$ , obtained by “subcalculation”;

–  $\Pi_{hrq}$  is the joint inclusion probability of PSUs  $hr$  and  $hq$

If the first stage design is stratified random sampling, the first term of  $v(\hat{T}_y)$  is

$$\sum_h \left(1 - \frac{n_h}{N_h}\right) \frac{n_h}{n_h - 1} \sum_{r \in s_{Ih}} (t_{hr} - \bar{t}_h)^2$$

where  $\bar{t}_h = (\sum_{r \in s_{Ih}} t_{hr})/n_h$ .

If the first stage design is approximately with replacement (not necessarily with uniform selection probabilities) within strata, the analogous formula to (†) is

$$v(\hat{T}_y) = \sum_h \sum_{r \in s_{Ih}} \frac{n_h}{n_h - 1} (t_{hr} - \bar{t}_h)^2$$

where  $n_h$  is the number of draws of PSUs in stratum  $h$ ; there is no second “within PSU” term.

No information about detail within the subsamples  $s_{hr}$  is needed.

Variance estimation in SUDAAN is built up from formulas (†) and (‡) for  $v(\hat{T}_y)$ .

Note: A *poststratification* option is available in DESCRIPT and RATIO. If  $\hat{T}_y$  is replaced by a poststratified estimator for  $T_y$ , then the  $y$  values used in † and ‡ would be replaced by residuals (from regression of  $y$  on the poststratification variables).

## OTHER KINDS OF QUANTITIES?

Survey population means, proportions, ratios, regression coefficients are expressed internally as functions of population totals:

$$\text{mean of } y = \frac{T_y}{N} = \frac{T_2}{T_1}$$

$$\hat{\beta} = \frac{T_4 - \frac{T_2 T_3}{T_1}}{T_5 - \frac{T_3^2}{T_1}}$$

where  $T_3 = \sum_{i=1}^N x_i$ ,  $T_4 = \sum_{i=1}^N y_i x_i$ , and  $T_5 = \sum_{i=1}^N x_i^2$ .

In general these estimands are of form

$$\theta = g(T),$$

and standard estimators take the form

$$\hat{\theta} = g(\hat{T}).$$

The errors of estimation can be ‘linearized’, then handled as errors in estimating a population total:

$$\hat{\theta} - \theta \simeq \sum_{\alpha} \frac{\partial g}{\partial T_{\alpha}} (\hat{T}_{\alpha} - T_{\alpha}) = \hat{T}_z - T_z$$

where  $z_i = \sum_{\alpha} (\partial g / \partial T_{\alpha}) y_{\alpha i}$ .

---

Population quantiles and GLM regression coefficients are expressed as solutions of *estimating equations*.

e.g. population median  $\theta_N$  solves

$$U(y, \theta_N) = \sum_{i=1}^N \phi(y_i - \theta_N) = 0,$$

where

$$\begin{aligned} \phi(y_i - \theta) &= 1 \text{ if } y_i < \theta \\ &= 0 \text{ if } y_i = \theta \\ &= -1 \text{ if } y_i > 0. \end{aligned}$$

The estimator comes from

$$\hat{U}(y, \hat{\theta}) = \sum_{i \in s} w_i \phi(y_i - \hat{\theta}) = 0.$$

---

e.g. Logistic regression:  $y_i = 1$  or  $0$

Model:

$$\begin{aligned}\mathcal{E}(y_i) &= \mu_i(\beta) \\ \text{logit}\mu_i(\beta) &= x_i'\beta;\end{aligned}$$

$y_i - \mu_i$  correlated within PSUs (clusters), independent between PSUs.

Population level estimating function system:

$$U(x, y, \beta) = \sum_h \sum_r D_r V_r^{-1}(Y_r - \mu_r(\beta))$$

$V_r$  is the covariance matrix of  $Y_r$  incorporating a “working correlation” structure.

Finite population coefficients  $\beta_N$  would solve the system  $U(x, y, \beta_N) = 0$ .

Sample estimate comes from a GEE system

$$\hat{U}(x, y, \hat{\beta}) = \sum_h \sum_{r \in s_{1h}} q_r D_r V_r^{-1}(Y_r - \mu_r(\hat{\beta})) = 0.$$

If  $V_r$  incorporated an “INDEPENDENT” working correlation structure, we might have

$$\hat{U}(x, y, \hat{\beta}) = \sum_{i \in s} w_i (y_i - \mu_i(\hat{\beta})) = 0.$$

In analytic studies, we think of  $\hat{\beta}$  as estimating  $\beta$  rather than  $\beta_N$ .

Where  $\hat{U}$  is smooth, the error of estimation can be “linearized”:

$$\hat{\beta} - \beta \simeq -J^{-1}\hat{U}(x, y, \beta)$$

where  $J$  is the expectation of the partial derivatives of the components of  $\hat{U}$  with respect to the components of  $\beta$ . This leads to versions of the so-called “sandwich estimator” of variance, which use formulae for the covariance matrix of  $\hat{U}$  from the design based theory for totals (Binder, 1983) or the generalized estimating equation theory (Zeger and Liang, 1986).

## IN GENERAL:

If no variance estimation method is specified, SUDAAN treats quantities to be estimated as descriptive, and variance (and covariance) estimates are computed by “Taylor Series Linearization”, taking the sampling design into account.

Instead of Taylor Series Linearization, we may request JACKKNIFE (Delete -1 or replicate weight) or BRR (“Balanced Repeated Replication”)

---

## WHAT ARE THE “DESIGN = ” CHOICES IN SUDAAN?

WR: (default) Within strata, PSU’s are taken to be selected “with replacement”, in a sequence of independent draws; the selection probability  $p_{hr}$  for PSU  $r$  in stratum  $h$  is the same from draw to draw. (For large strata, with and without replacement are very close, and  $\Pi_{hr} \simeq n_h p_{hr}$ .)

STRWR: (Single stage) stratified random sampling with replacement.

SRS: simple random sampling with replacement.

WOR: PSU’s are taken to be selected by simple random sampling without replacement; sampling in subsequent stages is by simple random sampling, with or without replacement.

UNEQWOR: PSU’s are selected without replacement and with unequal inclusion probabilities  $\Pi_{hr}$ ; sampling in subsequent stages is by simple random sampling, with or without replacement.

STRWOR: (Single stage) stratified random sampling without replacement.

(All the above implement variance estimation for complex quantities by Taylor series linearization.)

JACKKNIFE: Variance estimation by jackknife, taking the design to be approximately as in WR.

BRR: Variance estimation by BRR, taking the design to be approximately as in WR.

---

## THE RESAMPLING METHODS

Jackknifing and BRR make the “approximately with replacement” assumption and operate at the PSU level. The standard errors squared can be expressed as

$$v(\hat{\theta}) = \sum_{b=1}^B A_b (\hat{\theta}^{(b)} - \hat{\theta})^2$$

where  $\hat{\theta}^{(b)}$  is the estimate from the  $b$ th ‘resample’, and  $A_b$  is an adjustment constant for the  $b$ th resample. In the case of BRR,  $A(b)$  is  $1/B$ . For the delete-one jackknife, if the  $b$ th resample is formed by deleting PSU  $hr$ , then  $A_b$  is  $(n_h - 1)/n_h$ . The resamples can be implemented by supplying for each  $b$  an appropriate adjusted set of survey weights. e.g. for the jackknife, if the  $b$ th resample which involves deleting PSU  $hr$ , the adjusted weights are

$$\begin{aligned} w_i^{(hr)} &= 0 \text{ if } i \text{ is in PSU } hr \\ &= w_i n_h / (n_h - 1) \text{ if } i \text{ is in stratum } h \text{ but not PSU } hr \\ &= w_i \text{ if } i \text{ is not in stratum } h. \end{aligned}$$

Replicate weights are in the default case calculated automatically by SUDAAN. If the Replicate Weight Jackknife jackknife or BRR options are used, the variables containing user-supplied replicate weights may be named.

---

#### FOR ESTIMATING FROM SURVEY DATA, WHAT WOULD A SUDAAN PROGRAM LOOK LIKE?

The input file will have one record per sampling unit (e.g. household); every variable has a value in each record.

---

FOODEXP.SUD

PROC DESCRIPT DATA=TESTDAT PSUDATA=TESTPSU DESIGN=  
UNEQWOR PSU\_REC=6;  
NEST COL\_STR PSU\_ID SUB SEGMENT;  
TOTCNT POPPSU NSUB PSEG NPER;  
JOINTPROB PR1 PR2 PR3;  
SUBGROUP SIZE;  
LEVELS 12;  
TABLES SIZE;  
VAR FEXP;

TESTPSU.DBS

123456789012345678901234567890123

50041.75000000.50000000.50000000  
50042.50000000.75000000.50000000  
50043.50000000.50000000.75000000  
60061.59047500.34865581.22473035  
60062.34865581.66427500.27988120  
60063.22473035.27988120.47400000

TESTPSU.LAB

COL_STR	1N 2 0	Stratum ID
POPPSU	1N 2 0	Number of PSUs in stratum
PSU_ID	1N 1 0	PSU ID within COL_STR
PR1	1N 9 8	Joint probability for PSU 1 with PSU i
PR2	1N 9 8	Joint probability for PSU 2 with PSU i
PR3	1N 9 8	Joint probability for PSU 3 with PSU i

The input file TESTDAT called by FOODEXP.SUD would have variables including

COL STR: - stratum identifier  
PSU ID: - PSU identifier  
NSUB: - number of substrata in PSU  
SUB: - substratum identifier  
PSEG: - number of segments in whole substratum  
SEGMENT: - second stage unit identifier  
NHOUS: - number of households in segment  
SIZE: - number of people in household  
FEXP: - expenditure on food in reference week  
WTF: - survey weight

Commands in TESTPSU might set up variables such as

POPPSU: - number of PSUs in stratum  
PR1, PR2, PR3, . . . : - joint inclusion probabilities for the sample PSUs within the record's stratum

Commands

NEST: - links identifiers with sampling "stages"  
(Records must be sorted with respect to those variables, in order.)  
TOTCNT: - says that the variables listed contain the corresponding "population sizes".  
JOINTPROB: says that PR1, PR2, PR3 give the joint inclusion probabilities  
WEIGHT: identifies the variable whose values are the analysis or sample weights

---

## EXAMPLES

- PROC DESCRIPT (see manual)
  - PROC LOGISTIC (see manual)
  - PROC SURVIVAL
  - PROC LOGLINK
-

## PROC SURVIVAL

### THE PROPORTIONAL HAZARDS MODEL FOR TIME-TO-EVENT

The model:

$$h(t | z) = h_0(t) \exp(\beta' z)$$

$-h(t | z)dt$  is the probability an individual with covariate value  $z$  “fails” in  $(t, t + dt]$ , given survival up to time  $t$  –  $h_0(t)dt$  is that probability for an individual with  $z = 0$  (the “baseline hazard”)

The observations:

For individual  $j$ , we begin observing at  $t_{1j}$ ; we stop when a failure occurs ( $\delta_j = 1$ ) or when observation ceases for some other reason ( $\delta_j = 0$ ) at time  $t_{2j}$

The maximum partial likelihood equations for  $\beta$  are:

$$\hat{U}(\hat{\beta}) = \sum_{i \in s} w_i \delta_i \left\{ x_i - \frac{\hat{S}_1(t_{2i}, \hat{\beta})}{\hat{S}_0(t_{2i}, \hat{\beta})} \right\} = 0,$$

where

$$\hat{S}_0(t_{2i}, \hat{\beta}) = \sum_{j \in s} w_j I(t_{1j} < t_{2i} \leq t_{2j}) \exp(x_j' \hat{\beta}),$$

$$\hat{S}_1(t_{2i}, \hat{\beta}) = \sum_{j \in s} w_j I(t_{1j} < t_{2i} \leq t_{2j}) x_j \exp(x_j' \hat{\beta}).$$

The estimated robust variance-covariance matrix of  $\hat{\beta}$  has been derived by Binder (1992):

$$\text{Var}(\hat{\beta}) = J^{-1} V_U (J^{-1})',$$

where

$$J = \frac{\partial \hat{U}(\hat{\beta})}{\partial \hat{\beta}}$$

and  $V_U$  estimates the variance of an appropriately adjusted score function.

```

/*
Fitting the Cox model to the SIPP data
  (spells on the food stamps program)

C. Boudreau
Last updated: November 8, 2001.
*/

proc survival data="C:\Stats\Sudaan\food.stx" filetype=sasxport design=wr;
  title "Fitting the Cox model to the SIPP data"
    "(spells on the food stamps program)";

  weight WEIGHT;
  nest _one_ SUID;

  subgroups SEX RACE USSTATES; /* categorical variables */

  levels 2 4 44; /* number of levels for these categorical variables */

  strhaz USSTATES; /* stratified analysis */

  relevel SEX=1 RACE=1;

  event STATUS; /* censoring variable (failure=1, censored=0) */
  model TIME = SEX RACE AGE GRADE INCOME / ties = BRESLOW;

  print beta="BETA" sebeta="STDERR" t_beta="T:Beta=0"
    p_beta="p-value" covar="VAR(BETA)";

```

```
/*
Fitting the Cox model to the SIPP data
(spells on the food stamps program)
```

```
C. Boudreau
Last updated: November 8, 2001.
```

```
*/
```

```
Date: 11-08-2001                    Research Triangle Institute                    Page : 4
Time: 19:21:02                     The SURVIVAL Procedure                     Table : 1
```

```
Variance Estimation Method: Taylor Series (WR)
Dependent Variable: TIME
Censoring Variable: STATUS
Ties Handling: BRESLOW
Fitting the Cox model to the SIPP data
(spells on the food stamps program)
```

```
-----
```

Independent Variables and Effects	BETA	STDERR	T:Beta=0	p-value
-----				
SEX				
1	0.00	0.00	.	.
2	-0.14	0.06	-2.41	0.0161
RACE				
1	0.00	0.00	.	.
2	-0.17	0.12	-1.46	0.1450
3	0.41	0.29	1.40	0.1630
4	-0.05	0.35	-0.14	0.8878
AGE	-0.01	0.00	-3.07	0.0023
GRADE	0.01	0.01	0.77	0.4412
INCOME	0.00	0.00	4.40	0.0000
-----				

## PROC LOGLINK

### (REPEATED MEASURES EXAMPLE)

Example: epileptic seizures data

$Y_{ij}$  is number of events for subject  $i$  in time period  $j$ , which has length  $t_{ij}$ . A possible model:

$$E(Y_{ij}) = t_{ij}\lambda_{ij}$$

$$\log\lambda_{ij} = x'_{ij}\beta$$

$$\log E(Y_{ij}) = \log t_{ij} + x'_{ij}\beta$$

$t_{ij}$  is the “offset” variable.

$$\text{Var}(Y_{ij}) = \phi_{ij}E(Y_{ij})$$

The working correlation structure  $R$  can be INDEPENDENT (default, no clustering), or EXCHANGEABLE (corresponding to cluster random effects).

---

## REFERENCES

- SUDAAN Manual (2001)
- Binder, D. A. (1983) On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 51, 279 - 292.
- Binder, D. A. (1992) Fitting Cox’s proportional hazards model from survey data. *Biometrika* 79, 139 - 147.
- Boudreau, C. and Lawless, J. F. (2001) Survival analysis based on the proportional hazards model and survey data. Working Paper 2001-10, Department of Statistics and Actuarial Science, University of Waterloo.
- Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994) *Analysis of Longitudinal Data*. Oxford Science Publications.

- Godambe, V. P. (ed.) (1991) *Estimating Functions*. Oxford Science Publications.
- Kalbfleisch, J. D. and Prentice, R. L. (1980) *The Statistical Analysis of Failure Time Data*. Wiley, New York.
- Lawless, J. F. (2000). Event history analysis and longitudinal surveys. To appear in Skinner, C. J. (ed.) *Survey Analysis*. Wiley, New York.
- Lohr, S. L. (1999) *Sampling: Design and Analysis*. Duxbury.
- Thall, P. F. and Vail, S. C. (1990) Some covariance models for longitudinal count data with overdispersion. *Biometrics* 46, 657-671.
- Thompson, M. E. (1997) *Theory of Sample Surveys*. Chapman and Hall, London.
- Wolter, K. M. (1985) *Introduction to Variance Estimation*. Springer-Verlag, New York.
- Zeger, S. L. and Liang, K.-Y. (1986) Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42, 121-130.