サーベイデータ

ランダムサンプリングでないデータを分析する際は、サーベイデザインの加重、クラスタ、および層化を考慮してモデルを推定する必要があります。これらを無視して推定を行うと、推定値にバイアスがかかり、標準誤差を正しく計算できません。

Stata のサーベイデータ分析は、**svy** プリフィックスを用いて行います。**svyset** コマンドでサーベイデザインの構成を設定したのち、推定コマンドの前に「**svy**：」をつけて使用します。

サーベイデザインツール

- **svyset** コマンドを使用して、サーベイデザインの特性を識別する変数と、標準誤差の推定方法を指定します。
  一度設定すると、**svy** を用いたコマンドはこれらのデザイン指定を自動的に使用します。
- 次の 2 つの例が示すように、**svyset** を使用すると、広範囲の複雑なサンプリングデザインを利用できます。
  最初に、シンプルなシングルステージデザインを示し、次に複雑なマルチステージデザインを示します。

✧ 本文中のコマンドをコピーし、Stata のコマンドウィンドウに貼り付けて実行できます。
全ての操作のコマンドは、do ファイル surveydata.do にまとめられています。

例題 1：シングルステージデザインのサーベイデータ

- シングルステージのサーベイデザインでは、ストラタ（層）にまたがってクラスタサンプリングを行います。
- このとき、クラスタは復元を許さずにサンプリングします。
- Stata でサーベイデザインを利用する場合、ストラタ、PSU（クラスタ）、サンプリング加重、FPC (Finite population correction)の情報を設定します。
- 次の例では、実際の変数として **strata**、**su1**、**pw**、**fpc1** を用いて設定を行います。

データセットをインポートしてサーベイデザインの情報を設定します

```
use https://www.stata-press.com/data/r16/stage5a


svyset su1 [pweight=pw], strata(strata) fpc(fpc1)
```

次のように表示されます。

```
      pweight: pw
          VCE: linearized
  Single unit: missing
     Strata 1: strata
         SU 1: su1
        FPC 1: fpc1
```

- **svyset** は指定した変数の他に、デフォルトの標準誤差の推定方法が Taylor linearization であることを示しています。
- **svy** は 1 つのストラタ（singleton strata）に 1 つのサンプリングユニットしか存在しない場合、標準誤差に欠損値を報告します。

例題 2：マルチステージデザインのサーベイデータ

- 次のマルチステージデザインに基づいて収集された、アメリカの高校生の架空の身長や体重のデータがあります。
  - 第一ステージでは、郡は各州内で独立して選ばれました。
  - 第二ステージでは、選択した各郡内の学校が選ばれました。選ばれた各学校内で、すべての高校生がアンケートに回答しました。
- サーベイデザイン変数は次の通りです。

state	ストラタの ID
county	第一ステージのサンプリングユニット
ncounties	各州内の郡の総数
school	第二ステージのサンプリングユニット
nschools	各郡内の学校の総数
sampwgt	サンプリングされた各個人のサンプリング加重

データセットをインポートしてサーベイデザインの情報を設定します。

```
use https://www.stata-press.com/data/r16/multistage

svyset county [pw=sampwgt], strata(state) fpc(ncounties) || school, fpc(nschools)
```

次のように表示されます。

```
      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: state
         SU 1: county
        FPC 1: ncounties
     Strata 2: <one>
         SU 2: school
        FPC 2: nschools
```

サーベイデザインの情報を設定したデータセットを highschool.dta として保存します。

```
save highschool
```

これで、今後はサーベイデザインの情報を再設定必要する必要はありません。この新しいデータセットを読み込み直して、svyset を実行します。

```
clear

use highschool

svyset
```

次のように表示されます。

```
      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: state
         SU 1: county
        FPC 1: ncounties
     Strata 2: <one>
         SU 2: school
        FPC 2: nschools
```

- svyset でデザイン特性を指定したあと、svydescribe コマンドを使用して、サーベイデータの各ステージを確認できます。
- svydescribe はサンプリングユニット数、欠損データ、1 つのストラタ（singleton strata）に関する有用な情報を報告します。

例題 3：サーベイデータの記述

- svydescribe を使用して、高校生のサーベイデータセットの第一ステージの情報を確認します。
- 変数 weight を指定して、svydescribe で欠損値が含まれている層と、これが推定サンプルにどのように影響するかを確認します。

次のコマンドを実行します。

```
svydescribe weight
```

結果は次のようになります。

```
Survey: Describing stage 1 sampling units

      pweight: sampwgt
          VCE: linearized
  Single unit: missing
     Strata 1: state
         SU 1: county
        FPC 1: ncounties
     Strata 2: <one>
         SU 2: school
        FPC 2: nschools
```

Stratum	#Units included	#Units omitted	#Obs with complete data	#Obs with missing data	#Obs per included Unit min	mean	max
1	2	0	92	0	34	46.0	58
2	2	0	112	0	51	56.0	61
3	2	0	43	0	18	21.5	25
4	2	0	37	0	14	18.5	23
5	2	0	96	0	38	48.0	58
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
45	[illegible]	0	[illegible]	0	[illegible]	53.[illegible]	[illegible]
46	2	0	115	0	56	57.5	59
47	2	0	67	0	28	33.5	39
48	2	0	56	0	23	28.0	33
49	2	0	78	0	39	39.0	39
50	2	0	64	0	31	32.0	33
50	100	0	4,071	0	14	40.7	81
				4,071			

結果から、次のことがわかります。

- ストラタの数は 50
- 各ストラタは 2 つの PSU からなる
- PSU のサイズは様々である
- 総サンプルサイズ（高校生の人数）は 4,071 である
- 変数 **weight** には欠損データが存在しない

サーベイデータ解析ツール

- Stata の一連のサーベイデータコマンドは、通常の推定コマンドに **svy** プリフィックスを用いて行います。
- **svy** は、点推定でのサーベイデザイン特性および分散推定手法を考慮しながら、指定された推定コマンドを実行します。
  利用可能な分散推定手法は次の通りです。
  - BRR (balanced repeated replication)
  - bootstrap
  - jackknife
  - successive difference replication
  - first-order Taylor linearization
- デフォルトでは、**svy** は linearized variance estimator を使用して標準誤差を計算します。いわゆる first-order Taylor series linear approximation (Wolter 2007)に基づきます。

例題 4：母集団の平均の推定

mean コマンドに svy プリフィックスを使用して、高校生の平均体重を推定します。

```
svy: mean weight
```

結果は次のようになります。

```
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      50        Number of obs    =      4,071
Number of PSUs   =     100        Population size =  8,000,000
                                  Design df        =         50

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
      weight |   160.2863   .7412512      158.7974    161.7751
--------------------------------------------------------------
```

- ヘッダでは、第一ステージのストラタと PSU の数、サンプルサイズ、推定母集団サイズ、デザインの自由度を報告します。
- 結果の表には通常の mean コマンドの出力結果と同様に、推定平均とその標準誤差、信頼区間を表示します。

例題 5：サーベイ回帰

regress コマンドに **svy** プリフィックスを使用して、高校生の体重と身長の関係をモデル化します。

```
svy: regress weight height
```

結果は次のようになります。

```
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        50                  Number of obs     =      4,071
Number of PSUs     =       100                  Population size   =  8,000,000
                                                Design df         =         50
                                                F(   1,     50)   =     593.99
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2787

------------------------------------------------------------------------------
             |             Linearized
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      height |   .7163115   .0293908    24.37   0.000     .6572784    .7753447
       _cons |  -149.6183   12.57265   -11.90   0.000    -174.8712   -124.3654
------------------------------------------------------------------------------
```

- 例題 4 で **svy: mean** を使用したときのヘッダ要素に加え、**svy: regress** コマンドはモデルの *F* 検定と R^2 を出力します。
- 多くの Stata のモデルフィッティングコマンドは、係数がゼロであることを検定する *Z* 統計量を報告しますが、**svy** は常に *t* 統計量を報告し、デザインの自由度を使用して *p* 値を計算します。

例題 6：Cox 比例ハザードモデル

- 3 つの危険因子（喫煙状況、性別、居住地）を使用して肺がんの発生率をモデル化します。
- 今回のデータセットは長期的な健康調査のものです。
  the First National Health and Nutrition Examination Survey (NHANES I) (Miller 1973; Engel et al. 1978)、およびその 1992 年の疫学的追跡調査（NHEFS) (Cox et al. 1997)
- NHANES I 試験地点 1–65 および 66–100 のサンプルのデータを使用します。
  これらの地点に関連付けられた PSU およびストラタの変数を **svyset** します。
- 変数 **pweight** にはサンプリングしたデータの加重が用意されています。

データセットをインポートしてサーベイデザインの情報を設定します。

```
use https://www.stata-press.com/data/r16/nhefs

svyset psu2 [pw=swgt2], strata(strata2)
```

次のように表示されます。

```
      pweight: swgt2
          VCE: linearized
  Single unit: missing
     Strata 1: strata2
         SU 1: psu2
        FPC 1: <zero>
```

- 肺がんの情報は、1992 年の NHEFS インタビューデータから取得されました。
- 参加者の年齢を時間スケールとして使用します。
- 肺がんに罹患したことがなく、1992 年のインタビューで生存していた参加者のデータは除外します。
  肺がんに罹患したことがなく、1992 年のインタビューの前に死亡した参加者も、死亡年齢で除外します。

データセットを生存時間データとして宣言します。

```
stset age_lung_cancer [pw=swgt2], fail(lung_cancer)
```

次のように表示されます。

```
     failure event:  lung_cancer != 0 & lung_cancer < .
obs. time interval:  (0, age_lung_cancer]
 exit on or before:  failure
            weight:  [pweight=swgt2]

--------------------------------------------------------------------------------
     14,407  total observations
      5,126  event time missing (age_lung_cancer>=.)            PROBABLE ERROR
--------------------------------------------------------------------------------

      9,281  observations remaining, representing
         83  failures in single-record/single-failure data
    599,691  total analysis time at risk and under observation
                                                at risk from t =         0
                                     earliest observed entry t =         0
                                          last observed exit t =        97
```

- イベント時間が欠損している 5,126 の観測値があることは“probable error”であると **stset** は報告していますが、1992 年の NHEFS のドキュメントを参照すると、実際に完全な情報を持つ 9,281 人の参加者がいたことを確認できます。
- 喫煙状況は 1992 年の NHEFS インタビューデータから取得されました。NHANES I の一般的な病歴補足データを使用して、132 を除くすべての欠損値を記入しました。
- 喫煙状況は、以前の喫煙者と現在の喫煙者の別々の指標変数によって表されます。
  基本比較グループは非喫煙者です。
- 性別は 1992 年の NHEFS バイタルデータを使用して決定され、男性の指標変数によって表されます。
- 居住地情報は NHANES I の病歴アンケートから取得され、農村部および人口の多い（100 万人を超える）都市住宅の個別の指標変数によって表されます。
  基本比較グループは、人口が 100 万人未満の都市住宅です。

Cox 比例ハザードモデル分析を実行します。

```
svy: stcox former_smoker smoker male urban1 rural
```

結果は次のようになります。

```
(running stcox on estimation sample)

Survey: Cox regression

Number of strata   =        35                  Number of obs      =         9,149
Number of PSUs     =       105                  Population size    =   151,327,827
                                                Design df          =            70
                                                F(   5,     66)    =         14.07
                                                Prob > F           =        0.0000
```

_t	Haz. Ratio	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
former_smoker	2.788113	.6205102	4.61	0.000	1.788705	4.345923
smoker	7.849483	2.593249	6.24	0.000	4.061457	15.17051
male	1.187611	.3445315	0.59	0.555	.6658757	2.118142
urban1	.8035074	.3285144	-0.54	0.594	.3555123	1.816039
rural	1.581674	.5281859	1.37	0.174	.8125799	3.078702

- 上記の結果から、以前の喫煙者、現在の喫煙者ともに、非喫煙者よりも肺がんを発症するリスクが大幅に高いことがわかります。
- `svy: tabulate` を使用して、サーベイデータの一元表、二元表を作成し、二元分割表の独立性の検定(サーベイ調整済み)を実行できます。

例題 7：サーベイデータの二元表

- **svy: tabulate** を使用して、the Second National Health and Nutrition Examination Survey (NHANES II)(McDowell et al. 1981)のデータから、セルの比率とその標準誤差および信頼区間の二元表を作成します（サーベイデザインの特性は既に **svyset** で指定されています）。

データセットをインポートし、**svy: tabulate** に **format()**オプションを使用して、セル中の値と周辺値を小数点第 4 位まで表示します。

```
use https://www.stata-press.com/data/r16/nhanes2b

svy: tabulate race diabetes, row se ci format(%7.4f)
```

結果は次のようになります。

```
(running tabulate on estimation sample)

Number of strata   =        31                  Number of obs      =       10,349
Number of PSUs     =        62                  Population size    =  117,131,111
                                                Design df          =           31

------------------------------------------------------------------------------
1=white,        |
2=black,        |                diabetes, 1=yes, 0=no
3=other         |                0                1            Total
----------------+-------------------------------------------------------------
    White       |           0.9680            0.0320           1.0000
                |         (0.0020)          (0.0020)
                |  [0.9638,0.9718]   [0.0282,0.0362]
                |
    Black       |           0.9410            0.0590           1.0000
                |         (0.0061)          (0.0061)
                |  [0.9271,0.9523]   [0.0477,0.0729]
                |
    Other       |           0.9797            0.0203           1.0000
                |         (0.0076)          (0.0076)
                |  [0.9566,0.9906]   [0.0094,0.0434]
                |
    Total       |           0.9658            0.0342           1.0000
                |         (0.0018)          (0.0018)
                |  [0.9619,0.9693]   [0.0307,0.0381]
------------------------------------------------------------------------------
  Key:  row proportion
        (linearized standard error of row proportion)
        [95% confidence interval for row proportion]

  Pearson:
    Uncorrected   chi2(2)         =   21.3483
    Design-based  F(1.52, 47.26)  =   15.0056     P = 0.0000
```

- svy コマンド実行後に、全ての標準的な事後検定コマンド（例：estimates、lincom、margins、nlcom test、testnl）が使用可能です。

例題 8：平均比較

例題 2 の高校生のサーベイデータに話を戻します。性別変数のカテゴリ（男性と女性）によって識別された各部分母集団の体重の平均（単位：ポンド）を推定します。

```
use https://www.stata-press.com/data/r16/highschool

svy: mean weight, over(sex)
```

結果は次のようになります。

```
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =      50        Number of obs    =      4,071
Number of PSUs   =     100        Population size  =  8,000,000
                                  Design df        =         50

--------------------------------------------------------------
             |             Linearized
             |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
c.weight@sex |
        male |   175.4809   1.116802      173.2377    177.7241
      female |    146.204   .9004157      144.3955    148.0125
--------------------------------------------------------------
```

ここで、**test** コマンドを使用して、平均的な男性は平均的な女性よりも 30 ポンド重いという仮説検定を実行します。

```
test weight#1.sex - weight#2.sex = 30
```

結果は次のようになります。

```
Adjusted Wald test

 ( 1)  c.weight@1bn.sex - c.weight@2.sex = 30

       F(  1,    50) =    0.23
            Prob > F =    0.6353
```

例題 9：デザイン効果

estat effects を用いて平均推定のデザイン効果 DEFF および DEFT を求めます。

```
estat effects
```

結果は次のようになります。

Over	Mean	Linearized Std. Err.	DEFF	DEFT
c.weight@sex				
male	175.4809	1.116802	2.61016	1.61519
female	146.204	.9004157	1.7328	1.31603

Note: Weights must represent population totals for deff to be correct when using an FPC; however, deft is invariant to the scale of weights.

次に、**estat lceffects** を用いて平均推定値の差に対するデザイン効果 DEFF および DEFT を求めます。

```
estat lceffects weight#1.sex - weight#2.sex
```

結果は次のようになります。

```
 ( 1)  c.weight@1bn.sex - c.weight@2.sex = 0
```

Mean	Coef.	Std. Err.	DEFF	DEFT
(1)	29.27691	1.515201	2.42759	1.55768

Note: Weights must represent population totals for deff to be correct when using an FPC; however, deft is invariant to the scale of weights.