Part1. BRFSS (Behavioral Risk Factor Surveillance System) Overview

Go BRFSS website: http://www.cdc.gov/brfss/technical_infodata/surveydata.htm to obtain **data**, **codebook**, **questionnaire**, and other detailed information.

Characteristics of BRFSS system:

1. Collaborative project (by CDC (web: http://www.cdc.gov/) + other health & education agencies to monitor risk factors)
2. Diverse purposes for use (identify, measure, address, and propose)
3. Effective disease prevention & health promotion tool at state level

It is the largest continuously conducted telephone survey monitoring health behavioral risk factors on the state or local base.

Sampling Target for State BRFSS Datasets: the civilian, non-institutionalized adult (> aged 17) population (**exception**: the non-adult population aged 17 years and less, for example, in 2005 state BRFSS data for childhood asthma prevalence and childhood immunization for influenza)[a].

Two BRFSS downloadable datasets:

1. SMART (Selected Metropolitan/Micropolitan Area Risk Trends, especially, estimates for local metropolitan areas)
2. State BRFSS data

Three parts of BRFSS Questionnaire (by CDC and state health department together):

1. Core components (a standard set of questions by all states)
    a. Fixed core (every year, current health behaviors and demographics)
    b. Rotating core
        i. Offering topics on alternative years (e.g., hypertension, injuries, alcohol use, vaccinations, colorectal screening, and cholesterol in odd number year; physical activities, fruit and vegetable consumption, and weight control in even number year)
        ii. Emerging core (in case of "late-breaking" health issues and potential future use in the fixed core)
2. Optional modules[b] (by CDC about special topics such as indoor air quality or smoking cessation)
3. State added questions (flexible in changes, additions, and deletions of the questions at any time of the year without interference by CDC)

Quality Control of BRFSS Data:

1. Interviewer training and monitoring.
    a. Verification callbacks
    b. Monthly data editing
2. High, medium density stratification of a list-assisted random digit dialing (RDD) sampling

[a] http://www.cdc.gov/HealthyYouth/yrbs/index.htm for Youth Risk Behavior Surveillance System (YRBSS)
[b] http://apps.nccd.cdc.gov/BRFSSModules/ModByState.asp?Yr=2006

Part2. Obtaining Data

There will be ASCII and SAS format as an option you can download (I recommend to rather download SAS format data (SAS Transport Format)(click: *.exe). Why? Because you can keep variable names and labels.). After downloaded, CDBRFS04.XPT may be seen to you, which contains all states in the US.

Import data or Export to STATA format:
The simplest way is to use STATA for converting of SAS Xport data (xpt) into STATA:
* "`help fdause`" in STATA
* Import state BRFSS data 2000, 2002, 2004, 2006:

```
fdause "filename", clear
codebook, compact
```
for reviewing data or variables

Other Conversion Options (example using 2004):
1. Convert SAS Xport Transport File (xpt) into SAS data set (sas7bdat) using SAS command.

Step1. Double click of "CDBRFS04XPT.exe" for unzip and save in your data directory
Step2. Double click of "CDBRFS04.XPT" (which will be automatically saved in your SAS temporary work directory)
Step3. To save it as permanent file in your data directory (using SAS command as below)

```
libname perm 'W:\BRFSS\state.brfss.04';
data perm.brf04;
set work.Cdbrfs04;
proc contents varnum position data=perm.brf04; run;
```

* Note:
"libname" statement defines a directory where SAS data may be located.
"data" statement gives a name to the data.
"set" statement tells where the data is temporarily.
"proc contents" statement shows whether it works or not.

2. Convert SAS data set (sas7bdat) into STATA data set (*.dta):

```
libname in "W:\BRFSS\state.brfss.04";
proc export data=in.brf04 outfile="w:\BRFSS\state.brfss.04\brf04.csv"
dbms=csv replace;
run;

Now, here are the STATA commands to read in the ASCII file and save it.

cd "W:\BRFSS\state.brfss.04"
insheet using brf04.csv
save brf04
```

Select "Texas":
STATA command:
`keep if _state==48`
* By use of STATE FIPS CODE (variable name: _state; Texas=48)

Review the data:

`codebook, compact` (in STATA) or STATA command "`describe`"

Analyze the data:
Use "do" file of STATA (Refer to the programs)

Part3. Research Background (Problems & Significance)

Understanding the Downward Trend of Breast and Cervical Cancer Screening Rate in the US & Texas: Source: http://apps.nccd.cdc.gov/brfss/Trends/TrendData.asp

General question: Why does the downward trend in breast and cervical cancer screening rate[1] happen?

According to the following research questions, data for the demo will be selected.
1. Trends in the Screening Rate for Breast and Cervical Cancer in TX: 2000, 2002, 2004 & 2006.
2. *What Risk Factors Affect Screening Rate for Breast and Cervical Cancer Prevention Most After Adjusting for Confounders?*

How do you answer these research questions?

Data: 2000, 2002, 2004, 2006 Texas BRFSS data (Why 2001 & 2003 not included?)[1]

Outcome Variables (Dependent or Response Variables)

Breast Cancer Screening

1. Have You Ever Had a Mammogram (HADMAM)
2. How Long since Last Mammogram (HOWLONG)
3. Women aged 40+ that have had a mammogram in the past two years(_RFMAM2Y)**

Cervical Cancer Screening

1. Ever Had a Pap Test (HADPAP2)
2. How Long Since Last Pap Test (LASTPAP2)
3. Women aged 18+ that have had a pap test in the past three years (_RFPAP32)**

Risk Factors (Independent Variables)

1. *Insurance* (HLTHPLAN)
2. *Income* (_INCOMG)
3. *Education* (_EDUCAG)
4. *Cost Burden* (MEDCOST) (No variable in 2002)
5. *Primary Care Accessibility* (PERSDOC2)** (No obs. in 2000)

Confounders

1. *Age* (*AGE*)
2. *Health Status* (GENHLTH)
3. *Marriage* (MARITAL)
4. *Employment* (EMPLOY)
5. *Race/Ethnicity* (HISPANC2, RACE2)

** Available in 2002, 2004, 2006

Model: `xi: svy: logistic mam40 hlthplanc _incomg i._educag medcostc persdoc2c i.ageg genhlthc maritalc employc i.raceeth`

* Real variable names of STATA format data sets are listed in programming page.
* Decision on selecting of outcome variables, risk factors, and confounders is up to your own research questions.

Part4a. Understanding Survey Data

State BRFSS Data Sampling: based on
1. probability or
2. not (non-probability)

* Estimates by probability sampling are closer to the population with generalized (standardized) point estimates and confidence intervals (CIs). Simple random samples (SRS), systematic RS, cluster, multistage, and complex survey sample are part of probability sampling.

Three instrumental components in analyzing survey data reflecting survey design:
1. sampling weights (for considering the probability of selection of samples)
2. clustering (for more cost efficient in sampling than SRS)
3. stratification (for more effective in sampling than SRS)

* **Sampling weights** correct differential sampling rates (over-sampled or not), non-response adjustments, and post-stratifications adjustments (using census figures which are age, gender, races). **Clustering** is grouping of individuals (or per household) which needs to adjust for the variability of samples in the population. **Stratification** is grouped as strata which the sampling is supposed to be conducted independently across the strata.

BRFSS data = state-based complex survey design data using disproportionate stratified sampling (**DSS**=based on random digit dialing (**RDD**=sampling from listed and unlisted numbers (by the Bell Core Research (**BCR**) sampling frame) with computer-assisted telephone interviewing (**CATI**) systems).

Part4b. Weighting the BRFSS Data Set As Complex Survey Data

Use reference for general information on complex survey data and analysis (Aday, 2006)[2] or on-line reference: http://sru.soc.surrey.ac.uk/SRU43.html

What should be considered for appropriate point estimates of standard errors and CIs? No weighting leads to inaccurate inference of the results (bias the results) which are reflected by over/under estimated standard errors (Design effect (**DEFF**) tells whether more efficient (if DEFF<1) or less efficient (if DEFF>1) than SRS (DEFF=1).) General formula in weighting the state BRFSS data: (source: http://www.cdc.gov/brfss/technical_infodata/weighting.htm) are the following.

FINALWT = STRWT * 1/NPH * NAD * POSTSTRAT

* Note:
FINALWT is the final weight assigned to each respondent.
STRWT accounts for differences in the basic probability of selection among strata (subsets of area code/prefix combinations). It is the inverse of the sampling fraction of each stratum. There is seldom a complete correspondence between strata, which are defined by subsets of area code/prefix combinations, and regions, which are defined by the boundaries of government entities.
1/NPH is the inverse of the number of residential telephone numbers in the respondent's household.

NAD is the number of adults in the respondent's household.
POSTSTRAT is the number of people in an age-by-sex or age-by-race/ethnicity-by-sex category in the population of a region or a state divided by the sum of the preceding weights for the respondents in the same age-by-sex or age-by-race/ethnicity-by-sex category. It adjusts for noncoverage and nonresponse and forces the sum of the weighted frequencies to equal population estimates for the region or state.

Actual variables for the state BRFSS data set using STATA: STATA `svyset` Command:

svyset _psu [pw=**_finalwt**], strata(**_ststr**)

Based on Formula as

_finalwt=(_strwt)*(wt_prob)*(1)*(_poststr)

* Note:
_finalwt =provability weighting variable
_strwt =(nrecstr/nrecsel) = GEOWT
_raw =(numadult/_impnph)= (NAD/NPH)
wt_prob =_raw*1.5 if _denstr2=2,
wt_prob =_raw, otherwise
_poststr = age, gender, race within state population

where,
_raw = Raw weighting factor (compared with the density weighting factor)
nrecsel =Number of sample records selected from stratum
nrecstr =Number of telephone numbers in stratum sample
numadult =Number of adults in the household
_impnph =Number of imputed telephones
_denstr2 =Household density stratum code where,
1= High density and 2= Medium density
(The sampling ratio of high density numbers to medium is 1.5 to 1.0 in the density weighting)

_psu = primary sampling unit which reflects clustering as the annual sequence number "seqno"

_ststr= stratification variable which is a combination of state fips, geographic code, and density stratum code
"`help svyset`" in STATA for more details

* The rest of parts are programs in do file.

Reference:

1. Centers for Disease Control and Prevention (CDC). Use of mammograms among women aged >/= 40 years--united states, 2000-2005. *MMWR Morb Mortal Wkly Rep*. 2007;56:49-51.
2. Aday, Lu Ann. Cornelius,Llewellyn Joseph. *Designing and Conducting Health Surveys : A Comprehensive Guide.* 3rd ed. San Francisco: Jossey-Bass; 2006.

STATA Programs (using do file of STATA):

Data Obtaining:

```
* import data from sas.xpt:

fdause  "C:\Documents and Settings\mseo\Desktop\CDBRFS06.XPT"
fdause  "W:\BRFSS\state.brfss.04\CDBRFS04.XPT"
fdause  "W:\BRFSS\state.brfss.02\CDBRFS02.XPT"
fdause  "W:\BRFSS\state.brfss.00\CDBRFS00.XPT"

* obs: 06:355710(TX:6854) 04:303822(6317), 02:247964(6107), 00:184450(5018)
codebook  seqno _psu, com
* keep only TX
keep if _state==48
codebook  seqno _psu, com

save "W:\BRFSS\state.brfss.06\tx.brfss.06.dta"
save "W:\BRFSS\state.brfss.04\tx.brfss.04.dta"
save "W:\BRFSS\state.brfss.02\tx.brfss.02.dta"
save "W:\BRFSS\state.brfss.00\tx.brfss.00.dta"
```

Data Management:

```
********************************************************** SELECT TX
keep if _state==48

***SVYSET (state data)
svyset _psu [pw=_finalwt], strata(_ststr)

or

svyset, clear

********************************************************** DATA MGT.

*** GEN OUTCOME VARS:

gen hadmamc=.
replace hadmamc=1 if hadmam==1
replace hadmamc=0 if hadmam==2
* 1= YES , 0= NO

* dichotomize
gen howlongc=.
replace howlongc=0 if howlong<3
replace howlongc=1 if howlong>=3

label def howlong 0 "<2yrs" 1 ">=2yrs"
label val howlongc howlong

* 2002 "hadpap"
gen hadpap2c=.
replace hadpap2c=1 if hadpap2==1
replace hadpap2c=0 if hadpap2==2
* 1= YES , 0= NO

* 2000, 2002 "lastpap"

* dichotomize 2004

gen lastpap2c=.
replace lastpap2c=0 if lastpap2<=2
replace lastpap2c=1 if lastpap2>=3
```

```
* 2000 2002

gen lastpap2c=.
replace lastpap2c=0 if lastpap<=2
replace lastpap2c=1 if lastpap>=3

label def lastpap2 0 "<2yrs" 1 ">=2yrs"
label val lastpap2c lastpap2

*** 2000 no mam40, no pap18

gen mam40=.
replace mam40=1 if _rfmam2y==1
replace mam40=0 if _rfmam2y==2
* 1=YES , 0= NO in the past 2 yrs

gen pap18=.
replace pap18=1 if _rfpap32==1
replace pap18=0 if _rfpap32==2
* 1= YES, 0= NO in the past 3 yrs

* 2002 "_rfpap3y"

gen mam40=.
replace mam40=1 if _rfmam2y==1
replace mam40=0 if _rfmam2y==2

gen pap18=.
replace pap18=1 if _rfpap3y==1
replace pap18=0 if _rfpap3y==2

*** 2000 needs to create mam40, pap18 (refer to codebook)

gen mam40=.
replace mam40=1 if sex==2 & age>=40 & hadmam==1 & howlong<3
replace mam40=0 if sex==2 & age>=40 & hadmam==1 & (howlong>3 & howlong<6)
* 1=YES , 0= NO in the past 2 yrs

gen pap18=.
replace pap18=1 if sex==2 & age>=18 & hadpap2==1 & (lastpap2<=3)
replace pap18=0 if sex==2 & age>=18 & hadpap2==1 & (lastpap2>3 & lastpap<7)
replace pap18=0 if sex==2 & age>=18 & hadpap2==2
* 1= YES, 0= NO in the past 3 yrs

*** GENERATE hlthplanc, persdoc2c, medcostc

gen maritalc=.
replace maritalc=1 if marital==1
replace maritalc=0 if marital>=2 & marital<=6
* 1=married, 0=not

gen hlthplanc=.
replace hlthplanc=1 if hlthplan==1
replace hlthplanc=0 if hlthplan==2
* 1=YES with any health insurance, 0=NO

gen persdoc2c=.
replace persdoc2c=1 if persdoc2==1 | persdoc2==2
replace persdoc2c=0 if persdoc2==3
* 1=YES with personal doctors, 0=NO

* 2000 "persdoc"
gen persdoc2c=.
replace persdoc2c=1 if persdoc==1 | persdoc==2
replace persdoc2c=0 if persdoc==3
* 1=YES with personal doctors, 0=NO
```

```
gen medcostc=.
replace medcostc=1 if medcost==1
replace medcostc=0 if medcost==2
* 1=YES with NOT-use due to medical cost, 0=NO with use

*** OTHER VARS.

*** 2000 "hispanic", "race" instead of "hispanc2" "race2"
gen raceeth=.
replace raceeth=1 if hispanc2==2 & race2==1
replace raceeth=2 if hispanc2==2 & race2==2
replace raceeth=3 if hispanc2==2 & race2==3
replace raceeth=4 if hispanc2==1 & race2==8
replace raceeth=5 if race2>3 & race2<8 & hispanc2==2

* 2000
gen raceeth=.
replace raceeth=1 if hispanic==2 & race==1
replace raceeth=2 if hispanic==2 & race==2
replace raceeth=3 if hispanic==2 & race==6
replace raceeth=4 if hispanic==1 & (race==3| race==4| race==5)
replace raceeth=5 if hispanic==2 & (race==7 | race==8)

label def raceeth 1 "white" 2 "black" 3 "asian" 4 "hispanic" 5 "other"
label val raceeth raceeth

gen employc=.
replace employc=1 if employ==1 | employ==2
replace employc=0 if employc!=1

label def employc 1 "employed" 0 "unemployed"
label val employc employc


*** 2000, 2002 no _incomg or _educag

gen _incomg=.
replace _incomg=0 if income2<7
replace _incomg=1 if income2==7 | income2==8

gen _educag=.
replace _educag=1 if educa==1
replace _educag=2 if educa==2
replace _educag=3 if educa==3
replace _educag=4 if educa==4

gen ageg=.
replace ageg=1 if age>=18 & age<45
replace ageg=2 if age>=45 & age<65
replace ageg=3 if age>=65

label def ageg 1 "18-44" 2 "45-64" 3 "65+"
label val ageg ageg

gen genhlthc=.
replace genhlthc=1 if genhlth==1 | genhlth==2 | genhlth==3
replace genhlthc=0 if genhlth==4 | genhlth==5

label def genhlth 1 "good+" 0 "fair, poor"
label val genhlthc genhlth

* label _incomg, _educag

label def incomg 0 "<50T" 1 "50T+"
label val _incomg incomg

label def educag 1 "<high schl" 2 "grad high" 3 "attd coll" 4 "grad coll"
```

```
label val _educag educag

/* Model */

xi: svy: logistic  mam40  hlthplanc _incomg i._educag medcostc persdoc2c i.ageg
genhlthc maritalc employc i.raceeth
```

Data Analysis:

```
***SVYSET (state data)
svyset _psu [pw=_finalwt], strata(_ststr)

(or svyset, clear)

*2004
codebook hadmamc howlongc hadpap2c lastpap2c mam40 pap18 maritalc hlthplanc persdoc2c
medcostc ageg40 raceeth ///
employc ageg _incomg _educag genhlthc, com

*2002
no "medcostc"

*2000
no "persdoc2c"

*** outcome vars.
hadmamc howlongc hadpap2c lastpap2c mam40 pap18

*** risk factors
hlthplanc _incomg _educag medcostc persdoc2c

*** confounders
ageg genhlthc maritalc employc raceeth

*** run do file

svy: proportion hadmamc (or svy: tab hadmanc)
svy: proportion howlongc
svy: proportion mam40

svy: proportion hadpap2c
svy: proportion lastpap2c
svy: proportion pap18

svy: proportion hadmamc, over(ageg)
svy: proportion howlongc, over(ageg)
svy: proportion mam40, over(ageg)

svy: proportion hadpap2c, over(ageg)
svy: proportion lastpap2c, over(ageg)
svy: proportion pap18, over(ageg)
```

```
xi: svy: logistic  mam40  hlthplanc _incomg i._educag medcostc persdoc2c i.ageg  genhlthc
maritalc employc i.raceeth
```

```
xi: svy: logistic  pap18  hlthplanc _incomg i._educag medcostc persdoc2c i.ageg  genhlthc
maritalc employc i.raceeth

********** interaction term  (xi: ... i.(var)*(var)...) if needed,
********** mlogit for multinomical response variable, ologit for ordered response vars.
********** test if needed for the multinomial independent variables
```

Munseok Seo is working in Health Services Research Collaborative (HSRC) of UT SPH as senior research associate (www.sph.uth.tmc.edu/hsrc).