```
log using "D:\Jason\workshop\Key Stata commands\key stata command2.log", replace

*********************************************************************
* Example 1: Constructing variables using data in wide format
* Research question: how race and experience of violence predictd the number of depressive symptoms.
* Data: public data of Add Health at Wave III
* The number of respondents in the data: 6,504
* The numbers of respondents with valid values of racial backgrounds, experience of violence,
* or depressive symptoms change with the way these variables were constructed.
* The numbers of respondents in the analytic sample symptom also vary, depending on how variables were constructed
*********************************************************************


*************************************************************************************************
* Read in the data and check the number of respondents in the data and the format of data
*************************************************************************************************

use "D:\Jason\workshop\Key Stata commands\Add Health\depression.dta", clear
count
duplicates report aid

*************************************************
* Examine the variables of racial backgrounds
* Consider how the information of these variables should be aggregated

*************************************************

des h3od2 h3od4a h3od4b h3od4c h3od4d h3od6
sum h3od2 h3od4a h3od4b h3od4c h3od4d h3od6
tab1 h3od2 h3od4a h3od4b h3od4c h3od4d h3od6, mis
codebook h3od2 h3od4a h3od4b h3od4c h3od4d h3od6

*************************************************************************
* generate new race variables and code the missing values of these variables
*************************************************************************

clonevar hispanic = h3od2
clonevar white = h3od4a
clonevar black = h3od4b
clonevar native = h3od4c
clonevar asian = h3od4d
clonevar additional = h3od6


mvdecode hispanic white black native asian additional, mv(6=.\7=.\8=.\9=.)
tab1 hispanic white black native asian additional, mis


*************************************************************************************
* count the number of respondents with valid information on their racial backgrounds
*************************************************************************************

egen valid_race = rownonmiss(hispanic white black native asian additional)
tab1 valid_race, mis

list aid hispanic white black native asian additional valid_race in 1/20, nol


*************************************************************************************************
* check the combination of four racial backgrouns, including being white, black, native, and asian
*************************************************************************************************

egen race_temp1 = concat( white black native asian ), punct(_)
label variable race_temp1 "the combination of racial backgrounds"
tab1 race_temp1, mis


gen race_temp2 =.
label variable race_temp2 "the racial backgrounds, using four racial backgropund variables"
label define race_temp2  1 "1 white" ///
2 "2 black or African American" ///
3 "3 American Indian or N ative American" ///
4 "4 Asian or Pacific Islander"  ///
5 "5 hispanic"                   ///
6 "6 mixed racial backgounds"
label value race_temp2 race_temp2

replace race_temp2 = 1 if race_temp1 == "1_0_0_0"
replace race_temp2 = 2 if race_temp1 == "0_1_0_0"
replace race_temp2 = 3 if race_temp1 == "0_0_1_0"
replace race_temp2 = 4 if race_temp1 == "0_0_0_1"
replace race_temp2 = 6 if inlist(race_temp1, "0_1_1_0", "1_0_0_1","1_0_1_0","1_0_1_1" ,"1_0_1_0","1_1_0_0","1_1_0_1","1_
```

```
tab1 race_temp2, mis


****************************************************
* Incorporating the information of being hispanic
****************************************************


tab2 race_temp2 hispanic, mis

clonevar race_temp3 = race_temp2
replace race_temp3 = 5 if hispanic ==1

tab2 race_temp2 race_temp3 if hispanic ==1, mis
tab2 race_temp2 race_temp3 if hispanic ~=1, mis


*****************************************************************
* Incorporating the information of additional racial information
*****************************************************************

tab2 race_temp3 additional, mis

clonevar race_final = race_temp3
replace race_final = 1 if race_temp3 ==6 & additional ==1
replace race_final = 2 if race_temp3 ==6 & additional ==2
replace race_final = 2 if race_temp3 ==. & additional ==2
replace race_final = 3 if race_temp3 ==6 & additional ==3
replace race_final = 4 if race_temp3 ==6 & additional ==4

tab1 race_final, mis


***********************
* Violence variables
***********************


des h3ds18b h3ds18c h3ds18d h3ds18e h3ds18f h3ds18g

tab1 h3ds18b h3ds18c h3ds18d h3ds18e h3ds18f h3ds18g, mis

mvdecode h3ds18b h3ds18c h3ds18d h3ds18e h3ds18f h3ds18g, mv(6=.\8=.\9=.)


***************************************
* two ways of aggregating the items
***************************************

gen  sum_violence = h3ds18b + h3ds18c + h3ds18d + h3ds18e + h3ds18f + h3ds18g
egen rowtotal_violence = rowtotal (h3ds18b h3ds18c h3ds18d h3ds18e h3ds18f h3ds18g)
egen miss_violence = rowmiss (h3ds18b h3ds18c h3ds18d h3ds18e h3ds18f h3ds18g)



************************************
* compare these two new variables
************************************
sum sum_violence   rowtotal_violence
tab2 sum_violence  rowtotal_violence, mis

***************
* check the data
***************
list aid h3ds18b h3ds18c h3ds18d h3ds18e h3ds18f h3ds18g sum_violence rowtotal_violence miss_violence in 1/20, nol

sort miss_violence
list aid h3ds18b h3ds18c h3ds18d h3ds18e h3ds18f h3ds18g sum_violence rowtotal_violence miss_violence if inlist(miss_vio


****************************
* Depression variables
****************************

des h3sp5 h3sp6 h3sp7 h3sp8 h3sp9 h3sp10 h3sp11 h3sp12
tab1 h3sp5 h3sp6 h3sp7 h3sp8 h3sp9 h3sp10 h3sp11 h3sp12, mis

mvdecode h3sp5 h3sp6 h3sp7 h3sp8 h3sp9 h3sp10 h3sp11 h3sp12, mv(6=.\8=.\9=.)

gen  sum_depression = h3sp5 + h3sp6 + h3sp7 + h3sp8 + h3sp9 + h3sp10 + h3sp11 + h3sp12
egen rowtotal_depression = rowtotal (h3sp5 h3sp6 h3sp7 h3sp8 h3sp9 h3sp10 h3sp11 h3sp12)
egen miss_depression = rowmiss (h3sp5 h3sp6 h3sp7 h3sp8 h3sp9 h3sp10 h3sp11 h3sp12)

*******************
```

```
* check the data
********************
sum sum_depression   rowtotal_depression
tab2 sum_depression  rowtotal_depression, mis
sort miss_depression
list aid h3sp5 h3sp6 h3sp7 h3sp8 h3sp9 h3sp10 h3sp11 h3sp12 sum_depression rowtotal_depression miss_depression if inlist


*************************
* Analysis
*************************

reg sum_depression i.race_final sum_violence
sum sum_depression race_final sum_violence
count
count if sum_depression ~=.
count if sum_depression ~=. & race_final ~=.
count if sum_depression ~=. & race_final ~=. & sum_violence ~=.



reg rowtotal_depression i.race_final rowtotal_violence
sum rowtotal_depression race_final rowtotal_violence
count
count if rowtotal_depression ~=.
count if rowtotal_depression ~=. & race_final ~=.
count if rowtotal_depression ~=. & race_final ~=. & rowtotal_violence ~=.




*****************************************************************************************************************
* Example 2: NSFG
* Research question: how age and smoking is related to the probability of live birth.
* Data: public data of NSFG 2017-2019
* The number of respondents in the data: 3,709
* The number of pregnancy records: 10, 215.
* 3,662 respondents reported a total of 10,007 pregnancy with iive births
* 3,709 respondents have valid information on age, which have impact on 20,215 pregnancy records.
* 1,719 respondents have valid information on soming during pregnancy, which have impact on 2,293 pregnancy records
* The final analytic sample has 2,572 pregnancy records from 1,719 respondents.
*****************************************************************************************************************



use "D:\jason\workshop\Key Stata commands\nsfg\data\preg\example2.dta", clear

count
duplicates report caseid
duplicates report caseid pregordr
list caseid pregordr in 1/50, sepby(caseid)


************************************************
* Determine the number of respondents in the data
* This file has 3,709 female respondents and 10,215 pregnancy records
* respondents have an average of 2.75 pregnancies
************************************************

sort caseid pregordr
by caseid: gen preg_n = _n
by caseid: gen preg_N = _N

label variable preg_n "indicator of pregnancy"
label variable preg_N "total number of pregnancy of a respondent"

list caseid pregordr preg_n preg_N in 1/50, sepby(caseid) nol

tab1 preg_n, mis
sum preg_N if preg_n ==1


*****************************************
* The pregnancy outcomes
*****************************************


tab1  pregend1 pregend2, mis

gen livebirth1 = .
replace livebirth1 = 1 if inlist(pregend1,5,6)
replace livebirth1 = 0 if inlist(pregend1,1,2,3,4)
replace livebirth1 = . if inlist(pregend1,7,8,0)
label variable livebirth1 "outcome of a pregnancy"
```

```
label define livebirth 0 "not a livebirth" 1 "livebirth"
label value livebirth1 livebirth


gen livebirth2 = .
replace livebirth2 = 1 if inlist(pregend2,5,6)
replace livebirth2 = 0 if inlist(pregend2,1,2,3,4)
replace livebirth2 = . if inlist(pregend2,7,8,0)

label variable livebirth2 "additional outcome of a pregnancy"
labe value livebirth2 livebirth


tab2 livebirth1 livebirth2, mis


************************************
* Code the end outcome of pregnancy
************************************


********************
* Sum of the items
********************
gen sum_livebirth = livebirth1 + livebirth2
recode sum_livebirth (2=1)


************************
* the rowtotal function
************************

egen rowtotal_livebirth = rowtotal(livebirth1 livebirth2)
recode rowtotal_livebirth (2=1)


*****************************
* Use the replace commands
*****************************
gen code_livebirth = livebirth1
replace code_livebirth = 1 if livebirth1 ==0 & livebirth2 ==1

label value sum_livebirth livebirth
label value rowtotal_livebirth livebirth
label value code_livebirth livebirth




*************************************************************
* Compare differnt ways of constructin the LIVEBIRTH variable
*************************************************************

count
sum sum_livebirth rowtotal_livebirth code_livebirth

tab2 livebirth1 livebirth2, mis




**************************
* check the age variable
**************************

tab1 ager, mis
tab1 ager, mis nol


***********************
* check the smoke variable
***********************

tab1 postsmks, mis
tab1 postsmks, mis nol
clonevar smoke = postsmks
recode smoke (5=0) (8=.)
label define smoke 1 "was smoking" 0 "was not smoking"
label value smoke smoke

tab2 postsmks smoke, mis


******************
```

```
* Regression analysis
*******************

logit code_livebirth smoke ager

count
count if code_livebirth ~=.
count if code_livebirth ~=. & smoke ~=.
count if code_livebirth ~=. & smoke ~=. & ager ~=.


****************************************
* The number of respondents with valid information on livebirth
* A total of 10,007 livebirths came from 3,662 female respondents
****************************************

tab1 code_livebirth


gen livebirth_temp = 1 if code_livebirth~=.

sort caseid livebirth_temp
by caseid: gen valid_livebirth = _n

tab1 valid_livebirth
tab1 valid_livebirth if livebirth_temp ==1



list caseid pregordr code_livebirth livebirth_temp valid_livebirth in 1/50, sepby(caseid) nol
list caseid pregordr code_livebirth livebirth_temp valid_livebirth if inlist(caseid, 92033, 91773), sepby(caseid) nol



****************************************
* The number of respondents with valid information on smoking
* 2,293 pregnancy where mothers smoked during pregnancy.
* These pregnancies were from 1,719 female respondents.
****************************************


tab1 smoke


gen smoke_temp = 1 if smoke ~=.
sort caseid smoke_temp
by caseid: gen valid_smoke = _n

tab1 valid_smoke if smoke_temp ==1



**********************************************************************
* 3,709 female respondents reported age for 10,215 pregnancy records
**********************************************************************

tab1 ager, mis
sort caseid ager
by caseid: gen valid_ager = _n
tab1 valid_ager if ager ~=. , mis



*************************************************************************************************************
* valid cases of the analytic sample
* 1,719 female respondents povide valid information on live births, smoking, and age for 2,572 pregnancies
*************************************************************************************************************

gen valid_temp = 1 if code_livebirth ~=. & smoke ~=. & ager ~=.
tab1 valid_temp, mis


sort caseid valid_temp
by caseid: gen valid_all = _n
tab1 valid_all if valid_temp ==1, mis


log close
```