Eliminating Duplicate Cases in Stata[1]

```
* "h:\000\STATA_doc\DelDupCases.do"
* This is a sample program that eliminates duplicate cases from a datset.
* The sample data mimics data from CPS.
* People can be interviewed for up to three years.
* This researcher wants to save all of the cases that were interviewed
* in 2008, and any cases that were interviewed in 2007 and 2009
* who were not interviewed in 2008, and she wants only one case per person
* even if they were interviewed in multiple years.
** I have color coded this for you.
** My comments are in green, my commands to Stata are in blue, and the things
** stata has to tell me are in black.

** Get the data set.
use "h:\000\STATA_doc\test.dta", clear

** Show you the data as they are now .
list
     +----------------+
     | year   id   v1 |
     |----------------|
  1. | 2007    1    1 |
  2. | 2008    1    2 |
  3. | 2009    1    3 |
  4. | 2007    2    4 |
  5. | 2008    2    5 |
     |----------------|
  6. | 2009    2    6 |
  7. | 2007    3    7 |
  8. | 2009    4    8 |
     +----------------+

** create a variable named tyear.  If year = 2008 the value = 0.
** otherwise it will be 1. This variable will be used to sort
** so that we can keep ids for the year 2008 if they exist .
gen tyear = 1
replace tyear = 0 if year == 2008
(2 real changes made)

** Sort so that we can identify duplicate cases later.
sort id tyear

** Show you the data as they are now .
list
     +------------------------+
     | year   id   v1   tyear |
     |------------------------|
  1. | 2008    1    2       0 |
  2. | 2009    1    3       1 |
  3. | 2007    1    1       1 |
  4. | 2008    2    5       0 |
  5. | 2007    2    4       1 |
     |------------------------|
  6. | 2009    2    6       1 |
  7. | 2007    3    7       1 |
  8. | 2009    4    8       1 |
     +------------------------+
```

[1] Prepared by Patty Glynn, University of Washington, November 1, 2010. Thanks to Sara Vera for testing this.

```
** The following command creates a variable named ppdup that has the
** value of id for the case behind it.
** [_n-1] asks stata to look at the previous case .
** (similar to "lag" in SAS and SPSS) .

gen ppdup = id[_n-1]
(1 missing value generated)

** Show you the data as they are now .
list
     +---------------------------------+
     | year   id   v1   tyear   ppdup |
     |---------------------------------|
  1. | 2008    1    2       0       . |
  2. | 2007    1    1       1       1 |
  3. | 2009    1    3       1       1 |
  4. | 2008    2    5       0       1 |
  5. | 2009    2    6       1       2 |
     |---------------------------------|
  6. | 2007    2    4       1       2 |
  7. | 2007    3    7       1       2 |
  8. | 2009    4    8       1       3 |
     +---------------------------------+

** select cases where the id is not the same as the id in the previous case .
drop if ppdup == id
(4 observations deleted)

** Show you the data as they are now .
list
     +---------------------------------+
     | year   id   v1   tyear   ppdup |
     |---------------------------------|
  1. | 2008    1    2       0       . |
  2. | 2008    2    5       0       1 |
  3. | 2007    3    7       1       2 |
  4. | 2009    4    8       1       3 |
     +---------------------------------+



** All of the key commands not annotated and without "list" commands .

use "h:\000\STATA_doc\test.dta", clear
gen tyear = 1
replace tyear = 0 if year == 2008
sort id tyear
gen ppdup = id[_n-1]
drop if ppdup == id
```