Institute for Policy Research
Northwestern University
2040 Sheridan Road
Evanston, Illinois 60208-4100

NORTHWESTERN
UNIVERSITY

Introduction to Stata

Professor Christine Percheski

Online resources:

Professor German Rodriguez's Princeton website: http://data.princeton.edu/stata/default.html
UCLA's Academic Technology Services website: http://www.ats.ucla.edu/stat/stata/default.htm

Key commands:

clear

log using

capture log close

delimit ;

cd

use

describe

summarize

list

gen

replace

destring

label

tab

table

sort

drop

keep

merge

corr

save

regress

logit

Useful symbols

Arithmetic	Logical	Relational
+ add	! not (also ~)	= = equal
- subtract	| or	!= not equal (also ~ =)
* multiply	& and	< less than
/ divide		<= less than or equal
^ raise to power		> greater than
		>= greater than or equal

Examples of variable recoding

```
*creating dummy variables for race/ethnicity
gen nhwhite=0
replace nhwhite=1 if hispaneth==0 & race==1
label variable nhwhite “non-Hispanic white”

gen nhblack=0
replace nhblack=1 if hispaneth==0 & race==1
label variable nhblack “non-Hispanic black”

gen hispanic=0
replace hispanic=1 if hispaneth>=1
label variable hispanic “Hispanic”


*creating variable for number of persons in the household
*this variable adds together the number or kids and number of adults
gen numper=.
replace numper=nadults + nkids
label variable numper “Number of persons in household”
tab numper, m
```

More sophisticated code: A few examples

```
*creating a variable to indicate if a case has missing values on variable1, variable2 or variable3
egen missing=rowmiss(variable1 variable2 variable3)

*creating a series of dummy variables for each response category for the variable “city”
tab city, gen(city_)

*Setting the survey weights
svyset[pweight=pwt],strata(VARSTR)psu(VARPSU)

*Weighted mean of the outcome variable (poverty)
svy: mean poverty

* Weighted mean of the outcome variable by gender
svy: mean poverty, over(_female)
```

Example Code

```
*Code for IPR Stata training
*Creating dataset for exercise
*June 16, 2012
*Created by Christine
*This file calls in CPS 2011 data (corresponding to year 2010) and creates a dataset called classdemo.dta

clear
# delimit ;
capture log close;
set more off;

log using "Z:\IPRTrainingDemo_June2012", replace;
cd "Z:\Research\dissertation\womens_income\";
use "data\March2012downloads\MAR11.DTA", clear;

*Generating person idnum;
sort _hhid;
gen idnum=_n;

/* Race recode */
gen nrace=1 if _race==1 & _spneth==8;          * white;
replace nrace=2 if _race==2 & _spneth==8;      * black;
replace nrace=3 if _spneth>=3 & _spneth<8;     * Hispanic;
replace nrace=4 if nrace>3;                    * other;

*Working;
* 1= working, 0=not working;
gen working=.;
replace working=1 if _esr==1;
replace working=0 if _esr~=1 & _esr>0;

*Hours worked last week;
gen hrslwk=.;
replace hrslwk=hours if hours>=1;
replace hrslwk=0 if hours<=0;

*FTYR;
*Working ftyr means person worked 35 hours or more per week & 50 weeks or more per year;
gen ftyr=.;
replace ftyr=1 if _wkslyr>=50 & hours>=35;
replace ftyr=0 if _wkslyr<50 | hours<35;

*Earnings;
gen earnings=incwag;
replace earnings=. if aincwg1==1;

/* Education recode (children are missing) */;
gen educ=1 if grdatn>=31 & grdatn<=38;          * <HS;
replace educ=2 if grdatn>=39 & grdatn<43;       * HS or some college;
replace educ=3 if grdatn>=43 & grdatn<=46;      * BA or more;

*Kids in the household;
gen kids=0;
replace kids=1 if _child18>0;

*Respondent is a mother;
*This variable assumes that women living with minor children are mothers;
gen mother=0;
replace mother=1 if kids==1 & _female==1;
```

```
*Keep selected variables to be used in class exercise;
keep idnum wgt _female _relhd age nrace working hrslwk ftyr earnings educ kids mother state;

*Save data excerpt for class exercise;
save "Z:\IPR_training_CPSdata.dta", replace;

*Run descriptive statistics for data excerpt;
summ;
summ age hrslwk earnings, d;
summ earnings if age>18 & age<65;
summ hrslwk if age>18 & age<65;

*Creating a variable to indicate missing earnings data;
gen missing=0;
replace missing=1 if earnings==.;
tab missing, m;

*Dropping kids and the elderly;
drop if age<18 ;
drop if age>65;
tab age;

*Correlation between hours of work and earnings;
corr hrslwk earnings;


***REGRESSIONS;
*Regressing hours of work on earnings;
*Earnings is the outcome and hours of work is the predictor;
regress earnings hrslwk;

*Adding gender variable;
regress earnings hrslwk if _female==1;
regress earnings hrslwk if _female==0;
regress earnings hrslwk _female;

*Close log;
capture log close;

Stop!
```

Example Output

```
. *Code for IPR Stata training
. *Creating dataset for exercise
. *June 16, 2012
. *Created by Christine
. *This file calls in CPS 2011 data (corresponding to year 2010) and creates a dataset c
> alled classdemo.dta
.
. clear

. # delimit ;
delimiter now ;
. capture log close;

. set more off;

. log using "Z:\IPRTrainingDemo_June2012", replace;
-------------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  Z:\IPRTrainingDemo_June2012.smcl
  log type:  smcl
 opened on:  18 Jun 2012, 16:29:18

. cd "Z:\Research\dissertation\womens_income\";
Z:\Research\dissertation\womens_income

. use "data\March2012downloads\MAR11.DTA", clear;

. *Generating person idnum;
. sort _hhid;

. gen idnum=_n;

. /* Race recode */
> gen nrace=1 if _race==1 & _spneth==8;
(77948 missing values generated)

.            * white;
. replace nrace=2 if _race==2 & _spneth==8;
(23050 real changes made)

.        * black;
. replace nrace=3 if _spneth>=3 & _spneth<8;
(36622 real changes made)

.       * Hispanic;
. replace nrace=4 if nrace>3;
(18276 real changes made)

.                      * other;
. *Working;
. * 1= working, 0=not working;
. gen working=.;
(204983 missing values generated)

. replace working=1 if _esr==1;
(88801 real changes made)

. replace working=0 if _esr~=1 & _esr>0;
(67372 real changes made)
```

```
. *Hours worked last week;
. gen hrslwk=.;
(204983 missing values generated)

. replace hrslwk=hours if hours>=1;
(88801 real changes made)

. replace hrslwk=0 if hours<=0;
(116182 real changes made)

. *FTYR;
. *Working ftyr means person worked 35 hours or more per week & 50 weeks or more per yea
> r;
. gen ftyr=.;
(204983 missing values generated)

. replace ftyr=1 if _wkslyr>=50 & hours>=35;
(57352 real changes made)

. replace ftyr=0 if _wkslyr<50 | hours<35;
(147631 real changes made)

. *Earnings;
. gen earnings=incwag;

. replace earnings=. if aincwg1==1;
(1267 real changes made, 1267 to missing)

. /* Education recode (children are missing) */;
. gen educ=1 if grdatn>=31 & grdatn<=38;
(174689 missing values generated)

.            * <HS;
. replace educ=2 if grdatn>=39 & grdatn<43;
(85812 real changes made)

.         * HS or some college;
. replace educ=3 if grdatn>=43 & grdatn<=46;
(40743 real changes made)

.        * BA or more;
. *Kids in the household;
. gen kids=0;

. replace kids=1 if _child18>0;
(113340 real changes made)

. *Respondent is a mother;
. *This variable assumes that women living with minor children are mothers;
. gen mother=0;

. replace mother=1 if kids==1 & _female==1;
(59089 real changes made)

. *Keep selected variables to be used in class exercise;
. keep idnum wgt _female _relhd age nrace working hrslwk ftyr earnings educ kids mother
> state;

. *Save data excerpt for class exercise;
. save "Z:\IPR_training_CPSdata.dta", replace;
file Z:\IPR_training_CPSdata.dta saved

. *Run descriptive statistics for data excerpt;
```

```
. summ;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       state |    204983    55.31995    26.49914         11         95
         age |    204983    35.02181    22.16865          0         85
     _female |    204983    .5148037     .499782          0          1
      _relhd |    204983    2.805013    1.506672          1          7
         wgt |    204983    149334.2    93962.66       7348     904311
-------------+--------------------------------------------------------
       idnum |    204983      102492    59173.64          1     204983
       nrace |    204983    1.737242    1.042109          1          4
     working |    156173    .5686066    .4952724          0          1
      hrslwk |    204983    16.52163    20.85687          0         99
        ftyr |    204983    .2797891    .4488966          0          1
-------------+--------------------------------------------------------
    earnings |    203716    19485.01    41082.76          0    1259999
        educ |    156849    2.066618    .6696756          1          3
        kids |    204983    .5529239    .4971924          0          1
      mother |    204983    .2882629    .4529552          0          1



.  summ age hrslwk earnings, d;

                            age
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            3              0
10%            6              0       Obs              204983
25%           15              0       Sum of Wgt.      204983

50%           34                      Mean           35.02181
                        Largest       Std. Dev.      22.16865
75%           52             85
90%           65             85       Variance       491.4488
95%           74             85       Skewness       .2598875
99%           85             85       Kurtosis       2.093066

                           hrslwk
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs              204983
25%            0              0       Sum of Wgt.      204983

50%            0                      Mean           16.52163
                        Largest       Std. Dev.      20.85687
75%           40             99
90%           43             99       Variance       435.0089
95%           50             99       Skewness       .7623003
99%           67             99       Kurtosis       2.261043
```

```
                      earnings
------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs              203716
25%            0              0       Sum of Wgt.      203716

50%            0                      Mean           19485.01
                      Largest         Std. Dev.      41082.76
75%        29000        1099999
90%        59000        1099999       Variance       1.69e+09
95%        80000        1124999       Skewness        8.75611
99%       150000        1259999       Kurtosis        168.448



.  summ earnings if age>18 & age<65;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    earnings |    120329    31650.98    48290.76          0    1259999

.  summ hrslwk if age>18 & age<65;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      hrslwk |    121520    26.59917    20.99591          0         99



. *Creating a variable to indicate missing earnings data;
. gen missing=0;

. replace missing=1 if earnings==.;
(1267 real changes made)

. tab missing, m;

    missing |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |    203,716       99.38       99.38
          1 |      1,267        0.62      100.00
------------+-----------------------------------
      Total |    204,983      100.00



. *Dropping kids and the elderly;
. drop if age<18 ;
(58198 observations deleted)

. drop if age>65;
(20494 observations deleted)
```

```
. tab age;

        age |      Freq.     Percent        Cum.
------------+-----------------------------------
         18 |      3,319        2.63        2.63
         19 |      2,604        2.06        4.69
         20 |      2,521        2.00        6.69
         21 |      2,521        2.00        8.68
         22 |      2,368        1.88       10.56
         23 |      2,391        1.89       12.45
         24 |      2,321        1.84       14.29
         25 |      2,489        1.97       16.26
         26 |      2,704        2.14       18.40
         27 |      2,485        1.97       20.37
         28 |      2,547        2.02       22.38
         29 |      2,716        2.15       24.54
         30 |      2,884        2.28       26.82
         31 |      2,767        2.19       29.01
         32 |      2,798        2.22       31.23
         33 |      2,645        2.09       33.32
         34 |      2,780        2.20       35.52
         35 |      2,700        2.14       37.66
         36 |      2,744        2.17       39.83
         37 |      2,728        2.16       41.99
         38 |      2,833        2.24       44.24
         39 |      2,923        2.31       46.55
         40 |      3,275        2.59       49.14
         41 |      3,175        2.51       51.66
         42 |      2,976        2.36       54.01
         43 |      2,812        2.23       56.24
         44 |      2,794        2.21       58.45
         45 |      2,844        2.25       60.70
         46 |      3,077        2.44       63.14
         47 |      3,111        2.46       65.60
         48 |      3,023        2.39       68.00
         49 |      3,059        2.42       70.42
         50 |      3,058        2.42       72.84
         51 |      2,875        2.28       75.12
         52 |      2,766        2.19       77.31
         53 |      2,705        2.14       79.45
         54 |      2,634        2.09       81.54
         55 |      2,629        2.08       83.62
         56 |      2,493        1.97       85.59
         57 |      2,385        1.89       87.48
         58 |      2,263        1.79       89.27
         59 |      2,100        1.66       90.93
         60 |      2,156        1.71       92.64
         61 |      2,099        1.66       94.30
         62 |      1,936        1.53       95.84
         63 |      1,939        1.54       97.37
         64 |      1,867        1.48       98.85
         65 |      1,452        1.15      100.00
------------+-----------------------------------
      Total |    126,291      100.00

. *Correlation between hours of work and earnings;
. corr hrslwk earnings;
(obs=125065)

             |   hrslwk earnings
-------------+------------------
      hrslwk |   1.0000
    earnings |   0.4242   1.0000
```

```
. ***REGRESSIONS;
. *Regressing hours of work on earnings;
. *Earnings is the outcome and hours of work is the predictor;
. regress earnings hrslwk;

      Source |       SS       df       MS              Number of obs =  125065
-------------+------------------------------           F(  1,125063) =27450.10
       Model |  5.1802e+13     1  5.1802e+13           Prob > F      =  0.0000
    Residual |  2.3601e+14125063  1.8871e+09           R-squared     =  0.1800
-------------+------------------------------           Adj R-squared =  0.1800
       Total |  2.8781e+14125064  2.3013e+09           Root MSE      =   43441

------------------------------------------------------------------------------
    earnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hrslwk |   965.8774   5.829751   165.68   0.000     954.4512    977.3036
       _cons |   5780.309   194.2598    29.76   0.000     5399.563    6161.055
------------------------------------------------------------------------------

. *Adding gender variable;
. regress earnings hrslwk if _female==1;

      Source |       SS       df       MS              Number of obs =   64894
-------------+------------------------------           F(  1, 64892) =19795.13
       Model |  1.8103e+13     1  1.8103e+13           Prob > F      =  0.0000
    Residual |  5.9344e+13 64892   914506705           R-squared     =  0.2337
-------------+------------------------------           Adj R-squared =  0.2337
       Total |  7.7447e+13 64893  1.1935e+09           Root MSE      =   30241

------------------------------------------------------------------------------
    earnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hrslwk |   836.4914   5.945416   140.70   0.000     824.8384    848.1444
       _cons |   4158.454   178.3733    23.31   0.000     3808.843    4508.066
------------------------------------------------------------------------------

. regress earnings hrslwk if _female==0;

      Source |       SS       df       MS              Number of obs =   60171
-------------+------------------------------           F(  1, 60169) = 9952.93
       Model |  2.8687e+13     1  2.8687e+13           Prob > F      =  0.0000
    Residual |  1.7342e+14 60169  2.8822e+09           R-squared     =  0.1419
-------------+------------------------------           Adj R-squared =  0.1419
       Total |  2.0211e+14 60170  3.3590e+09           Root MSE      =   53687

------------------------------------------------------------------------------
    earnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hrslwk |   1010.818   10.13205    99.76   0.000     990.9589    1030.677
       _cons |   9328.248   370.4961    25.18   0.000     8602.074    10054.42
------------------------------------------------------------------------------
```

```
. regress earnings hrslwk _female;

      Source |       SS       df       MS              Number of obs =  125065
-------------+------------------------------           F(  2,125062) =14652.18
       Model |  5.4637e+13     2  2.7319e+13           Prob > F      =  0.0000
    Residual |  2.3317e+14125062  1.8645e+09           R-squared     =  0.1898
-------------+------------------------------           Adj R-squared =  0.1898
       Total |  2.8781e+14125064  2.3013e+09           Root MSE      =   43180

------------------------------------------------------------------------------
    earnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      hrslwk |   927.2163   5.878849   157.72   0.000     915.6939    938.7388
     _female |  -9668.002   247.9221   -39.00   0.000    -10153.92   -9182.079
       _cons |   11794.88    247.128    47.73   0.000     11310.51    12279.24
------------------------------------------------------------------------------

. *Close log;
. capture log close;

. Stop!

end of do-file
```