The bases of STATA

(the data are at: www.louischauvel.org/nhis)

LOUIS CHAUVEL

Sedulo curavi, humanas actiones non ridere, non lugare, neque detestari, sed intelligere.
I have striven not to laugh at human actions, not to weep at them, nor to hate them, but to understand them.
Baruch Spinoza, 1675/76 *Tractatus Politicus* (A Political Treatise)

"On peut avoir trois principaux objets dans l'étude de la vérité, l'un de la découvrir quant on la cherche; l'autre de la démontrer lorsqu'on la possède; la dernière de la discerner d'avec le faux quand on l'examine."
Blaise Pascal, *De l'esprit géométrique*

"We may have three principal objects in the study of truth : first, to discover it when we are seeking it; secondly, to demonstrate it, when we are in possession of it; and lastly, to distinguish it from falsehood, when we examine it"
Pascal / But the question is also "why?"….

This short document is a basic tutorial for STATA. It presents the bases of data management (opening a data file, selection of variables or of individuals, creating new variables, etc.), of descriptive statistics (frequency tables, basic summary statistics, two way tables, simple linear regressions, etc.). Later on we will introduce advanced statistics (panel regressions, duration models, etc.).

We will use a NHIS (National health interview series, the new name of the former IHIS Integrated health interview series) extract in order to test these methods.

(see http://www.louischauvel.org/nhis)

STATA & others statistical softs

STATA is not the simplest statistical software (SPSS or JMP are more user-friendly), but is less complicate than SAS or R. The main difficulty for beginners in STATA is the relative poverty of error messages in case of mistakes. But the help online are better than for its competitors. STATA is also "case sensitive": "GENDER", "gender" and "Gender" are three different variables. In formulae, sometimes spaces " " must be there and sometimes they must be avoided. Then, the entry cost could be relatively high for impatient people. But for R and Python people, Stata is rather simple.

STATA offers useful help online possibilities (type `help regression`, or `help logit`, also f.ex.), performing tools for programming and for reusing statistical results, powerful graphic utilities, etc. On top of that, a worldwide community of STATA geeks creates new statistical programs or procedures which are archived in Boston MA, and you simply have to type "**`ssc install apcd`**" for installing a new procedure for age period cohort models. Then, nowadays, STATA is a standard in the most important universities in the world. The best of the best softs is certainly "R", which is a freeware, but it is that complicate we can say learning STATA is a prerequisite before you learn "R".

See also **`ssc hot, n(100)`** to see all the fashionable ado programs recently downloaded and **`ssc new`** for the recent production.

How the data are organized?

To get how STATA works, let us consider the data structure. A standard file of stata data looks like an excel spreadsheet: a "flat" "rectangle" file we call it: in this table,

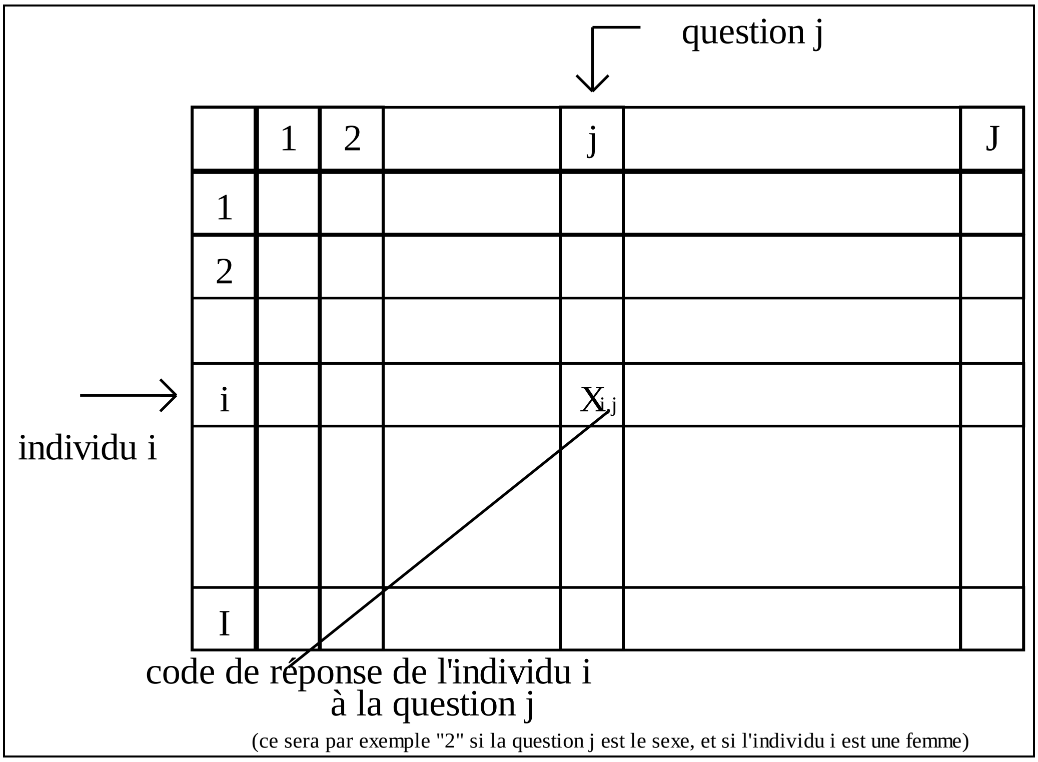
*each line is an individual (="observation" / "object" for SAS),

*each column is a variable.

In general, a value $X_{i,j}$ on line i and column j represents the answer of individual (="observation") i to the question j. The types of variable are diverse: we have "string" variables which contain characters such as letters; most variables are "numeric", and present numbers (with or without decimals); these numbers could be real measures ("metric" or "quantitative" variables, such as age in years, earnings in euros, weight in kg, etc.), ordinal data (1: very light to 10: very heavy), dichotomic (dummy) variables (0: no, 1: yes); or categorical (or polytomic) variables (1: French; 2: UK; 3: Spain; 4: Liechtenstein, etc.). The presentation is similar to that of Excel.

Our datafiles will be in STATA format (.dta), but many different strategies exist for the importation from different formats, including from excel files.

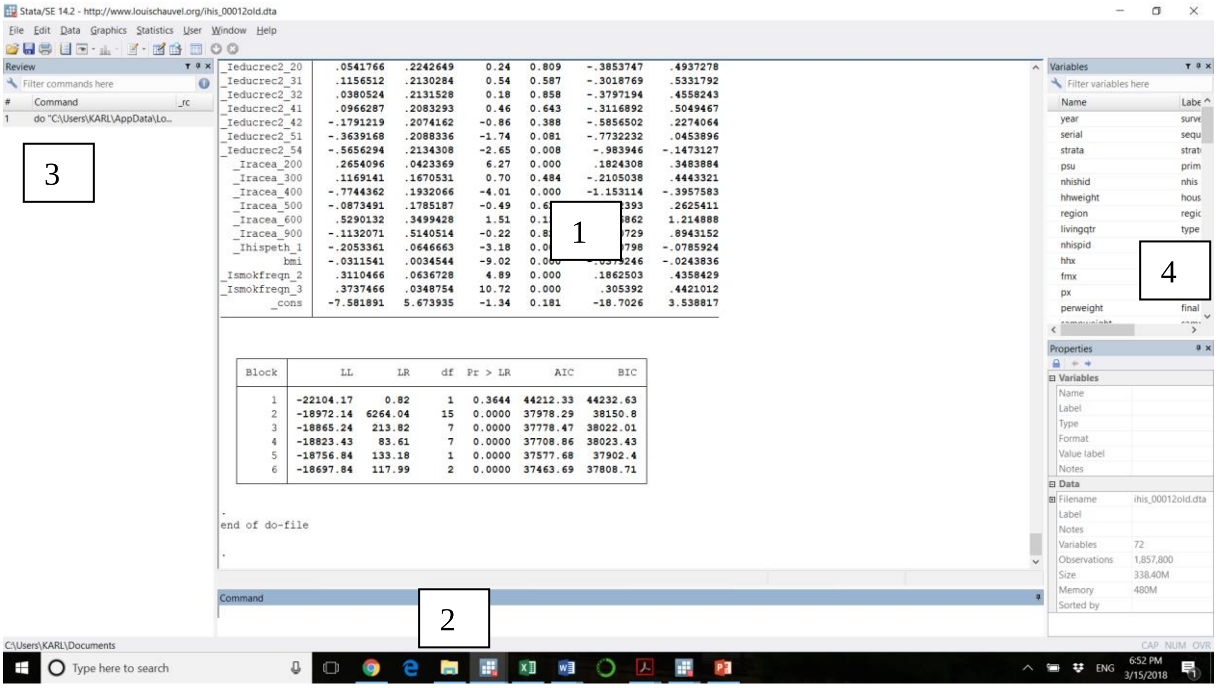
You want to see this dataset (like in SPSS): just type `edit` on the command line.


question j
1
2
j
J
1
2
i
I
individu i
$X_{i,j}$
code de réponse de l'individu i
à la question j
(ce sera par exemple "2" si la question j est le sexe, et si l'individu i est une femme)

Beginning with STATA

Opening STATA

We launch (my) STATA14, we see 4 windows: 1) the listing = results of analyses; 2) command line, where we can directly type our orders in STATA language; 3) the list of executed commands; 4) the list of the variables in the active STATA file.



#	Command	_rc
1	do "C:\Users\KARL\AppData\Lo...	

```
_Ieducrec2_20     .0541766   .2242649     0.24   0.809    -.3853747    .4937278
_Ieducrec2_31     .1156512   .2130284     0.54   0.587    -.3018769    .5331792
_Ieducrec2_32     .0380524   .2131528     0.18   0.858    -.3797194    .4558243
_Ieducrec2_41     .0966287   .2083293     0.46   0.643    -.3116892    .5049467
_Ieducrec2_42    -.1791219   .2074162    -0.86   0.388    -.5856502    .2274064
_Ieducrec2_51    -.3639168   .2088336    -1.74   0.081    -.7732232    .0453896
_Ieducrec2_54    -.5656294   .2134308    -2.65   0.008     -.983946   -.1473127
  _Iracea_200     .2654096   .0423369     6.27   0.000     .1824308    .3483884
  _Iracea_300     .1169141   .1670531     0.70   0.484    -.2105038    .4443321
  _Iracea_400    -.7744362   .1932066    -4.01   0.000    -1.153114   -.3957583
  _Iracea_500    -.0873491   .1785187    -0.49   0.6          393    .2625411
  _Iracea_600     .5290132   .3499428     1.51   0.1          862    1.214888
  _Iracea_900    -.1132071   .5140514    -0.22   0.8          729    .8943152
  _Ihispeth_1    -.2053361   .0646663    -3.18   0.0          798   -.0785924
          bmi    -.0311541   .0034544    -9.02   0.0          246   -.0243836
_Ismokfreqn_2     .3110466   .0636728     4.89   0.000     .1862503    .4358429
_Ismokfreqn_3     .3737466   .0348754    10.72   0.000      .305392    .4421012
        _cons    -7.581891   5.673935    -1.34   0.181    -18.7026    3.538817


  Block         LL         LR       df   Pr > LR        AIC        BIC

      1   -22104.17      0.82        1    0.3644   44212.33   44232.63
      2   -18972.14   6264.04       15    0.0000   37978.29    38150.8
      3   -18865.24    213.82        7    0.0000   37778.47   38022.01
      4   -18823.43     83.61        7    0.0000   37708.86   38023.43
      5   -18756.84    133.18        1    0.0000   37577.68    37902.4
      6   -18697.84    117.99        2    0.0000   37463.69   37808.71

.
end of do-file

.
```

Name	Labe
year	surve
serial	sequ
strata	strat
psu	prim
nhishid	nhis
hhweight	hous
region	regic
livingqtr	type
nhispid	
hhx	
fmx	
px	
perweight	final

Name	
Label	
Type	
Format	
Value label	
Notes	
Filename	ihis_00012old.dta
Label	
Notes	
Variables	72
Observations	1,857,800
Size	338.40M
Memory	480M
Sorted by	

C:\Users\KARL\Documents

HELP!

For help on a command (if you know its name F.ex. **`summarize`**), just type in the window #2 this: **`help summarize`** (in fact, **`he su`** is sufficient because Stata accepts you shorten command names). This command will open a window containing the help on "**summarize**" (this command gives means, standard deviations, etc. and all the simple statistics for quantitative variables).

On the web, you will find many different tutorials and websites to help you on Stata:

- The official Stata homepage (http://stata.com/)
- The fantastic UCLA site for statistics UCLA http://www.ats.ucla.edu/stat/stata/

To open a "dta" dataset

If a STATA data file (.dta) is in a directory :
"**C:\Documents and Settings\Pareto\Mes documents\site\**" on your disk c:\, the command:

```
use "C:\Documents and Settings\Pareto\Mes documents\site\USBMIext.dta ", clear
```

will open the data. Just be careful to your "path" c:\ etc. is the one on my own computer… not necessarily yours. You must find the appropriate path. Do not ask me for help on Mac: I know I don't know. Be aware old versions of stata can not open new stata datasets.

Quite often, the data you find are not in Stata, but in SAS, SPSS, etc. Many possibilities for translations exist. If it is too difficult: just ask the help online = chauvel@louischauvel.org Be aware that "R" has a fantastic library "foreign" where you can open the most important formats we know.

To open a "do" file = Stata syntaxes

You have two different possibilities to perform commands/analyses: the first one is with menus & mouse (generally for beginners and if you search for something new for you), the second with Stata programs (= syntaxes) that you can save (under the extension name .do), transform, adapt, replicate. This second strategy is the most professional.

If you want to open a "do" file, plz:

=> menu "window" / "do file editor" / "New do file" or type on the command line **doedit** "path and name of the file".

In this file, a very simple syntax could be :

```
clear all                /* close all former stata file  */
use "http://www.louischauvel.org/ihis_00012old.dta" , clear /* open internet file */
tabulate age                 /* give me plz the freq table of age */
```

To execute these lines please select them (with the mouse) and click on the appropriate icon (or type **ctrl + d**). The selected lines are the executed and the results appear on the appropriate window of STATA. You can re-execute all the instructions whenever you want.

NB : note that /* … */ means commentary. A star * at the beginning of a line means the line will not be executed (= commentary). The sign « /// » do the same thing, and is also useful when you need to write a long instruction on several lines.

```
tabulate AGE
```

gives the same results than:

```
tabulate ///
age
```

Read the data on the screen

When a data file is opened in STATA, if you want to check them, a solution is to type **`edit`** (or **`ed)`** . You will have an excel-like presentation of our data. You have the possibility to change the values in your file. This is very useful to check the quality/appearance of your data.

Using the codebook

Very often, the data you see on the screen via **`ed`** are “coded”, each value of a variable is a number; for the variable gender, 1 is the value for male or 2 for female, in general but sometimes other conventions are chosen. You generally need a “codebook” which gives the keys to understand your data file. In general, “good” data are well documented. Sometimes, the data goes with an attached codebook, and sometimes, the Stata data files are “formatted”, this means when you execute: **`codebook`** you have the complete list of the variables and values with no ambiguity. Sometimes, you receive “dirty” data, or missing information on several variables. In these more complicate cases, you must invest in cryptography...

Creating new variables

Our data files are raw material you must refine for having better statistical presentations. Generally speaking, you must create new variables with the original ones. This is f.ex. the case for the variable AGE that we do not use with all its details: we recode it as a categorical variable. A usual solution is to design a 5 years age groups variable (from 15 to 19 yo is coded 15, from 20 to 24 yo is coded 20, etc.). To do so, we can execute :

```
clear all
use "http://www.louischauvel.org/ihis_00012old.dta" , clear
tabulate age
generate ag5 = int(age/5)*5
recode ag5 (75/max=75) */ag5 is topcoded at 75, highest value for ag5/*
tabulate ag5
summarize ag5
```

<we will see later that **tabulate** (or **tab**) gives the frequency table of a categorical variable; **summarize** or **su** (see infra) gives the standard univariate statistics of a metric variable (= a quantitative variable which gives a measure); **su** variable **, detailed** give more details … >
Be careful: if the variable TOTO already exists, you can not execute again gen TOTO = expression. The command generate (or gen) is reserved to the creation of a new variable. To transform an existing variable (f.ex. in case of mistake in the expression), you must use “**replace**” and not “**generate**”. Other solution: drop the existing variable and reuse generate.

```
clear all
use "http://www.louischauvel.org/ihis_00012old.dta" , clear
generate ag5 = int(sex/5)*5
recode ag5 (75/max=75)
summarize ag5, detail
drop ag5
generate ag5 = int(age/5)*5
recode ag5 (75/max=75)
tab ag5
summarize ag5, detail
```

other solution with replace, which is easier than the drop, may be :

```
replace ag5 = int(age/5)*5
```

with **`generate`** and with **`replace,`** any kind of usual mathematic function can be used:

abs(var) : absolute value
exp(var) : exponential
ln(var) : “natural” (=Napierian) logarithm
log10(var) : decimal logarithm
sqrt(var) : square root

int(var) : integer
floor(var) : floor function
round(var,n) : the “modulo” operation
Etc.

These functions offer many possibilities of calculations of variables, such as the body mass index if you have weights and heights of populations … here in pounds and inches :

```
keep if weight>0 & weight<300
keep if weight>0 & height<90
gen bmi = weight*0.4536  / (height*0.0254 * height*0.0254 )
```

You have also an infinite diversity of mathematical, statistical, random and text functions…
An important feature is to use logical conditions (see below) to create variables :

```
gen obese = bmi>=30
gen overw = bmi<30 & bmi>=25
```

Recoding variables

Recoding variables is very useful for simplifying tables before publications. Recodes are useful too to create new variables:

```
recode bmi (min/20=1) (20/25=2) (25/30=3) (30/max=4), gen(bmi4)
tab ag5 bmi4, chi row
```

Selection of individuals by logic conditions

In many cases, we are interested in subpopulations, f.ex. when we work on the female labour force, male population and people at age 65 or higher are generally excluded (="dropped"); if one works on the "Greater London GL" in the UK, we will select the sole population in GL, and drop the other ones :

```
keep if age < 65 & sex == 2
```

which is equal to:

```
drop if age >= 65 | sex != 2
```

the usual relational operators and logical connectives are:

==	equal to (be careful: 2 "=")
!= ou ~=	not equal to
>	greater than
<	lower than
>=	greater than or equal to
<=	lower than or equal to
&	and
|	or (inclusive)

Generalities on commands and instructions in STATA

General syntax

The general syntax of a command is:

```
command [varlist] [if logical expression] [weight options] [, options]
```

<<<<< example : **tab ag10 if sex == 1 [iw= sampweight], generate(cx) >>>>>**

Between square brackets [] are optional arguments

`[varlist]` means you can give one variable or more

`[if expression]` the command is executed on a subgroup defined by the expression

`[weight options]` Stata offers many possible weighting options (without weighting, each individual costs for 1, but with weighting, an individual can cost for x, x being the weighting variable) (see later). The final options following a comma "," offer many possibilities that you can check by typing "help command" on the command line.

Example :

```
tab ag5 if sex == 2, generate(agefem)
```

this will give a frequency table of AG5 for females, and the option generate will create new variables AGEFEM1 to AGEFEMn which are n dummy variables (0/1) ; male (excluded of the command) will have a “missing value” “.” in the new variables.

VERY USEFUL IN GENERAL one of the best STATA feature :
After a command, type this:

```
return list
```

you will obtain there the list of temporary variables containing the statistical results. This example gives OLDPPL a dummy variable which is 1 for those who are 1.5 times older than the median, and 0 for the others, and also STDAGE, a so called “standardized” transformation of STDAGE, standardized because its mean is 0 and standard deviation 1:

```
summarize bmi, detail
return list
gen bigger= bmi>= r(mean) +2*r(sd)
tab bigger
```

Main commands for simple statistics

Here are the most important statistical instructions:

* the command **summarize** <variable> gives the main (basic) statistics such as means, standard deviations, etc. for quantitative (metric) variables; the option “`, detail`” proposes more results, percentiles, skewness and kurtosis, etc…. ;

```
summarize age, detail

. su age
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         AGE |     79244    45.66421     17.5638         18         96
```

* **tabulate** <1 variable> gives frequencies, percentages and cumulative percentages of a qualitative variable (« tri à plat », in French); the variable must not contain too many different values = it is not appropriate for incomes expressed in dollars in a large survey. The option **`generate (<newvariablename>)`** creates as many dummy variables (or “dichotomic” variables 0/1) as values in the tabulated variable.

```
tabulate bmi4
```

* **tabulate** < 2 variables> gives a cross table ; the options row and col give the line and column percentages, and the option chi2 give the khi-square test for independence between the variables.

```
tabulate ag5 bmi4, row nofreq chi
```

* **tabulate** < 2 variables>`, summarize(3rdvariable)`, gives the summary statistics (mean, freq and standard deviation) of the third variable (a metric one) in a cross table crossing the 2 first (categorical) variables :

```
tabulate ag5 bmi4 if ag5<70, summarize(bmi) nofreq nost
```

* **table** <list of max 4 var>, contents(<list of statistics you need and name of a variable >) this gives the statistics for each subgroup of observation which are the crossing of the subgroups of the pop

```
table ag5 sex if ag5<70, c(mean bmi median age)
```

* **tabstat** <quantitvar >, `by (`<categ variables >`)` statistics(<listofstats >) This give the statistics by groups defined by <categ variables > on the quantitative variable. A little bit complicate, plz consult => help tabstat ;

```
tabstat bmi ag5, by (sex) stat(mean)
```

* **pwcorr** <liste de variables> give a correlation matrix of the list of quantitative variables. The option covariance gives the covariances in place of correlations coefficients .

```
pwcorr height weight bmi age
```

bysort : repeating an instructions on diverse subgroups

The expression bysort <listofvariable> repeats an instruction separately for each subgroup of the list of variables :

```
bysort sex : pwcorr height weight bmi age
```

Example of comparing means with confidence intervals `ci`:

```
bysort sex: ci bmi
```

Usual statistical procedures

Tabulate with khi-2 test

For khi-square tests of cross tables, the now usual syntax:

```
tabulate educ bmi4, row nofreq chi
```

One factor anova (ie: Fischer test)

For a one factor anova (metric variable explained by a categorical variable) :

```
table ag5, c(mean bmi)
anova bmi ag5
table educ, c(mean bmi)
anova bmi educ
```

Two / three /multiple factors anova

For a two factors anova (metric variable explained by two categorical variables) :

```
anova bmi educ ag5
```

Then , simply execute this :

```
anova bmi educ ag5 racea
```

Simple and multiple linear regressions

The simple & multiple linear regression of books read are:

```
regress bmi age
su age
gen sdage = (age-r(mean))/r(sd)
gen sdagesq = age * age
su sdagesq
replace sdagesq = (sdagesq -r(mean))/r(sd)
regress bmi educ sdage sdagesq
```

Histograms with kernel & normal curves

```
histogram bmi, percent
histogram bmi, percent normal kdensity
```

XY scatter plots

```
preserve
collapse (mean) bmi ag10, by(educ)
twoway scatter  bmi ag10, mlabel(educ)
twoway  lfit bmi ag10  ||  scatter   bmi ag10, mlabel(educ)
restore
```

Computation of inequality measures

STATA proposes new statistical features for the computation of more exotic statistics such as inequality measures (Gini indexes and the like). The instruction **ineqdeco** by Staphen Jenkins is one of these. First todownload the program, execute this:

```
ssc install ineqdeco /* to be done once */
ineqdeco bmi
```

Multiple regressions with (one or more) categorical explanatory variables

For multiple regressions with categorical explanatory variables (and also metric ones), the instruction **`xi:`** (from STATA 11) you just have to type i. before each categorical explanatory variable

```
regress bmi i.ag10 i.racea i.educ i.sex
      Source |       SS           df       MS      Number of obs   = 1,693,184
-------------+----------------------------------   F(19, 1693164)  =   6364.55
       Model |  2648311.66        19  139384.824   Prob > F        =    0.0000
    Residual |  37080630.9 1,693,164  21.9002004   R-squared       =    0.0667
-------------+----------------------------------   Adj R-squared   =    0.0666
       Total |  39728942.5 1,693,183  23.4640571   Root MSE        =    4.6798

--------------------------------------------------------------------------------------------------------------
                                        bmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------------------------------+-----------------------------------------------------------------
                                       ag10 |
                                        30  |   1.373062   .0107133   128.16   0.000     1.352064    1.394059
                                        40  |    2.12782   .0113424   187.60   0.000     2.105589    2.150051
                                        50  |   2.426207   .0119794   202.53   0.000     2.402728    2.449686
                                        60  |   2.226793   .0127234   175.02   0.000     2.201856     2.25173
                                        70  |   1.555382   .0146891   105.89   0.000     1.526592    1.584172
                                            |
                                      racea |
                   black/african-american  |   1.616707    .010803   149.65   0.000     1.595534    1.637881
  aleut, alaskan native, or american indian |   1.468283   .0419335    35.01   0.000     1.386094    1.550471
                 asian or pacific islander  |  -1.556216   .0239319   -65.03   0.000    -1.603122    -1.50931
                                other race  |   .7621746   .0296166    25.73   0.000     .7041271    .8202222
   multiple race, no primary race selected  |   .7479034   .0929698     8.04   0.000     .5656858     .930121
                                   unknown  |  -.0096157   .0639595    -0.15   0.880    -.1349741    .1157427
                                            |
                                   educrec2 |
                      grade 1, 2, 3, or 4  |  -.0004062   .0566038    -0.01   0.994    -.1113477    .1105353
                         grade 5, 6, or 7  |  -.0383717   .0523604    -0.73   0.464    -.1409963    .0642528
                                   grade 8  |  -.3665374   .0521659    -7.03   0.000    -.4687808    -.264294
                       grade 9, 10, or 11  |  -.3561837   .0504094    -7.07   0.000    -.4549843   -.2573831
                                  grade 12  |  -.7401282   .0497477   -14.88   0.000    -.8376321   -.6426244
                  1 to 3 years of college  |  -.6798085   .0499872   -13.60   0.000    -.7777816   -.5818353
       4 years college/bachelor's degree  |  -1.295447   .0504026   -25.70   0.000    -1.394234   -1.196659
                                            |
                                        sex |
                                    female  |  -1.000306   .0072393  -138.18   0.000    -1.014495   -.9861172
                                      _cons |   25.12547   .0502099   500.41   0.000     25.02706    25.22388
--------------------------------------------------------------------------------------------------------------
```

Comparing nested models

Better than the comparison of R², the comparison of BICs is fine: the best model is that with lower BIC, and a difference of 4 is significant.

```
preserve
keep if missing(bmi +year  +ag5  +sex  +marstat   +educ  +race +hisp)==0
reg     bmi year i.ag5 i.sex i.marstat  i.educ i.race i.hisp i.smokf
xi: nestreg, lr: reg     bmi year (i.ag5 i.sex i.marstat) (i.educ) (i.race
i.hisp)
restore
```

This syntax create the successive models with the different blocks and then at the end:

```
+-------------------------------------------------------------+
| Block |        LL       LR     df  Pr > LR       AIC       BIC |
|-------+-----------------------------------------------------|
|     1 |  -5024263  78929.01     1   0.0000  1.00e+07  1.00e+07 |
|     2 |  -4984595  79335.44    15   0.0000   9969225   9969434 |
|     3 |  -4972619  23952.93     7   0.0000   9945286   9945582 |
|     4 |  -4958355  28527.76     7   0.0000   9916772   9917154 |
+-------------------------------------------------------------+
```

Choosing appropriate weightings for data

Many samples result from the standard model of “Simple Random Sampling” SRS which is a process of uniform probabilistic sampling of the population. In the sampled “universe” U of size N individuals (a country, f.ex.), each individual has a probability 1/p to be selected into the sample S of size n; p = N/n is the sampling rate. In SRS, the sampling rate is a constant over the universe. In the case of SRS, you can neglect the problem of weightings whatever the statistical procedure you use (regressions, etc…). The coefficients standard errors in models, confidence intervals foe statistics like means, etc. are computed accurately.

When the sampling rate is not a constant, you must tackle the problem of sampling, which could be quite complicate on STATA. The easiest way to do so is to find in the dataset (or compute it) a sampling rate weight, which is called also a “probability weight”.

In the IHIS example, you will find **`perweight`** which is this probability weight. By doing :

```
su perweight, d
```

you will see that the average perweight is 4326, which means that each American resident had a probability 1/4326 to have been selected. The interdecile ratio D9/D1 is 3.9, this means that we are far from a SRS. This means that the 10% more often selected subgroups where two times more selected in the sample than required; conversely, the 10% less selected would have to be two times more represented for having unbiased sample. You should use the probability weight **`perweight`** so that the bias of selection are corrected and the real size of the sample is preserved.

Compare the results of

```
regress bmi educ i.ag10 i.racea i.educ i.sex
regress bmi educ i.ag10 i.racea i.educ i.sex [pw=perweight]
```

The main problem is that the probability weights are appropriate and can be activated for all types of models; but we can not make use of probability weights with tables, ci, su, etc. And STATA is not satisfying on this aspect. Anyway, the pragmatic solution when you want to have descriptive statistics with tests of confidence intervals, is to make use of analytic weight aw with a “adjusted” or “standardized” weight, a transformation where the average weight is one. **`[aw=rsweight ]`** is the solution, but analytic weights are inactive for some statistics, such as chi-squares in tables.

```
ci mean bmi
```

More on confidence intervals

Just a basic reminder.

When we consider statistics computed from surveys, you should always ask this question: which are the limits of significance of my data? Is a computed mean of 50 mean 49 to 51 or 41 to 59? This issue is directly connected to the problem of sampling and then to the weighting strategies you choose for your data. In case of “Simple Random Sampling” SRS, the usual chi-square texts, confidence intervals, and models offer confidence intervals or significance diagnosis. In the other cases, you must find the appropriate weighting strategy.

Anyway, it is usual that you simply have a percentage f of people having property A and the size N of a SRS sample. This so-called “Gauss’ table” gives the limits of the confidence intervals at 95% for frequencies. Example : if candidate LJ in a N=1000 SRS receives 14% of voting intentions, candidate JMLP receives 16,5%, and candidate JC is at 18%, LJ can expect he will actually be the third.

frequeency f	Sample size N																				
(%)	1 000	1 200	1 400	1 600	1 800	2 000	2 500	3 000	3 500	4 000	4 500	5 000	6 000	7 000	8 000	9 000	10 000	20 000	30 000	50 000	100 000
2 or 98	0,89	0,81	0,75	0,70	0,66	0,63	0,56	0,51	0,47	0,44	0,42	0,40	0,36	0,33	0,31	0,30	0,28	0,20	0,16	0,13	0,09
3 or 97	1,1	0,98	0,91	0,85	0,80	0,76	0,68	0,62	0,58	0,54	0,51	0,48	0,44	0,41	0,38	0,36	0,34	0,24	0,20	0,15	0,11
4 or 96	1,2	1,1	1,0	0,98	0,92	0,88	0,78	0,72	0,66	0,62	0,58	0,55	0,51	0,47	0,44	0,41	0,39	0,28	0,23	0,18	0,12
5 or 95	1,4	1,3	1,2	1,1	1,0	0,97	0,87	0,80	0,74	0,69	0,65	0,62	0,56	0,52	0,49	0,46	0,44	0,31	0,25	0,19	0,14
6 or 94	1,5	1,4	1,3	1,2	1,1	1,1	0,95	0,87	0,80	0,75	0,71	0,67	0,61	0,57	0,53	0,50	0,47	0,34	0,27	0,21	0,15
8 or 92	1,7	1,6	1,5	1,4	1,3	1,2	1,1	0,99	0,92	0,86	0,81	0,77	0,70	0,65	0,61	0,57	0,54	0,38	0,31	0,24	0,17
10 or 90	1,9	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,95	0,89	0,85	0,77	0,72	0,67	0,63	0,60	0,42	0,35	0,27	0,19
12 or 88	2,1	1,9	1,7	1,6	1,5	1,5	1,3	1,2	1,1	1,0	0,97	0,92	0,84	0,78	0,73	0,69	0,65	0,46	0,38	0,29	0,21
14 or 86	2,2	2,0	1,9	1,7	1,6	1,6	1,4	1,3	1,2	1,1	1,0	0,98	0,90	0,83	0,78	0,73	0,69	0,49	0,40	0,31	0,22
16 or 84	2,3	2,1	2,0	1,8	1,7	1,6	1,5	1,3	1,2	1,2	1,1	1,0	0,95	0,88	0,82	0,77	0,73	0,52	0,42	0,33	0,23
18 or 82	2,4	2,2	2,1	1,9	1,8	1,7	1,5	1,4	1,3	1,2	1,1	1,1	0,99	0,92	0,86	0,81	0,77	0,54	0,44	0,34	0,24
20 or 80	2,5	2,3	2,1	2,0	1,9	1,8	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,96	0,89	0,84	0,80	0,57	0,46	0,36	0,25
22 or 78	2,6	2,4	2,2	2,1	2,0	1,9	1,7	1,5	1,4	1,3	1,2	1,2	1,1	0,99	0,93	0,87	0,83	0,59	0,48	0,37	0,26
24 or 76	2,7	2,5	2,3	2,1	2,0	1,9	1,7	1,6	1,4	1,4	1,3	1,2	1,1	1,0	0,95	0,90	0,85	0,60	0,49	0,38	0,27
26 or 74	2,8	2,5	2,3	2,2	2,1	2,0	1,8	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,98	0,92	0,88	0,62	0,51	0,39	0,28
28 or 72	2,8	2,6	2,4	2,2	2,1	2,0	1,8	1,6	1,5	1,4	1,3	1,3	1,2	1,1	1,0	0,95	0,90	0,63	0,52	0,40	0,28
30 or 70	2,9	2,6	2,4	2,3	2,2	2,0	1,8	1,7	1,5	1,4	1,4	1,3	1,2	1,1	1,0	0,97	0,92	0,65	0,53	0,41	0,29
33 or 67	3,0	2,7	2,5	2,4	2,2	2,1	1,9	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,1	0,99	0,94	0,66	0,54	0,42	0,30
36 or 64	3,0	2,8	2,6	2,4	2,3	2,1	1,9	1,8	1,6	1,5	1,4	1,4	1,2	1,1	1,1	1,0	0,96	0,68	0,55	0,43	0,30
40 or 60	3,1	2,8	2,6	2,4	2,3	2,2	2,0	1,8	1,7	1,5	1,5	1,4	1,3	1,2	1,1	1,0	0,98	0,69	0,57	0,44	0,31
43 or 57	3,1	2,9	2,6	2,5	2,3	2,2	2,0	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,0	0,99	0,70	0,57	0,44	0,31
46 or 54	3,2	2,9	2,7	2,5	2,3	2,2	2,0	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,1	1,0	0,70	0,58	0,45	0,32
50	3,2	2,9	2,7	2,5	2,4	2,2	2,0	1,8	1,7	1,6	1,5	1,4	1,3	1,2	1,1	1,1	1,0	0,71	0,58	0,45	0,32

find a more complete Gauss’ confidence interval at 95% there :
www.louischauvel.org/tabledegauss.doc

This table results from the general formula of p’s confidence interval :

p the proportion in the real universe of size N

f the measured frequency in the SRS sample of size n

When N>>n we have the 95% interval confidence of p: $p = f \pm 2\sqrt{\frac{f(1-f)}{n}}$

If n is close to N (N<10n), the general formula is ➔

$$\boxed{p = f \pm 2\sqrt{\frac{N-n}{N-1}\frac{f(1-f)}{n}}}$$

See the simplification when N>>n ; then the confidence interval is not dependent of N but simply of n.

For the mean, the standard formula for the ci at 95% of the mean M over the universe of size N, when we estimate the mean m and the standard deviation stdev of a variable on the SRS sample of size n, we have

$$\boxed{M = m \pm \frac{2\ \mathrm{stddev}}{\sqrt{n}}}$$

Computing and reading chi squares

Is the link between education and obesity significant? Is it the result of random sampling or reality in the universe :
“Null hypothesis” H0 = in the universe, the proportion of obese people is the same whatever the level of education.

	col j	Total l
lines i	$n_{i,j}$	$n_{i,.}$
total column	$n_{.,j}$	n.,. (or n)

$$\chi^2 = \sum_{i,j} \frac{\left(n_{i,j} - \frac{n_{i,.}n_{.,j}}{n}\right)^2}{\frac{n_{i,.}n_{.,j}}{n}}$$

$\frac{n_{i,.}n_{.,j}}{n}$ => **expected frequency in cell (i,j) under the null hypothesis**

```
tab educ  obese, chi row
           |        obese
  educrec2 |         0          1 |     Total
-----------+----------------------+----------
        12 |     1,928        853 |     2,781
           |     69.33      30.67 |    100.00
-----------+----------------------+----------
        13 |     3,819      1,620 |     5,439
           |     70.22      29.78 |    100.00
-----------+----------------------+----------
        14 |     4,151      1,514 |     5,665
           |     73.27      26.73 |    100.00
-----------+----------------------+----------
        15 |     2,686        673 |     3,359
           |     79.96      20.04 |    100.00
-----------+----------------------+----------
        16 |     1,370        299 |     1,669
           |     82.09      17.91 |    100.00
-----------+----------------------+----------
     Total |    13,954      4,959 |    18,913
           |     73.78      26.22 |    100.00

          Pearson chi2(4) = 190.8920   Pr = 0.000
```

Khi-2 = 190.9
degrees of freedom (df) = 4

degrees of freedom (df) = (number of lines-1)x(number of columns-1) <here (2-1).(2-1) = 1>
And now the diagnosis « table 1 different or not to the independence »
We have 0.000 probability of mistake (= 0.000% of risk) if we reject the null hypothesis H0
=> “table 1 is significantly different to independence” at the significance level of 0.05 (and even at 0.000)

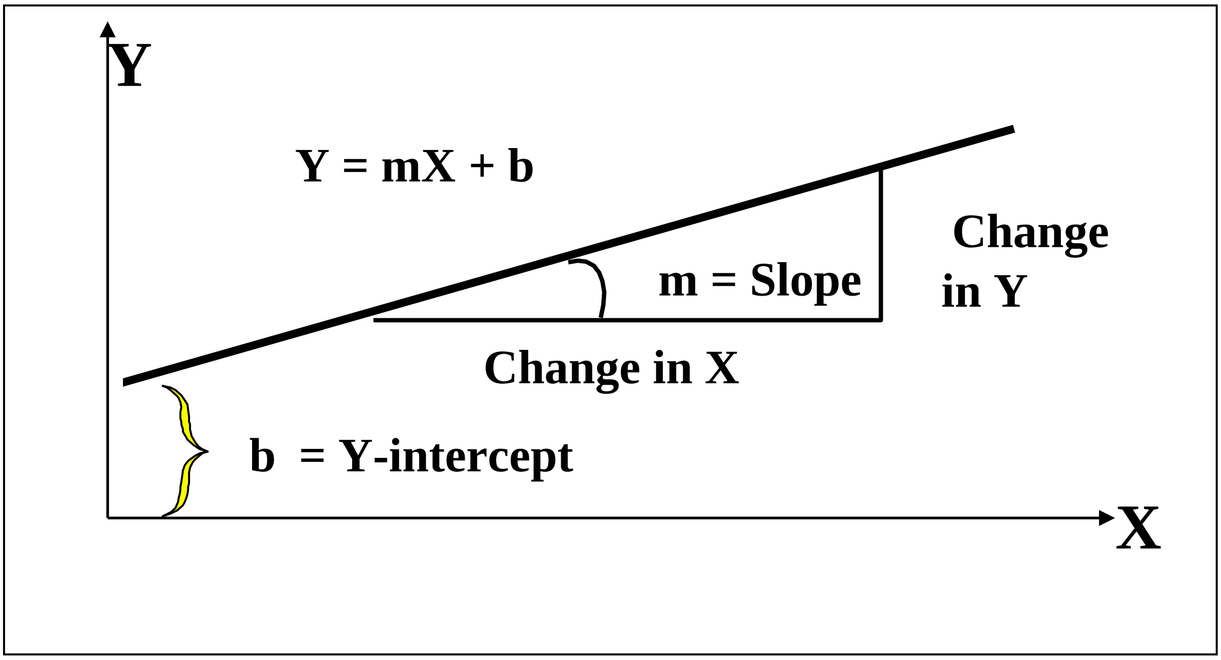
“chi-square probability” pr
of the test (“the p value”) =>

0.99 to 0.10	=> ☹
0.10 to 0.05	=> ?
0.05 to 0.1	=> ☺
0.01 to 0.001	=> ☺ ☺
0.000	=> ☺ ☺ ☺

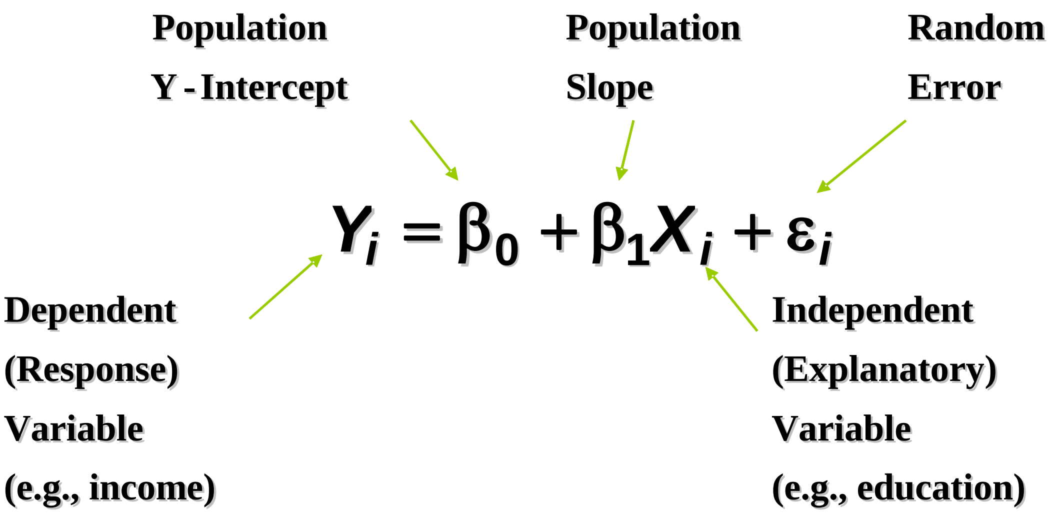
Assumptions on chi-square test
1. Random sample 2. Sample size (see 3.) 3. Minimum expected cell count (at least 5 indiv. per cell) 4. Independence (not a panel)

More on linear regressions (“OLS” = ordinary least square)

Still working



- Relationship Between Variables Is a Linear Function



Population Y - Intercept

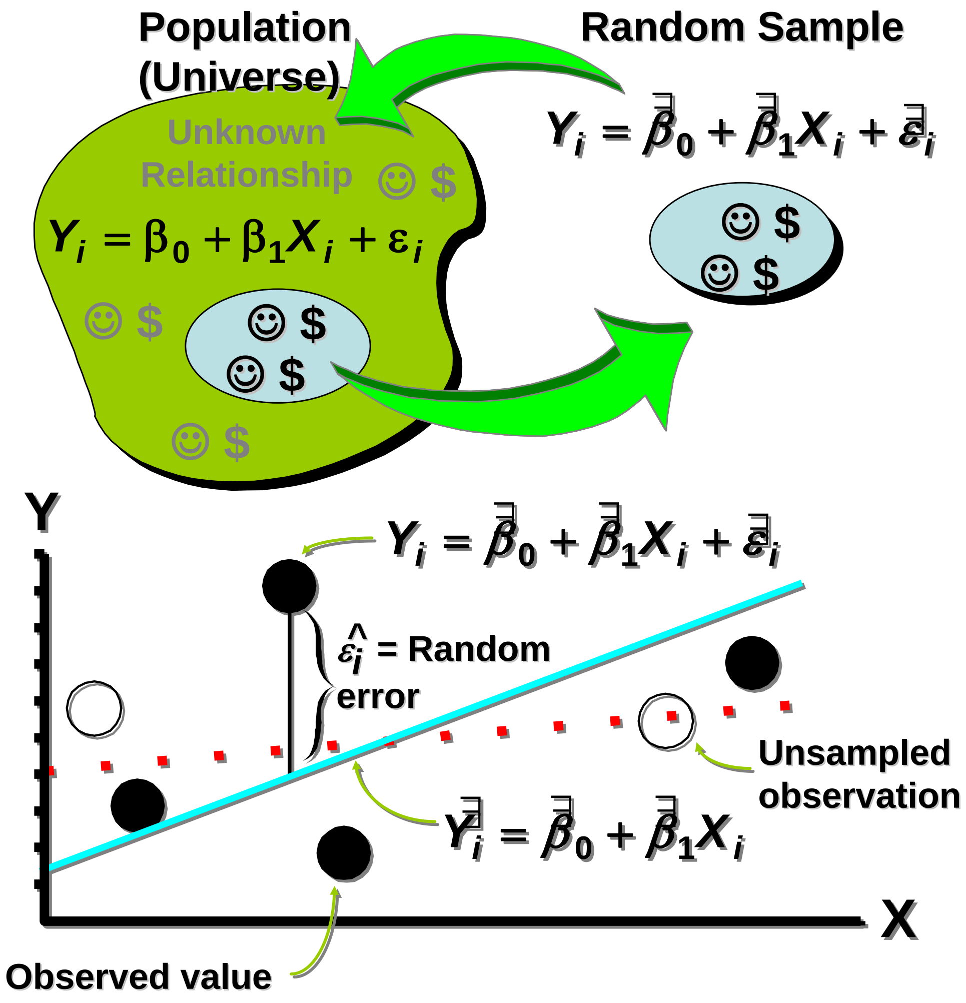
Population Slope

Random Error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent (Response) Variable (e.g., income)

Independent (Explanatory) Variable (e.g., education)


Population
(Universe)
Random Sample
Unknown
Relationship
$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i$
☺ $
Y
X
$\hat{\varepsilon}_i$ = Random
error
Unsampled
observation
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
Observed value

Prediction Equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Sample Slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Sample Y-intercept

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Proportion of Variation 'Explained ' by Relationship Between *X* & *Y*

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$

$$= \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 - \sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$

R^2= SS(Model)/SS(Total) => proportion of information of bmi explained by the model
Adjusted R^2 takes into account the number of explanatory variables (the more you have, the better the R^2)
MS=Mean square **MS(Model)/MS(Residual)** = Fisher's F, test for independence between the explained variable and the explanatory variable (in general, P value conclude to very high significance)

```
reg bmi i.ag5 sex i.educrec2

      Source |       SS       df       MS              Number of obs =   18913
-------------+------------------------------           F( 18, 18894) =   53.37
       Model |  24859.9757    18  1381.10976           Prob > F      =  0.0000
    Residual |  488924.131 18894  25.8772166           R-squared     =  0.0484
-------------+------------------------------           Adj R-squared =  0.0475
       Total |  513784.107 18912  27.1670953           Root MSE      =   5.087

------------------------------------------------------------------------------
         bmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ag5 |
         20  |   1.084086   .2752797     3.94   0.000      .544513    1.623659
         25  |   2.451234   .2701523     9.07   0.000     1.921712    2.980757
         30  |    3.19738   .2704688    11.82   0.000     2.667237    3.727523
         35  |   3.438657   .2700612    12.73   0.000     2.909313    3.968001
         40  |   3.356412   .2700453    12.43   0.000     2.827099    3.885725
         45  |   3.480183     .26918    12.93   0.000     2.952566      4.0078
         50  |   3.757646   .2703527    13.90   0.000     3.227731    4.287562
         55  |   4.057926   .2721516    14.91   0.000     3.524485    4.591367
         60  |   3.964893   .2784076    14.24   0.000     3.419189    4.510597
         65  |   3.979767   .2832091    14.05   0.000     3.424652    4.534883
         70  |   3.238529   .2948598    10.98   0.000     2.660578    3.816481
         75  |   2.535591   .3021028     8.39   0.000     1.943442    3.127739
         80  |   1.792443   .3133284     5.72   0.000     1.178291    2.406595
             |
         sex |  -.4182354   .0745287    -5.61   0.000    -.5643183   -.2721525
             |
    educrec2 |
         13  |  -.2966559    .118965    -2.49   0.013    -.5298379   -.0634738
         14  |  -.5450947     .11909    -4.58   0.000    -.7785218   -.3116677
         15  |  -1.803151   .1322925   -13.63   0.000    -2.062456   -1.543846
         16  |  -2.127477   .1588144   -13.40   0.000    -2.438768   -1.816187
             |
       _cons |   25.54755   .2789431    91.59   0.000      25.0008     26.0943
------------------------------------------------------------------------------
```

Assumptions for linear regressions:
* No excessive multicollinearity = test for the VIF All VIFs (variance inflation factor) must be below 10 http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm but vif less convincing for qualitative explanatory variables)
* Identifiability : less explanatory variable than observations, f.ex.
* No outliers
* Independence of errors
* variance is a constant (homoscedasticity, antonym = heteroscedasticity).

Saving predicted values and residuals

```
reg bmi i.ag5 sex i.educrec2
predict predbmi, xb
predict resibmi, residuals
```

this creates a variable predbmi with the predicted values, and resibmi with the residuals.

xb is optional

Some more on interactions

Testing interactions between var1 and var2 means that you assume the levels of var1 have an influence on the parameters of var2. Ex : education means different things for male and female population.

```
reg bmi i.ag5 sex##i.educrec2


      Source |       SS       df       MS              Number of obs =   18913
-------------+------------------------------           F( 22, 18890) =   46.98
       Model |  26654.9977    22  1211.59081           Prob > F      =  0.0000
    Residual |  487129.109 18890  25.7876712           R-squared     =  0.0519
-------------+------------------------------           Adj R-squared =  0.0508
       Total |  513784.107 18912  27.1670953           Root MSE      =  5.0782

------------------------------------------------------------------------------
         bmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         ag5 |
         20  |   1.097542   .2748141     3.99   0.000     .5588819    1.636203
(.................................................................)
         80  |   1.741442   .3128877     5.57   0.000     1.128154     2.35473
             |
       2.sex |   .5845152    .193839     3.02   0.003     .2045734    .9644571
             |
    educrec2 |
         13  |   .0913875   .1774383     0.52   0.607    -.2564074    .4391824
         14  |  -.0091458   .1775667    -0.05   0.959    -.3571924    .3389008
         15  |  -.7788379   .1942949    -4.01   0.000    -1.159673   -.3980025
         16  |  -1.118507   .2297144    -4.87   0.000    -1.568768   -.6682461
             |
sex#educrec2 |
       2 13  |  -.7058058   .2381583    -2.96   0.003    -1.172617   -.2389942
       2 14  |  -.9806234   .2368437    -4.14   0.000    -1.444858   -.5163886
       2 15  |  -1.890414   .2617264    -7.22   0.000    -2.403421   -1.377406
       2 16  |  -1.895247   .3154798    -6.01   0.000    -2.513615   -1.276878
             |
       _cons |   24.58001   .2747354    89.47   0.000     24.04151    25.11852
-----------------------------------------------------------------
```

When you compare with previous fit, R² is better. Difference in sex is inversed (female with no education have higher bmi than males of same level) educational effect (for male) is of lower intensity, but the interaction sex#educ show much stronger educational effects for females (this term would be non significant if men and women had the same educational effect.

ON STATA 10 : `sex*educ`

ON STATA 11 and after : `sex##educ` < sex#educ denotes the interaction without principal terms which are necessary in general>

Some more on logistic regressions

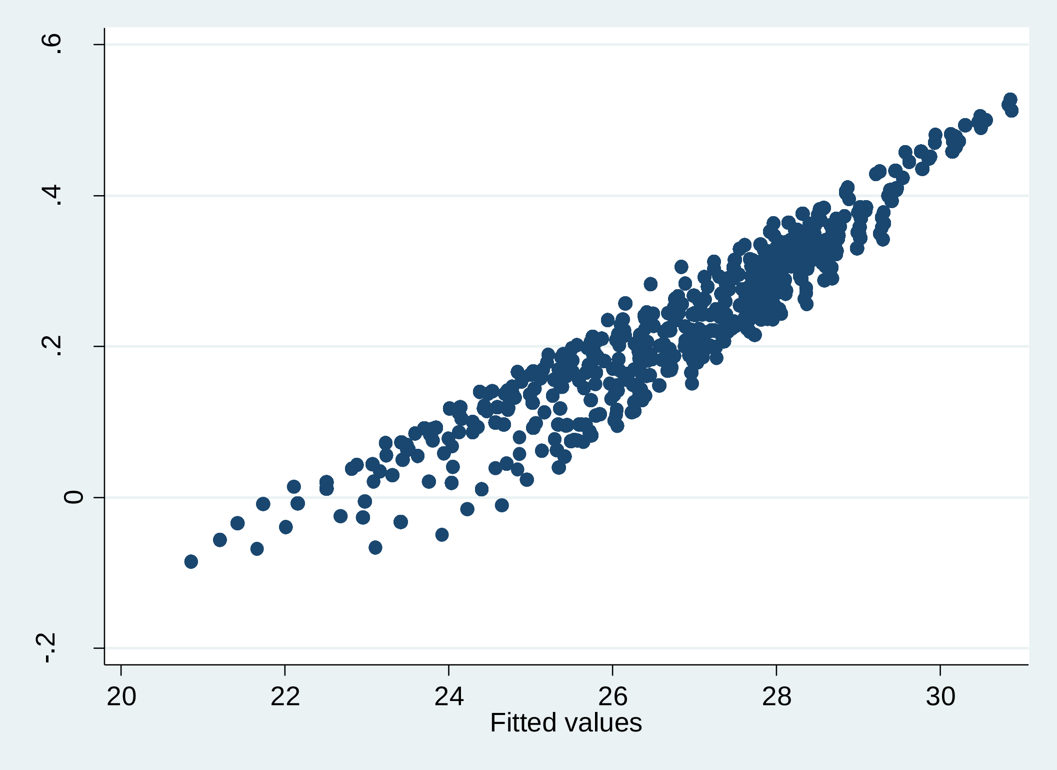
Regressions can not predict correctly qualitative explained variables (or binary 0/1 variables)

Example :

```
reg bmi i.ag5 sex##i.educrec2 sex##i.racea
predict predbmi
reg obese i.ag5 sex##i.educrec2 sex##i.racea
predict predob
twoway scatter predob predbmi, sort
```

You will observe that predicited values of obesity can be negative : impossible for a probability => the logistic regression copes with this issue.

Predicted values of bmi (x) and predicted values of obesity (y)



An example on obesity

- ➢ Predicting probabilities with linear regressions ("Ordinary Least Squares " OLS)

  ➔p(y="Obese") = a + b education + etc.

  ➔Missing the target => negative proba & proba>1

- ➢ The strategy of logistic regressions : predicting the "logit" = log(odds)

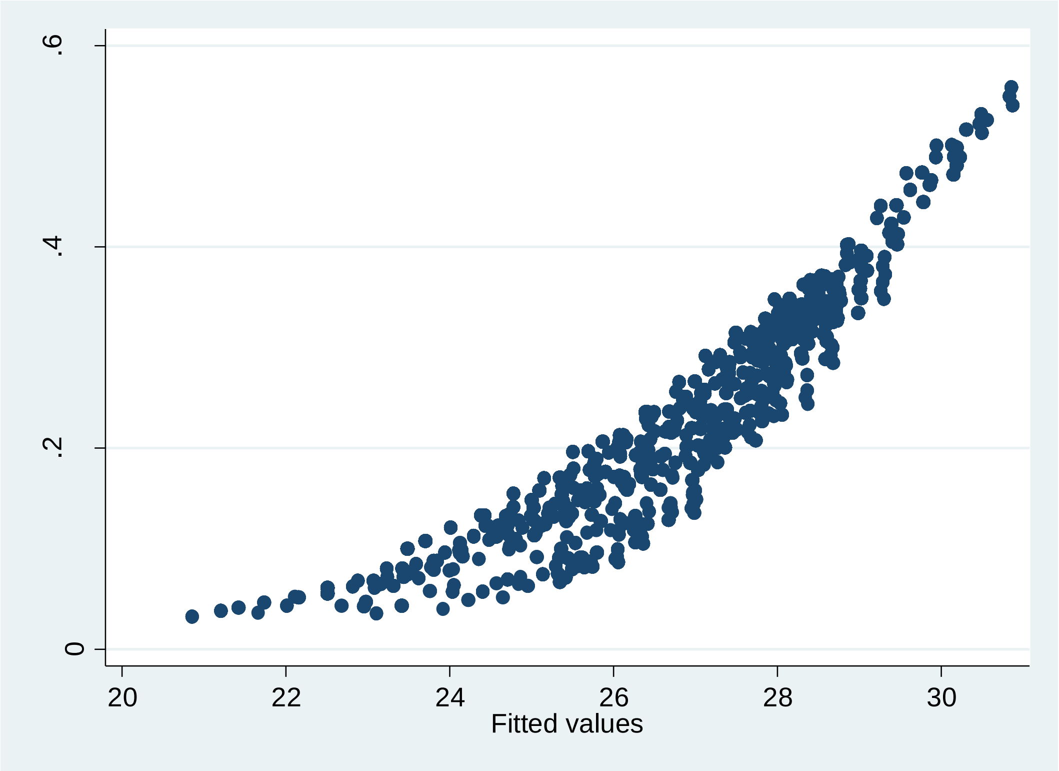
  ➔Remain inside the [0 ; 1] bracket Logit(p) = ln (p / (1-p))

$p(y = \text{"Obese"}) = \text{Logit}^{-1}(a + b \text{ education} + \text{etc})$

```
reg bmi i.ag5 sex##i.educrec2 sex##i.racea
predict predbmi
reg obese i.ag5 sex##i.educrec2 sex##i.racea
predict predob
twoway scatter predob predbmi, sort

logit obese i.ag5 sex##i.educrec2 sex##i.racea
predict predob2
twoway scatter predob2 predbmi, sort
```

Predicted values of bmi (x) and predicted2 values of obesity (y) by logit regression

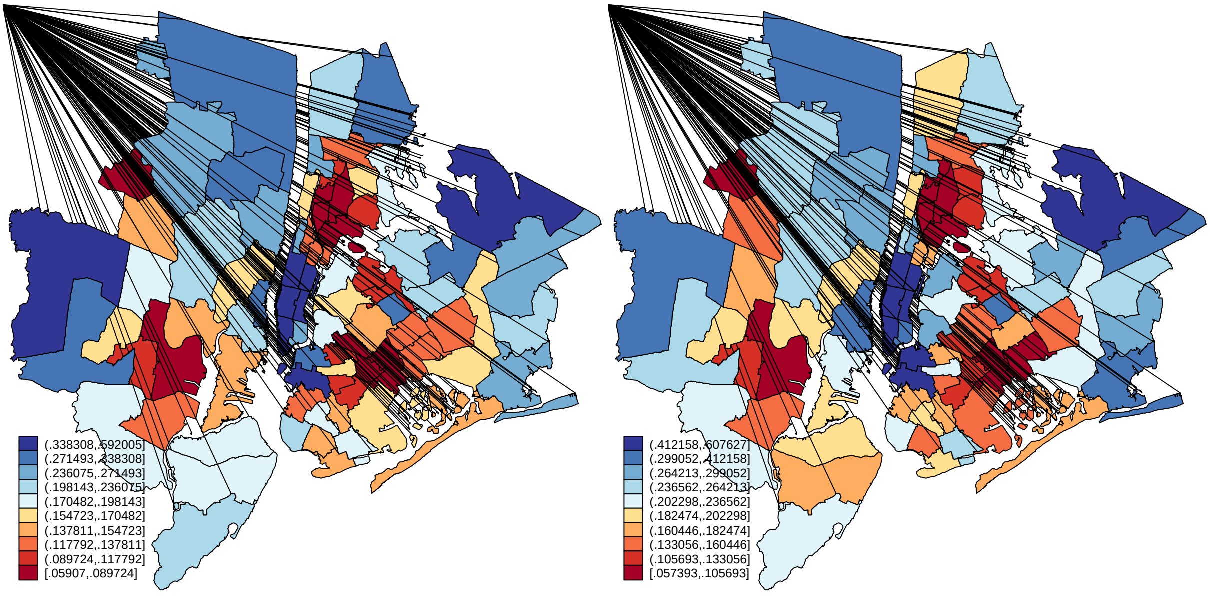


Some more on variable transformations

Working
Ladder gladder and qladder
Why income/wealth is to be used with logs?

Do you like cartography?

Working
http://www.louischauvel.org/pumanynj2005.htm



Proportions of experts managers and professionals in 2000 and 2008

Distributions and Inequality measures

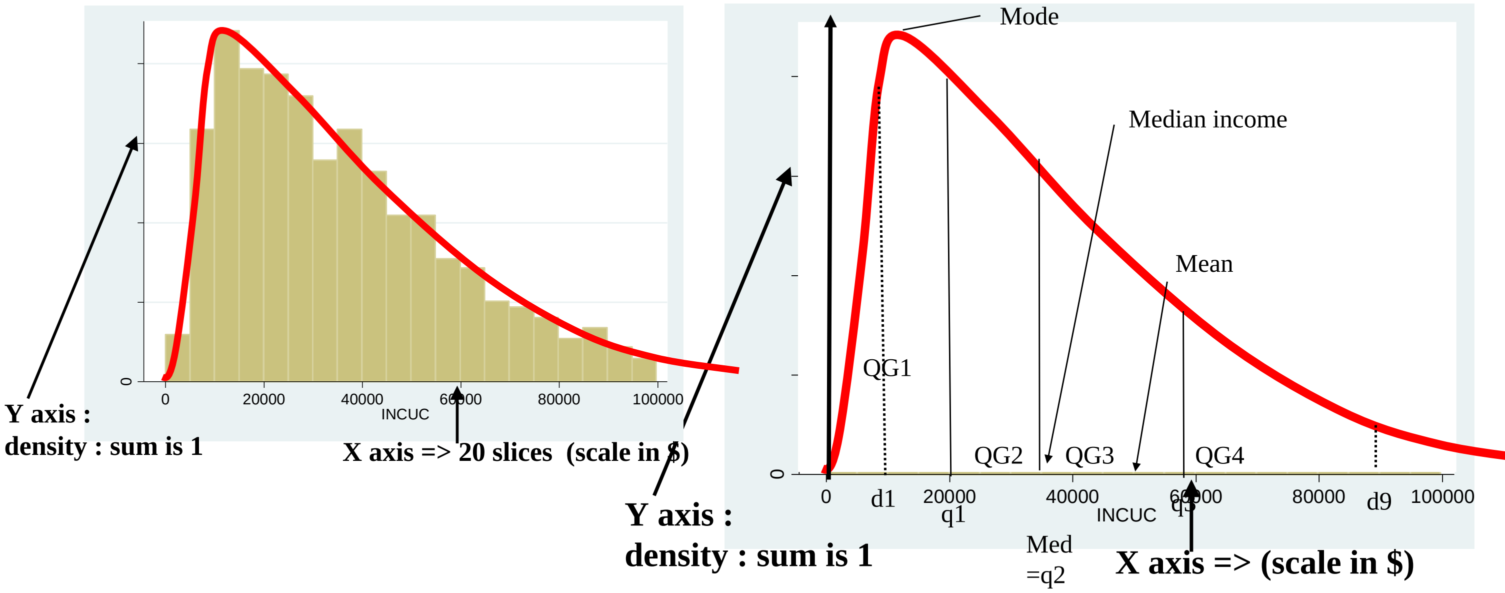
Working
http://www.louischauvel.org/scf2007b.do

Example of SCF Survey 2007 : Survey of Consumer Finances
http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html

20.000 households (=hhld), Complex weighting = highly stratified sample
Household income => transform in per CU income
Size of hhld => consumption unit = square root of hhld size
Household wealth => per hhld

Histograms of distributions

```
use "http://www.louischauvel.org/scf2007ext.dta", clear
gen UCN = 1                             /* calculation of number of capita in hhld */
replace UCN = UCN+1 if MARRIED==1
replace UCN = UCN+KIDS
gen INCUC = NORMINC / sqrt( UCN )     /* per UC income */
gen INTWGT = int(WGT)   /* per UC income */
histogram INCUC [fw=INTWGT] if INCUC <100000, bin(20)    /* first histogr */
summarize INCUC [fw=INTWGT] , d
```



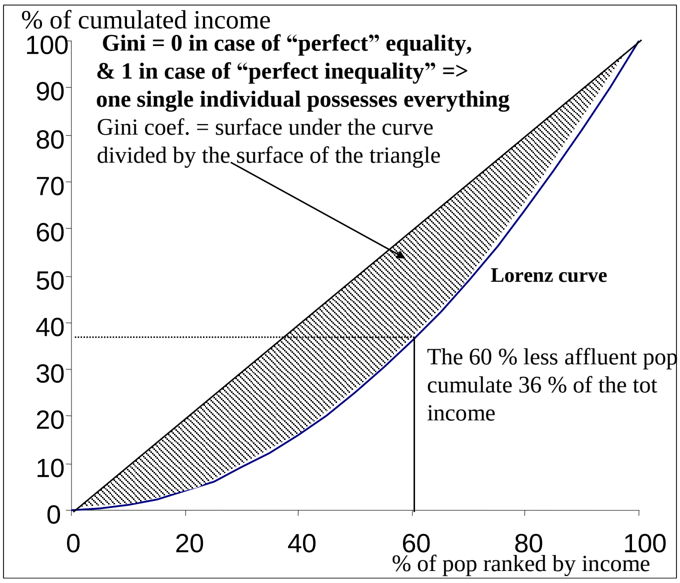
Y axis : density : sum is 1

X axis => 20 slices (scale in $)

Y axis : density : sum is 1

X axis => (scale in $)

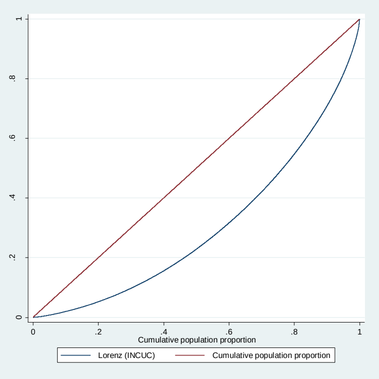
Lorenz curve & the Gini index:



Gini = 0 in case of "perfect" equality, & 1 in case of "perfect inequality" => one single individual possesses everything
Gini coef. = surface under the curve divided by the surface of the triangle

The 60 % less affluent pop cumulate 36 % of the tot income

```
use "http://www.louischauvel.org/scf2007ext.dta", clear
gen UCN = 1                          /* calculation of number of capita in hhld */
replace UCN = UCN+1 if MARRIED==1
replace UCN = UCN+KIDS
gen INCUC = NORMINC / sqrt( UCN )    /* per UC income */
gen INTWGT = int(WGT)   /* per UC income */
histogram INCUC [fw=INTWGT] if INCUC <100000, bin(20)   /* first histogr */
summarize INCUC [fw=INTWGT] , d
ssc install ineqdeco
ssc install glcurve
/* calculation of ginis interdeciles ratios atkinsons and entropy */
ineqdeco INCUC [fw=INTWGT]
glcurve  INCUC [fw=INTWGT] , lorenz pvar(xi) glvar(yi) /* the lorenz curve */
twoway (line yi xi, sort) (line xi xi, sort) , xsize(3) ysize(3) scale(.6)
```



Quantile regressions :

Iqreg is a quantile regression device. It fits the size of the range (and not the average value) of a quantitative variable. With incomes or wealth, you MUST work on the log value (you fit the ratios and not the absolute value that depends on the currency). The iqreg of LINCUC shows that age and education are not factors of increasing/decreasing inequality.
Non-married population is more unequal.

```
gen LINCUC = log(INCUC)    /* ln income */
xi: iqreg LINCUC  AGE i.EDCL i.MARRIED, quantiles(.1 .9)
```

```
i.EDCL            _IEDCL_1-4         (naturally coded; _IEDCL_1 omitted)
i.MARRIED         _IMARRIED_1-2      (naturally coded; _IMARRIED_1 omitted)
.9-.1 Interquantile regression                      Number of obs =     10537
  bootstrap(20) SEs                                 .90 Pseudo R2 =    0.1207
                                                    .10 Pseudo R2 =    0.1495
------------------------------------------------------------------------------
             |              Bootstrap
      LINCUC |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         AGE |   .0004344    .000847     0.51   0.608    -.0012258    .0020946
    _IEDCL_2 |  -.0141857   .0546681    -0.26   0.795    -.1213455    .0929741
    _IEDCL_3 |    .059525    .065677     0.91   0.365    -.0692144    .1882644
    _IEDCL_4 |  -.0491638   .0542337    -0.91   0.365    -.1554721    .0571445
 _IMARRIED_2 |    .193389    .024661     7.84   0.000     .1450487    .2417293
       _cons |     1.5492   .0689323    22.47   0.000     1.414079     1.68432
------------------------------------------------------------------------------
```