Formation STATA : Principaux éléments et commandes d'initiation au logiciel Stata

Presentation · July 2013

DOI: 10.13140/RG.2.2.30538.67527

1 author:

Abdeljaouad Ezzrari
Laboratoire de Statistique Appliquée à l'Analyse et la Recherche en Économie - (LASAARE)
17 PUBLICATIONS **78** CITATIONS

Formation STATA

Abdeljaouad EZZRARI,
Haut-Commissariat au Plan
Juillet 2013

Plan

- **Introduction**
- **Présentation du logiciel**
  - Présentation de l'interface
  - Comment transférer une base de données en format Stata
- **Fonctions et expressions**
- **Description des données**
  - Describe
  - List
  - codebook
  - lookfor

Plan

- **Les Extensions de fichiers Stata**
  - Fichier programme (*.do)
  - Fichier données (*.dta)
  - Fichier résultats (*.log ou smcl)

- **Commandes de gestion des variables**
  - Etiquetage des variables et des modalités
  - Création d'une nouvelle variable (quintile, décile, etc.)
  - Transformation d'une variable
  - Boucles

Plan

- **Fusion des bases des données**
  - Ajouter des variables
  - Ajouter des observations
  - Agréger des variables
- **Pondération**
- **Commandes de base de tabulation statistique**
  - Statistique descriptive et fréquence (variables qualitatives et quantitatives)
  - Tableaux de croisements (variables qualitatives)
  - Liaison entre les variables qualitatives et les variables quantitatives
  - Tests usuels

Plan

- **Graphiques dans Stata**
  - Histogramme
  - Diagramme en barre ou en secteurs
  - Nuages de points, etc.
- **Matrices dans Stata**
- **Régressions dans stata**
  - Moindres carrés ordinaires (MCO) et tests
  - Données de Panel
  - Econométrie des variables qualitatives

Plan

- Cartographie des indicateurs dans Stata
- Ajout de nouveaux modules Stata

Introduction

- Stata est un logiciel complet permettant l'analyse statistique et économétrique développé par Stata Corporation.

- C'est un logiciel particulièrement utilisé en épidémiologie et en économie.

- Ce logiciel est actuellement à la version 13.

- Il existe pour tous les systèmes d'exploitation (Windows, Linux, Mc, etc.).

Introduction

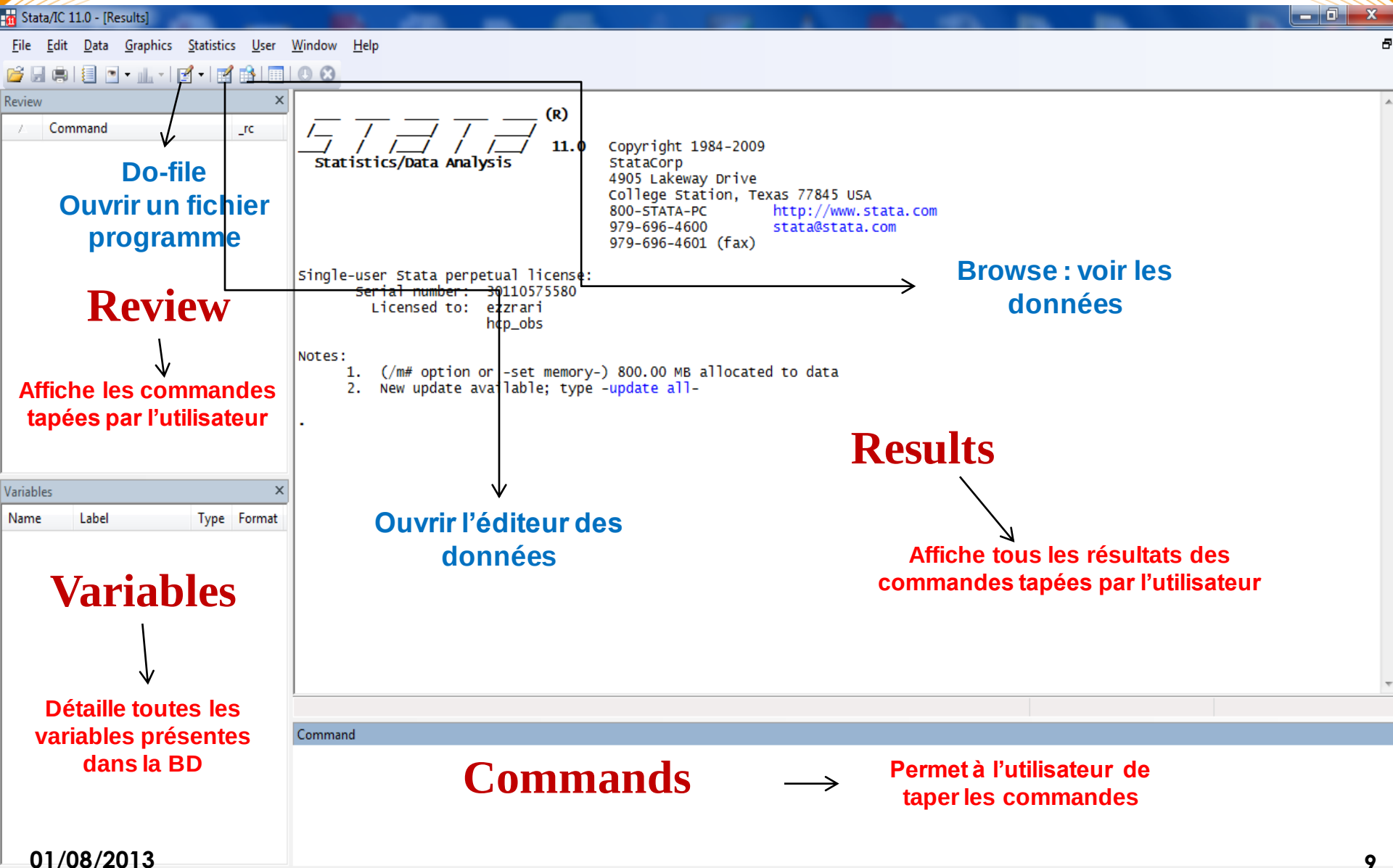
- **Mode de fonctionnement :**
  - Mode commande interactif
  - Mode Menu
  - Mode de programmation (fichiers .do)

- **C'est un logiciel assez flexible et complet. Possibilité de faire de programmation.**

- **Contrairement à d'autres logiciels (SAS, R, etc.), Stata a des problèmes pour gérer de très grosses bases de données.**

Présentation du logiciel

1- Présentation de l'interface



Do-file
Ouvrir un fichier programme

Review

Affiche les commandes tapées par l'utilisateur

Browse : voir les données

Ouvrir l'éditeur des données

Results

Affiche tous les résultats des commandes tapées par l'utilisateur

Variables

Détaille toutes les variables présentes dans la BD

Commands → **Permet à l'utilisateur de taper les commandes**

Présentation du logiciel

2- Lire ou transférer une BD au format Stata

- Si le fichier est déjà au format Stata, pour l'ouvrir il faut taper :

use "nom_fichier.dta", **clear** (ouvrir la totalité du fichier)

use var1 var2 var3 …. **using** "nom_fichier.dta", **clear** (n'ouvrir le fichier qu'avec les variables mentionnées var1 var2 var3…)

clear pour effacer le fichier de données déjà utilisé par Stata

Présentation du logiciel

2- Lire ou transférer une BD au format Stata

Si le fichier n'est pas au format Stata :

- utiliser le Stat Transfer : c'est un logiciel qui permet de convertir les données utilisables sous un autre format (Excel, SAS, R, Limdep, SPSS, etc.) au format Stata.

- Stata peut lire les données également sous format ASCII. Dans ce cas on utilise souvent les trois commandes suivantes :

  - **infile**
  - **insheet**
  - **infix**

Présentation du logiciel

2- Lire ou transférer une BD au format Stata

- on utilise **infile** si les données sauvegardées dans un fichier sont séparées par un espace, pour lire les données on utilise :

infile var1 var2 var3 ….. **using** "exercice1.prn" , **clear**

- on utilise **insheet** si les données sauvegardées dans un fichier sont séparées par des tabulations, pour lire les données on utilise

insheet var1 var2 var3 ….. **using** "exercice1.txt" , **clear** (le fichier ne contient pas les noms des variables)

insheet using "exercice1.txt" , **clear** (fichier contient les noms des variables)

Présentation du logiciel

2- Lire ou transférer une BD au format Stata

- on utilise **infix** s'il n'y aucune séparation entre les données. Dans ce cas, on aura besoin d'un autre fichier qui spécifie la disposition des données, c'est-à-dire un dictionnaire des variables.

  Exemple : On observe pour 4 ménages, 5 variables : identifiant (premier chiffre), milieu (second chiffre), âge du CM (deux suivants), revenu du ménage (5 chiffres qui suivent) et région (variable alphanumérique : dixième position) :

  113007000A

  225515000B

  314904500A

  423409000B

Pour lire les données on utilise :

infix identifiant 1 milieu 2 age 3-4 revenu 5-9 str region 10 **using** "classeur1.prn" , clear

Présentation du logiciel

2- Lire ou transférer une BD au format Stata

- Finalement après avoir lu les données, il faut les sauvegarder dans un fichier stata à l'aide de la commande :

save "nom_fichier ", **replace**

replace sert à remplacer le fichier s'il existe déjà

- Saisie manuelle des données (peu pratique) :

On utilise la commande **input**

Var. alphanumérique qui prend une position

input identifiant milieu age revenu str1 region
1 1 30 7000 A
2 2 55 15000 B
3 1 49 4500 A
4 2 34 9000 B
end

Fonctions et expressions

1. Opérateurs arithmétiques		2. Opérateurs de relation		3. Opérateurs logiques	
Addition	+	Supérieur Inférieur	> <	OU (alt gr + 6)	|
Soustraction	-	Supérieur ou égal	>=	ET	&
Multiplication	*	Inférieur ou égal	<=		
Division	/	Egal Egal (s'il y a if)	= ==		
Exposant	^	Différent	~= !=		

Fonctions et expressions

4. Fonctions		5. Expressions by, if et in
Racine carrée	sqrt	by : permet de répéter une commande pour chaque valeur (ou modalité) d'une variable donnée. Syntaxe générale pour by est : by variables : commande …
Exponentielle	exp	
Logarithme	log ln	if : permet de spécifier les conditions dans lesquelles une commande doit être exécutée. Syntaxe générale pour if est : commande …. if condition
Valeur Absolue	abs	
Partie entière	int	in : permet de spécifier les observations auxquelles s'applique une commande. Syntaxe générale pour in est : commande …. in intervalle

Description des données

Il y a plusieurs commandes qui permettent de décrire et de voir les données :

- **edit** : voir la base de données et permet de la modifier à la main

  edit ou **edit variables**

- **browse** : voir la base de données et ne permet pas de la modifier à la main

  browse ou **browse variables**

- **describe** : la commande describe permet de décrire les données de façon générale (format de la variable, label des modalités de la variable, label de la variable)

  describe : décrit toute la base

  describe variables : ne décrit que les variables indiquées

Description des données

insheet using "c:\formation_stata\exercice1.txt", clear

describe

```
Contains data
  obs:            15
 vars:             4
 size:            75
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
sexe            byte   %8.0g
age             byte   %8.0g
abonnement      byte   %8.0g
revenu          int    %8.0g
-------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved
```

describe age sexe

```
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
age             byte   %8.0g
sexe            byte   %8.0g
```

Description des données

- **list** : permet d'afficher la base de données ou un extrait de cette base dans la fenêtre des résultats

  list ou **list variables**

```
insheet using "c:\formation_stata\exercice1.txt", clear
```

list in 1/6 , voir la base de données pour uniquement les 6 premières observations

```
. list in 1/6

     +-------------------------------------+
     | sexe   age   abonne~t   revenu |
     |-------------------------------------|
  1. |    0    45          0     1234 |
  2. |    0    46          0     1250 |
  3. |    0    54          0     1400 |
  4. |    0    44          1     3500 |
  5. |    0    47          0     2600 |
     |-------------------------------------|
  6. |    1    47          1     2900 |
     +-------------------------------------+
```

Description des données

- **codebook** : permet de créer un dictionnaire des variables indiquant le nom de la variable, son label, son format, l'intervalle de ses valeurs, sa moyenne, son écart type, des quantiles (variable continue), fréquences des modalités et leurs labels (variable discrète) , etc.

```
insheet using "c:\formation_stata\exercice1.txt", clear
codebook sexe revenu
```

```
-------------------------------------------------------------------------------
sexe                                                            (unlabeled)
-------------------------------------------------------------------------------

                  type:  numeric (byte)

                 range:  [0,1]                        units:  1
         unique values:  2                    missing .:  0/15

            tabulation:  Freq.  Value
                            5  0
                           10  1

-------------------------------------------------------------------------------
revenu                                                          (unlabeled)
-------------------------------------------------------------------------------

                  type:  numeric (int)

                 range:  [1234,7000]                  units:  1
         unique values:  15                       missing .:  0/15

                 mean:   3356.27
            std. dev:    1743.76

           percentiles:        10%       25%       50%       75%       90%
                              1250      2350      3000      4900      6000
```

Description des données

❑ **lookfor** : c'est une commande qu'on utilise pour chercher les variables d'une grande base de données à partir des libellés des variables.
Le cas des Enquêtes DHS (noms des variables représentent le numéro des questions).

use "D:\D\ENPSF_2011\Household 1.dta", clear

lookfor eau

variable name	storage type	display format	value label	variable label
h308	double	%10.0g	h308	principale source d'eau que boivent les membre du menage
h309	double	%10.0g	h309	principale source d'eau utilisee
h310	double	%10.0g	h310	temps pour aller chercher de l'eau et revenir
h310a	double	%10.0g	h310a	qui se rend habituellement a la source d'eau
h311a	double	%10.0g	h311a	réservoir d'eau
h311y	double	%10.0g	h311y	ne reserve pas d'eau
h312b	double	%10.0g	h312b	y ajouter de l'eau de javel/chlore
h312y	double	%10.0g	h312y	ne traite pas l'eau
h324g	double	%10.0g	h324g	chauffe eau
h502c	double	%10.0g	h502c	zone contient l'eau stagnante
h502d	double	%10.0g	h502d	zone souffre d'une éruption des eaux usées

Extensions des Fichiers Stata

- **Fichier données** : c'est un fichier de données sous format stata avec l'extension **.dta** (les variables sont en colonnes et les individus sont en ligne).

- **Fichier programme** : c'est un fichier de commandes au format ASCII. Il permet à l'utilisateur de :
  - lancer plusieurs commandes Stata en une seule opération;
  - Garder une trace des commandes exécutées.

L'extension de ce programme est **.do**

On peut appeler un fichier do-file à partir du menu (do-file Editor) ou bien taper **doedit** dans la partie réservée aux commandes.

C'est un fichier de base dans Stata.

Extensions des Fichiers Stata

cd c:\formation_stata	/*spécifier le répertoire de travail*/
clear all	/*Effacer les fichiers existants et vider la mémoire*/
set memory 800m, permanent	/*permet d'augmenter la mémoire disponible*/
delimit	/*utile pour les commandes très longues et on souhaite revenir à la ligne*/
log using	/*ouvre un fichier résultat*/
cmdlog using	/*ouvre un fichier pour sauvegarder les commandes utilisées*/
log close	/*ferme le fichier résultat*/
cmdlog close	/*ferme le fichier des commandes*/

Extensions des Fichiers Stata

- **Fichier résultats** : c'est un fichier qui permet de stocker toutes les commandes exécutées ainsi que les résultats obtenus. Il y a deux types de fichiers :
  - un fichier en format smcl ouvrable uniquement sur le logiciel Stata
  - un fichier log ou txt ouvrable avec n'importe quel éditeur

commande utilisée est :

log using "nom_fichier ", **replace append** **(smcl)**

log using "nom_fichier.log", **replace append** **(fichier texte)**

log close (fermer) ; **log off** (suspendre) ; **log on** (reprendre)

Il y également une commande qui permet de sauvegarder uniquement les commandes exécutées sans résultats c'est : **cmdlog using**

Commandes de gestion des variables

Etiquetage des variables et des modalités :

Pour une meilleure description et une meilleure lecture des fichiers de données on affecte un label à chaque variable et à chaque modalité

- **Label des variables :**

label var var1 "nom de la variable"

- **Label des modalités :**

label define var1 1 "label1" 2 "label2" 3 "label3" …

label values var1 var1

label values var2 var1 (affecter les labels de la variable var1 à la variable var2)

Commandes de gestion des variables

Etiquetage des variables et des modalités :

Reprenant l'exemple précédent (dans un do-file) :

```
clear
cd c:\formation_stata
set mem 300m, permanent
capture log close
log using "resultats.log", replace
insheet using "exercice1.txt", clear
label var sexe "sexe de l'individu"
label var age "âge de l'individu"
label var abonnement "abonnement au téléphone"
label var revenu "revenu de l'individu"
label define sexe 1 "masculin" 0 "féminin"
label values sexe sexe
label define ouinon 1 "oui" 0 "non"
label values abonnement ouinon
save "base_données" , replace
log close
```

Commandes de gestion des variables

Création d'une nouvelle variable

Les principale commandes de création de variables sont : **generate** et **egen.**

La commande **egen** est une extension de la commande **generate**, elle est utilisée pour créer des variables avec des fonctions spécifiques.

Exemples :

```
gen var3=var1+var2                    /*addition*/
gen var4=5*var1                       /*multiplication*/
gen var6=var2/var1                    /*division*/
gen logvar=log(var)                   /*logarithme*/
gen  region=int(identifiant/10000)    /*partie entière*/
```

Commandes de gestion des variables

Création d'une nouvelle variable

Exemples :

```
use "ennvm07 ", clear
gen pauvrete=1 if (deptotp<=3834&milieu==1)|(deptotp<=3569&milieu==2)
replace pauvrete=0 if pauvrete==.   (. = missing)
ou gen pauvrete=(deptotp<=3834&milieu==1)  |  (deptotp<=3569&milieu==2)
```

Création des variables dymmies (dichotomiques)

```
gen var1=var2==1            /*(var1 est une variable dichotomique prenant la valeur 1 si
                              var2 est égale à 1, 0 sinon)*/
gen urbain=milieu==1
Ou bien
tabulate var2, gen(var)     /* créer des variables dichotomiques pour chaque modalité
                              de la var2*/
tabulate nivscol2, gen(niv_scolaire)
```

Commandes de gestion des variables

Création d'une nouvelle variable

```
egen var7=sum(var1)              /*somme de la variable1*/
egen var8=sd(var2)               /*écart type de la variable2*/
egen var10=rsum(var1 var2 var3)  /*somme des variables 1, 2 et 3*/

egen damm=rsum(alim habit habillement sante transport enseignement …) /*la dépense totale du ménage est la somme des différents groupes de dépenses*/
```

création des percentiles :

```
xtile quintile=deptotp, nq(5)                  /*quintile à l'échelle nationale*/
xtile quint_urb=deptotp if milieu==1, nq(5)    /*quintile au niveau urbain*/
xtile decile=deptotp, nq(10)                   /*décile à l'échelle nationale*/
xtile decile_rur=deptotp if milieu==2, nq(10)  /*décile au niveau rural*/
xtile percentile=deptotp, nq(100)              /*centile à l'échelle nationale*/
```

Commandes de gestion des variables

Création d'une nouvelle variable

sumdist deptotp

/*distribution des dépenses de consommation selon les déciles*/

sumdist deptotp, **ngp(5)**

/*distribution des dépenses de consommation selon les quintiles*/

sumdist deptotp, **lvar(l) pvar(p)**

/*distribution des dépenses de consommation selon les déciles tout en sauvegardant le % cumulé de la population et le % cumulé de la consommation totale*/

twoway (connect p p) (connect l p, sort)

/*Permet de faire la courbe de concentration de Lorenz*/

Commandes de gestion des variables

Gestion et Manipulation des variables

Il existe d'autres commandes relatives à la gestion des variables :

rename : permet de renommer la variable

rename anc_var new_var

drop : permet de supprimer une ou plusieurs variables

drop var1 var2 …. **in**, **if**

keep : permet de conserver dans le fichier les variables chiosies

keep var1 var2 var3 … **in**, **if**

sort : permet de trier le fichier selon des clés choisis

sort idt_men n_ordre

Commandes de gestion des variables

Gestion et Manipulation des variables

order : sert à ordonner les variables de la base

order idt_men n_ordre region province

aorder : sert à ordonner les variables de la base par ordre alphabétique

destring : permet de transformer une variable alphanumérique en variable numérique

destring region, g(c_region)

tostring : transformer variable numérique en une variable alphanumérique

tostring c_region, g(region)

encode : transformer variable alphanumérique en une variable numérique dont les modalités sont labelisées avec des chaines de caractère

encode region, g(c_region)

Commandes de gestion des variables

Transformation d'une variable

Il s'agit de recourir à des transformations des variables initiales à d'autres formes de variables selon l'usage.

Exemples :

-transformer l'âge en groupe d'âge; -

-transformer le type d'activité à 11 modalités (type d'activité détaillé) à un type d'activité à 3 ou 2 modalités (type d'activité agrégé).

```
recode age (0/14=1) (15/59=2) (60/max=3), g(groupe_âge) /*créer une autre variable*/

recode typeact (1=1) (2=2) (3/max=3), g(type_act_agr) /*créer une autre variable*/

recode var (1 2 3=1) (4/6 8=2) (7 9=3)     /*remplacer la variable existante*/
```

Commandes de gestion des variables

Boucles

Les boucles sont des programmes qui permettent de faire une seule manipulation des variables au lieu de plusieurs. Il y a deux commandes principales de boucles **: forvalues, foreach**

forvalues : on l'utilise si les variables contiennent des chiffres

Exemple : créer plusieurs fichiers (ENNVM07) relatifs aux 16 régions

```
forvalues i=1/16 {
use "ennvm07" if c_region==`i', clear
save "ennvm07_reg`i'", replace
}
```

foreach : on l'utilise pour toute autre variable

```
foreach var in sexe age etatmatr act_occ chomeur inactif lirecrir sans_niv f1 f2 second superir {
rename `var' `var'_cm
}
```

Fusion des Bases de données

Ajout des variables

L'objet est de fusionner deux bases de données contenant des individus en commun et des variables différentes.

Supposons qu'on dispose de deux bases de données (carte_démogaraphique) et (carte_emploi) de six individus et qu'on veut fusionner ces deux bases.

1- il faut s'assurer que les individus ont un **identifiant unique** dans les deux bases

2- **Trier** les deux bases selon cet identifiant

3- utiliser la commande **merge** dans stata pour la fusion

Fusion des Bases de données

Ajout des variables (one to one)

cartedemog.dta

ldmen	idind	sexe	age
01	0101	1	44
01	0102	2	38
01	0103	1	15
02	0201	2	36
02	0202	2	5
02	0203	1	8

emploi.dta

idmen	idind	sitac
01	0101	AO
01	0102	FF
01	0103	EE
02	0201	AO
02	0203	EE

```
cd c:\formation_stata
use "emploi.dta", clear
sort idind
save "emploi.dta", replace
use "cartedemog.dta", clear
sort idind
merge 1:1 idind using "emploi.dta"
```

Fusion des Bases de données

Ajout des variables (one to one)

Result	# of obs.	
not matched	1	
from master	1	**(_merge==1)**
from using	0	**(_merge==2)**
matched	5	**(_merge==3)**

idmen	idind	sexe	age	sitac	_merge
01	0101	1	44	AO	matched (3)
01	0102	2	38	FF	matched (3)
01	0103	1	15	EE	matched (3)
02	0201	2	36	AO	matched (3)
02	**0202**	**2**	**5**		**master only (1)**
02	0203	1	8	EE	matched (3)

Fusion des Bases de données

Ajout des variables (many to one) or (one to many)

cartedemog.dta

idmen	idind	sexe	age
01	0101	1	44
01	0102	2	38
01	0103	1	15
02	0201	2	36
02	0202	2	5
02	0203	1	8

logement.dta

idmen	typehab
01	Apprt
02	MM

```
cd c:\formation_stata
use "logement.dta", clear
sort idmen
save "logement.dta", replace
use "cartedemog.dta", clear
sort idmen
merge m:1 idmen using "legement.dta"
```

Fusion des Bases de données

Ajout des variables (many to one) or (one to many)

Result	# of obs.	
not matched	0	
matched	6	(_merge==3)

idmen	idind	sexe	age	typehab	_merge
01	0101	1	44	Apprt	matched (3)
01	0102	2	38	Apprt	matched (3)
01	0103	1	15	Apprt	matched (3)
02	0201	2	36	MM	matched (3)
02	0202	2	5	MM	matched (3)
02	0103	1	8	MM	matched (3)

Fusion des Bases de données

Ajout des variables

la commande mmerge est une extension de la commande merge et permet de faire la fusion des bases de données sans passer par le tri.

Exemple :

```
cd c:\formation_stata
use "cartedemog.dta", clear
mmerge  idind using "emploi.dta"      /*ajouter toutes les
                                        varaibles du fichier emploi*/

mmerge  idind using "emploi.dta", table ukeep(var1 var2) /*n'ajouter
que les variables var1 et var2 du fichier emploi*/
```

Fusion des Bases de données

Ajout des observations

Supposons qu'on dispose de deux bases de données, l'une pour le milieu urbain et l'autre pour le milieu rural et on veut les fusionner en une seule base. Il s'agit là d'ajout d'observations et la commande qu'on utilse dans Stata est **append.**

```
cd c:\formation_stata
use "fichier_urbain", clear
append using "fichier_rural"
save "fichier_national", replace
```

Fusion des Bases de données

Agréger des variables

Il s'agit de passer d'une base de données désagrégées à une base de données agrégées. En d'autres termes, il s'agit de remplacer la base de données utilsiée par une base de statistiques descriptives.

Supposons qu'on dispose des données par ménage sur la pauvreté et les niveaux de vie et nous voulons agréger les indicateurs de pauvreté et des niveaux de vie au niveau régional, la commande qu'on utilise dans Stata est : **collapse**

```
cd c:\formation_stata
use "ennvm07", clear
preserve                                    /*garder le fichier existant*/
collapse pauvreté deptotp (sum) pop=taille (count) men=up, by(c_region)
save "pauvreté_région", replace
restore                                     /*récupérer le fichier*/
```

Pondération

Plusieurs types de poids existent dans Stata :

- **fweight** : duplication de l'observation (le nombre d'individus représentés par une observation)
- **pweight** : c'est l'inverse du taux de sondage
- **aweight** : qd les observations sont des moyennes et le poids est le nombre d'éléments ayant servi au calcul de ces moyennes.

Un calcul sans biais de la précison des indicateurs calculés nécessite d'autres informations du plan d'échantillonnage : les strates, les unités primaires de sondage, etc.

Le plan d'échantillonnage se définit par la commande **svyset**

svyset up [pw=poids], strata(strate)

Pondération

Une fois le plan de sondage défini, le prefixe **svy** permet d'effectuer des estimations en estimant correctement la précision :

svy : **mean** var1
svy : **total** var1
svy : **ratio** var1/var2
svy : **reg** var1 var2

Pour les tabulations, on utilise généralement **fweight** (la variable relative à la pondération dans ce cas doit ête un nombre entier)

commande variables **[fw=fweight]**

Commandes de base da tabulation statistique

Statistique descriptive

Cas d'une variable quantitiave

```
summarize var1            /* N, mean, sd, min, max*/
summarize var1, detail    /* N, mean, sd, min, max, variance,
                             skewness, kurtosis, percentiles*/
tabstat var1              /* seulement la moyenne*/
tabstat var1, stat(n, mean, median, sd, var , min, max)      /*+ieurs statistiques*/
Tabstat var1, stat(mean, median) by(sexe)   /*+ieurs statitiques vetilées par
                                              une variable catégorielle*/
```

Exemple:

```
cd c:\formation_stata
use "ennvm07", clear
sum deptotp deptotm taille age [fw=coef_ind]
sum deptotp deptotm taille age [fw=coef_ind], detail
tabstat deptotp [fw=coef_ind]
tabstat deptotp [fw=coef_ind], stat(mean median min max N)
tabstat deptotp [fw=coef_ind], stat(mean median) by(milieu)
```

Commandes de base da tabulation statistique

Tableaux de croisement

```
tab var1 var2                    /*les n seulement d'un tableau de croisemen*/
tab var1 var2, row               /*les n+% lignes*/
tab var1 var2, row col           /*les n+%lignes+%colonnes*/
tab var1 var2, nofreq row col    /*%lignes+%colonnes*/

table var1 var2 var3             /*les n seulement d'une table à 3 entrées*/
```

Exemple:

```
cd c:\formation_stata
use "ennvm07", clear
tab sexe milieu [fw=coef_ind], col row
table sexe milieu etatmatr [fw=coef_ind], col row scol format(%12.0g)
```

Commandes de base da tabulation statistique

Liaison variables qualitatives et variables quantitatives

```
tab var1 var2, sum(var3) nofreq /*moyenne de v_quant vs 2 var qual*/

table var1 var2 var3, c (mean var4 median var4 ) row col scol
      /*les stat des d'une varaible quantitative en fonction de 3 var qualitatives*/
```

Exemple:

```
cd c:\formation_stata
use "ennvm07", clear
tab sexe milieu [fw=coef_ind], sum(deptotp) nofreq means
table milieu sexe etatmatr [fw=coef_ind], contents(mean p0_mon median p0_mon )
    row col scol
```

Commandes de base da tabulation statistique

Tests usuels

Test d'indépendance entre deux varaibles qualitatives

```
table var1 var2, chi2          /*relation d'indépendance entre deux variables*/
```

Test de corrélation de Pearson

```
corr var1 var2 var3            /*coefficient de corrélation entre les variables*/
pwcorr var1 var2 var3 …, sig   /*coef. corrél. entre les variables + degré de sig*/
```

Test de différences de moyennes

```
ttest  var1=var2               /*comparaison de la moyenne de 2 échantillons*/
ttest  var1=valeur             /*comparaison de la moyenne d'une variable*/
ttest  var1, by(var2)          /*comparaison de la moyenne de deux groupes*/
```

Exporter les tableaux statistiques

La commande logout permet d'exporter un tableau de résultats au format excel, word ou text.

```
logout, save(table1) excel replace : table var1 var2 ...   /*format Excel*/
logout, save(table1) word replace : table var1 var2 ...    /*format Word*/
logout, save(table1) tex replace : table var1 var2 ...     /*format texte*/
logout, save(table1) excle word ex replace : table var1 var2 ...   /*tous les formats*/
```

Exemple

```
cd c:\formation_stata
use "ennvm07", clear
logout, save(table1) excel word replace : tabstat deptotp alimsstabp habillement_p [fw=coef_ind], by(milieu)
```

Graphiques

Il existe plusieurs types de graphe selon la nature de la variable étudiée. La commande de base de graphe est : twoway suivie du type de graphique (line, bar, scatter, hist) et des variables (celles de l'axe vertical, ensuite celles de l'axe horizontal).

❑ Histogramme

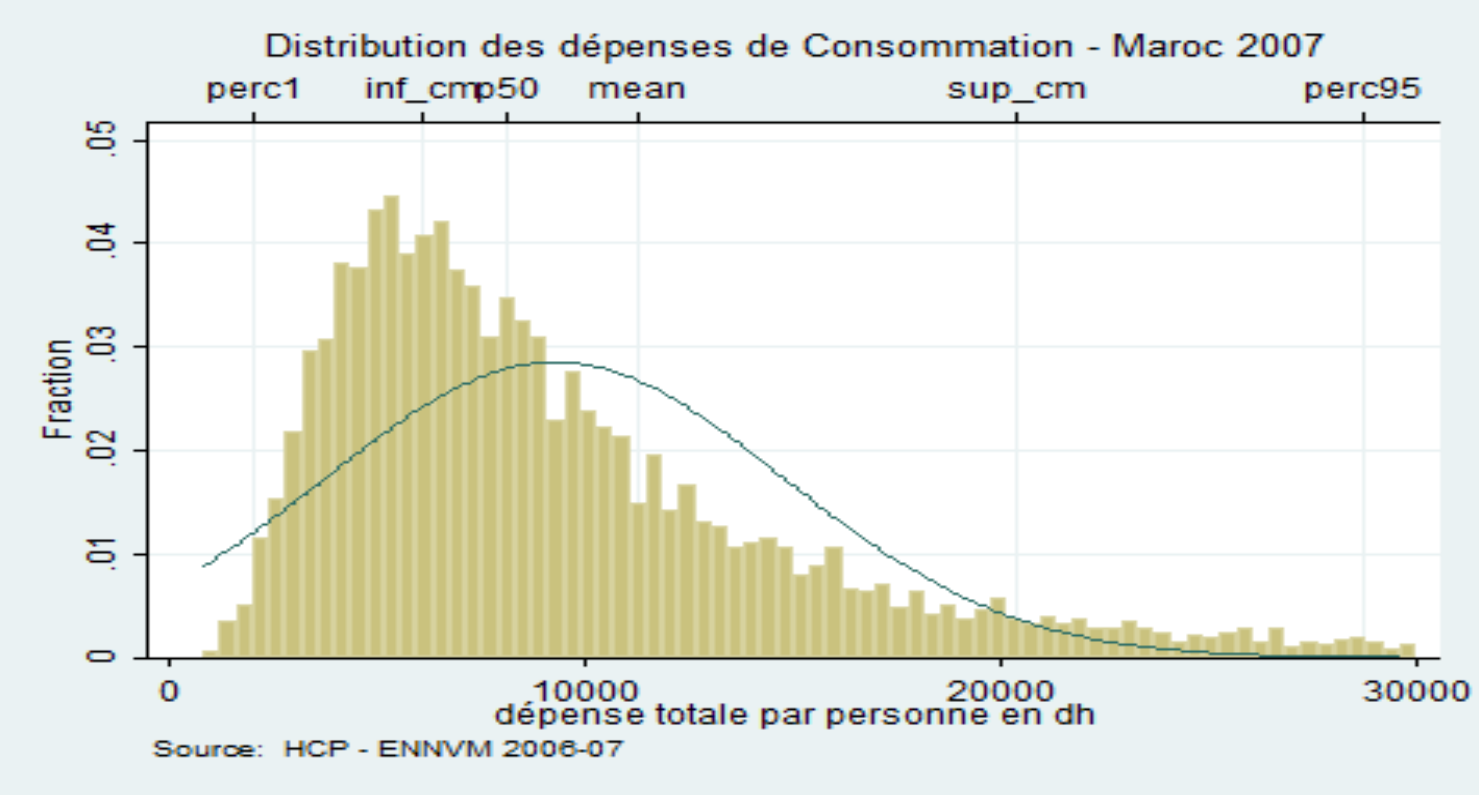
```
hist var1, width(20) start(50) fraction=freq
hist var1, width(20) start(50) fraction=freq normal      /*Ajouter la courbe de distribution normale*/
```

Exemple

```
cd c:\formation_stata
use "ennvm07", clear
hist deptotp [fw=coef_ind], width(2000) start(800) fraction normal
```

Graphiques

```
histogram deptotp [fw=coef_ind] if deptotp<=30000, fraction normal xaxis(1 2)
ylabel(0(0.01)0.05, grid) xlabel(8095 "p50" 2034 "perc1" 6071 "inf_cm" 11233 "mean"
20388 "sup_cm" 28688 "perc95" , axis(2) grid gmax) xtitle("", axis(2)) subtitle ("Distribution
des dépenses de Consommation - Maroc 2007") note("Source: HCP - ENNVM 2006-07")
```



Graphiques

❑ Diagramme en barre

Généralement la commande utilisée pour faire des graphiques en barre est graph bar

```
graph bar var1 , over(var2)          /*donne graphe de moyenne de la var1 en
                                     fonction de var2 (qualitative)*/ (vertical)
graph bar (median) var1, over(var2) /*donne graphe de la médiane de la var1 en
                                     fonction de var2 (qualitative)*/ (vertical)
graph hbar var1 , over(var2)         /*donne graphe de moyenne de la var1 en
                                     fonction de var2 (qualitative)*/ (horizontal)
```

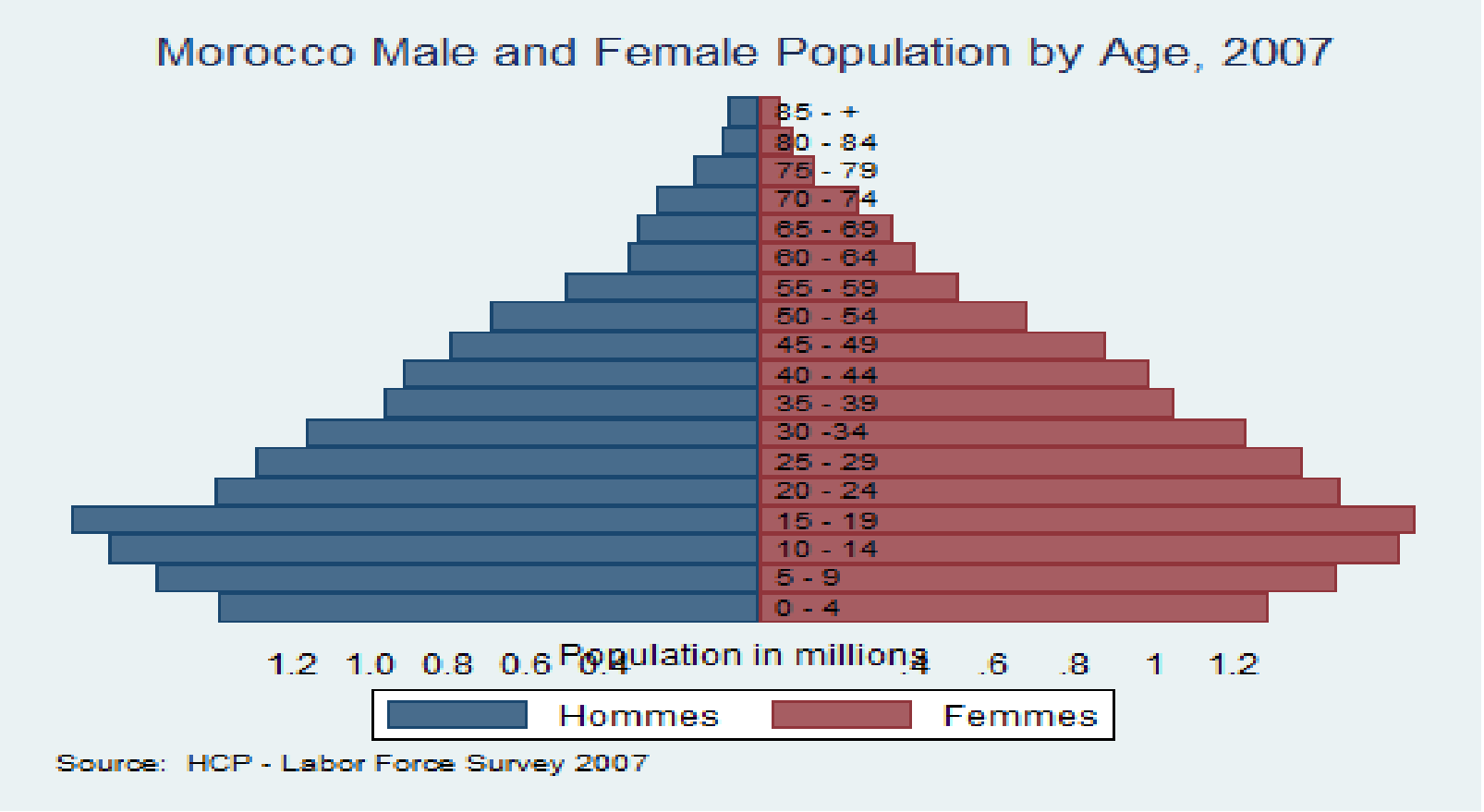
Exemple

```
cd c:\formation_stata
use "ennvm07", clear
graph bar deptotp [fw=coef_ind], over(nivscol2) ytitle(deptotp) title("les dépenses de consommation selon le niveau scolaire du CM", size(median)) note("Source: HCP – ENNVM 2006-07")
```

Graphiques

Diagramme en barre (Exemple de la pyramide des âges)

Voir le programme (**pyramide 2007.do**)



Graphiques

❑ Diagramme en secteur

La commande utilisée pour faire des Diagrammes en secteurs est graph pie

graph pie var1 , over(var2) /*donne graph en secteur de la var1 selon la var2*/

graph pie var1, over(var2) by(var3 total /*donne graph ventilé par la var3 de la var1 en fonction de var2 (qualitative)*/

Exemple

```
cd c:\formation_stata
use "ennvm07", clear
graph pie deptotp [fw=coef_ind], over(nivscol2) by(, title(Dépenses de consommation selon le milieur de résidence)) by(, note(Source: HCP – ENNVM 2006-07)) by(milieu, total)
```

Graphiques

❑ Nuage de points

Le nuage de points est utilisé lorsqu'on veut voir la liaison entre deux variables quantitatives et la commande utilsée est twoway scatter

```
twoway (scatter var1 var2)          /*donne graph de la var1 en fonct de la var var2*/

graph (scatter var1 var2) || (lfit var1 var2) /*donne en plus du graph de la var1 en
                                                fonction de la var2, la droite de régression */
```

Exemple

```
cd c:\formation_stata
use "ennvm07", clear
twoway (scatter coeff_budg_alim_pc deptotp) || (lfit coeff_budg_alim_pc deptotp), title("la relation entre le niveau de vie") subtitle("et le coefficient budgétaire de l'alimentaire") xlabel(50000 "5" 100000 "10" 150000 "15" 200000 "20")
```

Matrices

Création d'une matrice

- Pour créer une matrice dans Stata, on utilise la commande **matrix input**

matrix input A = (a,b,c,d\e,f,g,h\.....)
matrix input X = (1,2\3,4)

- Pour créer une matrice dans Stata à partir d'autres matrices, on utilise la commande **matrix define**

matrix define X = A + B + C
Ou
matrix X = A + B + C

Matrices

Transformation d'une matrice

- **Pour transférer une base de données en une matrice, on utilise la commande mkmat**

```
mkmat A B C D E, matrix(X)      /* créer une matrice X contenant les quatre
                                   variables de la base de données A, B, C et D*/
mkmat revenu                    /* créer une matrice ligne revenu contenant la
                                   variable revenu */
```

- **Pour transférer une matrice en une base de données on utilise la commande svmat**

```
svmat X                         /* transférer la matrice X en une base de données
                                   stata avec les lignes comme observations et les
                                   colonnes comme des variables*/
```

Matrices

Le calcul matriciel et d'autres utilisations

- **Le calcul matriciel dans stata se fait par des opérateurs arithmétiques tels +, - ou *, etc et par les fonctions matricielles de type inverse ou transposé.**

Exemple : dégager le vecteur de régression d'une variable Y en fonction d'un vecteur X:

```
matrix B = inv(X'X)X'Y        /*inv signifie l'inverse d'une matrice et ' est le symbole de la transposée d'une matrice*/
```

Autres utilsations

```
matrix dir        /*voir les différentes matrices utilisées dans le fichier de travail*/
matrix list       /*lister les matrices*/
matrix rename     /*renommer une matrice*/
matrix drop       /*supprimer une matrice*/
```

Régressions

MCO

Pour faire des régressions en MCO, on utilise la commande regress (reg) suivie de la variable dépendante, des variables indépendantes et le cas échéant aux options. La syntaxe générale est :

reg var_dep var_explicatives (if, in), options

Pour le cas de l'existence des variables qualitatives parmi les variables explicatives, il faut :

- soit créer des varibles dummy à partir de cette variable et introduire l'ensemble des variables créees sauf une (référence)
- soit utiliser la commande ci-dessous :

```
xi : reg dep_var var1 var2 i.var3 i.var4      /*var3 et var4 étant des variables qualitatives*/
reg dep_var var1 var2 i.var3 i.var4
```

La commande predict permet d'obtenir la valeur prédite (estimée) de la variable dépendante ainsi que les résidus de la régression

```
predict yhat, xb
predict residu, re
```

Régressions

❑ MCO

Pour les tests (normalité des résidus, hétéroscédasticité, endogénéité, etc.) il faut vous référer au (Manuel d'Initiation à Stata (Version 8), Kangni KPODAR : CERDI).

Exemple :

```
cd ..\Formation_Stata
use "ennvm07", clear
gen lndeptotp=ln(deptotp)
xi: reg lndeptotp taille age i.milieu i.nivscol2, robust
```

```
. xi: reg lndeptotp taille age i.milieu i.nivscol2, robust
i.milieu          _Imilieu_1-2        (naturally coded; _Imilieu_1 omitted)
i.nivscol2        _Inivscol2_1-4      (naturally coded; _Inivscol2_1 omitted)

Linear regression                               Number of obs =    7062
                                                F(  6,  7055) =  538.25
                                                Prob > F      =  0.0000
                                                R-squared     =  0.3725
                                                Root MSE      =  .56939
```

lndeptotp	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
taille	-.1037016	.0033316	-31.13	0.000	-.1102326	-.0971706
age	.0092073	.0005453	16.88	0.000	.0081384	.0102763
_Imilieu_2	-.2413459	.0143938	-16.77	0.000	-.2695621	-.2131297
_Inivscol2_2	.2253924	.0162557	13.87	0.000	.1935264	.2572584
_Inivscol2_3	.5890633	.0302782	19.45	0.000	.5297088	.6484177
_Inivscol2_4	1.121653	.0381733	29.38	0.000	1.046822	1.196484
_cons	9.16452	.0349956	261.88	0.000	9.095918	9.233122

Régressions

Données de Panel

La commande utilisée pour faire des régressions en panel est **xtreg.**

Attention : il faut déclarer que vous disposez des données en panel en mentionnant la varible individus et la variable temps.

```
xtset id tps

xtreg var_dep var_explicatives, fe      /*modèles à effets fixes*/

xtreg var_dep var_explicatives, re      /*modèles à effets aléatoires*/
```

Le test d'Hausman permet de choisir entre les deux modèles.

Régressions

Econométrie des variables qualitatives

Le choix de modèle à utiliser dépend de la nature de la variable expliquée.

Les variables qualitatives sont de deux types :

- **les variables dichotomiques : genre, milieu de résidence, vote, etc.. (variables à deux modalités)**
- **les variables polytomiques : ce sont des variables discrètes à plus de 2 modalités et il existe trois types de varaibles polytomiques :**
  - **les variables ordonnées (classes des dépenses, classes des revenus, degré de satisfaction, etc.)**
  - **les variables non ordonées (catégories socioprofessionnelles, le lieu de consultation médicale, le personnel consulté, etc.)**
  - **les varaibles séquentielles (le niveau de diplôme, etc.)**

Régressions

Econométrie des variables qualitatives

1) Si la variable expliquée est dichotomique (on utilise soit le modèle logit ou le modèle probit selon la distribution des aléas). Les commandes utilisées sont :

```
logit var_dep var_explicatives
logit var_dep var_explicatives, or        /*pour avoir les odd ratio*/
Ou
logistic var_dep var_explicatives
```

Les commandes de post-estimation sont les suivantes :

```
predict pscore, xb        /*les valeurs prédites*/
compute mfx, dydx         /*les effets marginaux*/
compute mfx, eyex         /*les élasticités*/
```

Etant donné que seuls les signes des coef. qui sont interprétables.

```
lstat                     /*le seuil pris par défaut est 0.5*/
lstat, cutoff(pr.)        /*possibilité de choisir un autre seuil*/
```

Exemple d'appliation : Insertion professionnelle des diplômes au Maroc (syntax_insertion.do et insert_dip.dta)

Régressions

Econométrie des variables qualitatives

2) Si la variable expliquée est polytomique ordonnée (on utilise soit le modèle logit ordonné ou le modèle probit ordonné selon la distribution des aléas). Les commandes utilisées sont :

oprobit var_dep var_explicatives, robust
ologit var_dep var_explicatives

Les commandes de post-estimation sont les suivantes :

predict pscore, xb /*les valeurs valeurs prédites de l'estimation*/
predict mod1 mod2 … /*les probabilités prédites de chaque modalité*/
mfx, predict (p outcome(0)) /*les effets marginaux de chaque modalité*/

Exemple d'appliation : Les déterminants de la pauvreté monétaire au Maroc : (syntax_déterminants_pauvreté.do et déterminants_pauvreté.dta)

Régressions

Econométrie des variables qualitatives

3) Si la variable expliquée est polytomique non ordonnée (on utilise soit le modèle logit conditionnel ou le modèle logit multinomial dit indépendant, ce dernier est le plus souvent utilisé). Les commandes utilisées sont :

```
clogit var_dep var_explicatives          /*pour le logit conditionnel*/
mlogit var_dep var_explicatives          /*pour le logit multinomianl*/
mlogit var_dep var_explicatives, base(i) /*permet de choisir la modalité de référence*/
```

Les commandes de post-estimation sont les suivantes :

```
mfx, predict (p outcome(1))              /*les effets marginaux de chaque modalité*/
```

Exemple d'appliation : Utilisation des services de santé : les déterminants du lieu consultation : (syntax_lieu_consultation.do et lieu_consultation.dta)

Régressions

Econométrie des variables qualitatives

4) Il y a d'autres modèles qu'on peut utiliser dans stata, notamment pour les données censurées. On utilise dans ce cas soit le modèle tobit ou les méthodes d'estimation en deux étapes (heckman ou hechprob)

Les commandes utilisées sont les suivantes

tobit var_dep var_explicatives, ul(.) ll(.) /*pour le logit conditionnel*/

heckman var_dep var_explicatives, select(variables_explicatives de l'équation de sélection) /*si la variable censurée est quantitative*/

heckprob var_dep var_explicatives, select(variables_explicatives de l'équation de sélection) /*si la variable censurée est dichotomique*/

Cartographie des indicateurs

Pour faire des cartes dans stata, il faut d'abord aller au site http://www.diva-gis.org/gdata et télécharger les données SIG de votre pays ou atres pays

Convertir les données sharpe en données stata par la commande shp2dta

shp2dta using XXX_adm1, database(region) coordinates(map) genid(id) gencentroids(center)

Pour faire la cartographie on utilise la commande :

spmap tx_pauvreté using map.dta, id(id) label(data("region.dta") label(VARNAME_1) xcoord(x_center) ycoord(y_center)) fcolor(Reds) title("Taux de pauvreté par région")

Ajout de nouveaux modules Stata

Pour installer des application récemment développées on utilse la commande ssc install.

- **ssc install logout** /*permet d'exporter des tableaux au format excel, word ou texte*/

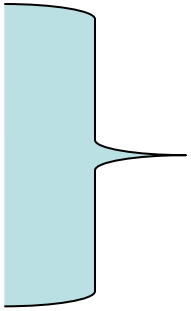
- **ssc install mmerge** /*permet de fusionner des bases sans passer par le tri*/

- **ssc install sumdist** /*permet d'avoir les statistiques détaillées de la distribution d'une variable quantitaitive selon les percentiles (distribution des dépenses)*/

-**ssc install spmap**

-**ssc install shp2dta**

-**ssc install mif2dta**



/*consiste à installer les commandes de la cartographie*/

Références Bibliographiques

-Aude Vescovo, IRD Afristat
«www.afristat.org/contenu/pdf/cera/Cours%20(A.%20Vescovo).pdf »

-Kangni KPODAR « Manuel d'initiation à Stata (version 8) :
«http://128.118.178.162/eps/prog/papers/0501/0501107.pdf»

-Estelle Ouellet : guide d'économétrie appliquée avec Stata :
« sceco.umontreal.ca/fileadmin/.../FAS/.../GuideEconometrieStata.pdf »

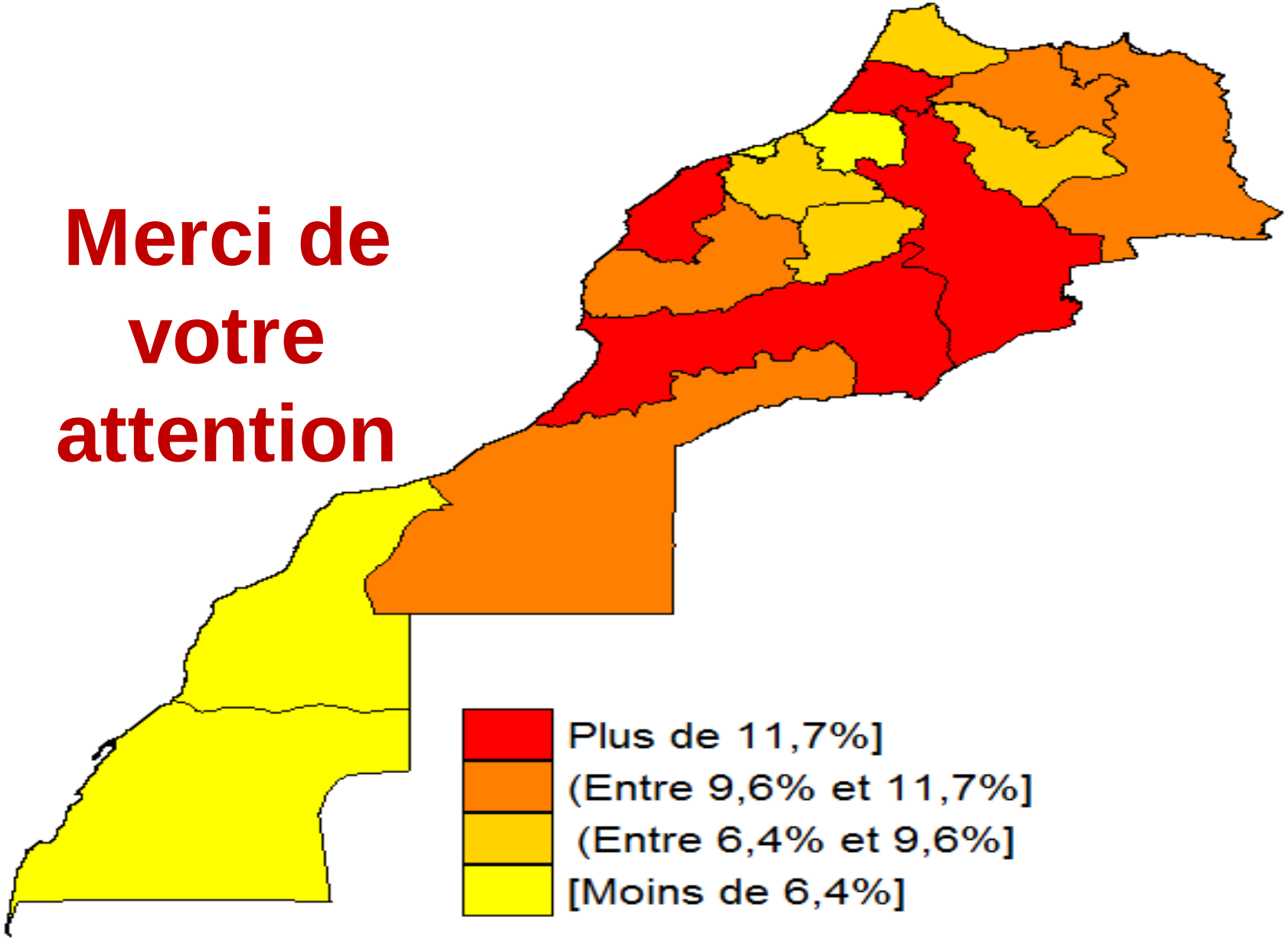
- Nicolas Couderc : Econométrie appliquée avec Stata :
«epi.univ-paris1.fr/.../com.univ.collaboratif.utils.LectureFichiergw?... »

-Olivier Cadot : Stata pour les nuls
« www.hec.unil.ch/ocadot/SECODEV_2008/Tools/Stata_nuls.pdf »

-Oscar Torres-Reyna : « Geeting Started in R~Stata, Notes on Exploring Data »
« dss.princeton.edu/training/RStata.pdf »

- http://www.ats.ucla.edu/stat/stata/library/GraphExamples/

Carte de la pauvreté régionale : 2007



Edité par le logiciel Stata

Merci de votre attention