**LECTURE 2**

**DATA OPERATIONS IN STATA, EXPLORATORY ANALYSIS OF LONGITUDINAL DATA, REGRESSION ANALYSIS WITH NONLINEAR SPECIFICATION**

In this lesson we will address some examples of constructing a database derived from multiple initial databases. There are two basic operations that can be done to combine different databases: we can “add observations” to an initial database, by deriving these additional observations from another database that contains the same variables as the starting data, but which refer to different subjects. In STATA this operation is called “append.” Or we can “add variables” to an initial database by exploiting variables that refer to the same subjects as the starting database, but are contained in a separate dataset. This operation is called “merge.”

We will also see that data can sometimes have a longitudinal structure, if, for example, we are able to observe the same subjects at different moments in time. We will then do some simple exploratory analysis of data with a longitudinal structure, testing some hypotheses.

We will then see how we can use the OLS regression method to model nonlinear relationships between the dependent variable and the independent variables. Finally, we will see how this changes the interpretation of the marginal effect of the regression.

The example in this lesson is based on the analysis of a database with a structure similar to the INPS administrative data referring to the population of employees in the private sector. These data are derived from UNIEMENS reports, which are compiled by employers and report the wages of each worker along with some of his or her characteristics (age, sex, etc.). These data have a matched employer-employee structure, meaning that for each worker we know which enterprise he or she works for. The data also have a longitudinal structure because UNIEMENS reports are sent annually.

In this lesson's example we will analyze fictitious data, but the original data can be studied using the Visitinps Scholar program, which allows us to analyze UNIEMENS reports in STATA format from 1974 through the most recent years.

A detailed description of each analysis performed in this lecture is provided below (in Italian). There are also comments in the .do and .log files of the lecture, which are provided in English.

1. **OPERAZIONI PRELIMINARI**

1) Creiamo nella memoria del computer una cartella di lavoro relativa a questa lezione

* Creiamo una cartella nella memoria del computer (per esempio “C:\Desktop\lezione\_2”). Questa sarà la cartella di lavoro in cui salveremo tutti i file relativi a questa lezione.
* Copiamo i file contenenti i database che analizzeremo in questa lezione nella cartella di lavoro che abbiamo creato

2) Apriamo il software STATA e iniziamo a scrivere un do file

* Scriviamo nel do file ed eseguiamo il comando **cd “C:\Desktop\lezione\_2”** (l’indirizzo contenuto nelle virgolette dipende da dove avete salvato la cartella di lavoro della lezione nel vostro computer). Questo comando indica a STATA che tutti i file che vogliamo utilizzare e tutti i file che vogliamo salvare si trovano in questa cartella.
* Eseguiamo il comando **cap log close** e successivamente il comando **log using lecture2.log, replace** Questo comando crea un file chiamato lecture3.log nella nostra cartella di lavoro. Questo file sarà un file di testo che riproduce tutti i risultati dei nostri comandi (registra tutto ciò che appare nella finestra dei risultati di STATA).

1. **ESERCIZIO: TEST DI IPOTESI E REGRESSIONE LINEARE UNIVARIATA**

a) Importiamo da excel il file “dati\_esercizio.xls” col comando **import excel "dati\_esercizio.xls", sheet("Sheet1") firstrow clear**. È anche possibile eseguire questa operazione utilizzando il menu a tendina di STATA, selezionando “import -> excel spreadsheet”

b) Elenchiamo le variabili con **des** e aggiungiamo dei label se necessario: **label var edu\_father "=1 se padre con diploma o laurea"**

c) Proviamo ad effettuare un t-test per verificare varie ipotesi sulle differenze medie di salario tra uomini e donne: **ttest salario\_mensile, by(donna)**. Ripetiamo il test per altre ipotesi

Immagine che contiene tavolo

Descrizione generata automaticamente

d) Proviamo a misurare l’effetto che ha l’invecchiamento di un anno sul salario: **reg salario\_mensile age**

**Immagine che contiene tavolo

Descrizione generata automaticamente**

1. **ANALISI ESPOLORATIVA DEL DATABASE RAPPORTI LAVORO E APPEND**

a) Apriamo il database “rapporti\_lavoro\_2000.dta” col comando “**use rapporti\_lavoro\_2000.dta, clear**”. Usiamo il comando “**des**” per controllare quali variabili contiene il database. Usiamo il comando “**tab anno**” per controllare a quali anni si riferiscono questi dati.

b) Adesso apriamo il database “rapporti\_lavoro\_2001.dta” col comando “**use rapporti\_lavoro\_2001.dta, clear**” e ripetiamo la stessa analisi. I due dataset contengono le stesse variabili, ma il primo si riferisce al 2000, il secondo si riferisce al 2001 (nel materiale della lezione c’è anche un terzo dataset analogo riferito al 1999).

c) Adesso aggiungiamo osservazioni al dataset “rapporti\_lavoro\_2001.dta” che è attualmente aperto nella memoria di STATA. Vogliamo in particolare aggiungere le osservazioni relative al 2000 e il 1999 contenute nei database “rapporti\_lavoro\_2000.dta” e “rapporti\_lavoro\_1999.dta” al database relativo al solo 2001, creando in questo modo dei dati con una struttura longitudinale. Eseguiamo i comandi “**append using rapporti\_lavoro\_2000.dta**” e “**append using rapporti\_lavoro\_1999.dta**”

d) Eseguiamo il comando “**des**” e successivamente il comando “**tab anno**”. Il numero di osservazioni e il numero di anni è aumentato? Adesso salviamo il database longitudinale che abbiamo creato col nome “panel\_rl.dta”. Eseguiamo il comando: **save panel\_rl.dta, replace**

e) Vogliamo testare se la percentuale di lavoratori a contratto determinato era maggiore nel 2001 rispetto al 2000. Per questa analisi le osservazioni relative al 1999 sono superflue, quindi elimino le osservazioni che si riferiscono al 1999 col comando “**drop if anno==1999**”. Proviamo a eseguire il comando “**tab anno**”. Come è cambiato il database?

Eseguiamo il comando “**tab anno, sum(tempo\_d)**” per calcolare la media in ciascun anno. Come interpreto questa media? Provate a eseguire i comandi “**tab tempo\_d if anno==2000**” e “**tab tempo\_d if anno==2001**”.

Le differenze nella proporzione di lavoratori a tempo determinato che osservo in anni diversi sono statisticamente significative? Per rispondere a questa domanda devo eseguire un test (formalmente un test di uguaglianza delle proporzioni in due campioni, dove i due campioni sono quello relativo al 2001 e al 2000). Eseguiamo il comando: **prtest tempo\_d, by(anno)**

1. **AGGIUNGERE VARIABILI COL COMANDO MERGE E ANALISI DI REGRESSIONE OLS CON SPECIFICAZIONI NON-LINEARI**

a) Apriamo il database “anagrafica\_soggetti.dta” con il comando “**use anagrafica\_soggetti.dta, clear**”. Utilizziamo il comando “**des**”. Questo database contiene informazioni aggiuntive sui soggetti presenti nel database lavoratori, in particolare due variabili (anno di nascita e sesso). Possiamo aggiungere queste variabili al database “rapporti\_lavoro\_2001.dta”? Per farlo dobbiamo utilizzare la chiave di aggancio “id\_soggetto” che identifica gli stessi individui nel database “rapporti\_lavoro\_2001.dta” e nel database “anagrafica\_soggetti.dta”

b) Apriamo il database “rapporti\_lavoro\_2001.dta” (**use rapporti\_lavoro\_2001.dta, clear**) e usiamo il comando “**des**”. Eseguiamo adesso il comando **merge 1:1 id\_soggietto using anagrafica\_soggetti.dta**. Col comando “**des**” notiamo che le variabili contenute in “anagrafica\_soggetti.dta” sono state aggiunte al database in uso. C’è anche una nuova variabile “\_m” che possiamo descrivere con **tab \_m**.

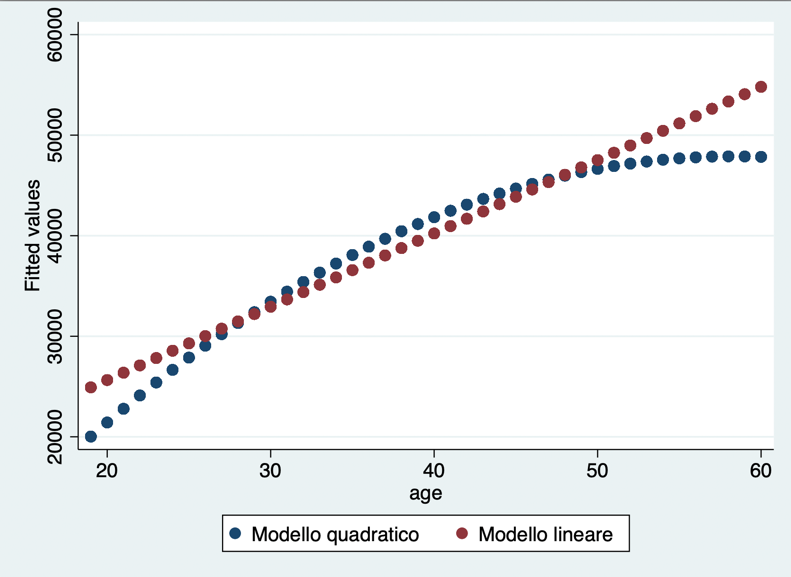
La variabile \_m indica le osservazioni per le quali c’era un id\_soggetto comune in “anagrafica\_soggetti.dta” e in “rapporti\_lavoro\_2001.dta” (*matched*), quelle per le quali un id\_soggetto in “rapporti\_lavoro\_2001.dta” non era presente in “anagrafica\_soggetti.dta” (*master*) e quelle in cui un id\_soggetto in “anagrafica\_soggetti.dta” non era presente in “rapporti\_lavoro\_2001.dta” (*using*). Teniamo solo le ossevazioni *matched* col comando **keep if \_m==3** e eliminiamo la variabile \_m con **drop \_m**

c) Generiamo la variabile “age” come differenza tra anno di nascita e anno corrente: **gen age = anno - anno\_nascita**. Eseguiamo poi la regressione: **reg retrib03 age**. Calcoliamo i valori predetti di “retrib03” per ciascun livello di “age” col comando **predict y\_hat** che genera una nuova variabile chiamata “y\_hat” coi valori predetti. Costruiamo uno scatterplot dei valori predetti per ciascun livello di età col comando **scatter y\_hat age**

d) Generiamo la variabile età al quadrato **gen age2=age^2** e eseguiamo la regressione **reg retrib03 age age2**. Il coefficiente associato a age2 è statisticamente significativo? Se sì, allora la relazione tra salari e età è descritta più accuratamente da una funzione non-lineare (quadratica in questo caso), rispetto alla funzione lineare che abbiamo calcolato al punto (c). Proviamo a disegnare la relazione tra salari ed età calcolata attraverso il polinomio quadratico col comando **predict y\_hat2** (che genera i valori del salario predetti dalla regressione per ciascun livello di età) e successivamente col comando

**twoway (scatter y\_hat2 age) (scatter y\_hat age), ///**

**legend(order(1 "Modello quadratico" 2 "Modello lineare"))**



e) Qual è l’effetto di invecchiare di un anno sui salari nel modello quadratico? L’effetto dipende dall’età. Possiamo calcolare quest’effetto a 39 anni col comando **ttest y\_hat2 if age==39|age==40, by(age)**. La differenza nelle medie tra y\_hat2 a 40 anni e y\_hat2 a 39 anni corrisponde all’effetto predetto di invecchiare un anno a 39 anni. Se calcoliamo questa differenza a 29 anni otteniamo un numero diverso: **ttest y\_hat2 if age==29|age==30, by(age)**. Nel modello lineare l’effetto dell’invecchiamento è costante per qualsiasi livello di età.

f) Un modello quadratico predice meglio la relazione salario età? Sappiamo di sì in base al modello di regressione. Possiamo indagare graficamente se questo è vero. Calcoliamo la media dei salari effettivi e delle predizioni per ciascuna età. Il seguente comando genera un nuovo database con una osservazione per età, in cui le variabili del nuovo database contengono le medie di retrib03 y\_hat y\_hat2 per età:

**collapse (mean) retrib03 y\_hat y\_hat2, by(age)**

Adesso disegnamo il livello della media di retrib03 y\_hat y\_hat2 per ciascuna età:

**twoway (connected retrib03 age, msize(vsmall)) ///**

**(connected y\_hat age, msize(vsmall)) ///**

**(connected y\_hat2 age, msize(vsmall))**

La linea blu è la media effettiva, quella rossa è la predizione lineare, quella verde è la predizione quadratica. Qual è la migliore?

