

Dati multi-fonte e analisi territoriali

Prof. Marco Tosi

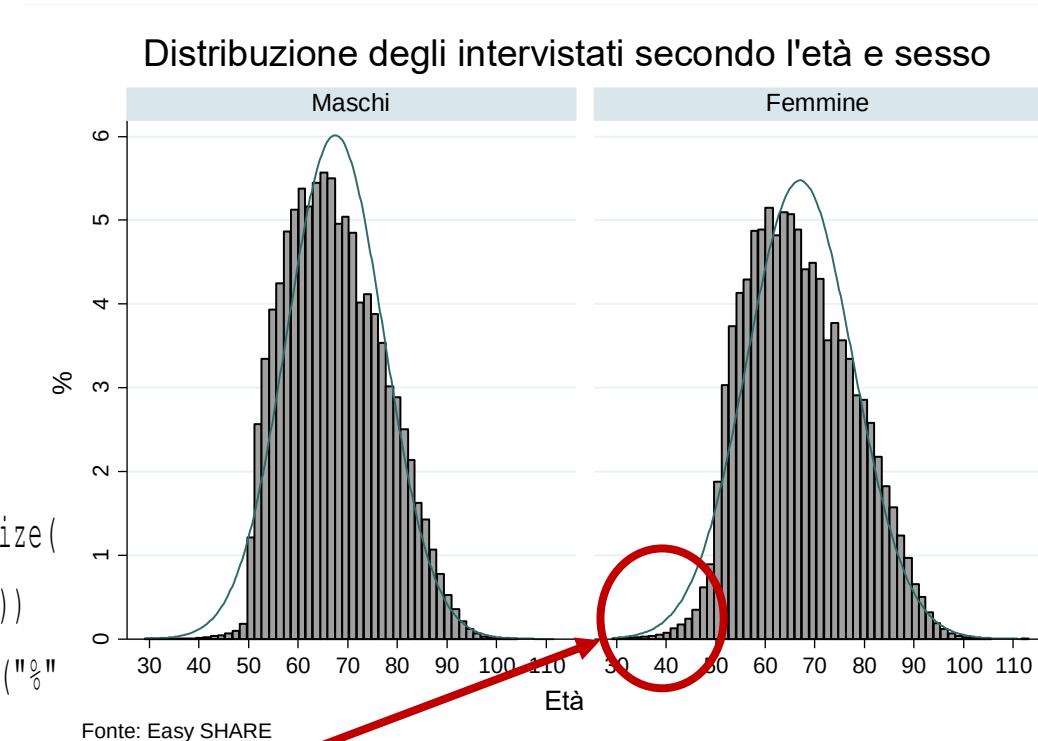
Dipartimento di Scienze Statistiche

Conoscere i dati:

(1) Popolazione di riferimento

- SHARE intervista i 50+ e i loro partner (che vivono assieme). Non sorprende che ci siano minori di 50.
 - Tuttavia: a quale popolazione vogliamo inferire i risultati?
 - Possibili implicazioni per lo studio delle differenze di genere?

```
. hist age if age>29, percent normal by(sesso, graphregion(c(white))) note("Fonte: Easy SHARE", size(> small)) legend(off) title("Distribuzione degli intervistati secondo l'età e sesso", color(black))> fcolor(gs10) lc(black) xlabel(30 40 50 60 70 80 90 100 110, format(%9.0f)) xtitle("Età") ytitle("%")> ) ylabel(0(1)6)
```



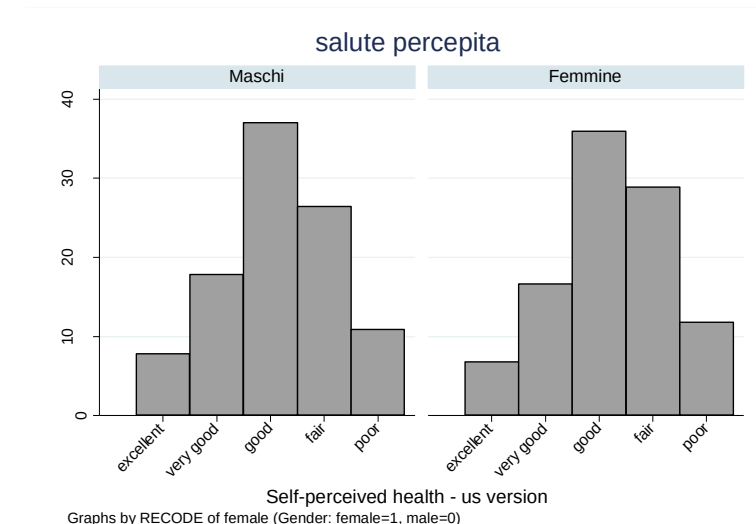
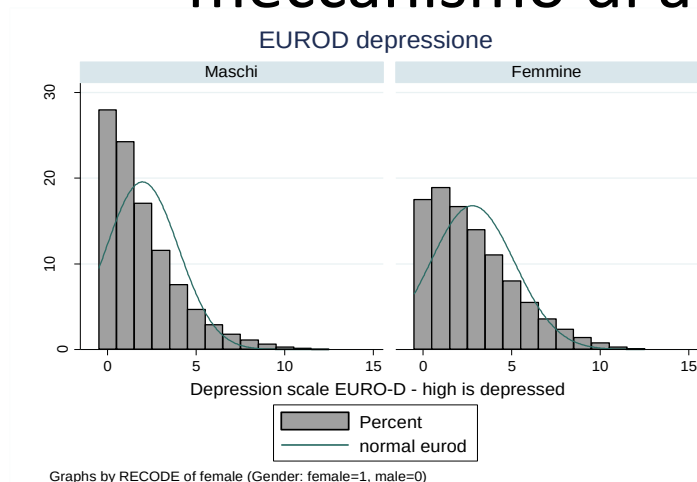
Conoscere i dati:

(2) misure

- Misure di salute percepita e auto-riportata dai rispondenti.

```
. hist sphus if sphus>0, dis percent by( Sesso, graphregion(c(white)) title("salute percepita")) x  
> label(1 "excellent" 2 "very good" 3 "good" 4 "fair" 5 "poor", angle(45)) fcolor(gs10) lcolor(bla  
> ck)
```

- Misure / Scale multi-item (eurod: scala su depressione). Distribuzione tipicamente skewed (in alcuni casi può essere zero-inflated, ossia in meccanismo di avere 0 sintomi è diverso dal N).



Conoscere i dati

(2) misure complesse

- Scale multi-item. Eurod Depressione è calcolata su 12 item 0-1 (alpha di Cronbach per la consistenza interna della scala)

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_{i=1}^k \sigma_{Y_i}^2}{\sigma_X^2} \right)$$

Test scale = mean(unstandardized items)							
Item	Obs	Sign	item-test corr.	item-rest corr.	interitem cov.	alpha	Label
euro_1	312494	+	0.6474	0.4943	.0223272	0.6726	RECODE of euro1 (Depression(part of EURO-D))
euro_2	311773	+	0.4205	0.2696	.0271578	0.7077	RECODE of euro2 (Pessimism (part of EURO-D))
euro_3	311825	+	0.4624	0.3671	.0274327	0.6986	RECODE of euro3 (Suicidality (part of EURO-D))
euro_4	311747	+	0.3502	0.2387	.0285349	0.7099	RECODE of euro4 (Guilt (part of EURO-D))
euro_5	312749	+	0.5422	0.3691	.02453	0.6949	RECODE of euro5 (Sleep (part of EURO-D))
euro_6	312322	+	0.4934	0.3860	.0266981	0.6948	RECODE of euro6 (Interest (part of EURO-D))
euro_7	312403	+	0.5140	0.3476	.025231	0.6978	RECODE of euro7 (Irritability (part of EURO-D))
euro_8	313031	+	0.4358	0.3250	.0274756	0.7013	RECODE of euro8 (Appetite (part of EURO-D))
euro_9	312395	+	0.5940	0.4313	.023491	0.6842	RECODE of euro9 (Fatigue (part of EURO-D))
euro_10	311855	+	0.4924	0.3468	.0259854	0.6975	RECODE of euro10 (Concentration (part of EURO-D))
euro_11	312112	+	0.3986	0.2632	.0276487	0.7078	RECODE of euro11 (Enjoyment (part of EURO-D))
euro_12	312422	+	0.5333	0.3766	.0249745	0.6930	RECODE of euro12 (Tearfulness (part of EURO-D))
Test scale					.0259574	0.7152	mean(unstandardized items)

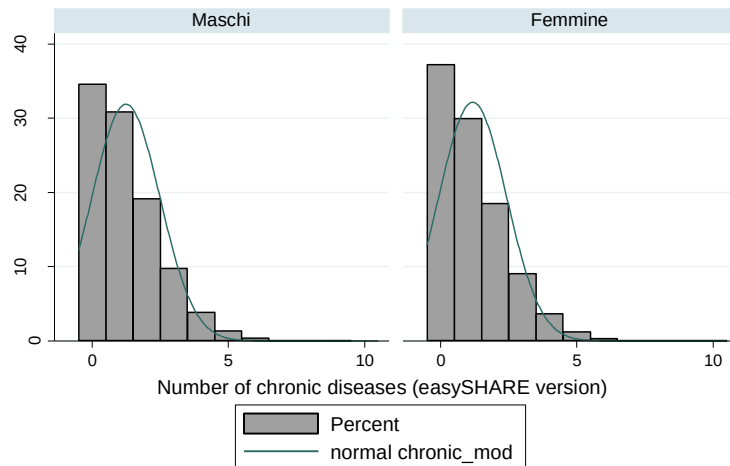
Accettabile

Conoscere i dati:

(2) altre misure di salute

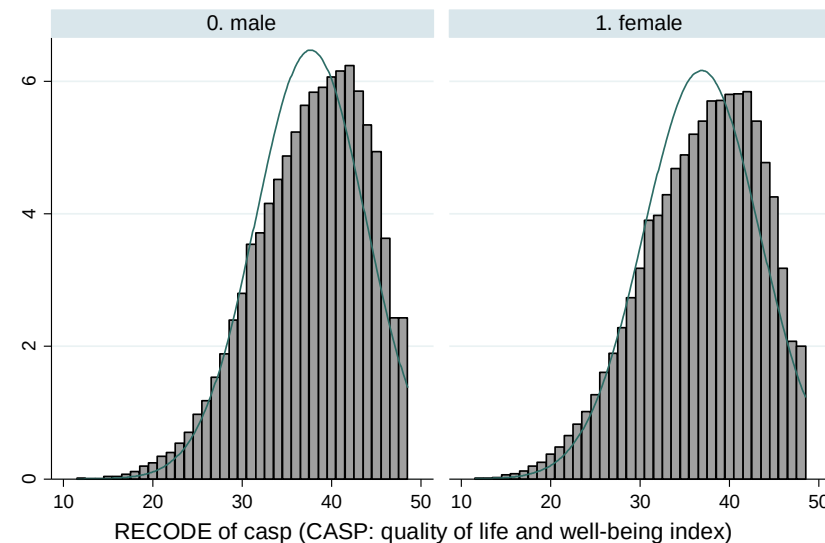
- Altre misure di salute presenti nei dati SHARE: numero di malattie croniche, CASP qualità della vita, forza della stretta della mano.

N of chronic diseases



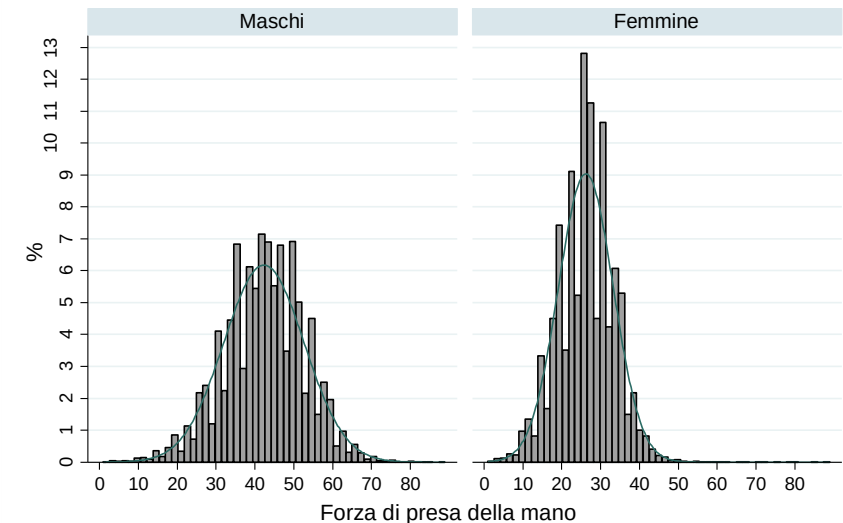
Graphs by RECODE of female (Gender: female=1, male=0)

quality of life (CASP)



Graphs by Gender: female=1, male=0

Distribuzione degli intervistati secondo il test fisico



Fonte: Easy SHARE

Obiettivo

Esempio: condizioni di salute dei pensionati

- Utilizziamo la forza della stretta della mano come variabile dipendente.
- Utilizziamo lo status occupazionale come variabile indipendente.
- Una associazione negativa, *ma... stiamo omettendo la variabile età*

```
. regress maxgrip Retired
```

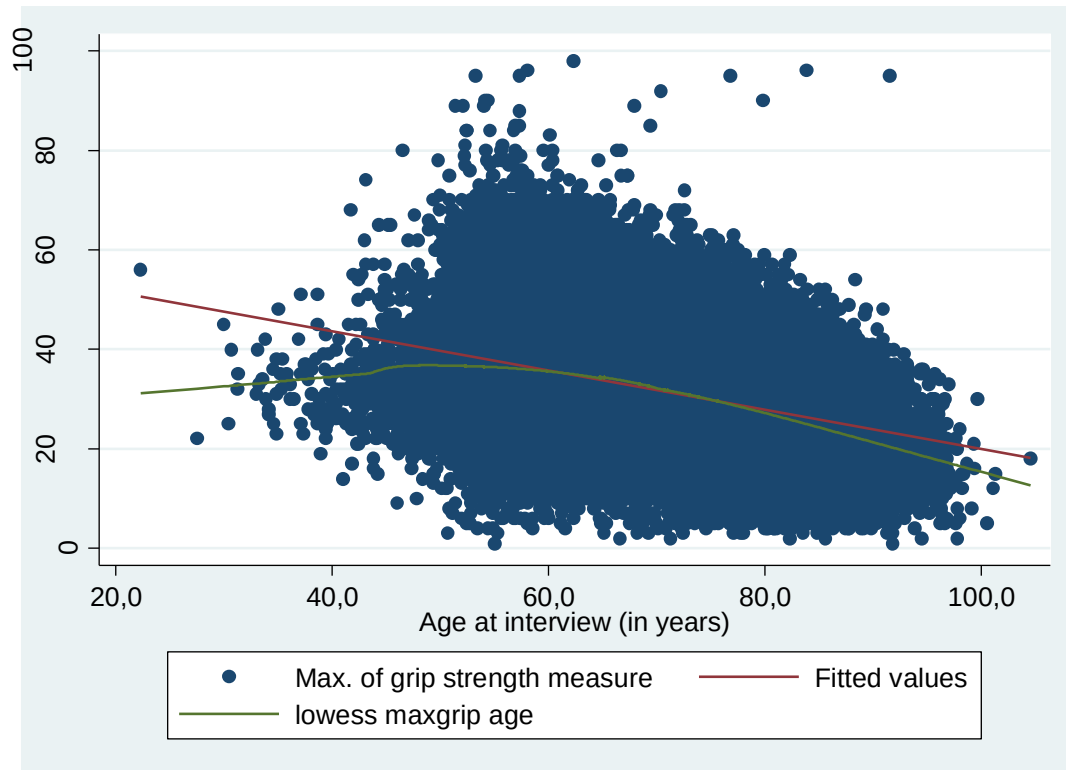
Source	SS	df	MS	Number of obs = 345381		
Model	777273.946	1	777273.946	F(1,345379) = 5710.43		
Residual	47011199.2345379	136.114817		Prob > F = 0.0000		
Total	47788473.2345380	138.364912		R-squared = 0.0163		
				Adj R-squared = 0.0163		
				Root MSE = 11.667		

maxgrip	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
Retired	-3.034915	.0401617	-75.57	0.000	-3.113631	-2.956199
_cons	35.0911	.0304615	1151.98	0.000	35.0314	35.15081

Scelta dei modelli e diagnostica

(1) Linearità

Regress maxgrip age



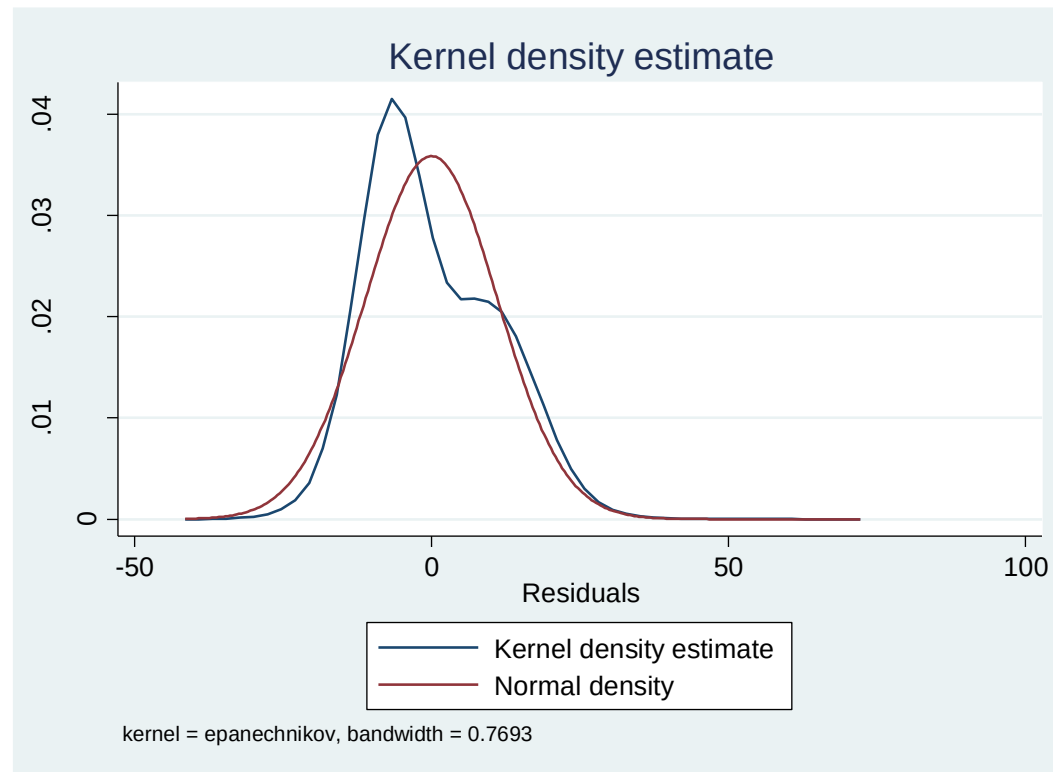
Nel caso di una associazione bivariata è abbastanza semplice testare la linearità.

(2) Normalità dei residui

Regress maxgrip age

predict r3,resid

kdensity r3,normal

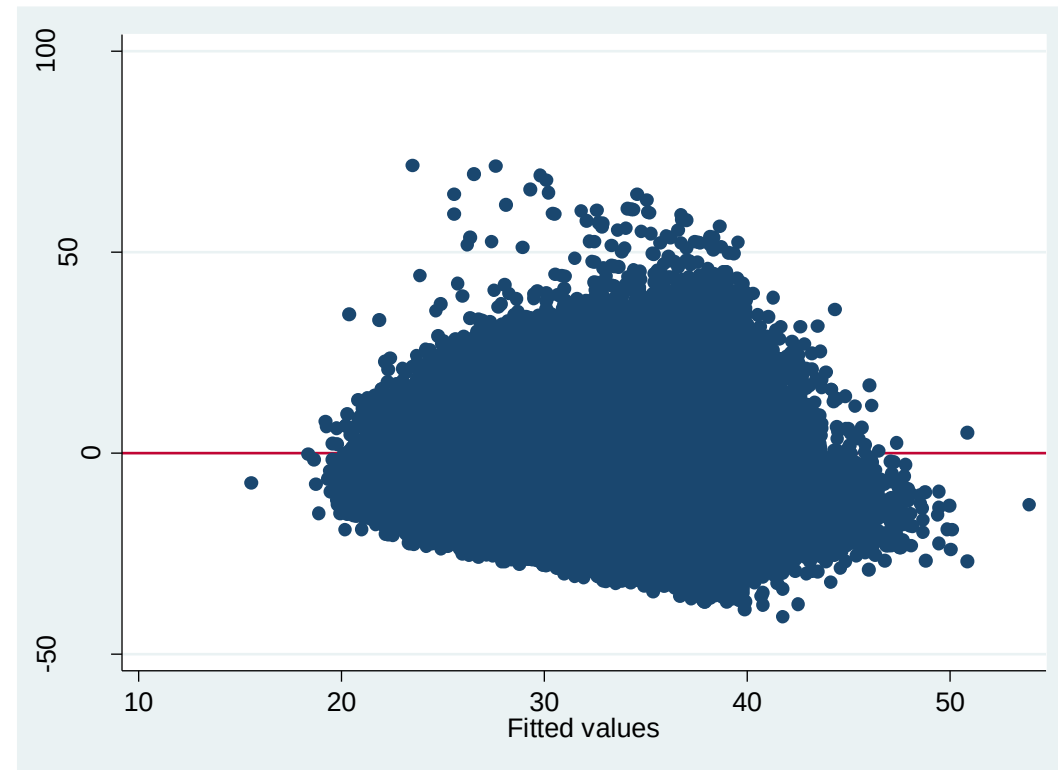


(3) Omoschedasticità

- Possiamo graficamente mettere in relazione i valori predetti dal modello con i residui stimati per controllare i pattern.

White Test per testare H_0 : Omoschedasticità

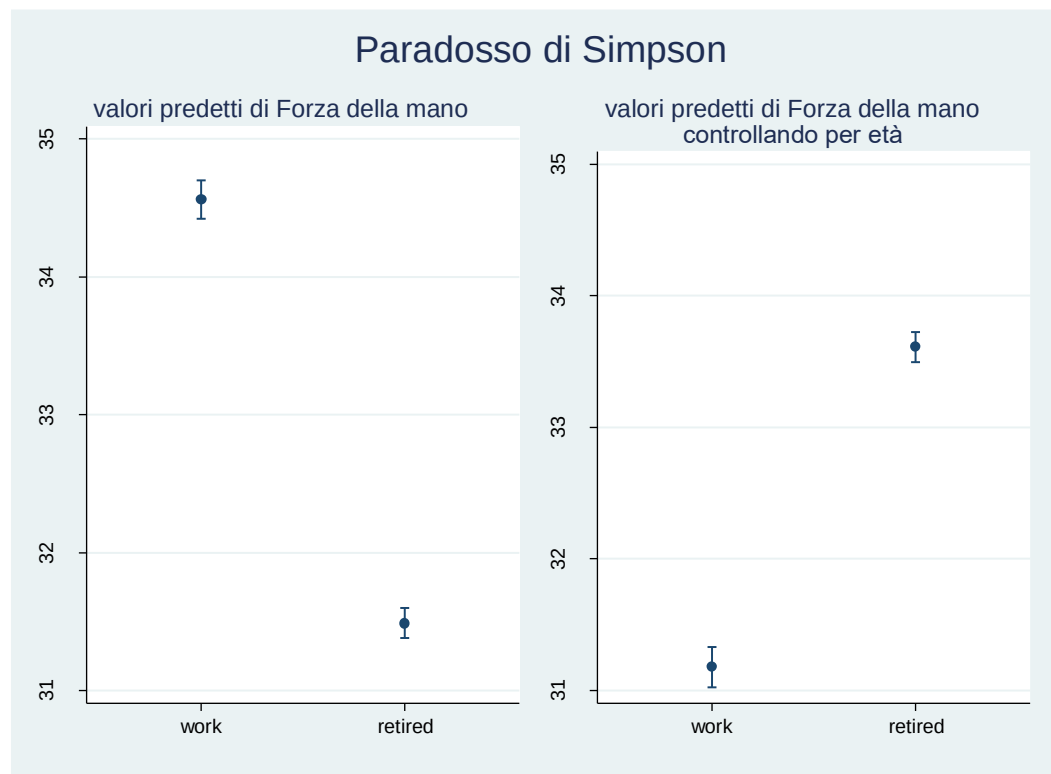
Source	chi2	df	p
Heteroskedasticity	6027.12	2	0.0000



Specificazione del Modello

Variabili omesse

Variabili non-osservate che introducono una distorsione delle stime.



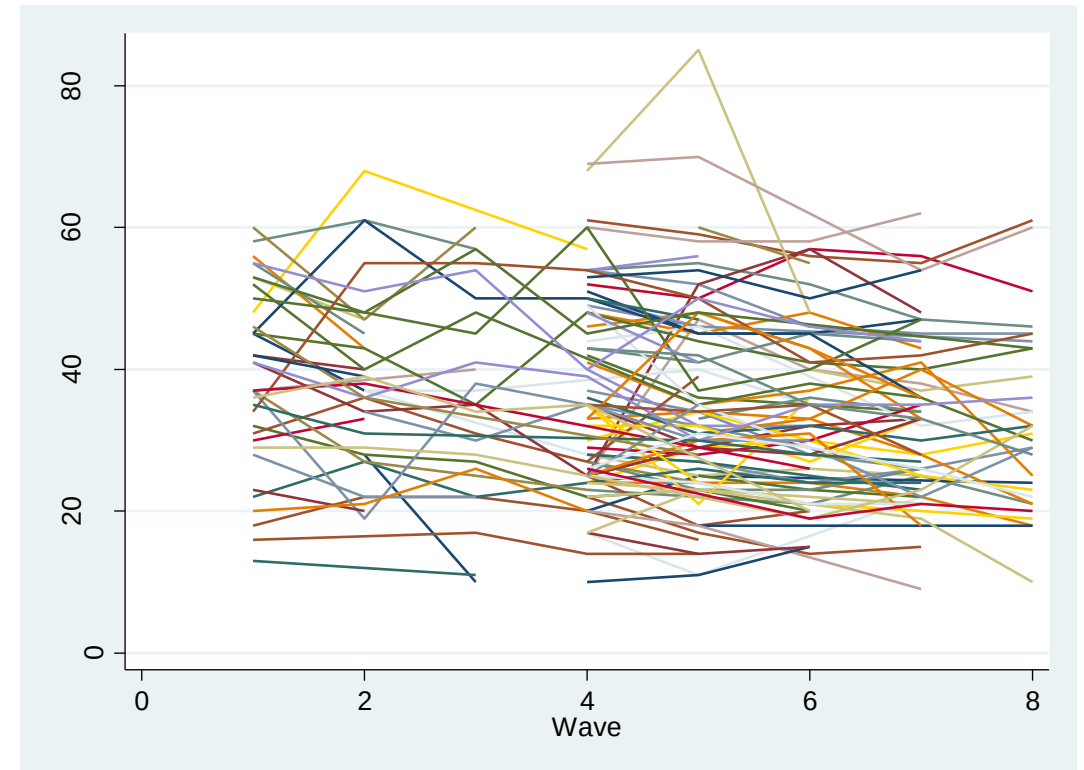
Specificazione del Modello:

Struttura longitudinale/ Panel dei Dati

```
xtset ID wave
```

```
xtline maxgrip if ID<500, overlay legend(off)
```

L'andamento nel tempo per i primi 500 individui mostra un andamento abbastanza stabile (e il re-sampling!)



Specificazione del modello

Struttura longitudinale

- I pattern di osservazioni nel tempo ci danno indicazioni sulla struttura longitudinale dei dati

```
. sort mergeid wave
. gen xx=_n if mergeid!=mergeid[_n-1]
(271985 missing values generated)
. bys mergeid: egen ID=min(xx)
. drop xx
```

	mergeid	ID	hhid	coupleid	wave	int_year
1	AT-000327-01	1	AT-000327-A	AT-000327-01-02	1	2004
2	AT-000327-01	1	AT-000327-A	AT-000327-01-02	2	2007
3	AT-000327-02	3	AT-000327-A	AT-000327-01-02	1	2004
4	AT-000327-02	3	AT-000327-A	AT-000327-01-02	2	2007
5	AT-000674-01	5	AT-000674-A		4	2011
6	AT-000674-01	5	AT-000674-A		5	2013
7	AT-000674-01	5	AT-000674-A		6	2015
8	AT-001215-01	8	AT-001215-A		4	2011
9	AT-001215-01	8	AT-001215-A		5	2013
10	AT-001215-01	8	AT-001215-A		6	2015
11	AT-001215-01	8	AT-001215-A		7	2017
12	AT-001215-01	8	AT-001215-A		8	2020

Distribution of T_i: min 5% 25% 50% 75% 95% max
 1 1 1 2 4 7 8

Freq.	Percent	Cum.	Pattern
9706	6.93	6.93	1.
9311	6.64	13.57	11
8784	6.27	19.84	...11111
7982	5.70	25.54	...1....
6921	4.94	30.48	1111
6545	4.67	35.15	1.....
5921	4.23	39.37	...1111.
5822	4.15	43.53	1...
4906	3.50	47.03	111.
74227	52.97	100.00	(other patterns)
140125	100.00		XXXXXXXX

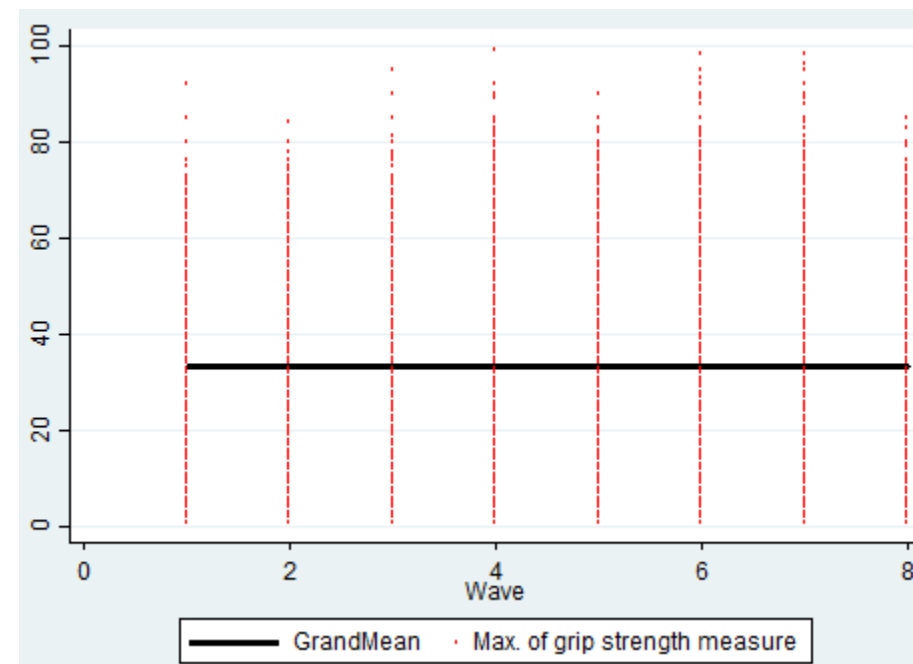
Specificazione del Modello:

Struttura longitudinale/ Panel dei Dati

xtset ID wave

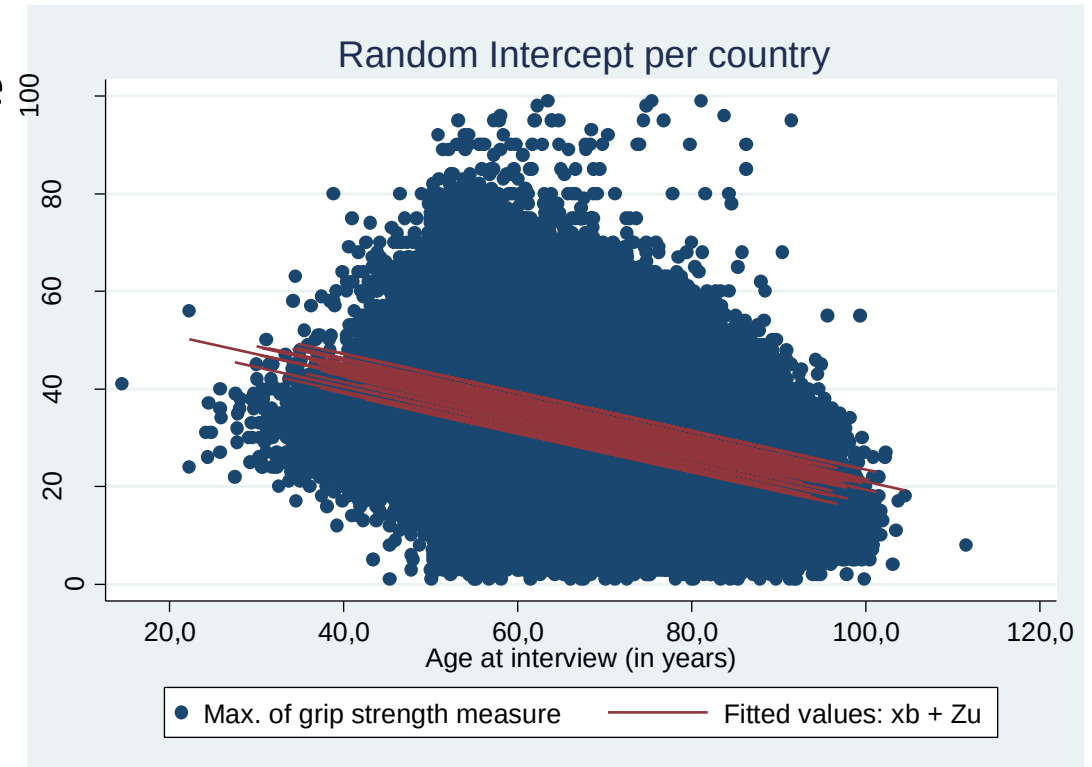
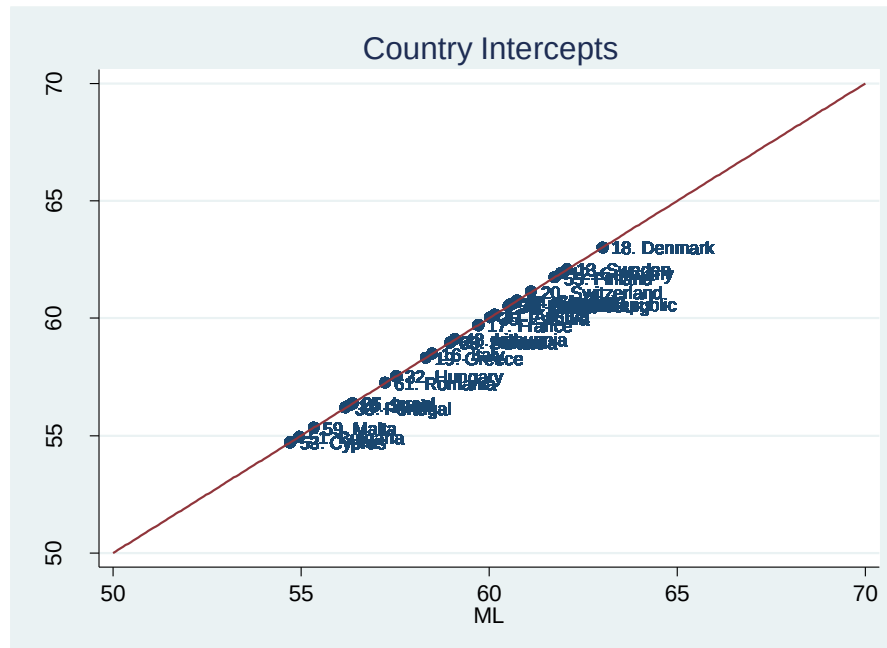
Xtmixed maxgrip

Possiamo tenere in considerazione della struttura longitudinale/ panel dei dati con un modello ad intercetta random in cui abbiamo una intercetta per ogni individuo (qui si mostra solo la media campionaria) e il termine di errore è libero di variare *tra* gli individui



Specificazione del Modello: Struttura gerarchica

Non abbiamo solo gli individui nel tempo ma le loro osservazioni sono (e possono dipendere da) all'interno di famiglie, regioni e paesi.



Regressione lineare OLS

```
. regress maxgrip age Retired i.country if wave ==7
```

Source	SS	df	MS	Number of obs = 68711		
Model	1316623.77	28	47022.2774	F(28, 68682) = 412.79		
Residual	7823705.24	68682	113.912018	Prob > F = 0.0000		
Total	9140329.01	68710	133.027638	R-squared = 0.1440		
				Adj R-squared = 0.1437		
				Root MSE = 10.673		

maxgrip	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.4562451	.0054332	-83.97	0.000	-.4668942	-.445596
Retired	2.13903	.1076294	19.87	0.000	1.928076	2.349983
country						
12. Germany	1.927765	.2719233	7.09	0.000	1.394796	2.460735
13. Sweden	2.059683	.2817019	7.31	0.000	1.507547	2.611818
15. Spain	-3.391408	.2644506	-12.82	0.000	-3.909731	-2.873086
16. Italy	-1.36306	.2675413	-5.09	0.000	-1.88744	-.8386791
17. France	-.4306083	.2804013	-1.54	0.125	-.9801944	.1189779
18. Denmark	3.163787	.2803875	11.28	0.000	2.614228	3.713346
19. Greece	-1.446247	.2904643	-4.98	0.000	-2.015557	-.8769376
20. Switzerland	1.345103	.3037176	4.43	0.000	.749817	1.940389
23. Belgium	.6353744	.2587903	2.46	0.014	.1281458	1.142603
25. Israel	-3.309676	.3348795	-9.88	0.000	-3.966039	-2.653312
28. Czech Republic	.2672323	.2676994	1.00	0.318	-.2574582	.7919228
29. Poland	-1.068751	.2604108	-4.10	0.000	-1.579156	-.5583468
31. Luxembourg	.6217767	.3890066	1.60	0.110	-.1406757	1.384229
32. Hungary	-2.862352	.3516646	-8.14	0.000	-3.551615	-2.17309
33. Portugal	-3.825569	.3910384	-9.78	0.000	-4.592004	-3.059134
34. Slovenia	.4688318	.2764983	1.70	0.090	-.0731044	1.010768
35. Estonia	.1702875	.2613348	0.65	0.515	-.3419282	.6825033
47. Croatia	.565671	.3048309	1.86	0.064	-.0317971	1.163139
48. Lithuania	-.7817736	.3170174	-2.47	0.014	-1.403127	-.16042
51. Bulgaria	-5.223263	.3225297	-16.19	0.000	-5.855421	-4.591106
53. Cyprus	-5.396848	.3776722	-14.29	0.000	-6.137085	-4.656611
55. Finland	1.735508	.3164554	5.48	0.000	1.115256	2.35576
57. Latvia	.5200704	.3369438	1.54	0.123	-.1403389	1.18048
59. Malta	-4.316746	.3690137	-11.70	0.000	-5.040013	-3.59348
61. Romania	-3.058842	.3135005	-9.76	0.000	-3.673303	-2.444382
63. Slovakia	-1.125472	.3260534	-3.45	0.001	-1.764536	-.4864079
_cons	62.80957	.3939348	159.44	0.000	62.03746	63.58168

Take-home message

- Conoscere le Fonti Dati (Survey of Health, Ageing, and Retirement in Europe – SHARE).
 - Qual'è la popolazione di riferimento?
 - Quale scopo ha l'indagine?
- Conoscere la Basi Dati.
 - Chi effettivamente è stato intervistato?
 - Qual è la struttura dei dati?
 - Quali contenuti/ variabili?
- Misurazione
 - Quali concetti (forza della mano) e come vengono misurati (dinamometro)
- Scelta di un modello e Diagnostica.
 - Qual è la natura della variabile dipendente?
 - Quali assunzioni? (es. OLS: Linearità, Normalità, Omoschedasticità, indipendenza degli errori, + corretta misurazione, specificazione del modello, casi influenti e collinearità)

Riassunto

Cose da ricordarsi per il corso

- Struttura longitudinale/panel dei dati:

W1	W2	Wave 3	Wave 4	Wave 5	Wave 6	Wave 7	Wave 8	SCS1	SCS2
2004	2006/07	2008/09	2011	2013	2015	2017	2019/2020	2020	2021

- Panel non bilanciato (pattern più frequente: 1 obs. wave 7):

30	30	Ireland	-	2007	2009/10 /11	-	-	-	-	2019/2020	2020	2021
----	----	---------	---	------	----------------	---	---	---	---	-----------	------	------

- Popolazione di riferimento: residenti con indirizzo valido (no ospedalizzazioni) di più di 50 anni e i loro partner (anche se < 50).
- Scopo dell'indagine è rilevare le condizioni di salute (varie misure).

Riassunto

Cose da ricordarsi per il corso

- Wave 3 (e in parte anche la 7) è retrospettiva.

Table 3: Structure of Wave 7 Questionnaire

Respondent type	Interview sections		No. of interviews
Respondent participated in Wave 3	Regular Panel		13.959
	DN CH PH BR CF MH HC EP GS SP FT CO AS HO HH AC EX IV		(18%)
Respondent did not participate in Wave 3	SHARELIFE	Regular Panel	62.561
	RC RP RA CC RE WQ DQ FS HS RH GL	(condensed)	(82%)

- Wave 4: un ri-campionamento di ampie dimensioni.

Riassunto

Cose da ricordarsi per il corso

- Tipi di rispondenti:

Table 4: Types of Respondents and Variable Names

Type of respondent	Variable name in cv_r and technical variables module	Name of filter in w1 & w2 questionnaire	Name of filter in w4, w5, w6, w7, w8 questionnaire
Family respondent	<i>fam_resp</i>	MN006_ (FAMILY RESPONDENT)	MN006_NumFamR
Financial respondent	<i>fin_resp</i>	MN007_ (FINANCIAL RESPONDENT)	MN007_NumFinR
Household respondent	<i>hou_resp</i>	MN008_ (HOUSEHOLD RESPONDENT)	MN008_NumHHR

- Tipi di file:

Table 5: Who Answers What in the Regular Modules?

Module	Name	All respondents ¹	Household Resp.	Financial Resp.	Family Resp.	non-proxy
CV_R	Coverscreen	cv-respondent only				
DN	Demographics	x				
SN	Social Networks	x				x
CH	Children				x	

Riassunto

Cose da ricordarsi per il corso

- Dati multi-fonte e territoriali sono intrinsecamente dati con una struttura gerarchica, in cui abbiniamo dati con unità macro e micro.
- Un random slope di un modello multilevel può essere rappresentato con mappe:

Ferraretto & Vitali (2023).
Parental socioeconomic status and age at leaving home in Europe:
Exploring regional differences, *Populations Space and Place*

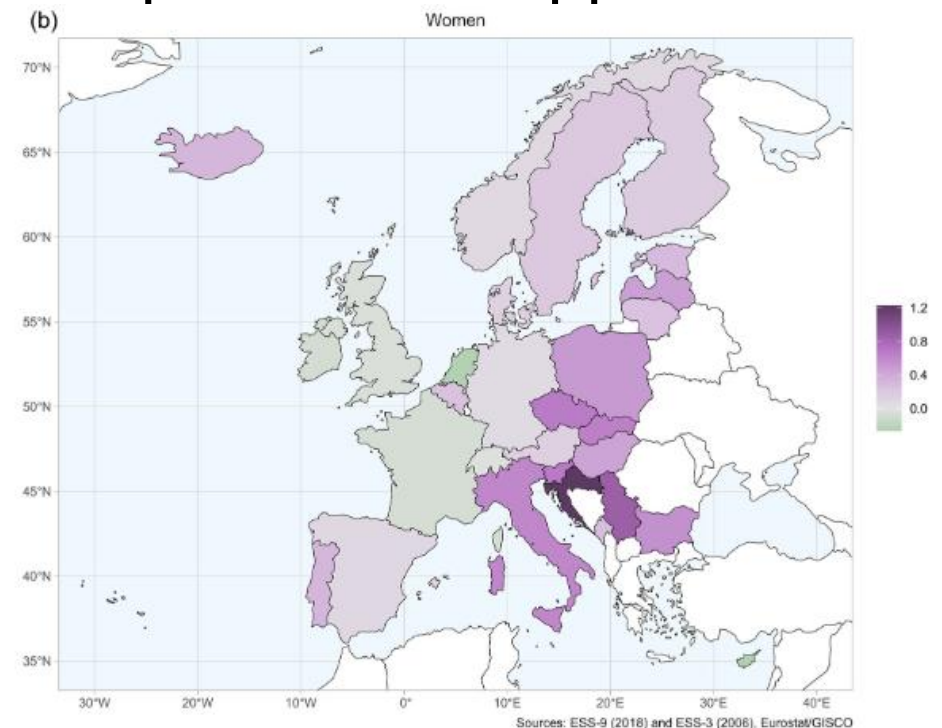


FIGURE 2 Random slope of parental SES on the age at leaving home, by country and gender. Estimates retrieved from M2, Table 5.

Riassunto

Cose da ricordarsi per il corso

- Benché i codici di STATA sono tutti disponibili su Moodle qualche comando di base va ricordato, ad esempio:
 - Come generare e ricodificare una variabile (gen, recode, replace)
 - Come settare i dati longitudinali/panel (xtset, xtdes, reshape)
 - Come riferirsi ad una osservazione immediatamente sopra o sotto ([_n-1])
 - Eliminare casi o variabili (drop, keep)
 - Grafici per vedere le distribuzioni (bar, hist, scatter)
 - Modelli di regressione (reg, posson, logit, probit, ologit, mlogit)
 - *E quando possiamo usare un modello piuttosto che un altro...*