Stata in the everyday life of a health economist

Pedro Pita Barros

NOVA School of Business & Economics
Healthcare Initiative Research

Stata and me: what I do?

- Research articles
- Books
- Data management
- Blog and other things

NOVA School of Business & Economics
Healthcare Initiative Research

Blog

- Need treatment of simple data, usually small samples
- Problem: they are in a variety of formats
- Solution: import from Excel files has become the easier solution – most data sources now have this possibility – Statistics Portugal, Pordata, AMECO / Eurostat, OECD Health Data
- Problem: formatting changes from time to time
- Solution: import to “my” Excel format, which is read by Stata easily

	A	B	C	D	E	F	G
1							
2	Quadro extraído em 17 de Setembro de 2015 (17:49:41)						
3	http://www.ine.pt						
4							
5							
6	**Período de referência dos dados**	**Óbitos de menos de 1 ano (N.º) por Local de residência; Mensal**					
7		**Local de residência**					
8		**Total**		**Portugal**		**Estrangeiro**	
9		T		PT		YY	
10		**N.º**		**N.º**		**N.º**	
11	**Junho de 2015**	19 //		18 //			
12	**Maio de 2015**	13 //		12 //			
13	**Abril de 2015**	23 //		23 //			
14	**Março de 2015**	18 //		18 //			
15	**Fevereiro de 2015**	21 //		21 //			
16	**Janeiro de 2015**	34 //		34 //			
17	Óbitos de menos de 1 ano (N.º) por Local de residência; Mensal - INE, Óbitos						
18							
19							
20	*Sinais convencionais:*						
21	*//: Dado preliminar*						
22							
23	Última atualização destes dados: 10 de setembro de 2015						

PORDATA

Óbitos de residentes em Portugal: total e no primeiro ano de vida

Indivíduo

Anos	Óbitos	
	Total	Menos de 1 ano
1960	95 007	16 576
1961	99 590	19 308
1962	96 864	17 300
1963	98 011	15 510
1964	96 878	14 974
1965	95 187	13 656
1966	100 088	13 379
1967	95 816	11 966
1968	94 661	11 917
1969	101 088	10 580
1970	93 093	10 027
1971	98 688	9 408
1972	90 315	7 234
1973	95 435	7 726
1974	96 928	6 515
1975	97 936	6 991
1976	102 027	6 244
1977	96 111	5 484
1978	96 194	4 878
1979	92 732	4 172
1980	⊥ 94 794	⊥ 3 839

NOVA School of Business & Economics | Healthcare Initiative Research

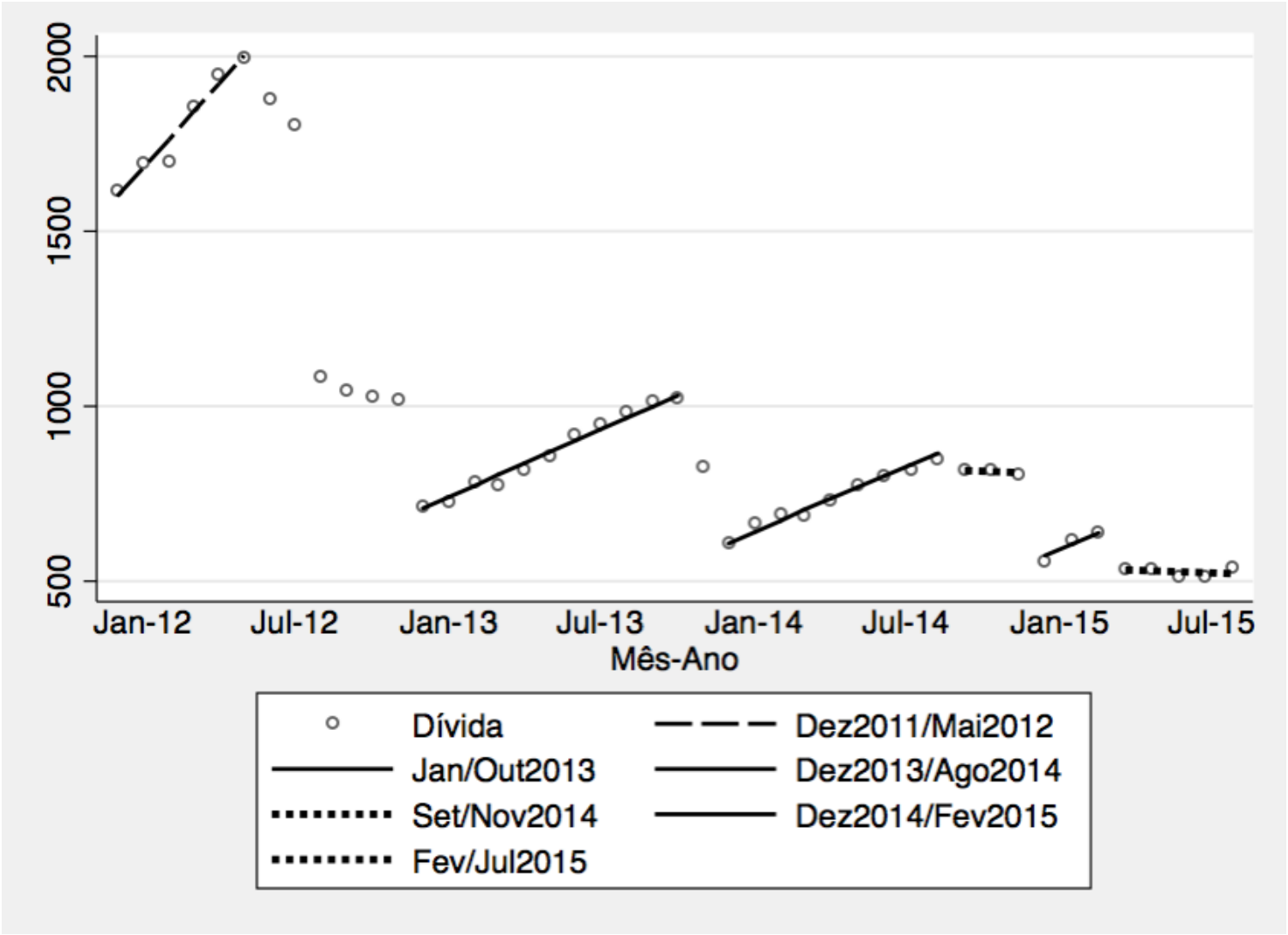
	A	B	C	D	E	F	G	H	I
1	AMECO RESULTS								
2									
3	Gross domestic product at current prices per head of population (HVGDP)								
4	Country	Unit	1960	1961	1962	1963	1964	1965	1966
5	Portugal	1000 EURO-F	0,05622	0,05806	0,06469	0,06657	0,07222	0,08437	0,09073
6									
7	Gross domestic product at current prices per head of population (HVGDP)								
8	Country	Unit	1960	1961	1962	1963	1964	1965	1966
9	Portugal	(1000 EUR)	0,37115	0,37929	0,42168	0,43392	0,47077	0,54992	0,59137
10									
11	Gross domestic product at current prices per head of population (HVGDP)								
12	Country	Unit	1960	1961	1962	1963	1964	1965	1966
13	Portugal	(1000 PPS)	0,44383	0,47046	0,53772	0,57933	0,6412	0,73678	0,80694
14									
15	Gross domestic product at 2010 reference levels per head of population (RVGDP)								
16	Country	Unit	1960	1961	1962	1963	1964	1965	1966
17	Portugal	1000 EURO-F	3,47544	3,571	3,91859	4,05266	4,2956	4,71906	4,9713
18									
19									
20									

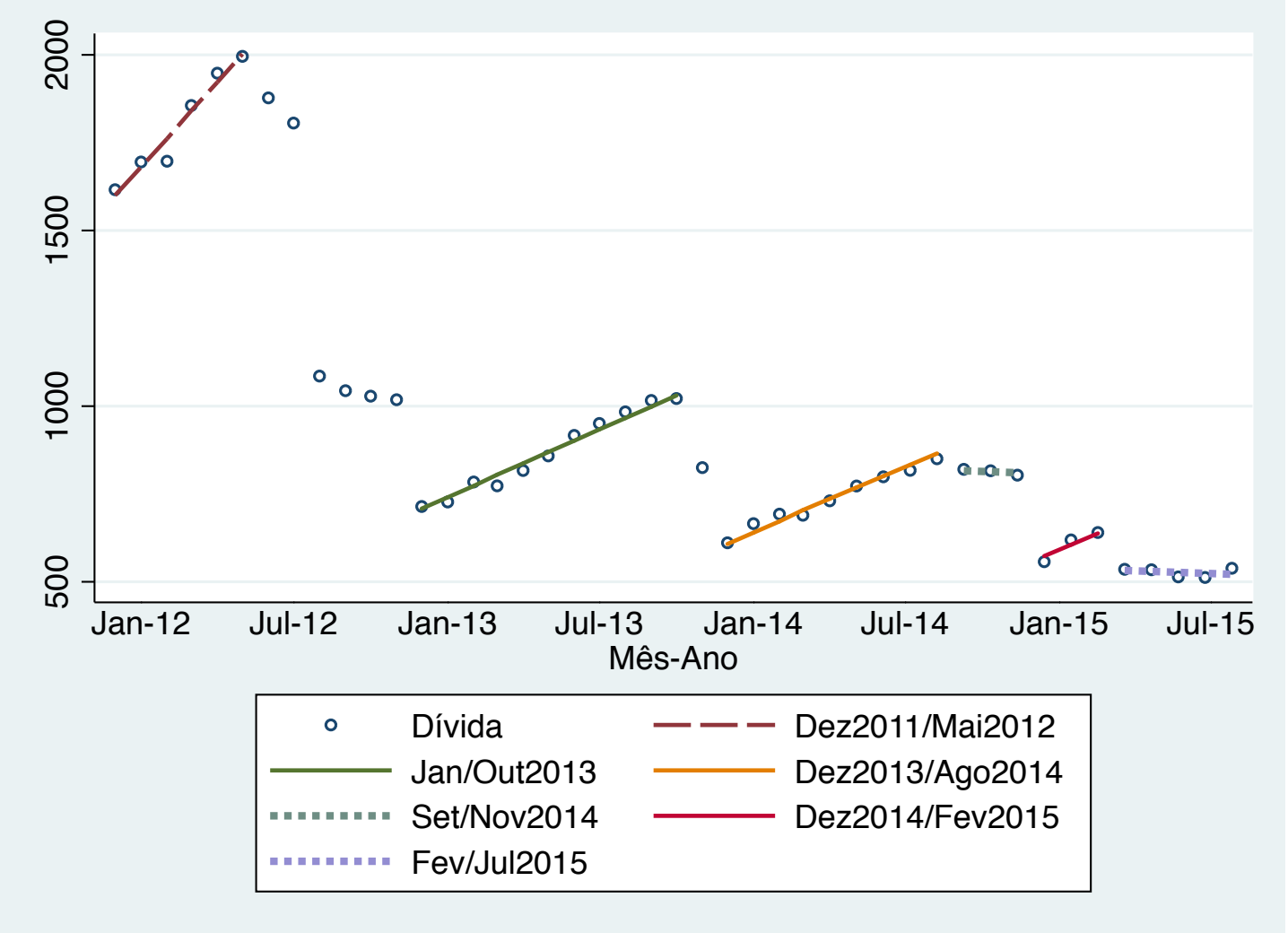
	A	B	C	D	E	F	G	H	I
1	OECD (2015), Suicide rates (indicator). doi: 10.1787/a82f3459-en (Accessed on 17 September 2015)								
2	https://data.oecd.org/healthstat/suicide-rates.htm								
3	Flags: .. Not available; | Break in series; e Estimated value; f Forecast value; x Not applicable; p Provisional data; s Strike;								
4	Information on data for Israel: http://oe.cd/israel-disclaimer								
5	Disclaimer and Terms and Conditions: http://www.oecd.org/termsandconditions/								
6									
7	Suicide rates, Total, Per 100 000 persons								
8									
9	Location	1960	1961	1962	1963	1964	1965	1966	1967
10	Portugal	11.7000	11.8000	11.1000	12.4000	12.6000	12.6000	12.6000	12.9000
11									
12									
13									
14									
15									
16									

NOVA School of Business & Economics | Healthcare Initiative Research

- Next step: export outputs
- Graph tools from Stata are great – high degree of flexibility; standard “schemes” are fine (though better in color than black/white)
- Export format is ok (png works well with wordpress)

NOVA School of Business & Economics
Healthcare Initiative Research


2000
1500
1000
500
Jan-12
Jul-12
Jan-13
Jul-13
Jan-14
Jul-14
Jan-15
Jul-15
Mês-Ano
Dívida
Dez2011/Mai2012
Jan/Out2013
Dez2013/Ago2014
Set/Nov2014
Dez2014/Fev2015
Fev/Jul2015



NOVA School of Business & Economics
Healthcare Initiative Research

- Exporting tables of regressions – could be easier – user-made commands help, but are not perfect for html/blog
- Print screens look nice graphically but not easy to read
- Wish: have direct export to html (the copy table as html does not work well in my Mac)

NOVA School of Business & Economics | Healthcare Initiative Research

Source	SS	df	MS	Number of obs	=	44
				F(15, 29)	=	415.98
Model	49035556.3	15	3269037.09	Prob > F	=	0.0000
Residual	227900.77	29	7858.64724	R-squared	=	0.9954
				Adj R-squared	=	0.9930
Total	49263457.1	44	1119624.02	Root MSE	=	88.649

divida_epe	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tend1	80.4159	21.19117	3.79	0.001	37.0751	123.7567
tend2	-190.5497	21.19117	-8.99	0.000	-233.8905	-147.2089
tend3	34.07061	8.452352	4.03	0.000	16.78361	51.35761
tend4	28.46714	11.44454	2.49	0.019	5.060433	51.87385
tend5	-8.370073	62.68432				
tend6	41.55665	62.68432				
tend7	-1.610583	28.03328				
ct1	554.7436	330.4509				
ct2	5406.715	457.0452				
ct3	-152.7956	254.9754				
ct4	825.0307	88.64901				
ct5	-430.7308	470.1556				
ct6	1206.637	2946.607				
ct7	-1472.327	3134.634				
ct8	613.8286	1514.316				

Source	SS	df
Model	49034084.1	12
Residual	229372.979	32
Total	49263457.1	44

divida_epe	Coef.	Std. Err.
tend1	80.4159	20.23844
tend2	-190.5497	20.23844
tend346	32.20296	6.45553
tend57	-2.737165	24.44024
ct1	554.7436	315.5943
ct2	5406.715	436.4971
ct3	-96.76603	195.341
ct4	825.0307	84.66348
ct5	-583.8994	266.177
ct6	941.8904	1149.731
ct7	-1004.643	326.4567
ct8	674.664	1320.316

	(1)	(2)
	divida_epe	divida_epe
tend1	80.42***	80.42***
tend2	-190.55***	-190.55***
tend3	34.07***	
tend4	28.47*	
tend5	-8.37	
tend6	41.56	
tend7	-1.61	
ct1	554.74	554.74
ct2	5406.71***	5406.71***
ct3	-152.80	-96.77
ct4	825.03***	825.03***
ct5	-430.73	-583.90*
ct6	1206.64	941.89
ct7	-1472.33	-1004.64**
ct8	613.83	674.66
tend346		32.20***
tend57		-2.74
N	44	44
R-sq	0.995	0.995

* p<0.05, ** p<0.01, *** p<0.001

NOVA School of Business & Economics
Healthcare Initiative Research

- Version:0.9
- StartHTML:0000000105
- EndHTML:0000001132
- StartFragment:0000000136
- EndFragment:0000001099
- <HTML><BODY><!--StartFragment--><TABLE BORDER><tr><td>. esttab m1 m2, b not

- (1) (2)
- divida_epe divida_epe

- tend1 80.42*** 80.42***
- tend2 -190.5*** -190.5***
- tend3 34.07***
- tend4 28.47*
- tend5 -8.370
- tend6 41.56
- tend7 -1.611
- ct1 554.7 554.7
- ct2 5406.7*** 5406.7***
- ct3 -152.8 -96.77
- ct4 825.0*** 825.0***
- ct5 -430.7 -583.9*
- ct6 1206.6 941.9
- ct7 -1472.3 -1004.6**
- ct8 613.8 674.7
- tend346 32.20***
- tend57 -2.737

- N 44 44

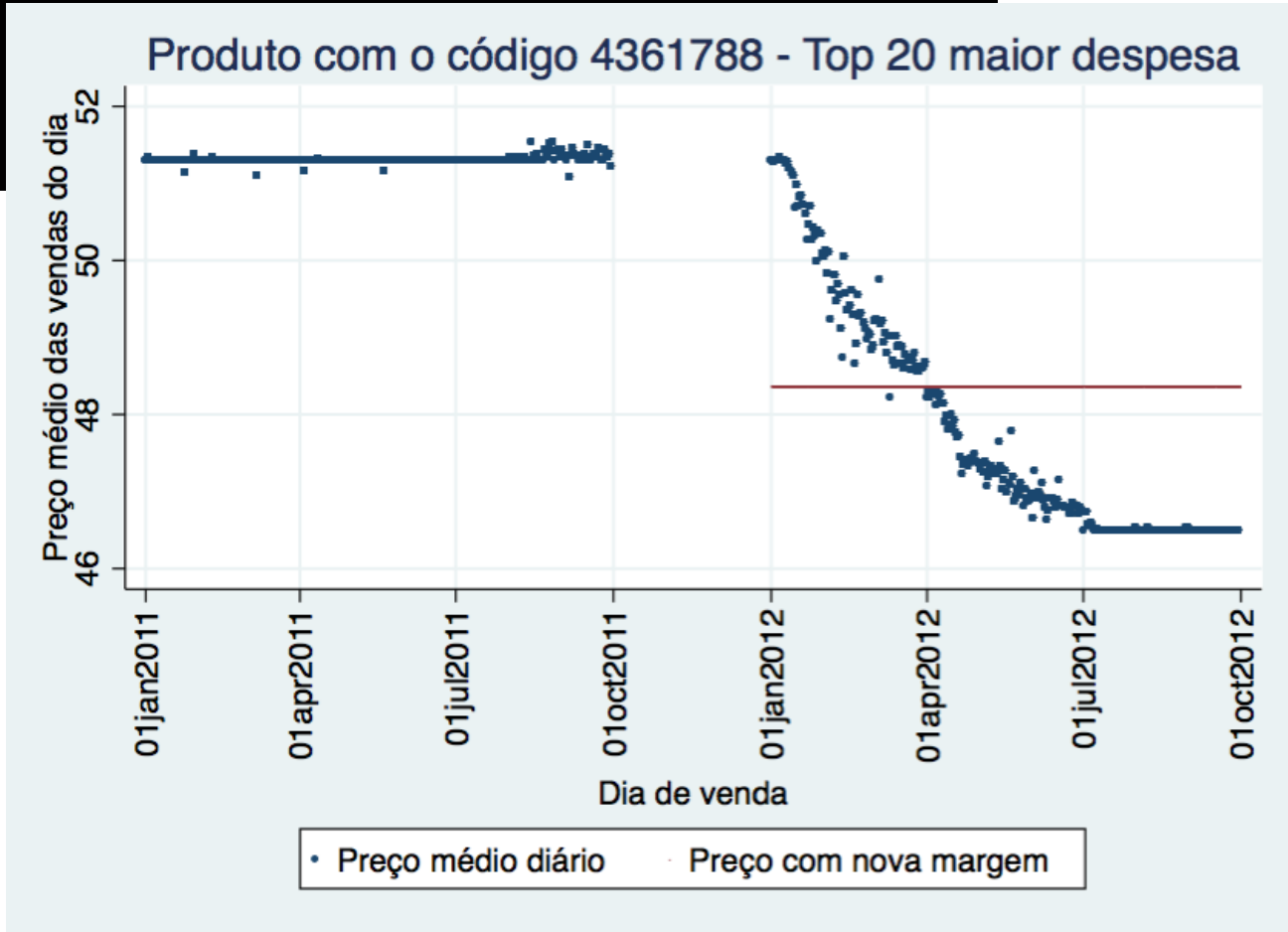
- * p<0.05, ** p<0.01, *** p<0.001
- </td></tr></TABLE><!--EndFragment--></BODY></HTML>

Data management

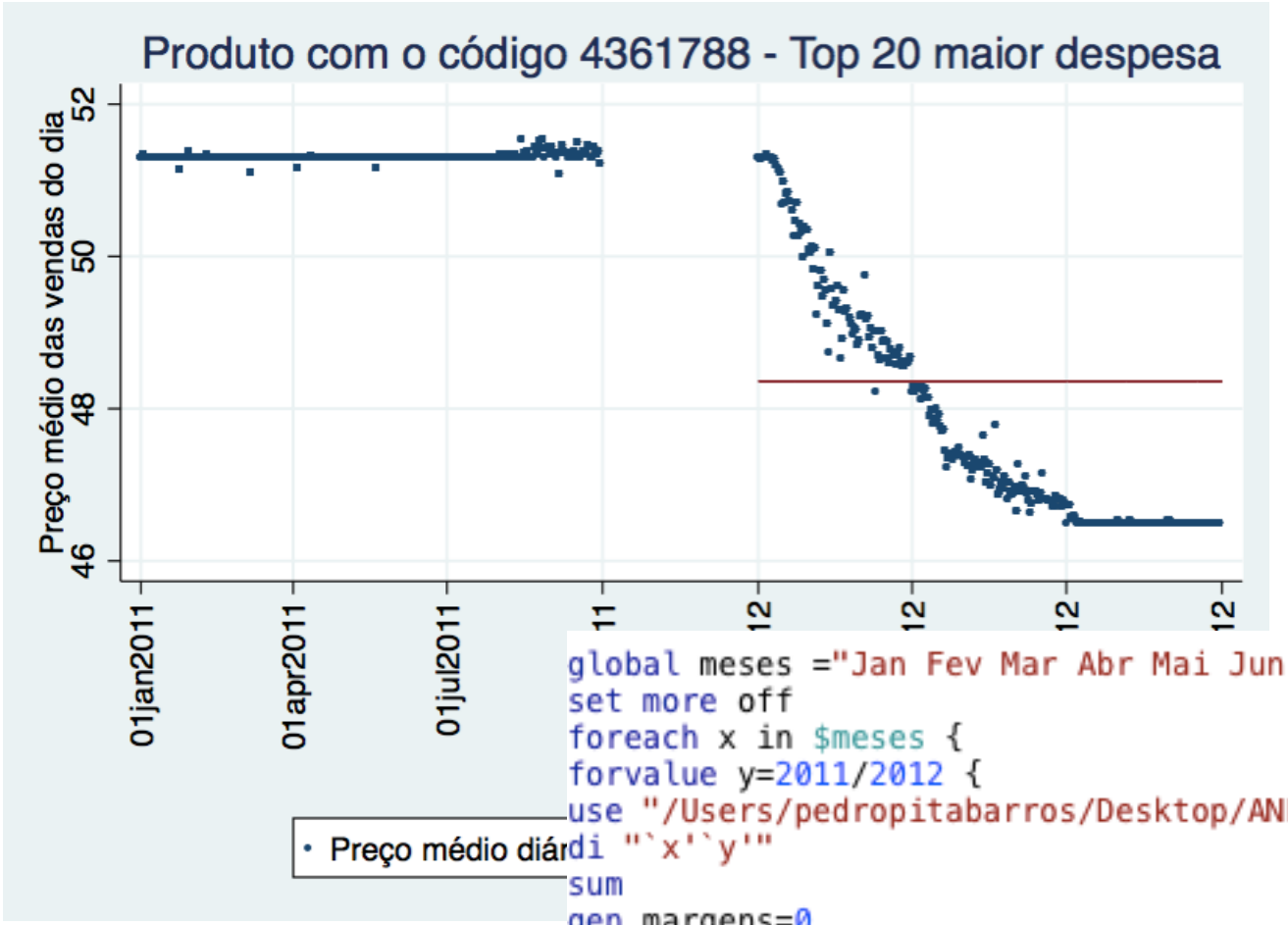
- Management of large data sets
  - Volume of data – 3,3 million observations per month, 24 months – using cycles in “do files”, size of joint database is very large (each file about 168-180 MB)
  - Joining repeated cross-sections – Ex: Standard Income and Living Conditions – Eurostat
  - Merging different data sets (ex: individual hospital admissions with residential data)

NOVA School of Business & Economics | Healthcare Initiative Research

Variable	Obs	Mean	Std. Dev.	Min	Max
cod_hmr	3,375,412	5063429	2898821	17311	1.00e+07
datavenda	0				
cod_prod	0				
qt	3,375,412	1.137234	2.162379	-119	900
pvp	3,375,412	1109.334	2470.368	-285720	999999
pago_utente	3,375,412	737.3289	3217.924	-993091	999999
cod_org	0				
valorcompa~o	1,487,838				
margens	3,375,412				



NOVA School of Business & Economics
Healthcare Initiative Research



Average price per day, each month from a different file

All data in a single file makes it very slow (close to 2,5 GB)

```
global meses ="Jan Fev Mar Abr Mai Jun Jul Ago Set Out Nov Dez"
set more off
foreach x in $meses {
forvalue y=2011/2012 {
use "/Users/pedropitabarros/Desktop/ANF-eixo1/`x'`y'.dta", clear
di "`x'`y'"
sum
gen margens=0
file open ano`y' using codigos`y'.txt, read write
file read ano`y' line
while r(eof)==0 {
    *display `line'
    quietly replace margens=1 if cod_prod=="`line'"
    file read ano`y' line
}
file close ano`y'
save "/Users/pedropitabarros/Desktop/`x'`y'_m.dta"
}
}
```

NOVA School of Business & Economics
Healthcare Initiative Research

Research papers

- Use of regression models available
- Flexibility for new likelihood functions
- New commands available
- User-built commands

NOVA School of Business & Economics
Healthcare Initiative Research

Example

Health and Health Care Demand
Effects of Double Coverage

by Pedro Pita Barros, Alberto Holly and Yevhen Pentsak

Version date : March 1, 2015

- Health – latent variable – 5 states
- Pharma consumption (yes/no)
- Visits to doctor – number of visits – count variable

NOVA School of Business & Economics | Healthcare Initiative Research

$$\begin{cases} y_1 \text{ such that } \mathcal{D}(y_1 \mid x_1, u_1^0) \text{ is the above mentioned Poisson distribution} \\ y_2^* = x_2'\beta_2^0 + \alpha_{21}^0 y_1 + u_2^0 \\ y_3^* = x_3'\beta_3^0 + \alpha_{31}^0 y_1 + \alpha_{32}^0 y_2 + u_3^0 \end{cases} \tag{4.3}$$

4.4. **The log-likelihood function.** The log-likelihood function for a sample with N observation may be written as

$$\mathcal{L}_N = \sum_{n=1}^{N} \sum_{k=0}^{\infty} \sum_{i=0}^{1} \sum_{j=1}^{5} z_{nkij} \log P_{nkij} \tag{4.32}$$

where

$$z_{nkij} = \begin{cases} 1 \quad \text{if } y_{n1} = k,\ y_{n2} = i \text{ and } y_{n3} = j \quad (k = 0, 1, \ldots, i = 0, 1 \text{ and } j = 1, \ldots, 5) \\ 0 \text{ otherwise} \end{cases}$$

$$P_{nkij} = \frac{1}{\sqrt{2\pi}\sigma_1} \int_{-\infty}^{+\infty} P(y_{n2} = i, y_{n3} = j \mid y_{n1} = k, u_{n1}) P(y_{n1} = k \mid u_{n1}) \exp\left(-\frac{u_{n1}^2}{2\sigma_1^2}\right) du_{n1}. \tag{4.36}$$

```
global var_pharma="gender income  age age2  schooling diabetes ashtma high_blood_p pai

global var_visits="gender income income2 public_sub private_sub age age2 schooling nor

global var_health="gender income public_sub private_sub age age2  schooling diabetes a


*************************************************************************
************Poisson Log-Normal three equation FIML model 12+24************
*************************************************************************

matrix X = (-6.01592556143,-5.25938292767,-4.62566275642,-4.05366440245,-3.52000681303
                0.224414547473,0.674171107037,1.12676081761,1.58425001096,2.0490035736
matrix W = (0.00000000000000001664368, 0.00000000000006584620243, 0.0000000030462542699
            0.000823692482688417000, 0.007048355810072670000, 0.037445470503230700000
            0.426931163868699000000, 0.286179535346443000000, 0.127739621784559000000
            0.000056886916364043700, 0.000021582457049023300, 0.000000040189711749410

matrix Z = (-0.0243502926634244, 0.0243502926634244, -0.0729931217877990, 0.0729931217

matrix T=(0.0486909570091397,   0.0486909570091397, 0.0485754674415034, 0.048575467441

capture program drop FIML_eq3
*************************************************************************
program define FIML_eq3
version 8
args lnf beta1 beta2 beta3 alpha21 alpha31 alpha32 rho12 rho13  sigmaP c1 c2 c3 c4
```

NOVA School of Business & Economics | Healthcare Initiative Research

```
forvalues i = 1/24 {

      replace `T20'=(-`beta2'-$ML_y1*`alpha21'-`arho12'*sqrt(2)* X[1,`i'])/`temp1'

      replace `T301'=(`c1'-`beta3'-$ML_y1*`alpha31'-`arho13'*sqrt(2)*X[1,`i'])/`temp2'
      replace `T302'=(`c2'-`beta3'-$ML_y1*`alpha31'-`arho13'*sqrt(2)*X[1,`i'])/`temp2'
      replace `T303'=(`c3'-`beta3'-$ML_y1*`alpha31'-`arho13'*sqrt(2)*X[1,`i'])/`temp2'
      replace `T304'=(`c4'-`beta3'-$ML_y1*`alpha31'-`arho13'*sqrt(2)*X[1,`i'])/`temp2'

      replace `T311'=(`c1'-`beta3'-$ML_y1*`alpha31'-$ML_y2*`alpha32'-`arho13'*sqrt(2)*X[1,`i'])/`temp
      replace `T312'=(`c2'-`beta3'-$ML_y1*`alpha31'-$ML_y2*`alpha32'-`arho13'*sqrt(2)*X[1,`i'])/`temp
      replace `T313'=(`c3'-`beta3'-$ML_y1*`alpha31'-$ML_y2*`alpha32'-`arho13'*sqrt(2)*X[1,`i'])/`temp
      replace `T314'=(`c4'-`beta3'-$ML_y1*`alpha31'-$ML_y2*`alpha32'-`arho13'*sqrt(2)*X[1,`i'])/`temp

          forvalues j = 1/64 {
replace `Tint01' = (T[1,`j']/2) *(1/(2*c(pi))* (`temp3')*exp(-0.5*(`T20'^2+`T301'^2-2*(Z[1,`j']/2+1/2

replace `Tint02' = (T[1,`j']/2) *(1/(2*c(pi))* (`temp3')*exp(-0.5*(`T20'^2+`T302'^2-2*(Z[1,`j']/2+1/2

replace `Tint03' = (T[1,`j']/2) *(1/(2*c(pi))* `temp3'*exp(-0.5*(`T20'^2+`T303'^2-2*(Z[1,`j']/2+1/2)*

replace `Tint04' = (T[1,`j']/2) *(1/(2*c(pi))* `temp3'*exp(-0.5*(`T20'^2+`T304'^2-2*(Z[1,`j']/2+1/2)*


replace `Tint11' =(T[1,`j']/2) *(1/(2*c(pi))* `temp3'*exp(-0.5*(`T20'^2+`T311'^2-2*(Z[1,`j']/2+1/2)*`

replace `Tint12' = (T[1,`j']/2) *(1/(2*c(pi))* `temp3'*exp(-0.5*(`T20'^2+`T312'^2-2*(Z[1,`j']/2+1/2)*

replace `Tint13' = (T[1,`j']/2) *(1/(2*c(pi))* `temp3'*exp(-0.5*(`T20'^2+`T313'^2-2*(Z[1,`j']/2+1/2)*

replace `Tint14' = (T[1,`j']/2) *(1/(2*c(pi))* `temp3'*exp(-0.5*(`T20'^2+`T314'^2-2*(Z[1,`j']/2+1/2)*
```

NOVA School of Business & Economics | Healthcare Initiative Research

```
matrix input visit= (0.3151542, 0.050743, -0.0181319, 0.1499232, 0.1724856, -0.0246328, 0.0003577, -0.(

matrix input pharma= (-1.270771, -0.1211893, 0.0367668, -0.000581, 0.0033042, -0.5776318, -0.3319239, -

matrix input hlth= (0.0434737, 0.2335649, 0.011538, 0.0847499, -0.0296402, 0.0001309, 0.0530095, -0.40

ml model lf FIML_eq3                (mu1: visit_doctor = $var_visits, noconst) ///
                        (mu2: pharma_use = $var_pharma, noconst) ///
                        (mu3: health = $var_health, noconst) ///
                                    /alpha21 /alpha31 /alpha32 /rho12 /rho13  /sigmaP /c1 /c2 /c3 /c4

ml init visit pharma hlth 0.1141355 -0.1073696 0.0506028 -0.795329 0.0056024 0.8664846 -3.739271 -2.65

ml maximize, search(off) difficult trace
```

About 11,000 observations; 2 iterations takes roughly 5 hours

NOVA School of Business & Economics | Healthcare Initiative Research

Results

TABLE 1. Comparison estimates - visits equation

	Unconstrained		constrained: $\rho_{23} = 0$		constrained: $\rho_{23} = \rho_{13} = 0$	
Visits	Coef.	Std.Err.	Coef.	Std.Err.	Coef.	Std.Err.
gender	0,429	0,084*	0,429	0,085*	0,430	0,085*
income	0,004	0,028	0,004	0,028	0,004	0,029
public sub	0,156	0,037*	0,157	0,037*	0,155	0,037*
private sub	0,175	0,066*	0,176	0,066*	0,173	0,066*

NOVA School of Business & Economics | Healthcare Initiative Research

Pharma use	Unconstrained		constrained: $\rho_{23} = 0$		constrained: $\rho_{23} = \rho_{13} = 0$	
	Coef.	Std.Err.	Coef.	Std.Err.	Coef.	Std.Err.
gender	-1,338	0,087*	-1,341	0,086*	-1,353	0,086*
income	-0,119	0,029*	-0,118	0,029*	-0,120	0,029*
age	0,033	0,002*	0,033	0,002*	0,032	0,002*
age2	-0,001	0,000*	-0,001	0,000*	-0,001	0,000*
schooling	0,005	0,004	0,005	0,004	0,006	0,004
public sub	-0,023	0,038	-0,021	0,038	-0,032	0,038
private sub	-0,002	0,069	-0,002	0,069	0,000	0,068

TABLE 3. Comparison estimates - health status equation

Health status	Coef.	Std.Err.	Coef.	Std.Err.	Coef.	Std.Err.
gender	0,209	0,087*	0,175	0,081*	0,206	0,081*
income	0,231	0,025*	0,227	0,025*	0,237	0,025*
public sub	-0,006	0,033	-0,015	0,032	0,008	0,032
private sub	0,065	0,059	0,064	0,059	0,059	0,059
alpha31	-0,119	0,013*	-0,106	0,007*	-0,098	0,007*
alpha32	0,104	0,059	0,043	0,028	0,121	0,027*

Our interest: role of subsystems – impact on number of visits, but not pharma or health directly

NOVA School of Business & Economics | Healthcare Initiative Research

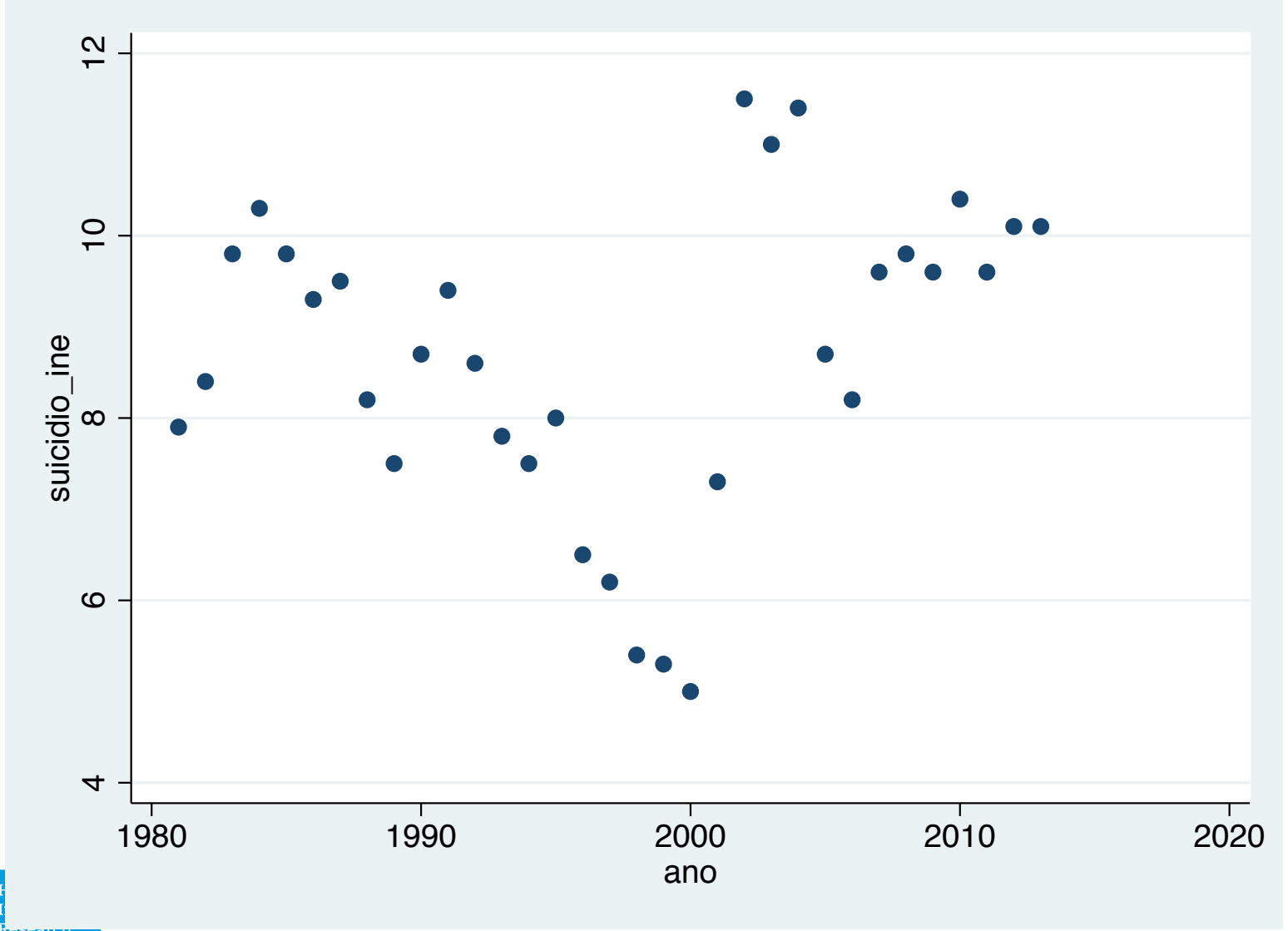
- Problems: a “do file” that runs in Stata 13 does not in Stata 14 – not easy to find out why…
- Look for more efficient coding to reduce time to solve the problem

New commands

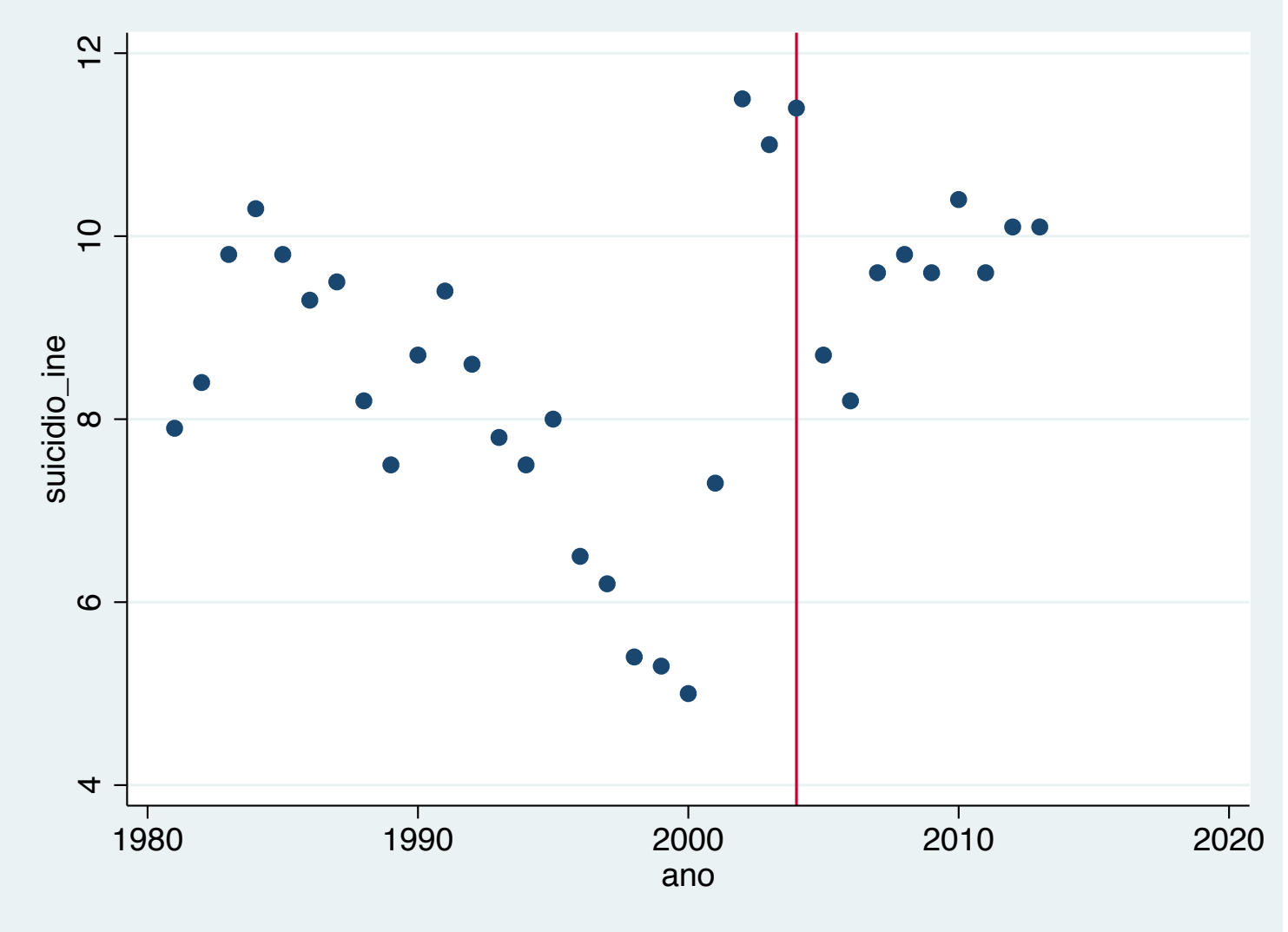
- Example:
  - Treatment effects – matching estimator (new command)
  - Krls – user command for machine-learning technique to fit non-linear function
  - Traditional regression
- Suicides during the crisis period

NOVA School of Business & Economics | Healthcare Initiative Research

What happened



After 2004 (5 years each side)



```
Treatment-effects estimation                   Number of obs      =         10
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =          1
```

suicidio_ine	AI Robust Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
crise (1 vs 0)	.75	.5040337	1.49	0.137	-.237888	1.737888

Source	SS	df	MS		
Model	.441005099	2	.22050255	Prob > F	0.7954
Residual	6.5239949	7	.931999272	R-squared	0.0633
Total	6.965	9	.773888889	Adj R-squared	-0.2043
				Root MSE	.9654

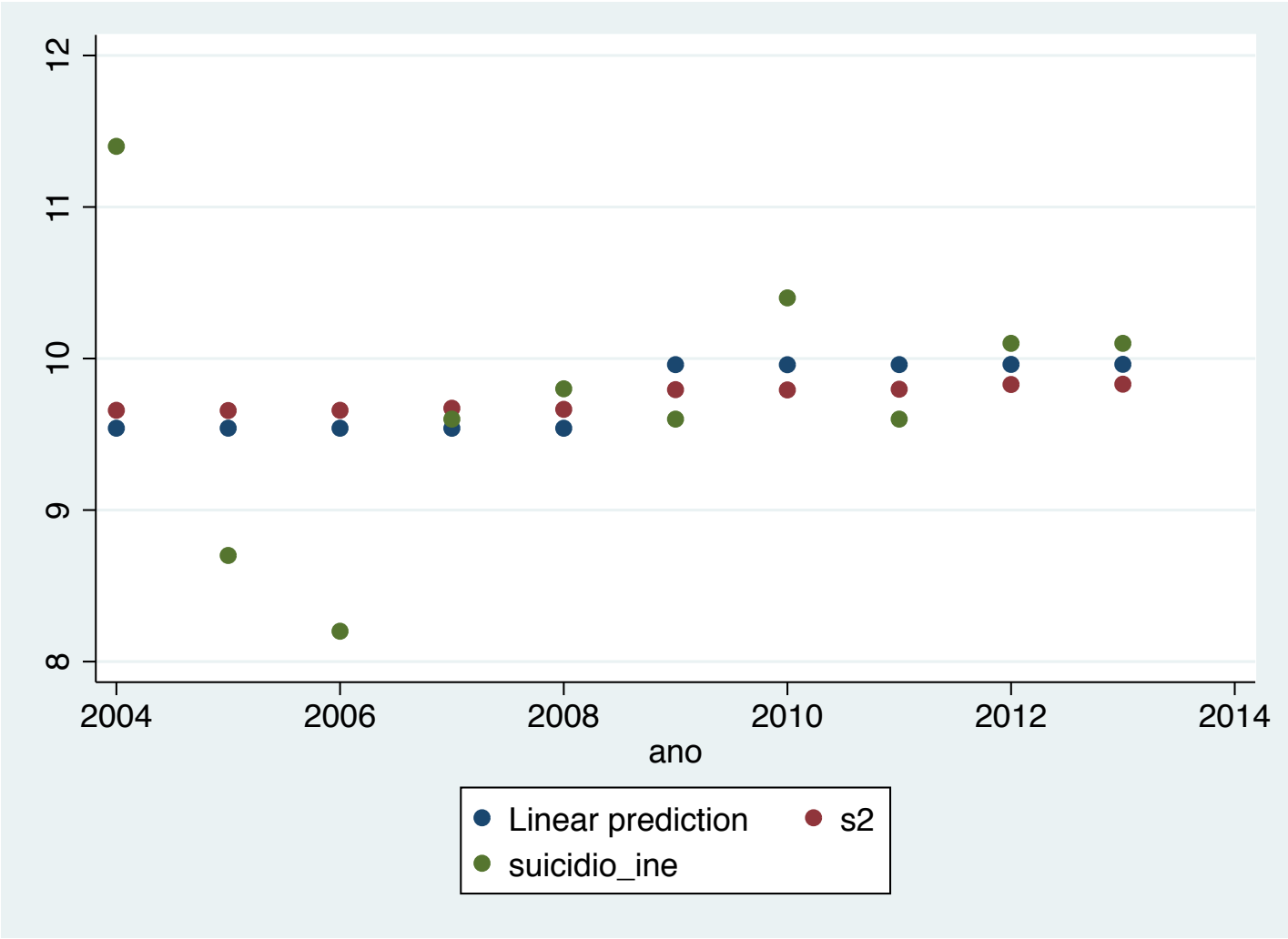
suicidio_ine	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nova	-.0328985	14.06491	-0.00	0.998	-33.29114	33.22534
crise	.4182112	.9785804	0.43	0.682	-1.895764	2.732186
_cons	9.540854	.5653784	16.88	0.000	8.203946	10.87776

```
Po                                                                    10
                                                                    .922
                                                                    .01
                                                Sigma          =       2
                                                Eff. df        =   1.068
                                                R2             =  .03967
                                                Looloss        =   9.309
```

suicidio_ine	Avg.	SE	t	P>|t|	P25	P50	P75
nova	.246557	1.00417	0.246	0.812	-.228982	.220407	.507675
*crise	.115	.205448	0.560	0.591	.110739	.12176	.131342

* average dy/dx is the first difference using the min and max (i.e. usually 0 to 1)

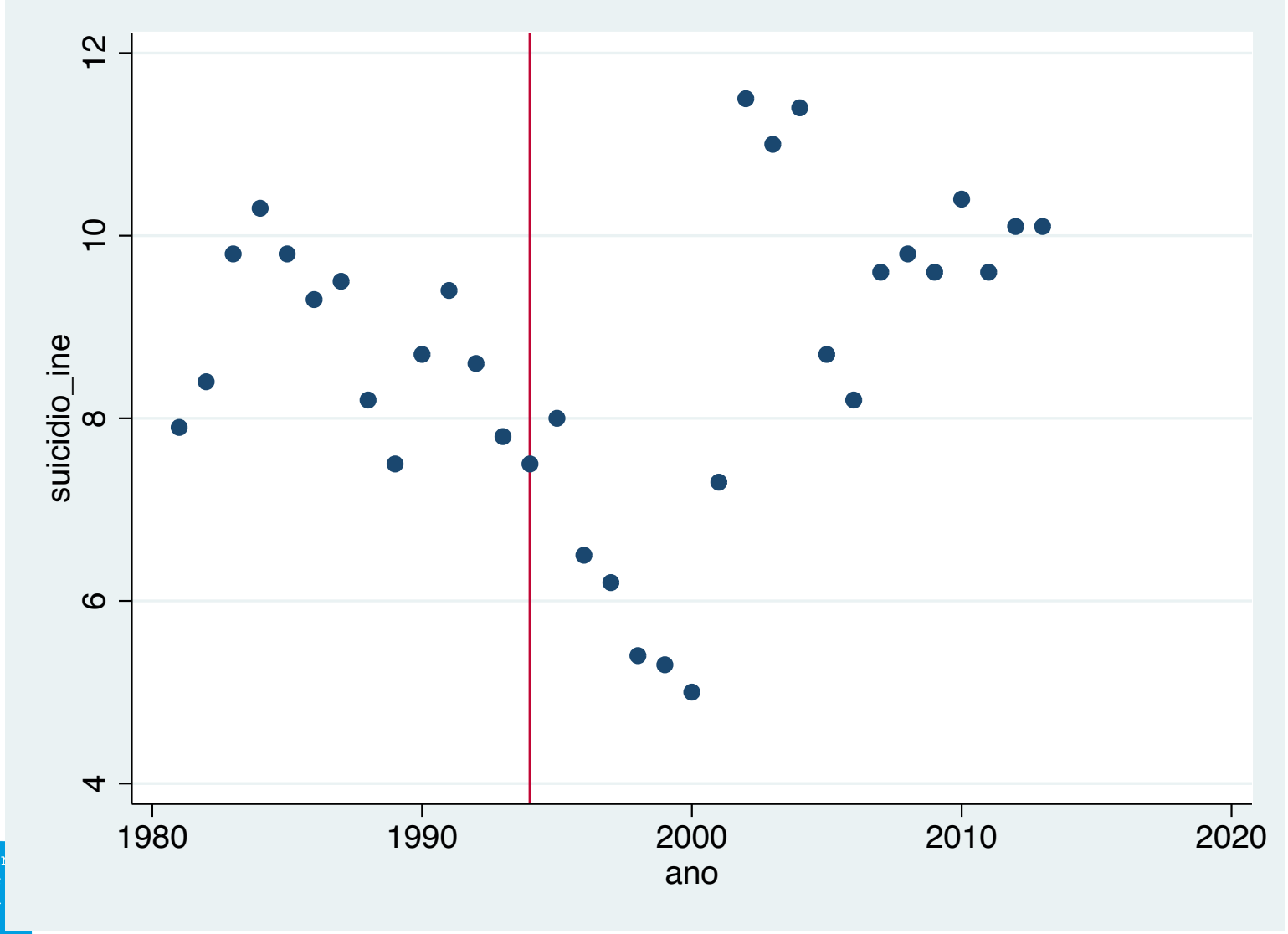
NOVA School of Business & Economics | Healthcare Initiative Research



Non-significant differences – the three methods are roughly giving the same info

NOVA School of Business & Economics
Healthcare Initiative Research

Go back 10 years: since 1994



NOVA School of Business & Economics
Healthcar Initiative Research

```
Treatment-effects estimation                   Number of obs      =         47
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =          1
```

suicidio_ine	AI Robust Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
crise (1 vs 0)	1.502128	.3245476	4.63	0.000	.8660261	2.138229

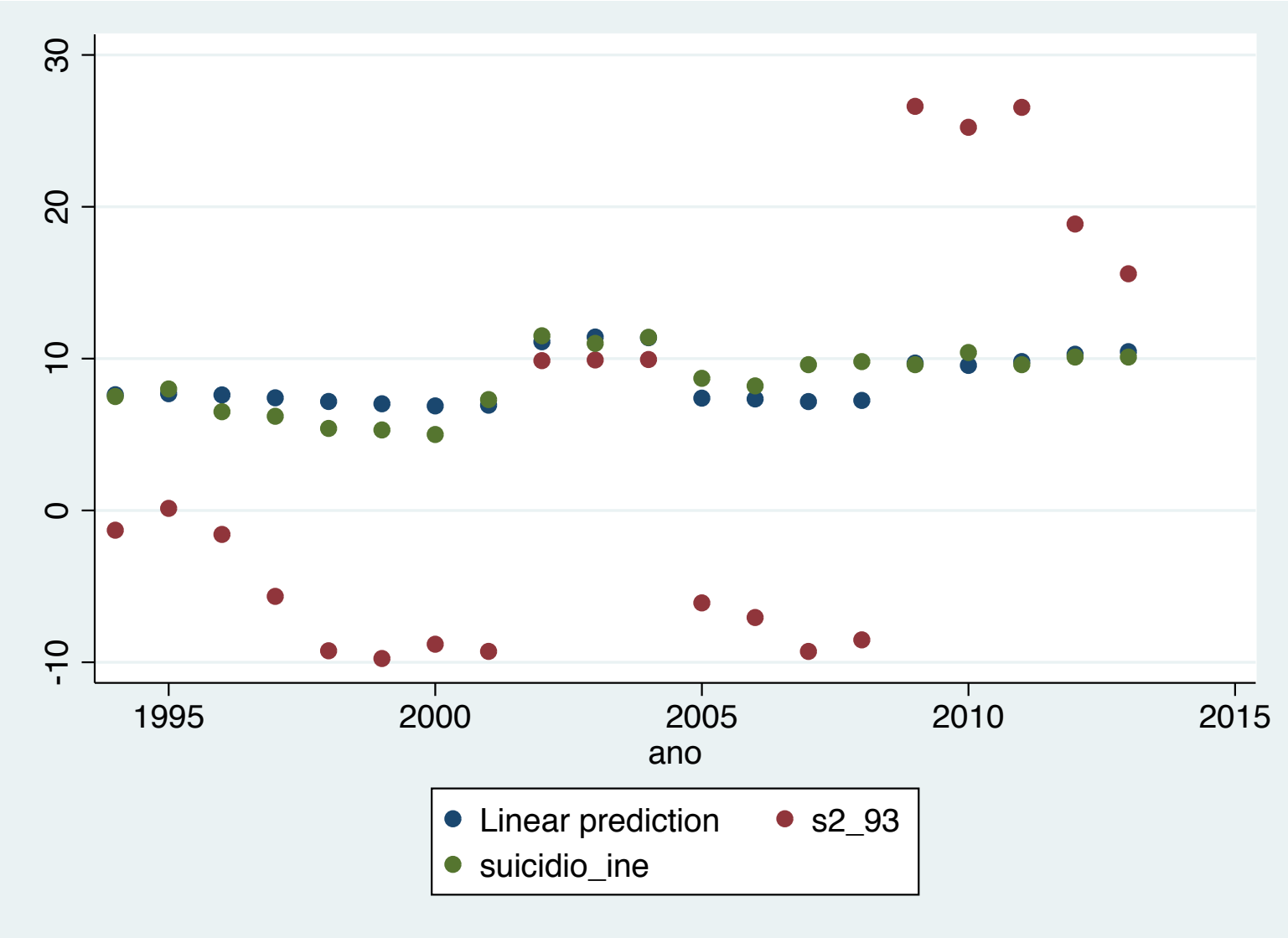
suicidio_ine	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
nova	-11.67562	12.95757	-0.90	0.381	-39.14445	15.79321
crise	2.026338	1.008443	2.01	0.062	-.1114649	4.16414
report	4.031454	.865829	4.66	0.000	2.195978	5.866929
_cons	7.601864	.5179894	14.68	0.000	6.503776	8.699953

suicidio_ine	Avg.	SE	t	P>|t|	P25	P50	P75
nova	-8.49304	9.22985	-0.920	0.370	-36.0908	1.39009	16.2799
*crise	1.67243	.840599	1.990	0.063	2.1125	2.2406	2.39262
*report	2.43551	.709319	3.434	0.003	.932195	3.19907	3.84897

```
* average dy/dx is the first difference using the min and max (i.e. usually 0 to 1)
```

Impact of crisis – there are differences but not too different

NOVA School of Business & Economics | Healthcare Initiative Research



We look at predicted effect – the krls is wildly non-linear (unreasonable so)

NOVA School of Business & Economics
Healthcare Initiative Research

Stata in the life of a health economist

- Helpful in many ways
- Could be improved to interact with new social media
- Works very well for management of large data sets (speed can be an issue)
- Flexible enough to accommodate new models / likelihood functions
- Great to test different technique using both official and community-based commands

NOVA School of Business & Economics
Healthcare Initiative Research