Institut Català de la Salut
Direcció d'Atenció Primària Metropolitana Nord
Unitat de Suport a la Recerca

IDIAP Jordi Gol

Multiple Imputation

Una possible solució a la presència dels missings

Aplicació en la cohort ARTPER

Guillem Pera

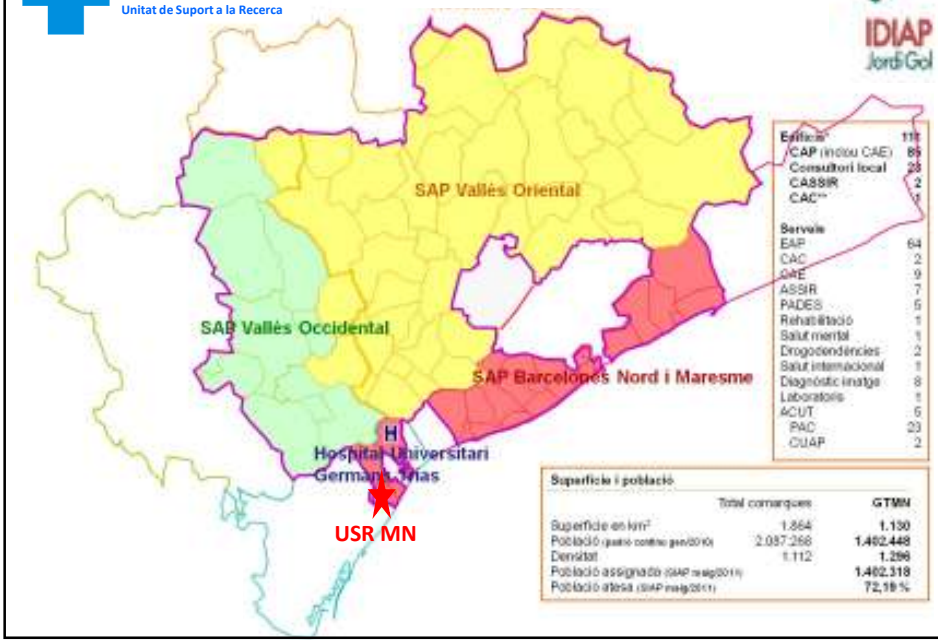
USR Metropolitana Nord, IDIAP Jordi Gol

UAB Universitat Autònoma de Barcelona
Servei d'Estadística Aplicada

25 d'abril de 2013

Institut Català de la Salut
Direcció d'Atenció Primària Metropolitana Nord
Unitat de Suport a la Recerca

IDIAP Jordi Gol



Enquesta…

- Què feu quan teniu missings en un estudi?
- Quina mena de missings teniu? (dades repetides, relacionats amb altres variables, ...)
- Quants missings teniu?
- Quines tècniques useu per evitar l'efecte dels missings?
- Amb quin software?

Missings: el problema

La presència de missings provoca:

- Disminució de la potència
  al tenir menys n
- Biaix
  si el fet de ser missing es relaciona amb alguna altra variable

Missings: el problema

L'efecte perniciós dels missings dependrà:

- De la seva quantitat
- De si la seva presència és aleatòria
- De la importància* de la variable en el nostre estudi [*entenguis importància com la necessitat (estadística o no) d'incloure una variable]

Missings: les solucions

NO HI HA SOLUCIÓ!!

Els mètodes estadístics per imputar missings no es poden comprovar, ja que NO DISPOSEM de les dades reals.

Podem fer:

- simulacions (eliminar dades i veure com es comporten els mètodes estadístics al imputar els missings)
- comparar-nos amb dades externes (prevalença, associacions,...)
- anàlisi de sensibilitat

Missings: les solucions

Algunes propostes:

- Anàlisi de casos complets (CCA)
- Creació d'una categoria "missing"
- Imputació de la mitjana (o mediana)
- Imputació del darrer valor conegut (LVCF) (longitudinal)
- Imputacions (regressió, etc)
- ...

Missings: patrons

- MCAR: Missing completely at random
- MAR: Missing at random
- MNAR: Missing not at random

Missings: patrons
Exemple

- Volem saber la relació entre tensió arterial (TA) i malalties cardiovasculars (MCV)
- Alguns participants no s'han pres la TA (=missing)
- Y=variable d'interès (=tensió arterial), X=covariables (=edat), R=indicador [0=missing, 1=observat]

Missings: patrons
Missing completament aleatori (MCAR)

- La probabilitat de ser missing NO depèn ni de les dades observades ni de les no observades
- La TA no s'ha observat en el 10% dels pacients perquè un dia es va fer malbé l'aparell de mesura
- Els pacients amb TA mesurada tenen una TA similar als pacients missing
- $P(R=1|Y,X)=P(R=1)$
- La probabilitat de tenir TA missing és independent del tot

Missings: patrons
Missing aleatori (MAR)

- La probabilitat de ser missing depèn de les dades observades PERÒ NO de les no observades
- La TA no s'ha observat en el 10% dels pacients perquè la majoria dels menors de 40 anys de l'estudi no es van fer l'examen de salut
- Els pacients amb TA mesurada deuen tenir una TA superior als pacients missing, ja que aquests són més joves
- $P(R=1|Y,X)=P(R=1|X)$
- La probabilitat de tenir TA missing depèn de l'edat
- Per a subjectes de la mateixa edat, TA és MCAR

Missings: patrons
Missing no aleatori (NMAR)

- La probabilitat de ser missing depèn de les dades no observades
- La TA no s'ha observat en el 10% dels pacients perquè els que tenien valors més alts tenien mal de cap i no s'han presentat a la consulta
- El fet de no tenir TA mesurada depèn de la pròpia TA
- $P(R=1|Y,X)=P(R=1|Y)$ $[f(Y|X,R=0 \neq f(Y|X,R=1)]$
- La probabilitat de tenir TA missing depèn de TA

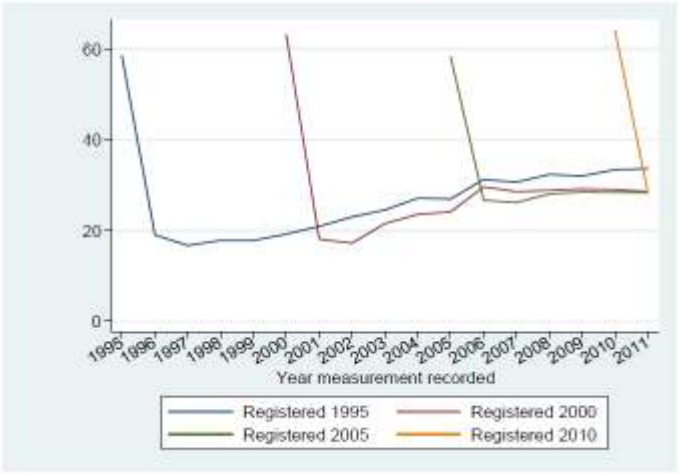
Missings: patrons

- MCAR: Missing completely at random
- MAR: Missing at random
- MNAR: Missing not at random

Normalment NO PODEM COMPROVAR AQUESTES HIPÒTESIS

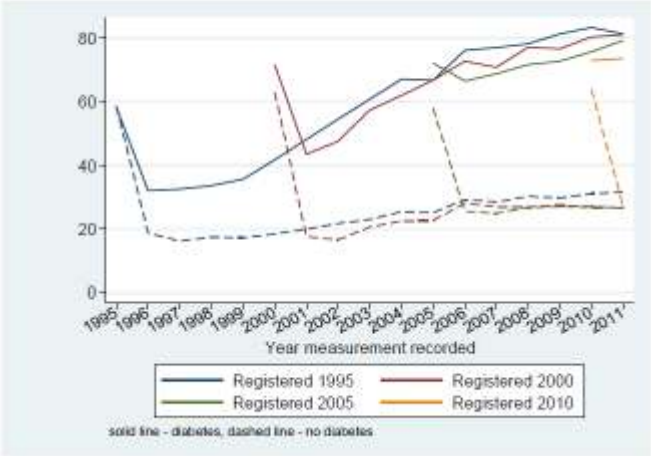
Els patrons poden diferir entre variables
És important pensar quin és el patró i raó de ser dels missings

Recording of weight



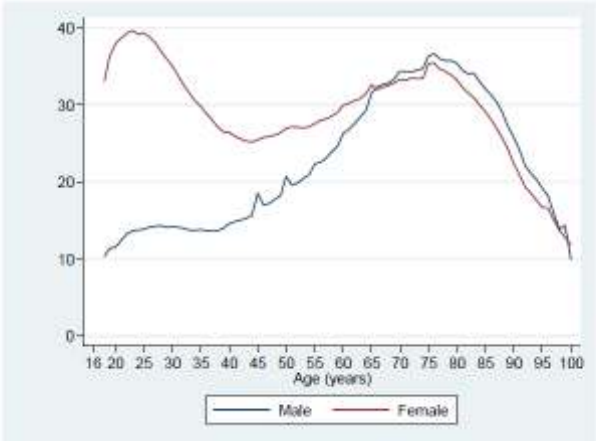
by Irene Petersen & Jonathan Bartlett

Recording of weight in diabetics and non-diabetics



by Irene Petersen & Jonathan Bartlett

Recording of weight by age and gender



by Irene Petersen & Jonathan Bartlett

Missings: les solucions

Algunes propostes:

- Anàlisi de casos complets (CCA)
- Creació d'una categoria "missing"
- Imputació de la mitjana (o mediana)
- Imputació del darrer valor conegut (LVCF) (longitudinal)
- Imputacions (regressió, etc)
- ...

Missings: les solucions

Anàlisi de **casos complets** (CCA):

- Pèrdua de potència
- Biaix (excepte MCAR)

Missings: les solucions

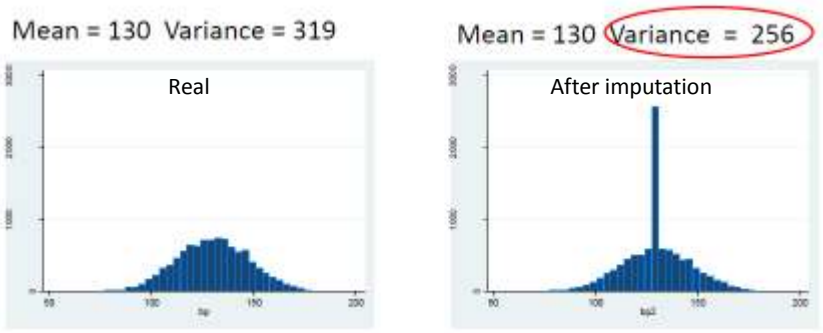
Creació d'una **categoria "missing"**:

- Agrupació en una sola categoria d'individus de categories (reals) diferents
- Biaix (sever) en qualsevol direcció
- No serveix ni per ajustar
- TOTALMENT DESACONSELLABLE

Missings: les solucions

Imputació de la mitjana:

- Infraestimació de la variança
- No apropiat per a variables categòriques
- Biaix (excepte MCAR)



by Irene Petersen & Jonathan Bartlett

Missings: les solucions

Imputació del darrer valor conegut (**LVCF**):

- Assumpció molt forta de que les mesures són constants
- Per tant, possible biaix i infraestimació de la variança

Missings: les solucions

Imputació per regressió:

- Infraestima variabilitat (no té en compte la incertesa que suposa la presència d'un missing)
- Possible biaix (menys que els anteriors mètodes segurament)
- Acceptable si el missing és MAR i afecta només a la variable resposta (trials)

Multiple Imputation (MI)

MI és una tècnica estadística flexible, basada en la simulació.

Consta de 3 passos:

1. **IMPUTATION:** Crea M còpies de les dades, reemplaçant els missings per imputacions, basant-se en les dades observades (**imputation model**).

2. **COMPLETED-DATA ANALYSIS:** Analitza el nostre model d'interès (tècniques estàndard) M cops (1 cop per cada dataset (**estimation model**)).

3. **POOLING:** Combina els M resultats del pas anterior en un de sol.

Multiple Imputation (MI)

- Té en compte la incertesa provocada per la imputació, gràcies a la iteració [IC més amplis]

- Òbviament els resultats dels models d'estimació difereixen ja que també difereixen els datasets.

- El model infereix correctament ja que l'estimador es basa en les dades observades i en la distribució de les imputades donades les observades (MAR).

- La relació que s'estudia en l'estimation model s'ha d'incloure, d'alguna forma, en l'especificació dels possibles valors de les imputacions en l'imputation model.

- No importa si imputem variables explicatives o resposta.

Multiple Imputation (MI)

- La relació que s'estudia en l'estimation model s'ha d'incloure, d'alguna forma, en l'especificació dels possibles valors de les imputacions en l'imputation model.

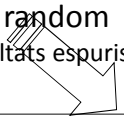
Si això no es respecta, tindrem biaix cap a nul en els estimadors.

Aquest biaix dependrà del % de missings i de la força de l'associació entre les variables implicades.

Per a patir biaixos significatius aquesta associació ha de ser forta i la proporció de missings considerable.

Multiple Imputation (MI). Perills.

- Ometre la variable resposta del imputation model
  Molts cops els missings estan en les variables explicatives. Però la relació d'aquestes amb la resposta cal tenir-la en compte.
- No normalitat
  Es pot controlar amb transformacions o models ad-hoc
- **Missing at random**
  El model d'imputació ha d'incloure TOTES les variables del model d'anàlisi , les variables que provoquen MAR i totes les que es relacionin amb la variable afectada.
- Problemes computacionals
- Missing NOT at random
  MI pot donar resultats espuris en aquest cas. El biaix pot ser major que el de CCA.



En el cas que MI i CCA donin resultats molt diferents cal explorar perquè, pensar si podem estar en MNAR i publicar ambdós resultats.

Multiple Imputation (MI)

- MI es va popularitzant
- Però encara es reporta molt malament (no s'indica el % de missings, quins models d'imputació, quantes iteracions, no es discuteix el patró de missings, etc)

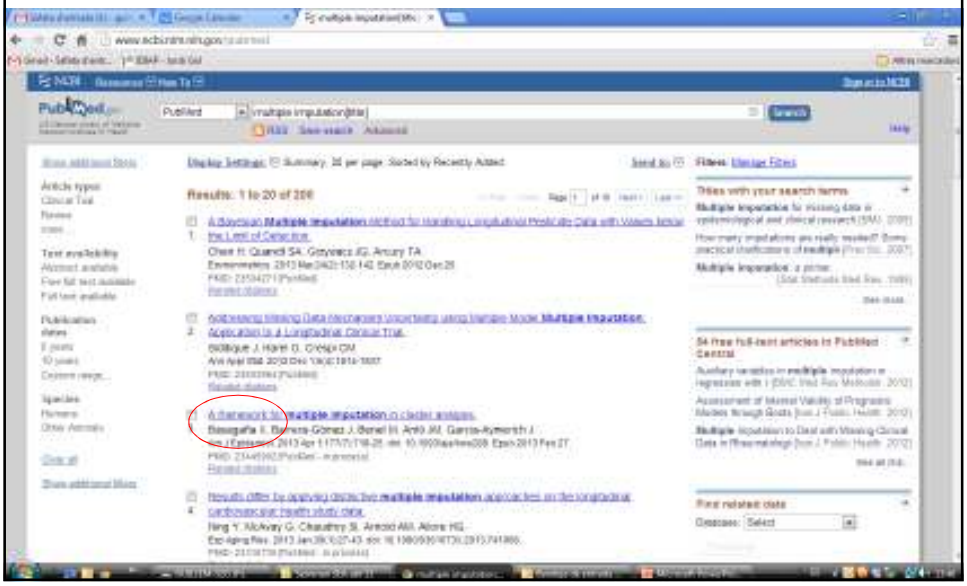
Notes pràctiques

- Valorar el patró de missings
- Intentar que els missings siguin MAR
- Definir un bon imputation model (que tingui en compte l'estimation model i MAR, així com les característiques de les variables (ordinals, valors negatius, potències, etc)) i que inclogui variables correlacionades amb les que tinguin missing
- Si l'estimation model té interaccions, efectes no lineals, etc, l'imputation model també (FCS)
- Triar un nombre suficient de rèpliques (quan més alt (>20) millor)
- Comparar els resultats amb CCA i interpretar

Bibliografia

- Jonathan AC Sterne et al. *Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.* BMJ 2009. [pros i contres de MI]
- Stef van Buuren. *Multiple imputation of discrete and continuous data by fully conditional specification.* Stat Methods Med Res 2007. [més teòric: chained equations vs. joint modelling]
- Louise Marston et al. *Issues in multiple imputation of missing data for large general practice clinical databases.* Pharmacoepidemiol Drug Saf 2010. [exemple]
- StataCorp. *Stata Multiple-Imputation reference manual.* Stata: Release 12. Statistical Software. College Station, TX: StataCorp LP. 2011. [manual d'Stata]

Bibliografia



L'estudi ARTPER

- www.artper.org
- Cohort poblacional de 3786 persones >50 anys del Barcelonès i Maresme
- Objectius principals:
  - Prevalença AP i associació amb altres FRCV
  - Factor pronòstic d'AP per a MCV i mortalitat
  - Incidència d'AP i altres FRCV
- Baseline: 2006-08
- Seguiment: continu (e-CAP, CMBDH, mortalitat, trucades cada 6 mesos, 2n tall presencial 2012)

Objectiu (exemple MI)

- **Relacionar la presència d'AP amb el nivell d'activitat física de lleure**

- AP=valor d'ITB<0,9 [≥1,4 exclosos]
- n=3.551
- Activitat física de lleure (hores al dia, dies al mes=hores/setmana. Contínua)

Al principi de l'estudi no es va registrar. 546 missings (15%)

Variables (exemple MI)

VARIABLES ORIGINALS			Missings	
AP	**Variable resposta**	0 Sa, 1 AP	0	
HSAFLL	**Activitat física de lleure**	Contínua (hs/setmana)	546	15%
EDAT		Anys	0	
SEXE		0 Home, 1 Dona	0	
MODAF	**Ha modificat la seva AF en els darrers 10 anys**	1 ha augmentat molt, 2 ha augmentat, 3 cap canvi, 4 ha disminuït, 5 ha disminuït molt	43	1%
GLIC	**Glicèmia**	mg/dl	1806	51%

286 individus amb AP (8.1%)

Patró missing

- **És HSAFLL MCAR?**

Al principi de l'estudi no es va registrar. 546 missings (15%)

Ho sembla...
... però cal comprovar-ho!!!

```
. mi set wide
. mi register imputed hsafl
. set seed 250413
. tab _mi_miss sexe, chi2 col
```

_mi_miss	sexe home	dona	Total
0	1,416 88.67	1,589 81.32	3,005 84.62
1	181 11.33	365 18.68	546 15.38
Total	1,597 100.00	1,954 100.00	3,551 100.00

Pearson chi2(1) = 36.4444 Pr = 0.000

HSAFLL té més missings entre les dones. Cal introdüir la variable SEXE a l'imputation model

(single) Imputation model

```
. mi impute truncreg hsafll ap edat sexe, add(20) ll(0)

Univariate imputation                       Imputations =       20
Truncated regression                               added =       20
Imputed: m=1 through m=20                        updated =        0

Limit: lower =            0             Number truncated =      241
       upper =         +inf                          left =      241
                                                     right =        0
```

Variable	Observations per *m*			
	Complete	Incomplete	Imputed	Total
hsafllleure	3005	546	546	3551

```
(complete + incomplete = total; imputed is the minimum across m
 of the number of filled-in observations.)

.
end of do-file
```

Imputation model

```
. summ hsafll _1_-_20
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hsafllleure	3005	7.203113	6.106695	0	67.5
_1_hsaflle~e	3551	7.172834	6.009987	0	67.5
_2_hsaflle~e	3551	7.175257	6.006257	0	67.5
_3_hsaflle~e	3551	7.282981	6.051299	0	67.5
_4_hsaflle~e	3551	7.167775	6.002409	0	67.5
_5_hsaflle~e	3551	7.254544	6.037292	0	67.5
_6_hsaflle~e	3551	7.337092	6.160405	0	67.5
_7_hsaflle~e	3551	7.349584	6.102762	0	67.5
_8_hsaflle~e	3551	7.239658	6.059511	0	67.5
_9_hsaflle~e	3551	7.337618	6.146307	0	67.5
_10_hsafll~e	3551	7.129289	5.976243	0	67.5
_11_hsafll~e	3551	7.235492	6.024347	0	67.5
_12_hsafll~e	3551	7.386353	6.141568	0	67.5
_13_hsafll~e	3551	7.383189	6.184231	0	67.5
_14_hsafll~e	3551	7.270245	6.093506	0	67.5
_15_hsafll~e	3551	7.228972	6.036023	0	67.5
_16_hsafll~e	3551	7.319965	6.088639	0	67.5
_17_hsafll~e	3551	7.247781	6.039502	0	67.5
_18_hsafll~e	3551	7.209316	6.035757	0	67.5
_19_hsafll~e	3551	7.157074	6.030617	0	67.5
_20_hsafll~e	3551	7.154094	5.990593	0	67.5

Imputation model

```
. summ hsafll _1_-_20 if _m==0
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hsafllleure	3005	7.203113	6.106695	0	67.5
_1_hsaflle~e	3005	7.203113	6.106695	0	67.5
_2_hsaflle~e	3005	7.203113	6.106695	0	67.5
_3_hsaflle~e	3005	7.203113	6.106695	0	67.5
_4_hsaflle~e	3005	7.203113	6.106695	0	67.5
_5_hsaflle~e	3005	7.203113	6.106695	0	67.5
_6_hsaflle~e	3005	7.203113	6.106695	0	67.5
_7_hsaflle~e	3005	7.203113	6.106695	0	67.5
_8_hsaflle~e	3005	7.203113	6.106695	0	67.5
_9_hsaflle~e	3005	7.203113	6.106695	0	67.5
_10_hsafll~e	3005	7.203113	6.106695	0	67.5
_11_hsafll~e	3005	7.203113	6.106695	0	67.5
_12_hsafll~e	3005	7.203113	6.106695	0	67.5
_13_hsafll~e	3005	7.203113	6.106695	0	67.5
_14_hsafll~e	3005	7.203113	6.106695	0	67.5
_15_hsafll~e	3005	7.203113	6.106695	0	67.5
_16_hsafll~e	3005	7.203113	6.106695	0	67.5
_17_hsafll~e	3005	7.203113	6.106695	0	67.5
_18_hsafll~e	3005	7.203113	6.106695	0	67.5
_19_hsafll~e	3005	7.203113	6.106695	0	67.5
_20_hsafll~e	3005	7.203113	6.106695	0	67.5

Imputation model

```
. summ hsafll _1_-_20 if _m==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hsaflleure	0				
_1_hsaflle~e	546	7.006191	5.449275	.0167904	31.39901
_2_hsaflle~e	546	7.021949	5.422886	.0211766	37.49305
_3_hsaflle~e	546	7.722549	5.722267	.0107456	28.44278
_4_hsaflle~e	546	6.973286	5.393492	.0054002	28.82313
_5_hsaflle~e	546	7.537601	5.636957	.039393	27.01628
_6_hsaflle~e	546	8.074465	6.403761	.0026856	37.05117
_7_hsaflle~e	546	8.155709	6.023126	.0015128	30.95472
_8_hsaflle~e	546	7.440791	5.794226	.0818962	31.38698
_9_hsaflle~e	546	8.077889	6.314503	.0114371	35.26545
_10_hsafll~e	546	6.722987	5.186275	.0015738	32.35209
_11_hsafll~e	546	7.413693	5.55117	.0135646	31.7802
_12_hsafll~e	546	8.394843	6.24006	.0092697	32.71806
_13_hsafll~e	546	8.374269	6.511899	.0629378	42.61243
_14_hsafll~e	546	7.639719	6.012524	.0232201	34.17892
_15_hsafll~e	546	7.371293	5.634392	.02366	28.91562
_16_hsafll~e	546	7.963079	5.952828	.0097891	31.95087
_17_hsafll~e	546	7.493618	5.654418	.0142911	32.14428
_18_hsafll~e	546	7.243455	5.634546	.0181617	27.89192
_19_hsafll~e	546	6.903691	5.591911	.0005858	32.53796
_20_hsafll~e	546	6.88431	5.30333	.0000922	27.4847

Estimation model

```
. mi estimate, dots: logit ap hsafll edat sexe

Imputations (20):
  .........10.........20 done

Multiple-imputation estimates                   Imputations     =         20
Logistic regression                             Number of obs   =       3551
                                                Average RVI     =     0.0388
                                                Largest FMI     =     0.1340
DF adjustment:   Large sample                   DF:     min     =    1083.30
                                                        avg     = 3250588.13
                                                        max     = 8185974.12
Model F test:       Equal FMI                   F(   3,20871.8) =      51.33
Within VCE type:          OIM                   Prob > F        =     0.0000
```

ap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsaflleure	-.0301449	.0124206	-2.43	0.015	-.0545161	-.0057737
edat	.0822256	.0071976	11.42	0.000	.0681186	.0963326
sexe	-.8380656	.133865	-6.26	0.000	-1.100437	-.575694
_cons	-7.371945	.5097624	-14.46	0.000	-8.371062	-6.372829

Estimation model

```
. mi estimate, dftable

Multiple-imputation estimates                   Imputations     =         20
Logistic regression                             Number of obs   =       3551
                                                Average RVI     =     0.0388
                                                Largest FMI     =     0.1340
DF adjustment:   Large sample                   DF:     min     =    1083.30
                                                        avg     = 3250588.13
                                                        max     = 8185974.12
Model F test:       Equal FMI                   F(   3,20871.8) =      51.33
Within VCE type:          OIM                   Prob > F        =     0.0000
```

ap	Coef.	Std. Err.	t	P>|t|	DF	% Increase Std. Err.
hsaflleure	-.0301449	.0124206	-2.43	0.015	1083.3	7.36
edat	.0822256	.0071976	11.42	0.000	8185974.1	0.08
sexe	-.8380656	.133865	-6.26	0.000	335038.7	0.38
_cons	-7.371945	.5097624	-14.46	0.000	4480256.4	0.10

Estimation model

```
. mi estimate, vartable nocitable

Multiple-imputation estimates                   Imputations     =         20
Logistic regression

Variance information
```

	Imputation variance Within	Between	Total	RVI	FMI	Relative efficiency
hsaflleure	.000134	.000019	.000154	.152651	.134032	.993343
edat	.000052	7.5e-08	.000052	.001526	.001524	.999924
sexe	.017785	.000129	.01792	.007588	.007537	.999623
_cons	.259323	.00051	.259858	.002064	.00206	.999897

Estimation model

```
. mi estimate: logit ap hsafll edat sexe

Multiple-imputation estimates                   Imputations     =         20
Logistic regression                             Number of obs   =       3551
                                                Average RVI     =     0.0388
                                                Largest FMI     =     0.1340
DF adjustment:   Large sample                   DF:     min     =    1083.30
                                                        avg     = 3250588.13
                                                        max     = 8185974.12
Model F test:       Equal FMI                   F(   3,20871.8) =      51.33
Within VCE type:          OIM                   Prob > F        =     0.0000
```

ap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsaflleure	-.0301449	.0124206	-2.43	0.015	-.0545161	-.0057737
edat	.0822256	.0071976	11.42	0.000	.0681186	.0963326
sexe	-.8380656	.133865	-6.26	0.000	-1.100437	-.575694
_cons	-7.371945	.5097624	-14.46	0.000	-8.371062	-6.372829

```
. logit ap hsafll edat sexe, nolog

Logistic regression                             Number of obs   =       3005
                                                LR chi2(3)      =     152.97
                                                Prob > chi2     =     0.0000
Log likelihood =  -689.9607                     Pseudo R2       =     0.0998
```

ap	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hsaflleure	-.0313743	.0134025	-2.34	0.019	-.0576427	-.0051059
edat	.0859726	.008376	10.26	0.000	.069556	.1023892
sexe	-.9784195	.1589514	-6.16	0.000	-1.289959	-.6668804
_cons	-7.740686	.5948184	-13.01	0.000	-8.906508	-6.574863

(multiple) imputation model

```
. misstable patterns hsaflleure glicMI modaf estudis, freq

  Missing-value patterns
    (1 means complete)
```

Frequency	Pattern 1	2	3	4
1,393	1	1	1	1
1,442	1	1	1	0
258	1	1	0	0
249	1	1	0	1
75	1	0	1	0
67	1	0	1	1
15	0	1	1	1
13	1	0	0	0
11	0	1	1	0
11	1	0	0	1
6	0	1	0	0
6	0	1	0	1
2	0	0	0	1
2	0	0	1	1
1	0	0	0	0
3,551				

```
Variables are  (1) modaf  (2) estudis  (3) hsaflleure  (4) glicMI
```

```
. mi impute chained (regress) glicMI (truncreg, ll(0)) hsaflleure (ologit) modaf estudis = edat sexe ap,add(20)

Conditional models:
            modaf: ologit modaf i.estudis hsaflleure glicMI edat sexe ap
          estudis: ologit estudis i.modaf hsaflleure glicMI edat sexe ap
       hsaflleure: truncreg hsaflleure i.modaf i.estudis glicMI edat sexe ap , ll(0)
           glicMI: regress glicMI i.modaf i.estudis hsaflleure edat sexe ap

Performing chained iterations ...

Multivariate imputation                     Imputations =       20
Chained equations                                 added =       20
Imputed: m=1 through m=20                       updated =        0

Initialization: monotone                     Iterations =      200
                                                burn-in =       10

           glicMI: linear regression
       hsaflleure: truncated regression
            modaf: ordered logistic regression
          estudis: ordered logistic regression
```

	Observations per *m*			
Variable	Complete	Incomplete	Imputed	Total
glicMI	1745	1806	1806	3551
hsaflleure	3005	546	546	3551
modaf	3508	43	43	3551
estudis	3380	171	171	3551

(complete + incomplete = total; imputed is the minimum across *m* of the number of filled-in observations.)

```
. mi estimate: logit ap edat sexe hsaflleure glicMI i.modaf i.estudis

Multiple-imputation estimates                   Imputations     =         20
Logistic regression                             Number of obs   =       3551
                                                Average RVI     =     0.1368
                                                Largest FMI     =     0.4006
DF adjustment:   Large sample                   DF:     min     =     124.26
                                                        avg     = 1102818.94
                                                        max     = 3828931.50
Model F test:       Equal FMI                   F(  10,10782.1) =      17.85
Within VCE type:          OIM                   Prob > F        =     0.0000
```

ap	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
edat	.0749889	.0074665	10.04	0.000	.0603548	.089623
sexe	-.8562371	.1415123	-6.05	0.000	-1.133637	-.5788376
hsaflleure	-.0214534	.0148384	-1.45	0.151	-.0507989	.0078921
glicMI	.0104909	.002323	4.52	0.000	.0058931	.0150887
modaf						
2	1.737626	1.053416	1.65	0.099	-.3270309	3.802284
3	1.589586	1.021352	1.56	0.120	-.4122292	3.591401
4	1.78817	1.022668	1.75	0.080	-.2162226	3.792562
5	2.261608	1.024396	2.21	0.027	.2538285	4.269387
estudis						
2	-.2854247	.2258668	-1.26	0.206	-.7282658	.1574165
3	-.977909	.5682026	-1.72	0.085	-2.09205	.1362322
_cons	-9.625488	1.220385	-7.89	0.000	-12.01772	-7.233255

CCA

```
. logit ap edat sexe hsaflleure glicMI i.modaf i.estudis, nolog
note: 1.modaf != 0 predicts failure perfectly
      1.modaf dropped and 47 obs not used

note: 5.modaf omitted because of collinearity

Logistic regression                             Number of obs   =       1346
                                                LR chi2(9)      =      94.99
                                                Prob > chi2     =     0.0000
Log likelihood = -295.91873                     Pseudo R2       =     0.1383

------------------------------------------------------------------------------
          ap |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        edat |    .068255   .0128592     5.31   0.000     .0430513    .0934586
        sexe |  -1.187829   .2524404    -4.71   0.000    -1.682603   -.6930549
  hsaflleure |  -.0173571   .0193307    -0.90   0.369    -.0552446    .0205304
      glicMI |   .0108787   .0027662     3.93   0.000     .0054571    .0163003
             |
       modaf |
          1 |          0  (empty)
          2 |   .0179104   .4721069     0.04   0.970    -.9074022     .943223
          3 |  -.7415508   .2995775    -2.48   0.013    -1.328712   -.1543897
          4 |   -.300027   .2917828    -1.03   0.304    -.8719107    .2718566
          5 |          0  (omitted)
             |
     estudis |
          2 |  -.7524321   .3344429    -2.25   0.024    -1.407928   -.0969361
          3 |  -.9453168   .6959169    -1.36   0.174    -2.309289    .4186552
             |
       _cons |  -6.734488   1.106718    -6.09   0.000    -8.903615   -4.565361
------------------------------------------------------------------------------
```

Comparativa (simulació)

Efecte de la glicèmia sobre AP

		1000x						
	n	β	SE	IC95%		p	Ampl. IC	β/β_{real}
Dades reals	3512	7.9	1.6	4.7	11.0	0.0000	6.3	1
CCA	1745	9.7	2.2	5.4	14.0	0.0000	8.6	1.23
Mean imputation	3512	9.7	2.2	5.5	14.0	0.0000	8.5	1.23
Regression imputation	3512	15.0	2.1	10.8	19.2	0.0000	8.4	1.91
MI	3512	10.8	2.3	6.3	15.3	0.0000	9.0	1.37

Efecte sobre la variable principal d'estudi (HSAFLL)

		100x						
	n	β	SE	IC95%		p	Ampl. IC	β/β_{real}
Dades reals	3512	-2.05	1.48	-4.97	0.87	0.17	5.85	1
CCA	1346	-1.74	1.93	-5.52	2.05	0.37	7.58	0.846
Mean imputation	3527	-2.05	1.49	-5.00	0.89	0.17	5.89	1.001
Regression imputation	3527	-2.17	1.50	-5.13	0.80	0.15	5.93	1.056
MI	3551	-2.15	1.48	-5.08	0.79	0.15	5.87	1.046

Altres anàlisis

- Existeixen molts escenaris en que es pot aplicar MI:
  - Dades longitudinals (coming soon!), supervivència, clusters, models jeràrquics, patrons monotons, subgrups i restriccions, transformacions de variables, interaccions, no-linials, ...

•Stata dóna cobertura a la majoria d'aquests escenaris (altres: R, SAS, REALCOM, MLwiN, ...)

Agraïments

L'ús d'aquestes tècniques i part de la presentació estan inspirats en el curs

Missing data and new methods for multiple imputation of longitudinal electronic health records

Irene Petersen (University College London)

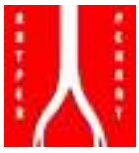
Jonathan Bartlett (London School of Hygiene and Tropical Medicine)

Agraïments

Anna Espinal

UAB Universitat Autònoma de Barcelona — Servei d'Estadística Aplicada

Maite Alzamora & grup ARTPER



Moltes gràcies per la vostra atenció!

Guillem Pera
Unitat de Suport a la Recerca Metropolitana Nord
CAP II Santa Coloma
Carrer Major 49-53, planta 1ª
08921 Santa Coloma de Gramenet
Telèfon: 93 462 86 05
gpera.bnm.ics@gencat.cat
www.idiapjgol.org