

MULTIPEL IMPUTATION

-

**Ett sätt att hantera problemet
med "missing data"**

Pär-Ola Bendahl

IKVL, Avdelningen för Onkologi

Lunds Universitet

Par-Ola.Bendahl@med.lu.se

Översikt

- Introduktion till problemet
- Enkla metoder att hantera "missing data"
- Multipel imputation
- Exempel och jämförelse med andra metoder

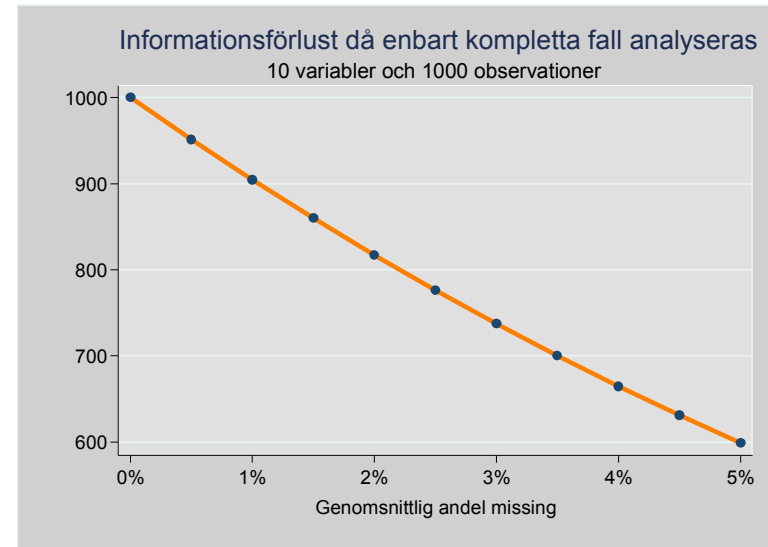


Introduktion till problemet

Missing data – Ansats 1

Complete case analysis

id	ålder	metastasfri_5y	hist_grad	tum_storl	upa	her2_pos	ki67
1	45	0	3	20			85
2	36	0	3	15	0,08	Nej	35
3	47	0	1	26	0,08	Nej	3,5
4	46	0	2	27	0,05	Nej	15
5	44	1	2	7	0,04	Nej	35
6	48	0	2	10	0,22	Nej	15
7	50	0	1	13	0,12	Nej	3,5
8	39	0	3	21	0,08	Nej	25
9	45	0	3	17	0,14	Nej	45
10	42	0	1	23	0,15		
11	35	0	1	18		Nej	0,5
12	43	0	1	22	0,81	Nej	0,5
13	47	0	1	18	0,36	Nej	0,5
14	40	0	3	18	0,09	Ja	25
15	30	0	3	15	0,55	Nej	65



- Default i de flesta programpaket
- Ineffektivt utnyttjande av data
- Kan leda till bias i skattningar och SE

A low-angle photograph of a bare tree with many thin, dark branches reaching upwards against a clear, light blue sky. The tree is on the left side of the frame, and its branches spread out across the upper half of the image.

Varför saknas data?

Ex: Saknade blodtrycksvärden

- Sköterskan glömde att ta blodtrycket
- Nedtecknat blodtryck oläsbart
- Blodtryck tas på alla äldre men bara på ett urval av de yngre
- Mätning endast om vissa symptom
- Mätning underkänd
 - Dåligt kalibrerat mätinstrument

Spelar det någon roll varför data saknas?

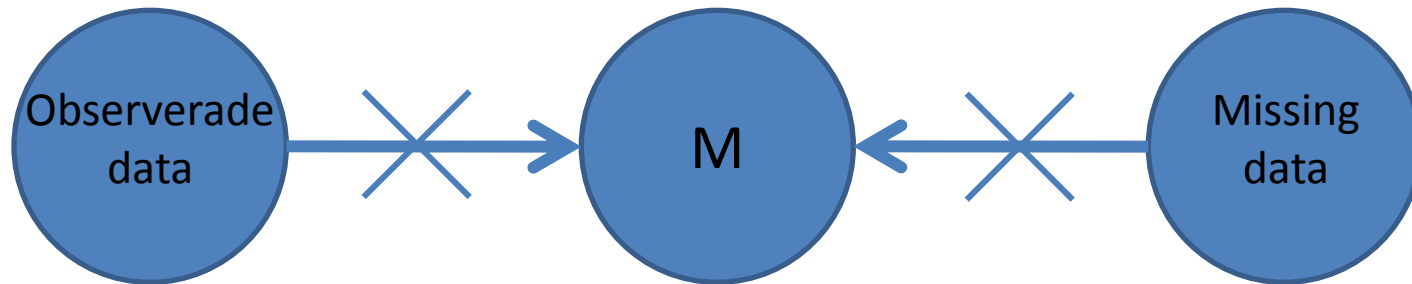


Missing data-mekanismer

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

Missing completely at random (MCAR)

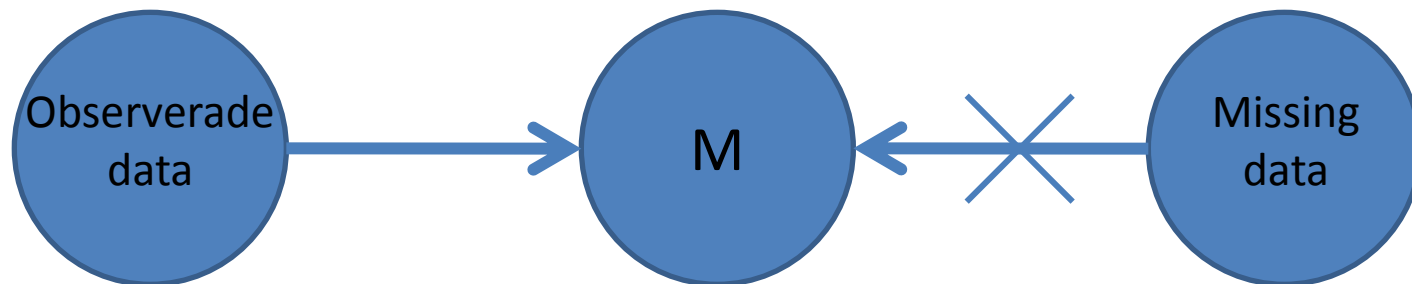
- $M = P(\text{data missing})$ beror varken på observerade eller saknade data.



- Ex: Blodprov som tappats i golvet

Missing at random (MAR)

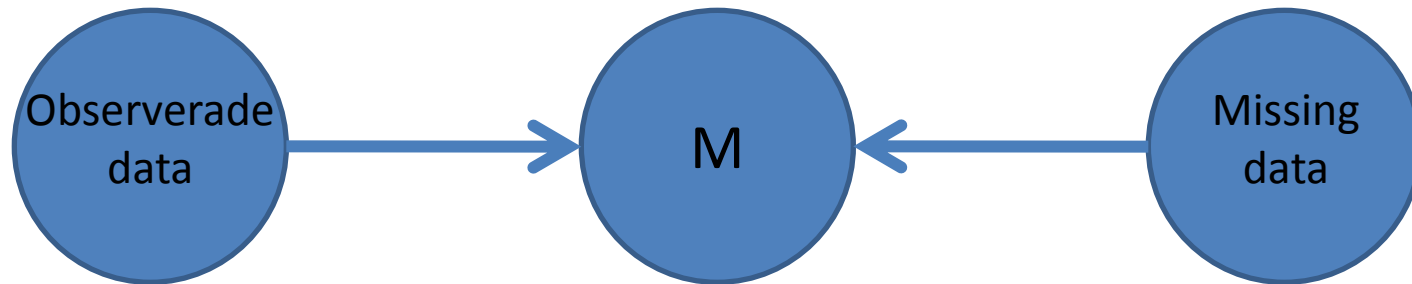
- $M = P(\text{data missing})$ beror på observerade data men **INTE** på det okända saknade mätvärdet



- Ex: Högre sannolikhet att blodtrycket mäts på äldre individer. Ålder mäts.

Missing not at random (MNAR)

- $M = P(\text{data missing})$ beror på både observerade data och det okända saknade mätvärdet



- Ex: Individer med högt blodtryck har högre sannolikhet att komma med i stickprovet än andra individer av samma ålder

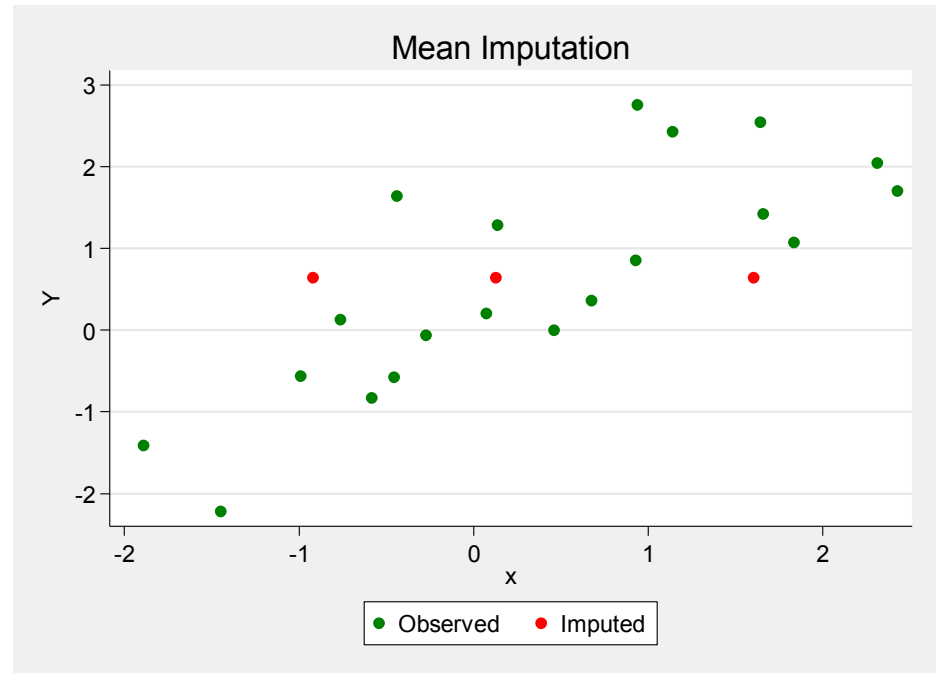
Hur skiljer man på MCAR, MAR och MNAR

- Att beskriva data räcker inte
- Kännedom om datainsamlingsprocessen nödvändig
- MCAR vs MAR
 - Undersök mönster i "missing data"
 - Kan vissa variabler prediktera "missingness"
- MAR vs MNAR
 - Mätning av fler förklarande variabler (från MNAR mot MAR)

Enkla imputationsmetoder



Mean Imputation



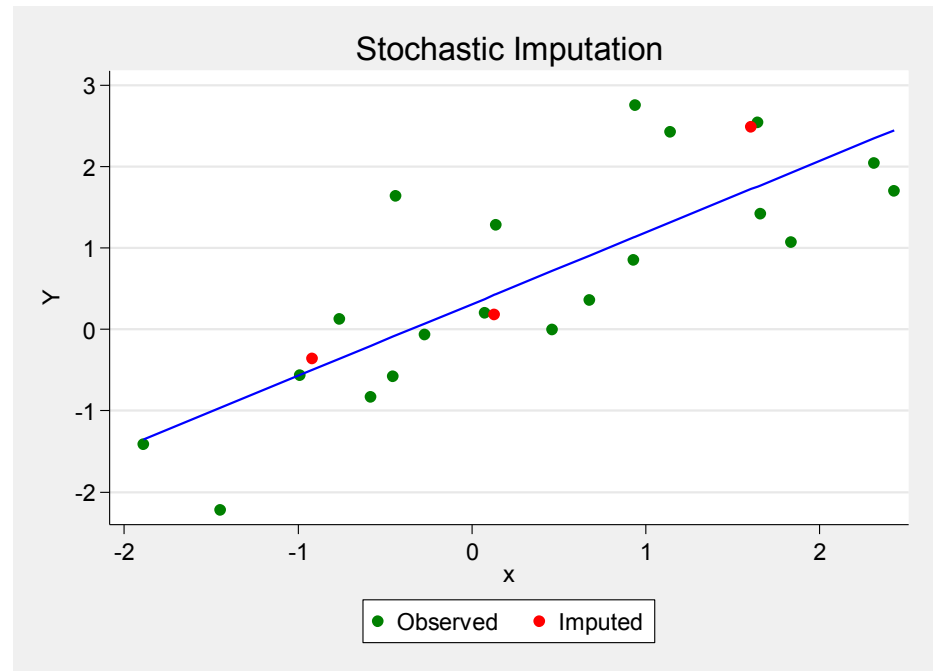
- Ersätt alla saknade värden för Y med medelvärdet för Y
 - Variansen i Y underskattas
 - Associationer kan förvrängas
 - Bias i parameterskattningar och SE (speciellt under MNAR)

Regression Imputation



- Anpassa en regressionsmodell till fullständigt observerade data
- Använd denna för att prediktera saknade värden
 - Bättre
 - Associationer mellan variabler bibehålls
 - Korrelationer överskattas

Stokastisk imputation



- Addera ett slumpmässigt normalfördelat fel till det predikterade värdet
 - Ännu lite bättre

A close-up photograph of pink cherry blossoms and green leaves, with sunlight filtering through the foliage, creating a warm, golden glow. The background is a soft, out-of-focus green.

Multipel Imputation

Idé

- Utnyttja korrelationsstrukturen i data
- Använd stokastisk imputation för att skapa ***m*** fullständiga imputerade dataset
- Analysera varje dataset separat
- Kombinera resultaten (Rubins regler)



MICE

2012-03-22

Multiple Imputation by Chained Equations

- Van Buuren et al. 1999
- Implementerat i Stata av Patrick Royston (uvis/ice)

Univariat: X_1 har missing values (förutsätts vara MAR)
 $X_2, \dots, X_k$ fullständigt observerade
Anpassa regression av X_1 på $X_2, \dots, X_k$ (fullständiga data)
Imputera de saknade X_1 -värdena genom slumpmässig
draging från fördelningen för icke observerade data givet
observerade data
Normalfördelade slumpfel förutsätts
Proper imputation – Beaktar både estimations- och prediktionsfel

MICE - Multivariat

- Missing values för mer än en variabel
 1. Fyll i alla hål **helt slumpmässigt** (initialisera)
 2. Univariat Stok. Imp. – Regression av X_1 på $X_2, \dots, X_k$
Univariat Stok. Imp. – Regression av X_2 på övriga
...
Univariat Stok. Imp. – Regression av X_k på övriga
 3. Upprepa cykeln i steg 2 cirka 10 gånger. Detta ger **ett** imputerat dataset
 4. Upprepa steg 1-3 **m** gånger (**m** = 3-50)

Regressionsmodeller i MICE

- Kontinuerliga variabler imputeras med **Linjär Regression**
 - Variabeltransformation kan vara nödvändig
- Binära variabler – **Logistisk regression**
- Kategoriska data > 2 klasser – **Multinomial logistisk regression**
- Ordnade kategoriska data – **Ordinal logistisk regression**
- **mi impute chained** i Stata 12

Några kommentarer

- Antagande om **multivariat normalfördelning**
 - **mi impute mvn** i Stata 12
 - Fungerar sämre än MICE för kategoriska data
- **Predictive Mean Matching (PMM)**
 - **mi impute pmm** i Stata 12
 - Semiparametrisk metod
 - Prediktera "missing value" med regressionsmodell
 - Vilka verkliga observationer ligger närmast denna prediktion (kNN)?
 - Välj slumpmässigt en av dessa
 - Fördel: Aldrig orimliga prediktioner
 - Fungerar ofta bra vid skeva fördelningar

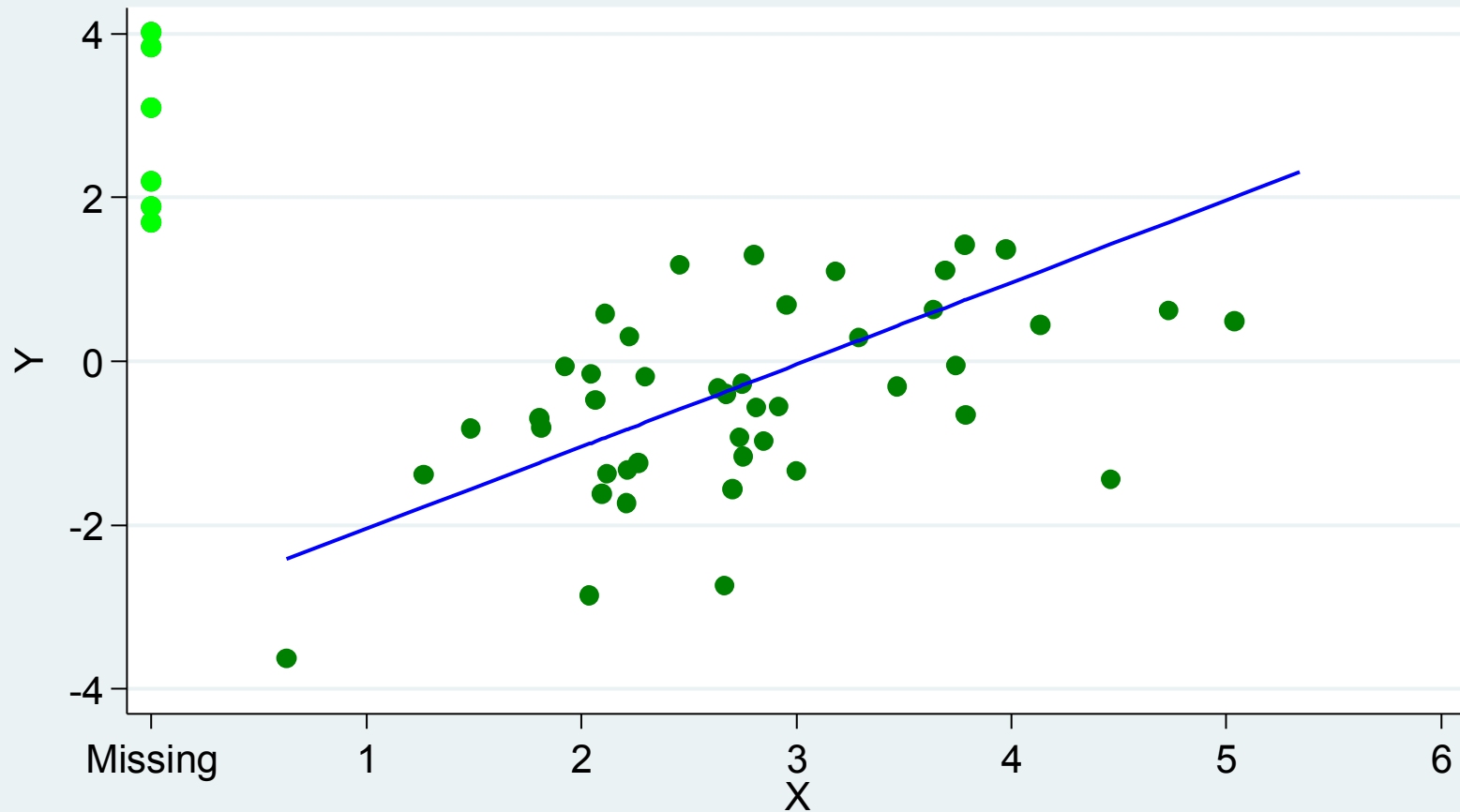
A close-up photograph of several large, dark rocks covered in a thick, vibrant green moss. The moss has a textured, carpet-like appearance. In the background, there are blurred, dry, brownish branches and foliage, suggesting a forest or woodland setting. The lighting is soft and natural, highlighting the texture of the moss.

Bör utfallsvariabeln Y betraktas som en variabel i mängden vid imputation av kovariatvärden?

Svaret är ja!

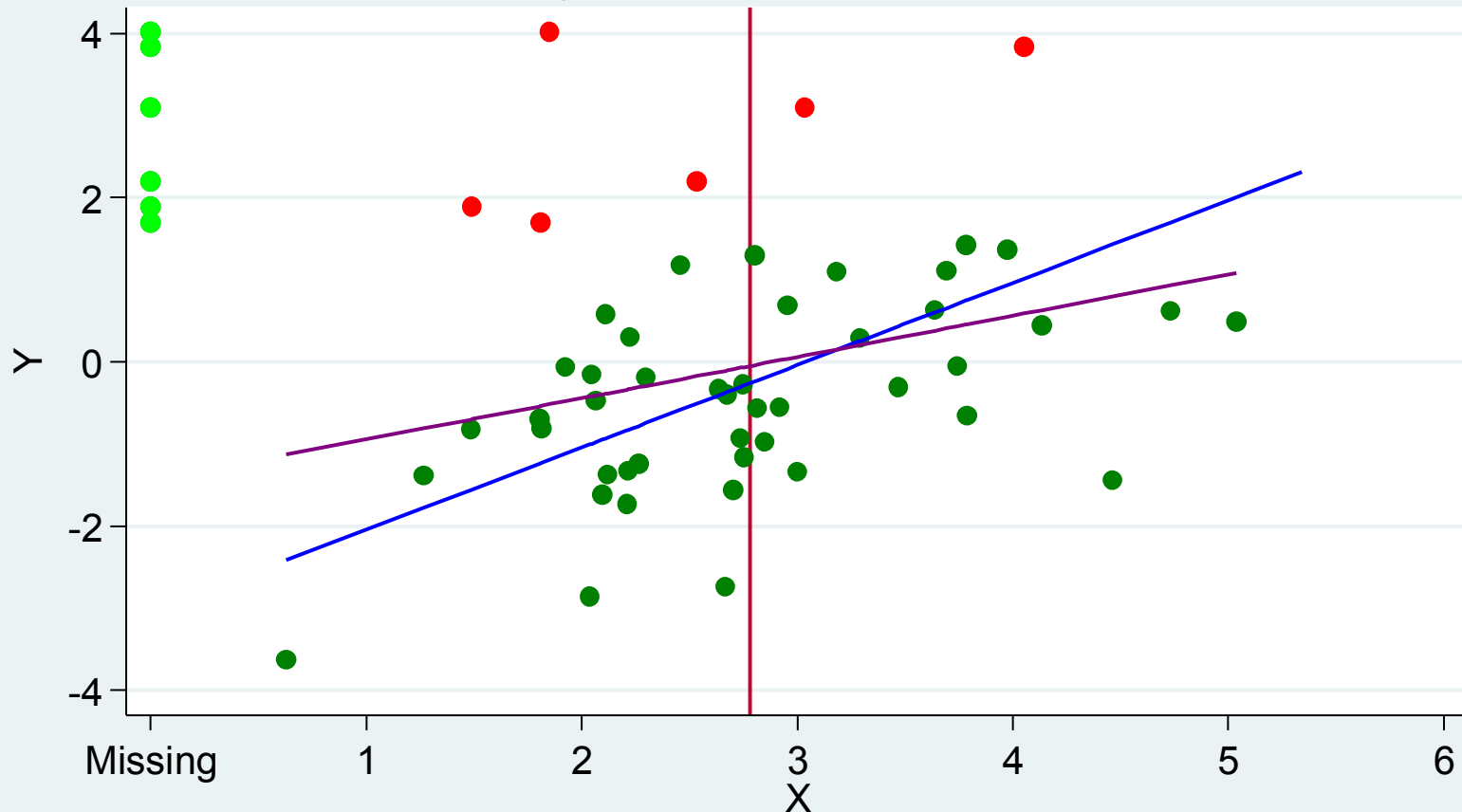
- Om utfallsvariabeln Y ej med i imputationsmodellen så underskattas associationen mellan kovariaterna och utfallsvariabeln
- Ingen association alls mellan Y och kovariaterna i den del av data som imputerats om Y inte beaktas vid imputation
- Associationen späds alltså ut vid felaktig imputation
- Kan leda till bias i parameterskattningar
- Moons et al. 2006 – bra exempel

Observerade data: X MAR



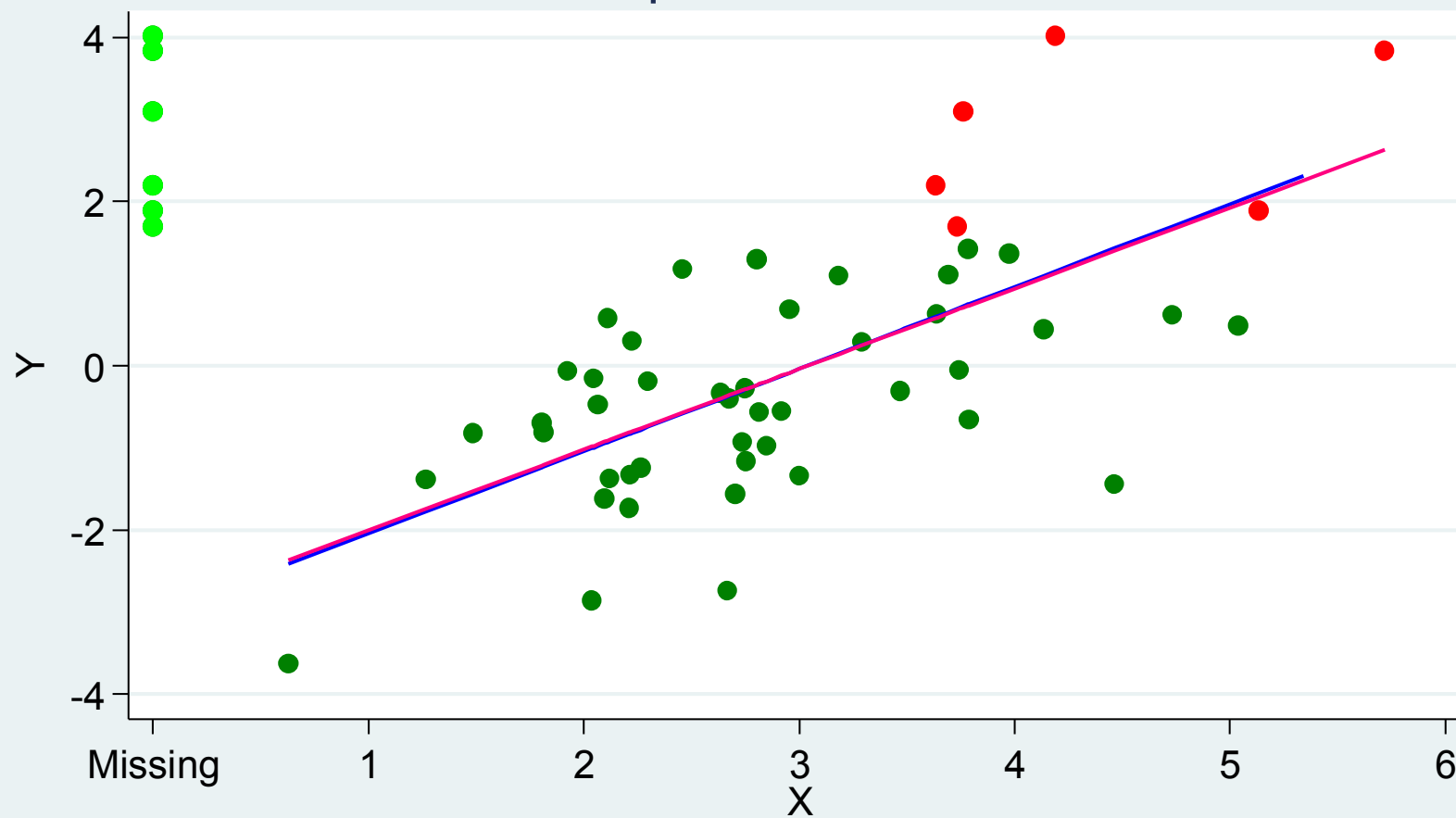
- Observerade data
- Sann regression
- Observationer för vilka X saknas

Y ej i imputationsmodellen



- Observerade data
- Imp. utan hänsyn till Y
- Sann regression
- Biased regression

Y i imputationsmodellen



● Observerade data

● Imp. med hänsyn till Y

— Sann regression

— Unbiased regression

Hur bör överlevnadsdata imputeras?

2012-03-22

29



Congenialitet

- Imputationsmodellen bör ”matcha” analysmodellen
- Om data skall analyseras med Cox-regression bör man ta med Nelson-Aalen-skattningen av den kumulativa hazarden och händelseindikatorn i imputationsmodellen
- Idé: imputera en gång per projekt

Ett exempel ...



Bröstcancer – tid till fjärrspridning

- N=1314
- 303 har fått fjärrmetastas
- Prognosfaktorer
 - Antal positiva lymfkörtlar
 - Diagnosålder (år)
 - Tumörstolek (mm)
 - Histologisk grad (1-3)
 - Östrogenreceptorstatus (+ vs -)

Resultat av Cox-regression

Faktor	HR	95% CI	P
Antal positiva körtlar	1.065	1.049-1.082	<0.001
Ålder	0.979	0.969-0.988	<0.001
Storlek	1.018	1.010-1.026	<0.001
Grad	1.899	1.543-2.338	<0.001
ER-status + vs -	0.833	0.647-1.073	0.16

Nu simulerar jag olika bortfallsmönster för ER beroende på grad

MCR

- 5% bortfall för ER i grad 1-2 och 20% i grad 3
- Grad kan prediktera "ER-missingness"
- Complete case analysis

Före bortfall (n=1314)

Efter bortfall (n=1092)

Faktor	HR	95% CI	P	HR	95% CI	P
Antal positiva körtlar	1.018	1.010-1.026	<0.001	1.061	1.043-1.079	<0.001
Ålder	0.979	0.969-0.988	<0.001	0.976	0.965-0.988	<0.001
Storlek	1.018	1.010-1.026	<0.001	1.026	1.016-1.037	<0.001
Grad	1.899	1.543-2.338	<0.001	1.752	1.384-2.219	<0.001
ER-status + vs -	0.833	0.647-1.073	0.16	0.858	0.634-1.162	0.32

Justera för missingness i analysen?

- Inför missing som en tredje ER-kategori
- Parametrisera ER-effekten som 2 dummyvariabler
 - ER+ vs ER- respektive ER-missing vs ER-
- Därmed kan vi få med alla obs i analysen

Före bortfall				Efter justering för bortfall		
Faktor	HR	95% CI	P	HR	95% CI	P
Antal positiva körtlar	1.018	1.010-1.026	<0.001	1.066	1.049-1.083	<0.001
Ålder	0.979	0.969-0.988	<0.001	0.979	0.969-0.988	<0.001
Storlek	1.018	1.010-1.026	<0.001	1.017	1.010-1.025	<0.001
Grad	1.899	1.543-2.338	<0.001	1.927	1.546-2.401	<0.001
ER-status + vs -	0.833	0.647-1.073	0.16	0.890	0.660-1.200	0.44

MICE

- 20 imputerade dataset
- Imputation med logistisk regressionsmodell
- Nelson-Aalen skattningen och händelseindikatorn med i modellen

Före bortfall

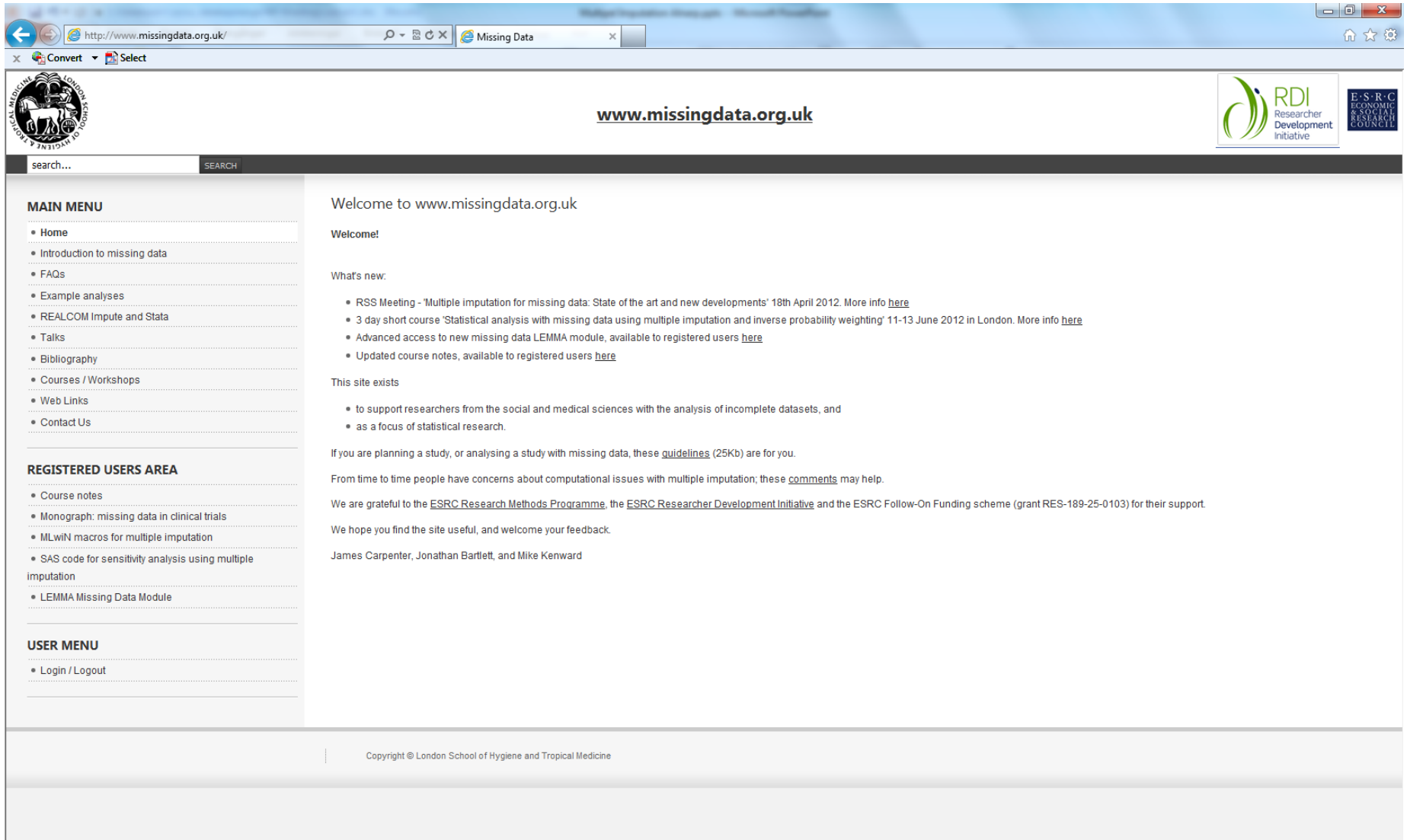
Skattningar efter MI

Faktor	HR	95% CI	P	HR	95% CI	P
Antal positiva körtlar	1.018	1.010-1.026	<0.001	1.065	1.049-1.082	<0.001
Ålder	0.979	0.969-0.988	<0.001	0.979	0.969-0.988	<0.001
Storlek	1.018	1.010-1.026	<0.001	1.017	1.010-1.026	<0.001
Grad	1.899	1.543-2.338	<0.001	1.905	1.542-2.354	<0.001
ER-status + vs -	0.833	0.647-1.073	0.16	0.837	0.627-1.118	0.23

Sammanfattning

- Lär känna dina data
 - Studera mönster i missing data
- Imputera inte slentrianmässigt
 - Väl genomtänkt congenial imputation
- Jämför imputerade och observerade värden
- Jämför MI-analys med complete case-analys
- Alla variabler i analysmodellen skall vara med i imputationsmodellen – även den beroende

Mer info om MI



The screenshot shows a web browser window with the URL <http://www.missingdata.org.uk/>. The browser's address bar shows the URL, and the page title is "Missing Data". The website's header includes the London School of Hygiene & Tropical Medicine logo on the left, the URL www.missingdata.org.uk in the center, and logos for the RDI (Researcher Development Initiative) and ESRC (Economic and Social Research Council) on the right. Below the header is a search bar with the text "search..." and a "SEARCH" button. The main content area is divided into two columns. The left column contains a "MAIN MENU" with links to Home, Introduction to missing data, FAQs, Example analyses, REALCOM Impute and Stata, Talks, Bibliography, Courses / Workshops, Web Links, and Contact Us. Below this is a "REGISTERED USERS AREA" with links to Course notes, Monograph: missing data in clinical trials, MLwiN macros for multiple imputation, SAS code for sensitivity analysis using multiple imputation, and LEMMA Missing Data Module. At the bottom of the left column is a "USER MENU" with a link to Login / Logout. The right column contains a "Welcome to www.missingdata.org.uk" message, a "Welcome!" greeting, a "What's new:" section with four bullet points about RSS meetings, short courses, advanced access to the LEMMA module, and updated course notes, a "This site exists" section with two bullet points about supporting researchers and statistical research, a paragraph about guidelines for planning a study, a paragraph about computational issues with multiple imputation, a paragraph about gratitude to ESRC funding schemes, a paragraph about feedback, and a list of names: James Carpenter, Jonathan Bartlett, and Mike Kenward. The footer of the website contains the text "Copyright © London School of Hygiene and Tropical Medicine".

http://www.missingdata.org.uk/

Missing Data

Convert Select

 www.missingdata.org.uk  

search... SEARCH

MAIN MENU

- Home
- Introduction to missing data
- FAQs
- Example analyses
- REALCOM Impute and Stata
- Talks
- Bibliography
- Courses / Workshops
- Web Links
- Contact Us

REGISTERED USERS AREA

- Course notes
- Monograph: missing data in clinical trials
- MLwiN macros for multiple imputation
- SAS code for sensitivity analysis using multiple imputation
- LEMMA Missing Data Module

USER MENU

- Login / Logout

Welcome to www.missingdata.org.uk

Welcome!

What's new:

- RSS Meeting - 'Multiple imputation for missing data: State of the art and new developments' 18th April 2012. More info [here](#)
- 3 day short course 'Statistical analysis with missing data using multiple imputation and inverse probability weighting' 11-13 June 2012 in London. More info [here](#)
- Advanced access to new missing data LEMMA module, available to registered users [here](#)
- Updated course notes, available to registered users [here](#)

This site exists

- to support researchers from the social and medical sciences with the analysis of incomplete datasets, and
- as a focus of statistical research.

If you are planning a study, or analysing a study with missing data, these [guidelines](#) (25Kb) are for you.

From time to time people have concerns about computational issues with multiple imputation; these [comments](#) may help.

We are grateful to the [ESRC Research Methods Programme](#), the [ESRC Researcher Development Initiative](#) and the ESRC Follow-On Funding scheme (grant RES-189-25-0103) for their support.

We hope you find the site useful, and welcome your feedback.

James Carpenter, Jonathan Bartlett, and Mike Kenward

Copyright © London School of Hygiene and Tropical Medicine



Tack till:
MRC, Cambridge
Linn - för foton
Er - för visat intresse