Puuduvad andmed

Biostatistika 10. loeng

Puuduvaid andmeid otsimas.
Juhuslike arvude generaator (2024).

Elulemusanalüüs – üks variant analüüsida puuduvaid vaatluseid sisaldavat andmestikku

Eluiga	Jälgimisaeg
8.9	1.8
7.9	1.1
1.7	1.7
0.8	0.8
3.3	2.3
3.8	1.9
2.5	2.5
…	

Tegelik keskmine: **6.3**

Teadaolevate eluigade keskmine: **1.9**

Jälgimisaegade keskmine: **2.0**

Elulemusanalüüsi abil hinnatud keskmine: **6.2**

Puuduvad andmed üldisemalt...

age	bmi	hyp	chl
2	22.7	1	187
1	NA	1	187
1	NA	NA	NA
1	20.4	1	113
2	22.0	1	238
1	NA	NA	NA
1	29.6	1	NA
3	21.7	1	206
	.....		

Puuduvad andmed võivad tekitada nihet tulemustes – elulemusanalüüsis puudusid meil suurema tõenäosusega nende inimeste eluead, kes elasid kauem. Küsitlusuuringus võivad oma kaalu aga suurema tõenäosusega märkimata jätta näiteks need, kes oma kaalu häbenevad; jne.

Mudelite headuse võrdlemisel on vajalik, et katsetatakse mudeleid samade vaatluste peal. Kui aga lisame mudelisse tunnuse, mille väärtustest testandmestikul osad mõõtmised on puudu – siis me ei saa nende inimeste jaoks prognoose leida ja rikkama mudeli prognoosivõimet lihtsamaga adekvaatselt võrrelda

Mõnede statistikameetodite – näiteks peakomponentanalüüs, mitmene regressioonanalüüs jne kasutamine võib muutuda peaaegu võimatuks, kui kord on ühes tunnuses üks väärtus puudu, siis puudub teise tunnuse väärtus jne …

Puuduvate andmete liigitamisest

Palk	Elukoht	…
~~2070~~	Pärnu	…
~~2720~~	Pärnu	…
1450	Antsla	
1260	Türi	
1750	Tallinn	
2700	Tallinn	
1300	Kohila	
1320	Oandu	

→

M
1
1
0
0
0
0
0
0

Puuduvate andmete liigitamisest

Palk	Elukoht	...
~~2670~~	Pärnu	...
~~2720~~	Pärnu	...
1450	Antsla	
1260	Türi	
1750	Tallinn	
2700	Tallinn	
1300	Kohila	
1320	Oandu	

X_{obs} = *Elukoht, Vanus, Sugu*

X_{mis} = *Palk, Lemmikpoliitik, ravitulemus*

$P(M|X_{obs}, X_{mis}) = P(M)$ MCAR
$P(M|X_{obs}, X_{mis}) = P(M|X_{obs})$ MAR
$P(M|X_{obs}, X_{mis}) \neq P(M|X_{obs})$ NMAR

MCAR – täiesti juhuslik puudumine

Ka kõigi andmete teadmine ei aitaks mul otsustada, millised neist võiksid puududa andmebaasist

$P(M \mid Palk, Elukoht, ...) = P(M)$

ehk

$P(M \mid X_{obs}, X_{mis}) = P(M)$

MCAR – täiesti juhuslik puudumine

Ka kõigi andmete teadmine ei aitaks mul otsustada, millised neist võiksid puududa andmebaasist

$P(M \mid Palk, Elukoht, ...) = P(M)$

ehk

$P(M \mid X_{obs}, X_{mis}) = P(M)$

Näide 1

valim

vanus	sugu	perekonnaseis	kaal	pikkus
22	1	3	52	164
21	2	1	88	181
20	1	1	80	172
19	1	1	47	164
19	1	1	61	168
19	1	1	51	165
19	1	1	53	160
20	1	1	70	170
22	1	3	52	167
20	1	1	46	159
.....			.....	

Juhusliku valimi korral see, kas keegi on jäänud valimist välja (M=1) või sattus valimisse (M=0) ei sõltu ühestki tunnusest – seega on tegemist täiesti juhusliku puudumisega.

MCAR – täiesti juhuslik puudumine

Ka kõigi andmete teadmine ei aitaks mul otsustada, millised neist võiksid puududa andmebaasist

$P(M \mid Palk, Elukoht, ...) = P(M)$

ehk

$P(M \mid X_{obs}, X_{mis}) = P(M)$

Näide 2

kosmilised kiired

vanus	sugu	perekonnaseis	kaal	pikkus
22	1	3	52	164
21	2	1	88	181
20	1	1	80	172
19	1	1	47	164
19	1	1	61	168
19	1	1	51	165
19	1	1	53	160
20	1	1	70	170
22	1	3	52	167
20	1	1	46	159
.....			.....	

MCAR – täiesti juhuslik puudumine

Ka kõigi andmete teadmine ei aitaks mul otsustada, millised neist võiksid puududa andmebaasist

$P(M \mid Palk, Elukoht, ...) = P(M)$

ehk

$P(M \mid X_{obs}, X_{mis}) = P(M)$

See kas mingi bait kettal/arvuti mälus sai tabamuse kosmilise kiirega (M=1) või mitte (M=0) ei sõltu seal baidis säilitatavast informatsioonist – järelikult tegemist MCAR puudumisega.

Näide 2

vanus	sugu	perekonnaseis	kaal	pikkus
22	1	3	52	164
21	2	1	88	181
20	1	1	80	172
19	1	1	47	164
19	1	1	61	168
19	1	1	51	165
19	1	1	53	160
20	1	1	70	170
22	1	3	52	167
20	1	1	46	159
.....		.....		

MCAR – täiesti juhuslik puudumine

Ka kõigi andmete teadmine ei aitaks mul otsustada, millised neist võiksid puududa andmebaasist

$P(M \mid X_{obs}, X_{mis}) = P(M)$

Statistikatarkvara (enamasti) eeldab, et puuduvate andmete tekkemehhanismiks on MCAR.

Kui MCAR, siis
- olemasolevate vaatluste keskmine ≈ kõigi vaatluste keskmine
- olemasolevate vaatluste hajuvus ≈ kõigi vaatluste hajuvus
- korrelatsioonikordajad kõigil andmetel/olemasolevatel andmetel sarnased
- puuduvatest andmetest tingitud ebatäpsust lihtne kirjeldada (standardviga, usaldusintervallid, ...)

Näide 2

vanus	sugu	perekonnaseis	kaal	pikkus
22	1	3	52	164
21	2	1	88	181
20	1	1	80	172
19	1	1	47	164
19	1	1	61	168
19	1	1	51	165
19	1	1	53	160
20	1	1	70	170
22	1	3	52	167
20	1	1	46	159
.....		.....		

MAR – juhuslik puudumine

Andmebaasis esineb tunnus/tunnused, mis võimaldavad kirjeldada puudumistõenäosuse muutumist

$P(M \mid Palk, Elukoht, ...) = P(M \mid Elukoht)$

ehk

$P(M \mid X_{obs}, X_{mis}) = P(M \mid X_{obs})$

$P(M \mid Elukoht, Pikkus, haigus) = P(M \mid Elukoht)$

$P(M \mid Elukoht=Tartu) = 1/4$
$P(M \mid Elukoht \neq Tartu) = 3/4$

mõõdame Tartus palju inimesi (nad on käepärast võtta), kaugematest kohtadest vähem – kasutame kihtvalimit

Näide

Elukoht	vanus	sugu	pikkus	haigus
Tartu	20	1	167	0
Tartu	21	2	192	0
Tartu	19	1	157	0
Tartu	64	1	161	1
Valga	56	2	176	1
Valga	40	1	157	0
Valga	8	2	146	0
Valga	76	1	180	1
Põlva	52	2	185	1
Põlva	10	2	105	0
Põlva	89	1	151	1
Põlva	4	1	100	0

MAR – juhuslik puudumine

Andmebaasis esineb tunnus/tunnused, mis võimaldavad kirjeldada puudumistõenäosuse muutumist

$P(M \mid Palk, Elukoht, ...) = P(M \mid Elukoht)$

ehk

$P(M \mid X_{obs}, X_{mis}) = P(M \mid X_{obs})$

Alternatiiv: ääremaade sidekanalid kehvemad, seega sealt andmete ülekandmisel esineb rohkem vigu või tõrkeid. Seega ka Kihnu vaatlused suurema tõenäosusega puudu kui näiteks Tartus tehtud mõõtmised.

Sellisel juhul ei sõltu vaatluse puudumine mitte mõõdetud väärtusest, kuid Kihnus tehtud mõõtmised – näiteks õhu puhtuse vaatlused – võivad olla süstemaatiliselt teistsugused kui Tartus tehtud mõõtmised.

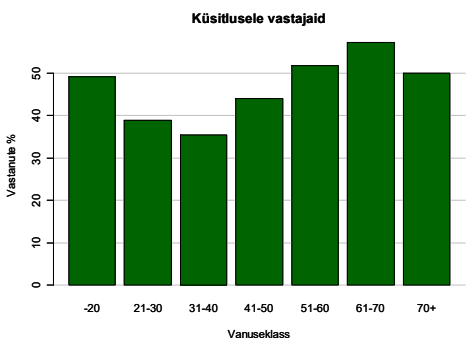
Näide

Elukoht	vanus	sugu	pikkus	haigus
Tartu	20	1	167	0
Tartu	21	2	192	0
Tartu	19	1	157	0
Tartu	64	1	161	1
Valga	56	2	176	1
Valga	40	1	157	0
Valga	8	2	146	0
Valga	76	1	180	1
Põlva	52	2	185	1
Põlva	10	2	105	0
Põlva	89	1	151	1
Põlva	4	1	100	0

Enesehinnang tervisele

Vastamise protsent:

Mehed	38%
Naised	50%



NMAR – mittejuhuslik puudumine

Andmetesse auke tekitav protsess on mittejuhuslik, kui andmete kadumine on otseselt seotud nende väärtustega (ja need väärtused pole teiste, olemasolevate andmete abil prognoositavad):

$$P(M \mid Ravitulemus, Sugu, ...) \neq P(M \mid Sugu, ...)$$

ehk

$$P(M \mid X_{mis}, X_{obs}) \neq P(M \mid X_{obs})$$

Millal on tegemist NMAR-ga?

Inimene otsustab, kas varjata oma palka, oma palganumbri põhjal. Inimese kohta kogutud andmete põhjal pole võimalik täpselt prognoosida tema tegelikku palka;

Valime valimisse need inimesed, kelle palganumbrid annavad meile soovitava tulemuse.

Uuringust puuduvad kehva tervisega inimesed (sest nad ei tulnud uuringu päeval kohale), korjatud informatsiooni põhjal pole võimalik inimese tervislikku seisundit võimalik täpselt tuvastada

Kas MCAR, MAR või NMAR?

- MCAR vs MAR. Kontrollitav olemasolevate andmete põhjal.
- MAR vs NMAR. Pole kontrollitav olemasolevate andmete põhjal. Otsuse tegemiseks vaja uurida/teada, kuidas ja miks tekivad puuduvad andmed.

Mida teha puuduvate andmetega?

- Unusta inimesed (firmad, vallad,...) kelle andmetes esineb puuduvaid väärtuseid. *Complete Case Analysis, Listwise deletion*
- Andmete analüüsimisel ja statistikute leidmisel arvesta puuduvaid andmeid ja puuduvate andmete tekkepõhjuseid
  Full Likelihood, *Sensitivity Analysis, ...*
- Aukude täitmine ehk imputeerimine
  Mean, Hot-Deck, Multiple Imputation, ...

Ignoreerin puuduvaid väärtuseid?

Statistik	Puuduvaid väärtuseid sisaldav valim	Perfektne valim	Tegelik väärtus
keskmine	1012	1317	1328
UI	(997..1027)	(1304..1329)	
sd	608	630	629
r	0.79	0.77	0.77
UI	(0.78..0.80)	(0.76..0.78)	

Imputeerimine ehk aukude täitmine

- keskmise või mediaani kasutamine;
- puuduva väärtuse prognoosimine regressioonmudeli abil;
- puuduv väärtus asendatakse juhusliku arvuga;
- ....

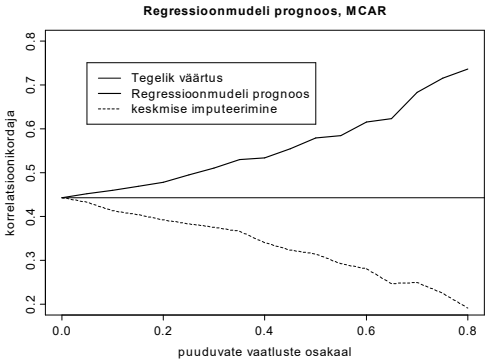
Asendan keskmisega?

Statistik	Puuduvad väärtused asendatud keskmisega	Perfektne valim	Tegelik väärtus
keskmine	1012	1317	1328
UI	(1003..1021)	(1304..1329)	
sd	468	630	629
r	0.6	0.77	0.77
UI	(0.59..0.61)	(0.76..0.78)	

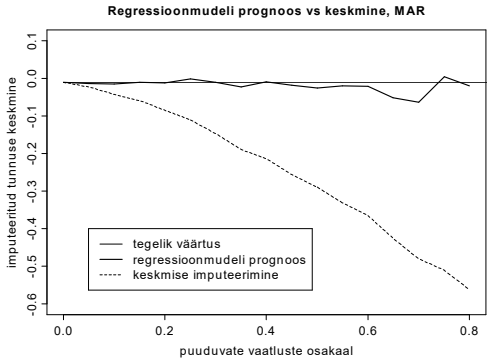
Asendan keskmisega?

Statistik	Puuduvad väärtused asendatud prognoosiga	Perfektne valim	Tegelik väärtus
keskmine	1313	1317	1328
UI	(1301..1325)	(1304..1329)	
sd	621	630	629
r	0.79	0.77	0.77
UI	(0.78..0.80)	(0.76..0.78)	

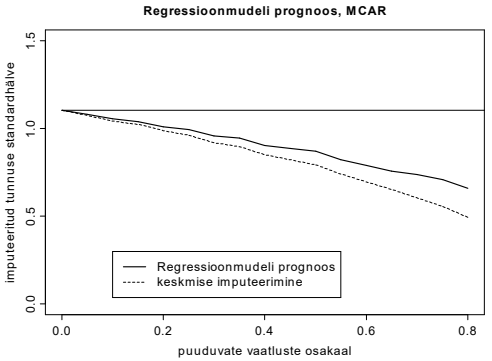
Regressioonimudeli prognoos, seos tunnuste vahel



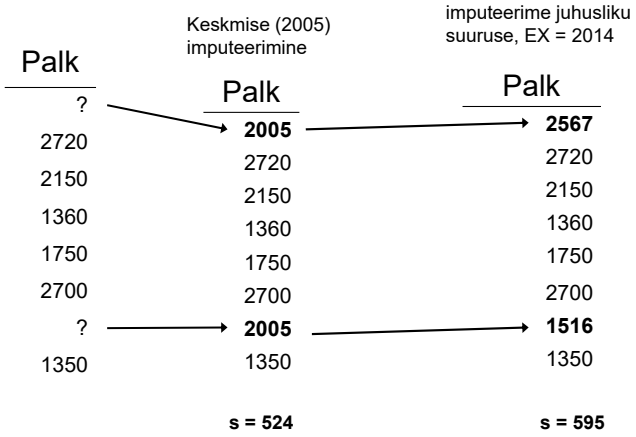
Regressioonmudeli prognoos, MAR



Regressioonimudeli prognoos, hajuvus



Augu täitmine juhusliku väärtusega



Probleem:
Peaksime teadma palganumbrite hajuvust, selleks et korrektselt juhuslikke palkasid genereerida…

Puuduvate väärtusteta valimi pealt hinnang
s=640

Mitmene imputatsioon I

(Multiple imputation)

1. Leitakse, milline on puuduvat väärtust sisaldava tunnuse tinglik jaotus, tingimusel et on teada nende tunnuste väärtused, mida me teame

$$P(\textit{palk} \mid \textit{piirkond}, \textit{haridus}, \ldots)$$

2. Augu täiteks genereerime juhusliku suuruse jaotusest

$$\textit{palk}_i \sim P(\textit{palk} \mid \textit{piirkond}, \textit{haridus}, \ldots)$$

Mitmene imputatsioon II

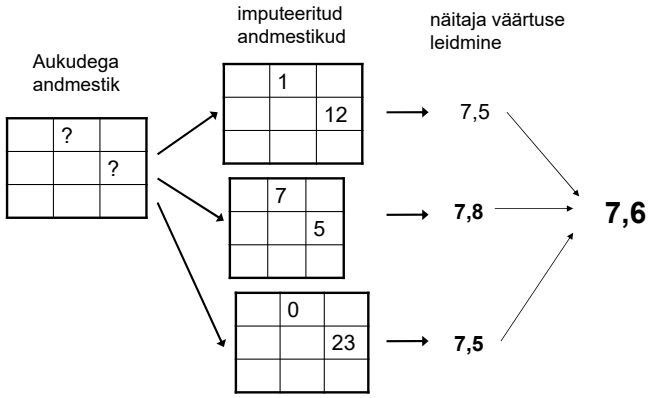
Erinevad juhuslikud suurused viivad erinevate “täidetud” andmestikeni.

Enamasti tekitatakse mitu (vähemalt 3) erinevat andmestikku.

Soovitava näitaja (keskmine, dispersioon,...) väärtus leitakse eraldi iga tekitatud andmestiku jaoks.

Lõplik näitaja väärtus selgub erinevate andmestike pealt leitud tulemuste keskmistamisel.

Mitmene imputatsioon III



Mitmene Imputatsioon IV

$D(Y) = E_X D(Y|X) + D_X E(Y|X)$

eri andmestike pealt leitud hinnangute dispersioonide keskmine

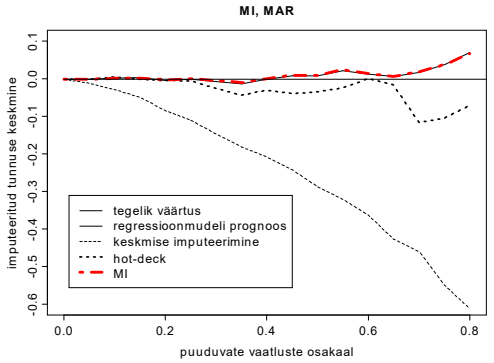
eri andmestike pealt leitud hinnangute dispersioon

$$\hat{s}^2(\hat{\beta}) = \overline{\hat{s}_i^{\,2}(\hat{\beta})} + (1+1/m)s^2(\hat{\beta}_i)$$

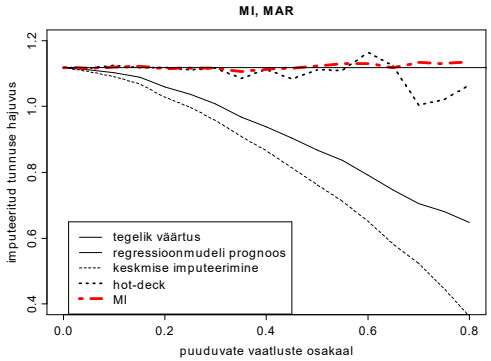
vahel lisatakse „parandusliige“ kompenseerimaks seda, et oleme imputeerinud /auke täitnud hinnatud jaotusest – ja seega võivad augud olla täidetud veidi „valesti“.

$$\begin{aligned}
\mathrm{D}(Y) &= \mathrm{E}(Y^2) - [E(Y)]^2 \\
&= \mathrm{E}_X\left[\mathrm{E}(Y^2|X)\right] - [\mathrm{E}_X\left[E(Y|X)\right]]^2 \\
&\qquad \mathrm{D}(Y|X) = \mathrm{E}(Y^2|X) - [E(Y|X)]^2 \\
&\qquad \mathrm{E}(Y^2|X) = \mathrm{D}(Y|X) + [E(Y|X)]^2 \\
&= \mathrm{E}_X\left[\mathrm{D}(Y|X) + [E(Y|X)]^2\right] - [\mathrm{E}_X\left[\mathrm{E}(Y|X)\right]]^2 \\
&= \mathrm{E}_X\left[\mathrm{D}(Y|X)\right] + \mathrm{E}_X\left[[E(Y|X)]^2\right] - [\mathrm{E}_X\left[\mathrm{E}(Y|X)\right]]^2 \\
&= \mathrm{E}_X\left[\mathrm{D}(Y|X)\right] + \mathrm{D}_X\left[E(Y|X)\right]
\end{aligned}$$

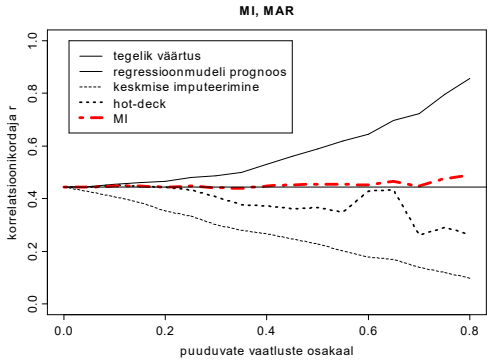
keskmine ja summa, MI, MAR



hajuvus, MI, MAR



Seosed tunnuste vahel, MI, MAR



Asendamine sarnase inimese väärtusega (Hot-Deck / PMM)

Asenda puuduv väärtus võimalikult sarnase inimese (firma, valla,...) teadaoleva väärtusega.

Milline on sarnaseim inimene (firma, vald,...) ?

Variant sarnaseima valikuks:
Loo mudel tunnuse prognoosimiseks mis sisaldab puuduvaid väärtuseid (*Y*). Arvuta olemasolevaid tunnuseid kasutades prognoositud *Y*-tunnuse väärtused nii puuduvate *Y*-tunnuse väärtustega vaatlustele kui ka nendele vaatlustele, mille puhul *Y*-tunnuse väärtust teame. Loe lähedaseimaks (sarnaseimaks) selline vaatlus (sellised vaatlused), mille puhul *Y*-tunnuse prognoosid tulid sarnased (Predicitve Mean Maching, PMM).

Mitmene imputeerimine R-is

- Mitmed lisamoodulid (mice, mi, amelia, …)

```
library(mice)


imp=mice(andmed, meth=c("polyreg", "pmm", "logreg",
                                    "norm","",""), m=1000)
complete(imp, action=1)
fit=with(imp, glm(Y~factor(X1)+factor(X2)+…))
summary(pool(fit))
```

Sensitiivsusanalüüs

Küsitlusuuringu abil uuriti suguhaiguse levikut seksuaaltöötajate seas. Saadi järgmised tulemused (1 = on esinenud, 0 = pole olnud, ? = keeldus vastamast):

1; 1; 0; 1; 0; ?; 0; 0; ?; 1

Milline on suguhaiguste levimus?

Sensitiivsusanalüüs

1; 1; 0; 1; 0; ?; 0; 0; ?; 1

Võimalikud variandid:

1; 1; 0; 1; 0; 0; 0; 0; 0; 1 Pr = 40%
1; 1; 0; 1; 0; 0; 0; 0; 1; 1 Pr = 50%
1; 1; 0; 1; 0; 1; 0; 0; 0; 1 Pr = 50%
1; 1; 0; 1; 0; 1; 0; 0; 1; 1 Pr = 60%

Ignorantsuspiirkond: 40%...60%
95%-Uncertainty region: 0.12…0.88