Bortfallsproblematik ur ett metodperspektiv

Daniel Thorburn

Surveyföreningen

2011-05-27

Olika metodaspekter

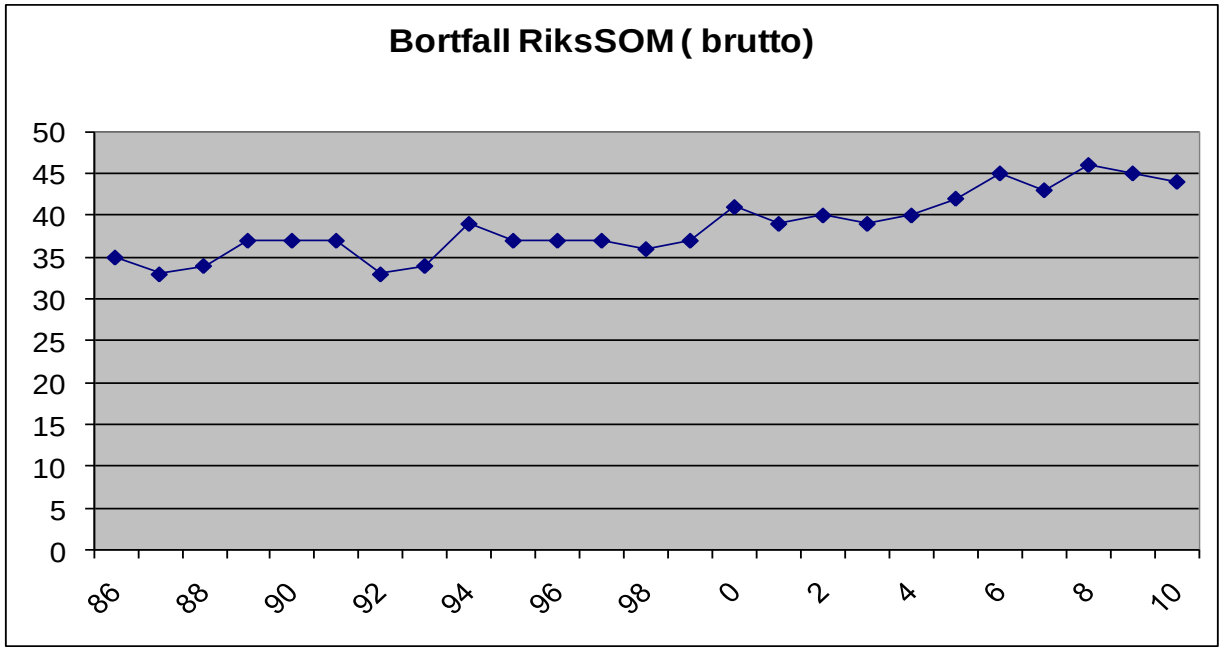
- Bortfall versus andra fel
- Psykologi – varför svarar man? (inte?)
- Åtgärder vid insamling (förebygg!)
- Bortfallsuppföljning hur mäta kvaliteten
- Vad göra med ett stickprov med bortfall?
- Modellering (t.ex latent class, ML, etc)

1. Total Survey Design

- Bortfall kontra ramfel.
  - Ex: genom att ändra ramen kan man definiera bort bortfall (t ex företag med mer än 10 anställda; individer i hushåll med fast telefon).

- Bortfall kontra mätfel
  - Ex: genom att definiera om variabeln (Vill ha opinionen den 10 maj, men accepterar att frågan gäller intervjudatum 1-30 maj: minnesfel; proxy-intervjuer).
  - Multimode – avpassa mät och urvalsteknik efter ip

- Bortfall kontra aktualitet
- …



(”Nettobortfallet” är ungefär 5-7% lägre)

Bortfallsklimat

SCBs ökning är förvånansvärt stor stor. Under samma tid har både bortfall och vägrarandel enligt Jan ökat med 200 % för AKU. SOM använder postenkät med telefonuppföljning (Multimode)

Om cirka 15 år är bortfallet i båda undersökningarna lika stora

Den här bilden tror jag mer speglar uppgiftslämnarklimatet.

Bortfallsandel bra mått – men säger inte allt

- Modernt bland metodstatistiker är att säga Välj inte att minimera/redovisa bortfallet – försök minimera totalfelet. Det är inte säkert att högre svarsfrekvens leder till mindre totalfel.

- Speciellt om man använder samma metod.

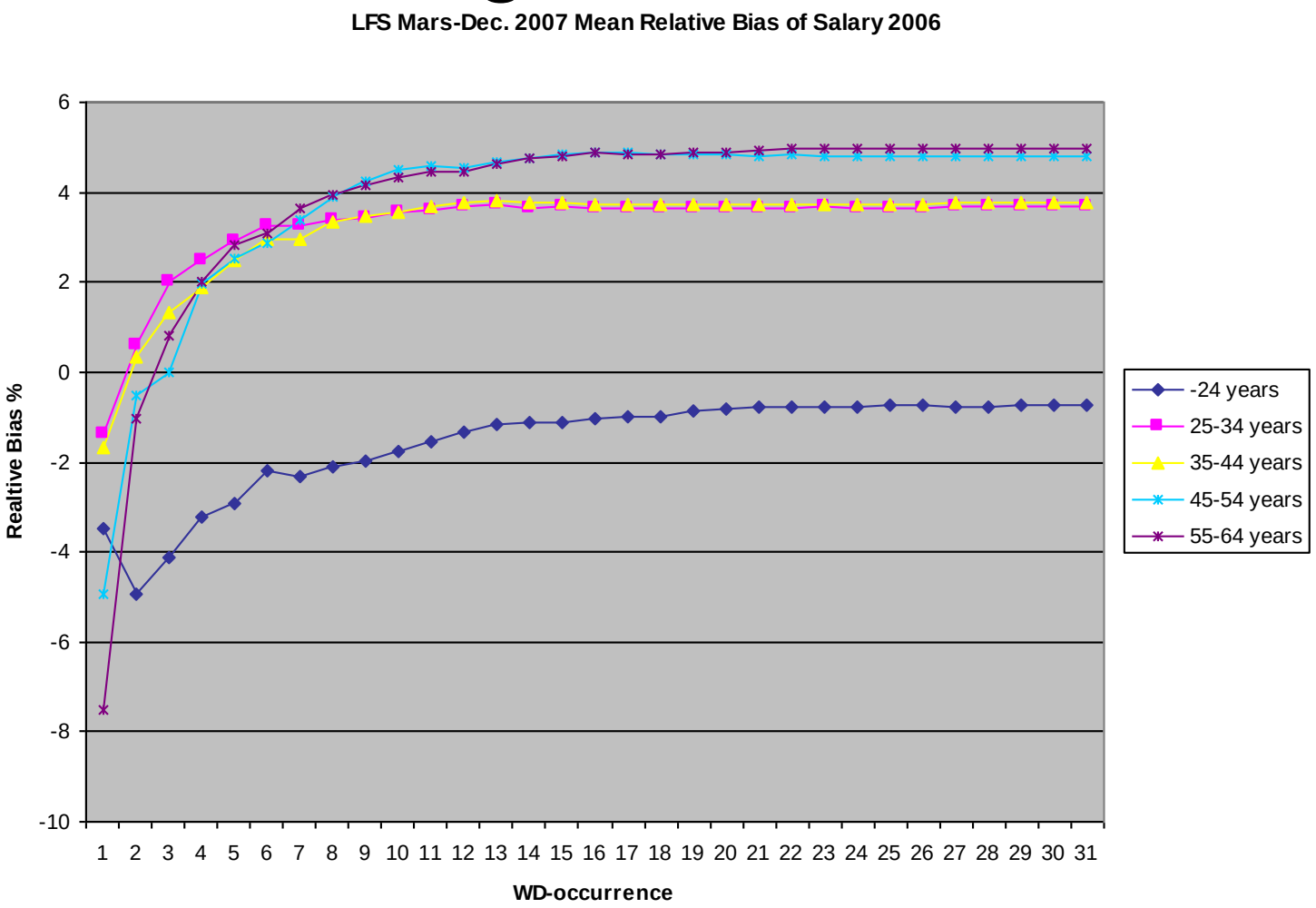
- Ofta bättre att följa upp med hjälp av andra metoder, t ex telefonuppföljning på postenkät. Leder ofta till både högre svarsfrekvens men framförallt mindre bias eftersom olika metoder når olika grupper.

- Multimode – jfr Åke och Jan.

Bortfallsandel bra mått – men säger inte allt

- Bortfallsredovisning (Surveyföreningen bortfallssnurra), Den är en del av kvalitetsredovisningen
- Redovisa tänkbara effekter Dela upp på undergrupper (kön ålder etc). Allt behöver inte redovisas men man bör ha en egen uppfattning.
- Gör bortfallsstudier (jfr Åke)
- Kvalitetskontroll (Ha alltid en viss uppfattning om totalfelet. Kanske från tidigare bortfallstudier)

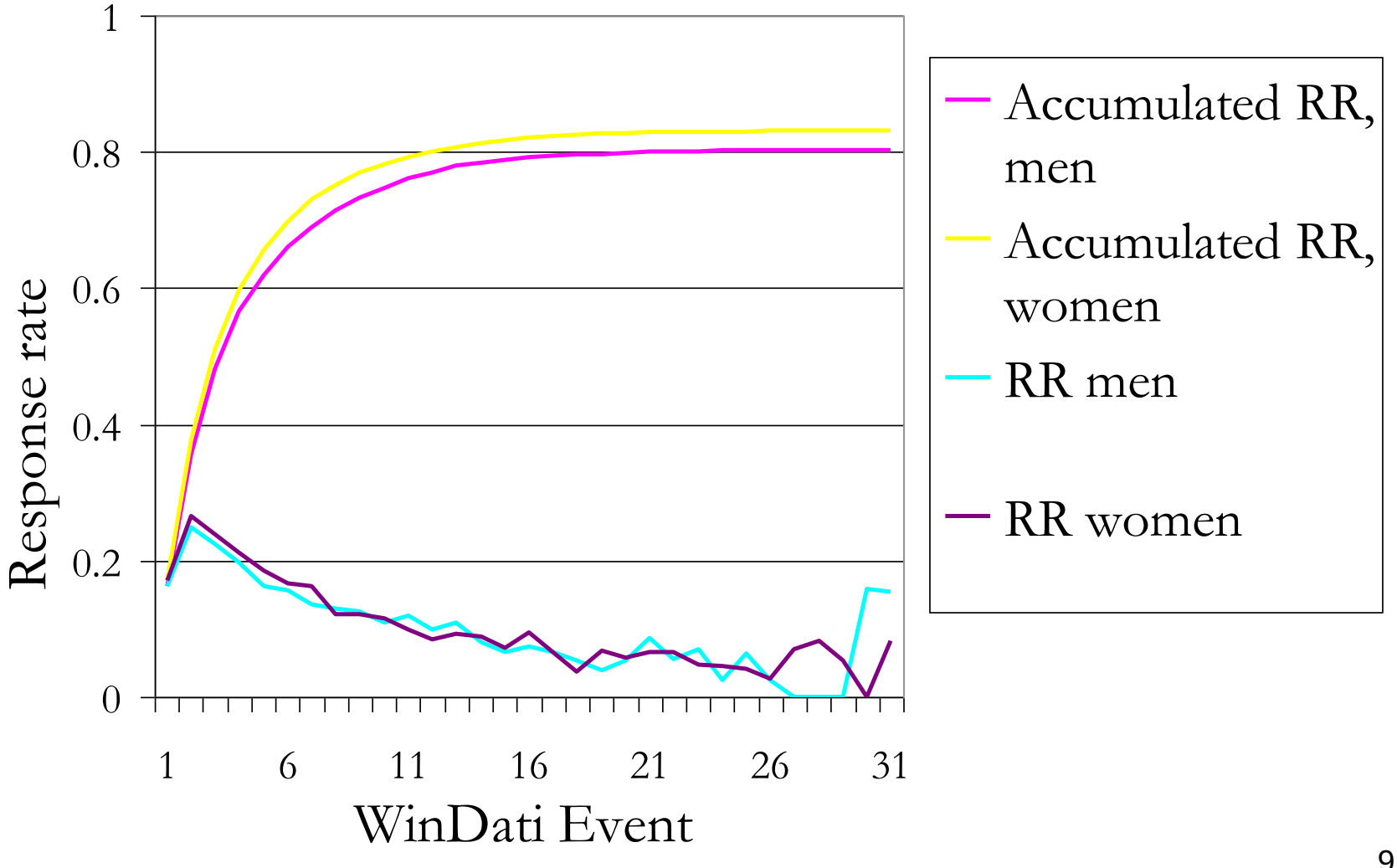
Mean relative bias of salary after age, 2006



Bortfall eller totalfel

- Överförenklad modell.
  - Tre grupper Stugsittare, Vanliga knegare, A-lagare.
  - Den mellersta gruppen har högst löner. Först får man underskattningen, sedan stiger det förbi noll för att sluta med kraftigt positiv bias.
  - Man kan åstadkomma unbiased skattning genom att stoppa halvvägs men vill man verkligen det?
  - Ofta är urvalsstorleken bestämd för att uppnåe en viss varians
- Vad händer vid jämförelser mellan grupper/över tiden.
  - Ofta hänvisas till att om man använder samma metod så blir bortfallsfelet detsamma och jämförelser går bra över tiden och mellan grupper
  - Men det har inte alltid stöd. Se t ex Åkes data eller mina för jämförelser mellan grupper

Response Rates, April 2007



2. Vad göra när data redan finns och där finns bortfall?

- Utan hjälpinformation kan man inte göra mycket (finns lite modellansatser och lite om användning av bortfallsuppföljning vid tidigare us).
- Men med hjälpinformation i ramen så …

- Studerad variabel, Y
- Bakgrundsinformation (kön, ålder, bostadsort, …, X
- Svarsindikator, R

Rubins indelning

- Missing completely at random (Ignorable non-response) MCAR – glöm bortfallet; låtsas att det erhållna stickprovet var det avsedda. ((Y,X) ^ R)

- Missing at Random, MAR, Ungefär ”all behövlig information finns i X” eller ”R och Y är oberoende givet X”. Då kan man skatta totalen vändevärdesriktigt genom att använda X **på rätt sätt**. (Y ^ R | X). (Det man brukar ”anta”)

- Not missing at random (NMAR) Då måste man modellera bortfallsmekanismen. (Bortfallet beror direkt av Y)

(Formell definition av MAR

- För **alla** tänkbara delmängder, Z, av de observerbara värdena sådana att P(r = Z) > 0 (r är den observerade delen av stickprovet, s) måste gälla att

- $Y_{U\text{-}Z}$ och $r_{U\text{-}Z}$ är oberoende givet Y_Z och X)

- I praktiken har man aldrig MAR, men ibland ligger man rätt nära

- Det viktiga är att välja rätt hjälpvariabler och att utnyttja dem på rätt sätt. Tag gärna med lite för många. (Bra vid register som vi har i Sverige)

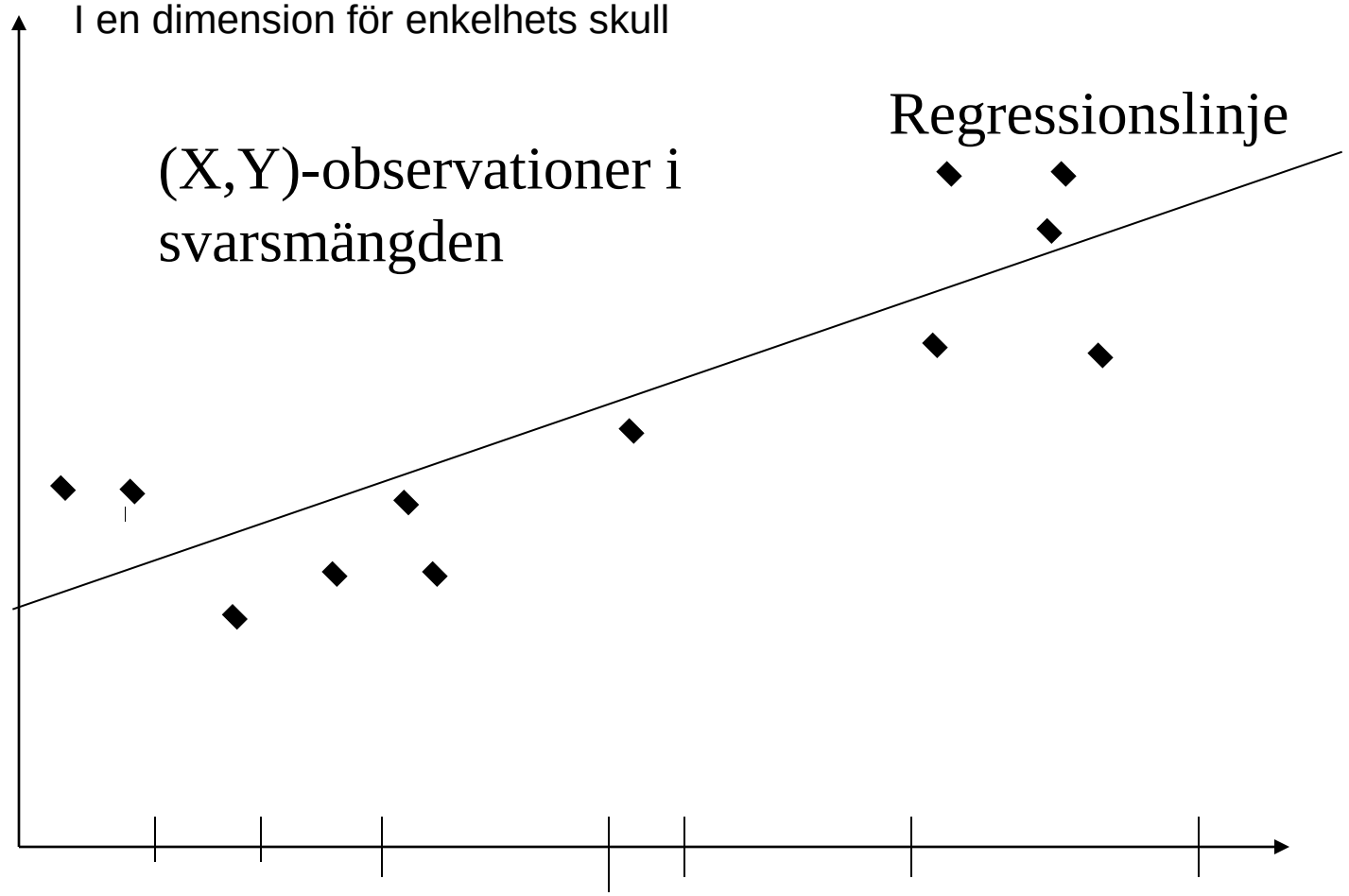
- Fundera t ex på transformationer (t ex logaritmera något av x, klassindela gärna x, avgör om samspel bör tas med.

Två vanliga sätt

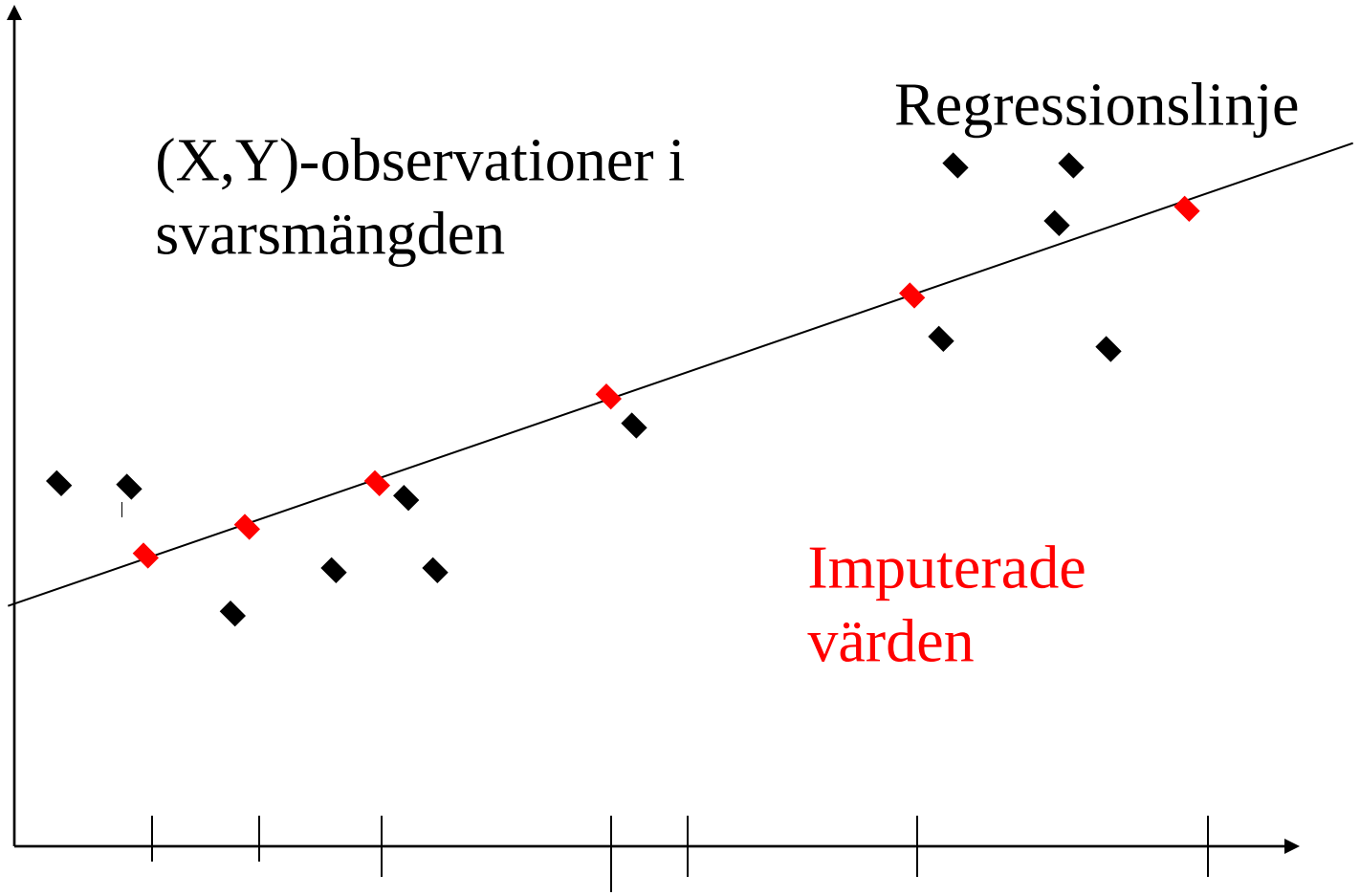
- Avsedd skattning $t_{yS} = \Sigma_{i€S}\, \omega_i\, y_i \sim \Sigma_{i€S}\, y_i/\pi_i$

- Omvägning
  - Ändra vikterna $t_{yR} = \Sigma_{i€R}\, \omega^*_i\, y_i$
  - t ex $\omega^*_i = \omega_i$ /P*(i € r)
- Imputering
  - Förutsäg alla saknade data, t ex linjär regression $y^*_i = f(S,X_i) \sim a^* + b^*X_i$
  - Sätt in $t_{yI} = \Sigma_{i€R}\, \omega_i\, y_i + \Sigma_{i€S-R}\, \omega_i\, y^*_i$

Exempel på modellimputering



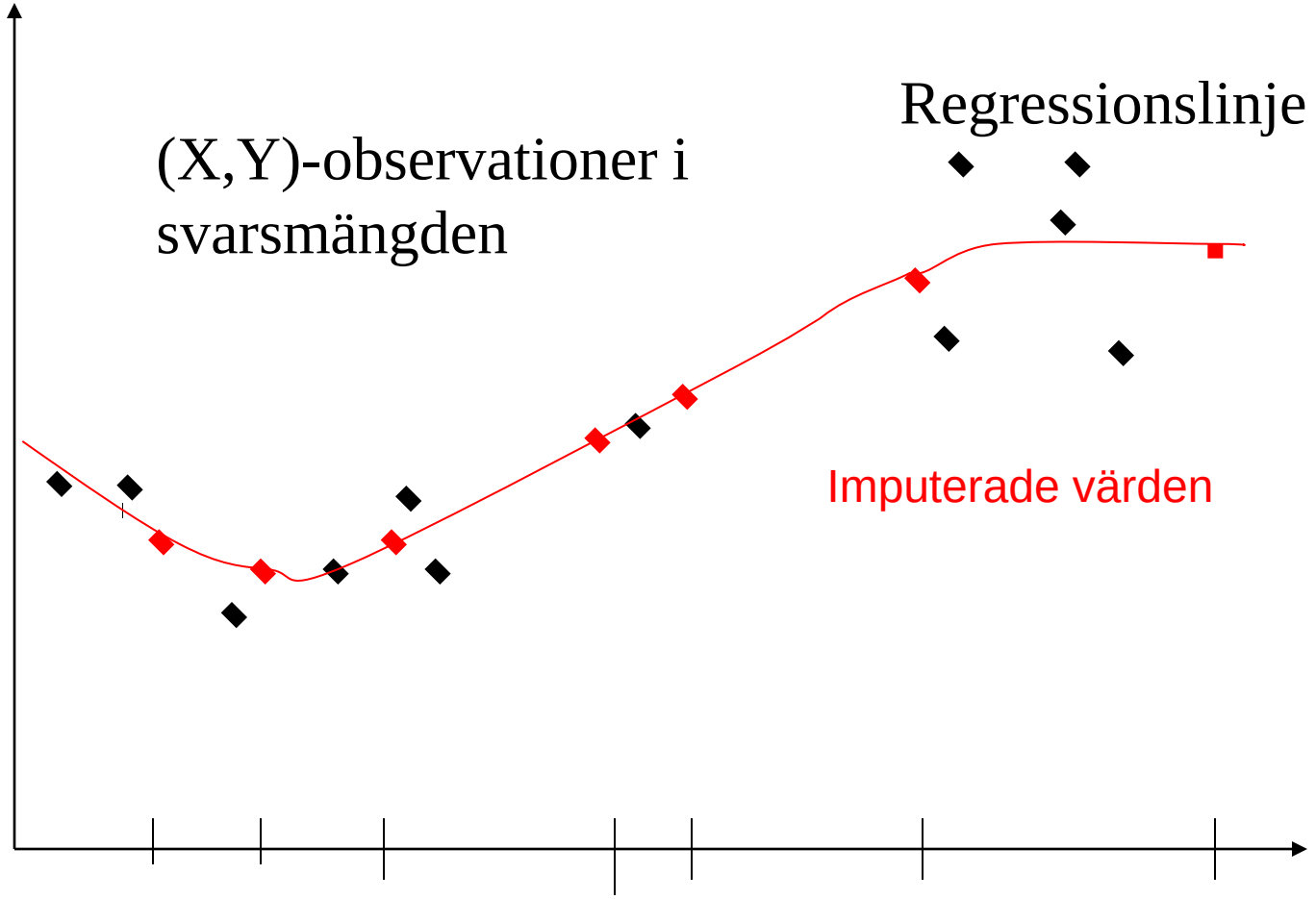
Exempel på modellimputering



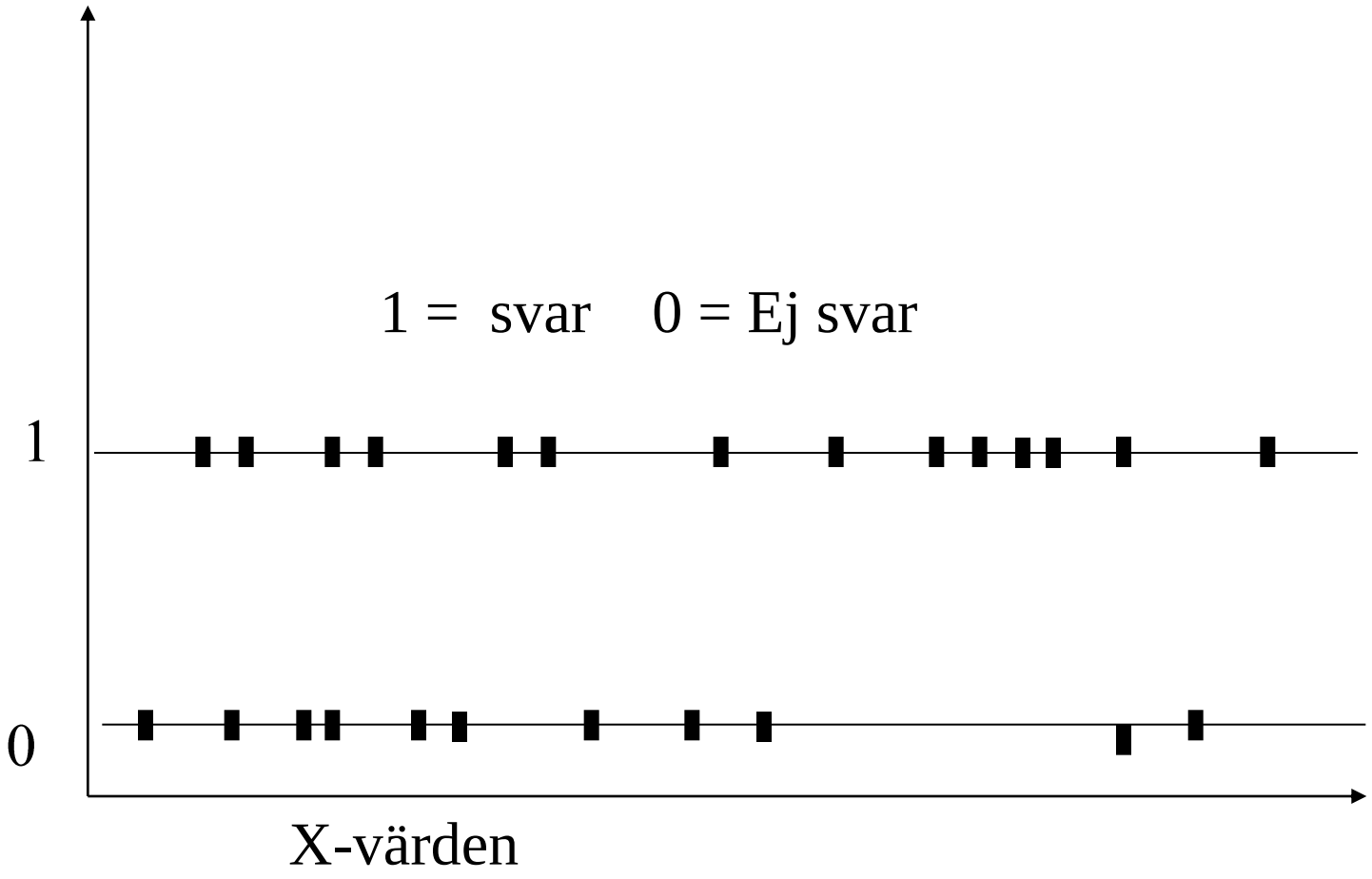
hjälpvariabler för bortfallet

- (Lite besvärligt med variansskattningar)
- Inte alltid lämpligt med rät linje
- Kanske en andragradskurva
- eller efterstratifiera dvs använd en styckvis konstant kurva
- Numera forskas det en del på icke-parametrisk kurvanpassning med t ex splinefunktioner eller kärnskattningar.

Exempel på modellimputering

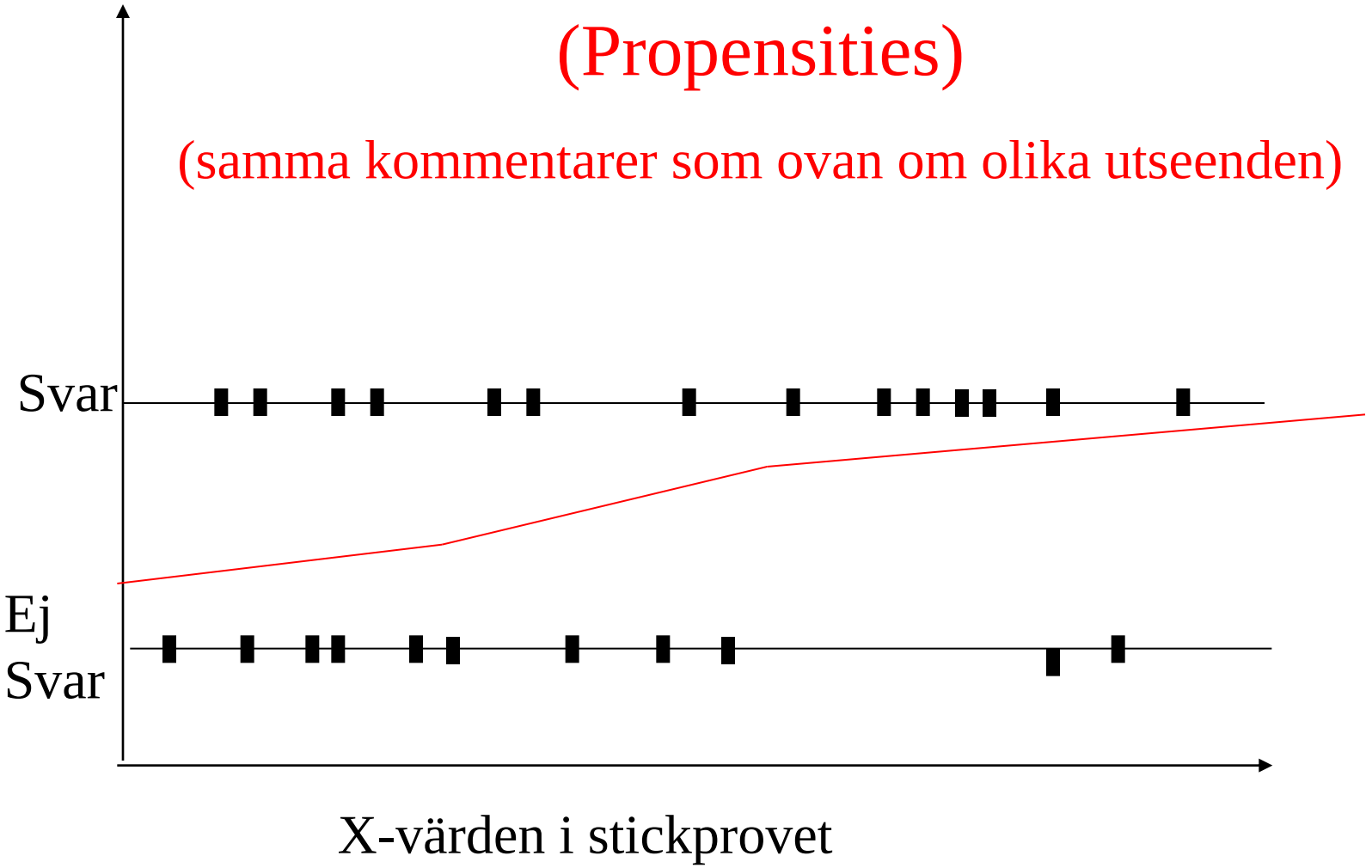


Svarssannolikhet (Propensity)



Skattade svarssannolikheter (Propensities)

(samma kommentarer som ovan om olika utseenden)



- Massor av metoder/namn

- Omviktning
  - Efterstratifiering
  - Regressionskattningar
  - RHG-grupper (Response homogeneity classes)
  - Omviktningsklasser
  - Kalibrering
  - Raking eller IPF
  - Propensity scores
  - Omviktning med skattade svarssannolikheter
  - Uppskattade genom frågor …

- Imputering
  - Efterstratifiering
  - Medelvärdesimputering
  - Regressionsskattningar
  - Plus slumpfel
  - Nearest neighbour
  - Modell donator
  - Verklig donator
  - Hot Deck
  - Cold deck
  - Multipel imputering
  - Massimputering
  - Imputering till populationsnivån …

men görs det vettigt blir det ungefär samma sak

men görs det vettigt blir det ungefär samma sak
och de flesta statistiker är vettiga

men görs det vettigt blir det ungefär samma sak
och de flesta statistiker är vettiga

- Okorrigerad bias (i medlet och endimensionell x)
  - cov(R,Y) / svarsandel
- Korrigerad bias
  - cov(R,Y) - cov(R,X)cov(X,Y)/var(X) / svarsandel
  - vid flerdimensionell X: (cov(R,Y) – $\Sigma_{y,X}\ \Sigma_{X,X}^{-1}\ \Sigma_{y,R}$) / svarsandel
- Dvs hjälpvariabeln måste ha samband både med bortfallsmönstret och den studerade variabeln för att ha någon effekt

(Exakt formel vid regressionsskatningar)

För- och nackdelar

- Omvägning endast ett värde måste sättas in vid omvägning (vid imputering ersätts alla saknade variabler)

- Imputering tar hänsyn till sambandet mellan hjälpvariabel och studerad variabel och ger normalt lägre slumpfel (men svårare att skatta variansen bra)

- Båda metoderna kan dock kombineras med t ex regressionsskattningar för att minska slumpfelet. (Kalibrering är en metod som gör båda stegen i ett)

- Även under MAR är det svårt att exakt skatta variansen. (utom vid multipel imputering som dock kräver stor datakraft). Bra approximationer finns i de flesta fall.

Hur uppskatta bortfallsfelet effektivare?

- Bortfallsbias vid skattning utan användning av hjälpinformation
  - m hjälpvariabler (kön, ålder, inkomst, …) (hjälp kan vara vet totalen och frågar i svarsmängden).
  - Skatta totalen för hjälpvariablerna från stickprovet och beräkna relativ bias genom att jämföra med känd total för alla dessa m.
  - Den relativa biasen för den intressanta variabeln har förmodligen ungefär samma fel som dessa

Hur uppskatta bortfallsfelet effektivare?

- Bortfallsbias vid skattning utan användning av hjälpinformation
  - m hjälpvariabler (kön, ålder, inkomst, …) (hjälp kan vara vet totalen och frågar i svarsmängden).
  - Skatta totalen för hjälpvariablerna från stickprovet och beräkna relativ bias genom att jämföra med känd total för alla dessa m.
  - Den relativa biasen för den intressanta variabeln har förmodligen ungefär samma fel som dessa

- Men om man med skattningen redan använt hjälpinformation t ex för kompensationsvägning
  - Uppskatta relativ bias för dessa m vid en skattning som utnyttjar de m-1 andra.
  - Den relativa biasen för den intressanta variabeln givet de m andra har förmodligen ungefär samma fel som dessa

Hur representativt är ett stickprov?

- Finns några olika förslag
- R-indikatorer bygger på att se hur mycket vikterna ändras vid omviktning
- t ex skatta svarssannolikheterna för alla i stickprovet $p^*(x_i)$
- Beräkna $Var(p_x^*(X)) = (1/n)\Sigma_{x€S} (p_x^*(X) – p^*)^2$
  (skrivet för lika urvalssannolikheter, och där p* är totalbortfallet)
- Sätt R’ = $1 – Var(p_x^*(X)) / (p^*(1-p^*))$
- eller R” = $1 – 4\ Var(p_x^*(X))$ eller …
- En nackdel är att denna definition alltid ökar med mängden hjälpinformation eller bättre när hjälpinformationen sämre. Men används t ex för att jämföra olika planer och också effekten av ansträngningar att öka svarsandelen.

Tack, för ert tålamod!

Tack, för ert tålamod!

-

En oersättlig tillgång vid all bortfallsbekämpning!