# Korrektion för bortfall

Rikard Gard / Metodstatistiker
Örebro Universitet / SCB
2019

# Kort om mig

# Kort om mig


- Utbildning i Retorik och Statistik
- 1,5 år som telefonintervjuare
- 3,5 år som statistiker / prognosmakare
- 3 år som metodstatistiker

# Inledning och bakgrund

# Syftet med uppsatsen

- Utforska k-NN metoden i differensestimatorn för att se om den kan minska bortfallsskevheten för fritidsfiskeundersökningen

- Uppsatsen heter Design-based and Model-assisted Estimators Using Machine Learning Methods
  - -Exploring the k-Nearest Neighbor method applied to data from the Recreational Fishing Survey

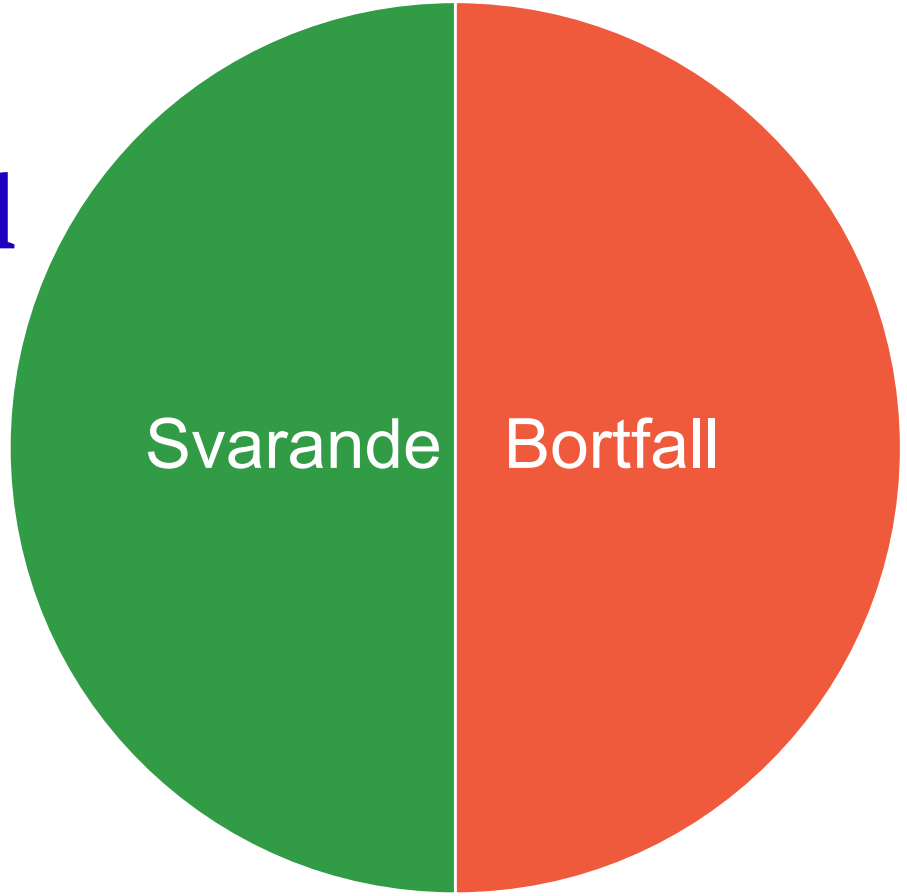
SCB

# Inledning

- Hämtat inspiration från Breidt och Opsomers artikel om differens-estimatorn och maskininlärning
- Använt mig av design och data från fritidsfiskeundersökningen
- Utforskat och utvecklat K närmsta grannarna-metoden (k-NN) inom ramen för den designbaserade och modellassisterade differens-estimatorn
- Beräknat bortfallsskevhet, punkt- och variansskattningar för olika varianter av estimatorer och jämfört dom

SCB

# Bakgrund om bortfall




- Bortfall är ett stort problem
- Speciellt inom fritidsfiskeundersökningen
- Selektionsbias (Orsak kan vara att de som gillar undersökningens ämne gärna svarar och blir överrepresenterade)

# Bakgrund om bortfall

| Panel | Period 1 | Period 2 | Period 3 |
|---|---|---|---|
| 10 No fishing / Nonresponse | 38 % | NA | NA |
| 10 Fished | 66 % | NA | NA |
| 11 No fishing / Nonresponse | 43 % | 40 % | NA |
| 11 Fished | 68 % | 62 % | NA |
| 12 No fishing / Nonresponse | 41 % | 38 % | 33 % |
| 12 Fished | 75 % | 68 % | 63 % |
| 13 No fishing / Nonresponse | NA | 40 % | 38 % |
| 13 Fished | NA | 76 % | 66 % |
| 14 No fishing / Nonresponse | NA | NA | 32 % |
| 14 Fished | NA | NA | 69 % |
| New sample | 44 % | 41 % | 41 % |

SCB

# Data och design

**1. Fiskade du i Sverige under 2017?**

*Hit räknas allt fritidsfiske med spö, nät, bur m.m. som sker utan yrkesfiskelicens.*
*Med en dag menas en dag då du fiskat, oavsett hur länge eller om du fått någon fångst eller inte.*

- ☐ Ja, 1 dag
- ☐ Ja, 2–5 dagar
- ☐ Ja, 6–10 dagar
- ☐ Ja, 11–15 dagar
- ☐ Ja, mer än 15 dagar

- ☐ Nej

---

**2. a) Fiskade du i Sverige under januari–april 2018?**

- ☐ Ja
- ☐ Nej

**b) Fiskade du i Sverige under maj–augusti 2018?**

- ☐ Ja, 1 dag
- ☐ Ja, 2–5 dagar
- ☐ Ja, 6–10 dagar
- ☐ Ja, 11–15 dagar
- ☐ Ja, mer än 15 dagar

*Var god gå till fråga 4 på sida 3!*

- ☐ Nej

SCB

# Data

- Fritidsfiskeundersökningen

- Ett år består av tre omgångar:
  - > januari – april (Etapp 1) 2800 personer
  - > maj – augusti (Etapp 2) 5400 personer
  - > september – december (Etapp 3) 2800 personer

- Varje omgång består av fyra urval
  - Tre paneler
  - Ett nytt urval

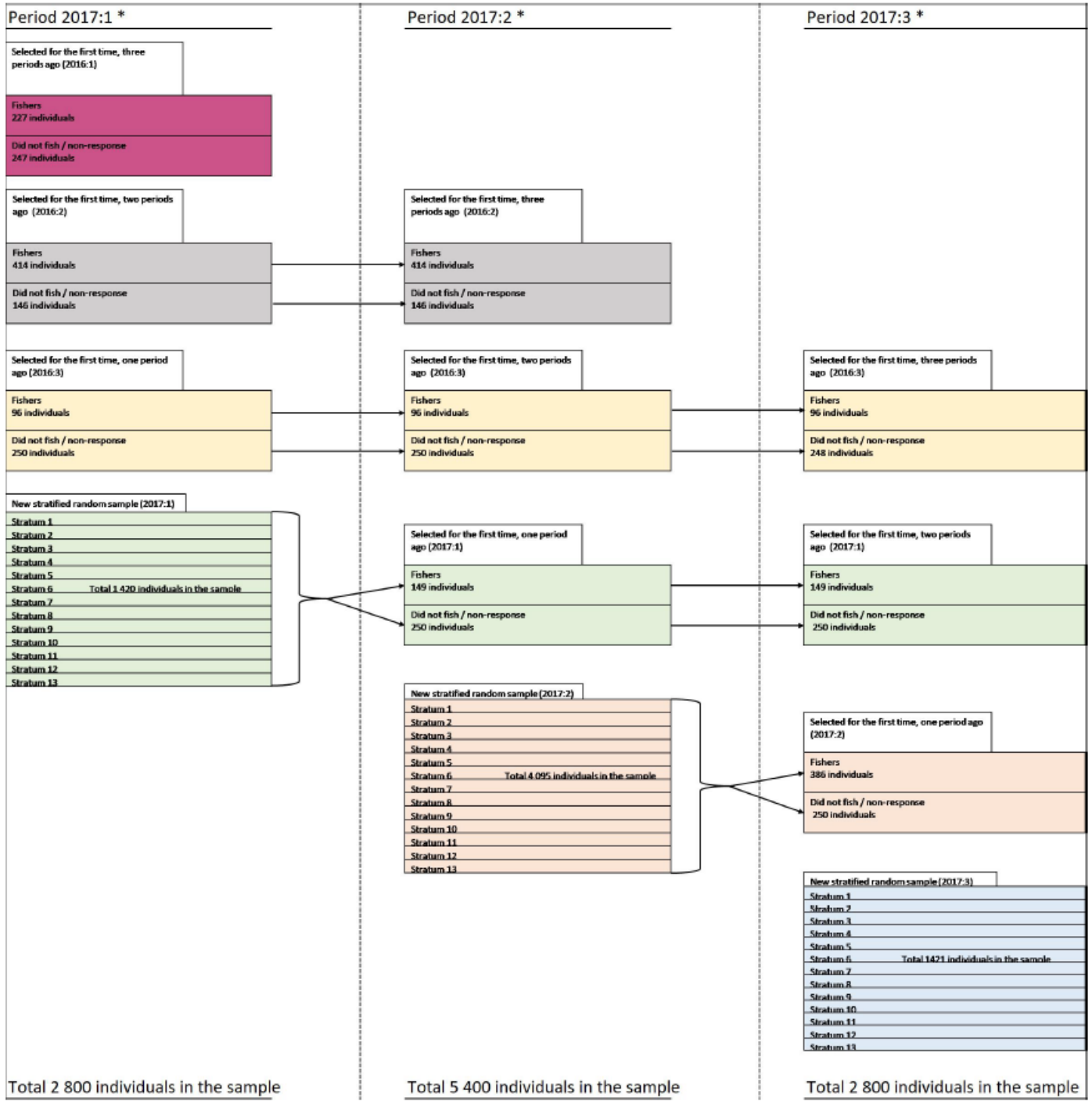
SCB

# Data

- Responsvariabel (Målvariabel) – inkomst från IoT (Inkomst och taxeringsregistret)

- Förklaringsvariabel (Hjälpvariabler):
  - Ålder (Numerisk, Ålder vid årets slut)
  - Kön (Kategorisk, Kvinna / Man)
  - Utbildning (Kategorisk, 3 nivåer)
  - Region (Kategorisk, Folkbokföring, 8 nivåer)
  - Födelselandgrupp (Kategorisk, Sverige eller utomlands)
  - Storstad (Kategorisk, Folkbokföring, 2 nivåer)

# Design


Period 2017:1 *
Period 2017:2 *
Period 2017:3 *
Selected for the first time, three periods ago (2016:1)
Fishers
227 individuals
Did not fish / non-response
247 individuals
Selected for the first time, two periods ago (2016:2)
Fishers
414 individuals
Did not fish / non-response
146 individuals
Selected for the first time, three periods ago (2016:2)
Fishers
414 individuals
Did not fish / non-response
146 individuals
Selected for the first time, one period ago (2016:3)
Fishers
96 individuals
Did not fish / non-response
250 individuals
Selected for the first time, two periods ago (2016:3)
Fishers
96 individuals
Did not fish / non-response
250 individuals
Selected for the first time, three periods ago (2016:3)
Fishers
96 individuals
Did not fish / non-response
248 individuals
New stratified random sample (2017:1)
Stratum 1
Stratum 2
Stratum 3
Stratum 4
Stratum 5
Stratum 6 Total 1 420 individuals in the sample
Stratum 7
Stratum 8
Stratum 9
Stratum 10
Stratum 11
Stratum 12
Stratum 13
Selected for the first time, one period ago (2017:1)
Fishers
149 individuals
Did not fish / non-response
250 individuals
Selected for the first time, two periods ago (2017:1)
Fishers
149 individuals
Did not fish / non-response
250 individuals
New stratified random sample (2017:2)
Stratum 1
Stratum 2
Stratum 3
Stratum 4
Stratum 5
Stratum 6 Total 4 095 individuals in the sample
Stratum 7
Stratum 8
Stratum 9
Stratum 10
Stratum 11
Stratum 12
Stratum 13
Selected for the first time, one period ago (2017:2)
Fishers
386 individuals
Did not fish / non-response
250 individuals
New stratified random sample (2017:3)
Stratum 1
Stratum 2
Stratum 3
Stratum 4
Stratum 5
Stratum 6 Total 1421 individuals in the sample
Stratum 7
Stratum 8
Stratum 9
Stratum 10
Stratum 11
Stratum 12
Stratum 13
Total 2 800 individuals in the sample
Total 5 400 individuals in the sample
Total 2 800 individuals in the sample

# Design

- Varje nytt urval har 13 stratum

- Varje panel har 2 stratum
  - De som fiskade i första urvalet
  - De som inte fiskade / de som var bortfall i första urvalet

- Paneldesign där en person kan vara med 1 eller 4 etapper

# Teori och metod

# Differensestimatorn och arbetsmodellen

$$DIFF(y,\hat{m}) = \sum_{k\in U} \hat{m}(\mathbf{x}_k) + \sum_{k\in r} \frac{y_k - \hat{m}(\mathbf{x}_k)}{\pi_k^{nr}} = DIFF(y,m) + (remainder) \qquad (5)$$

- Vad den här formeln säger är att differensestimatorn och en arbetsmodell baserat på ett urval är lika med differensestimatorn med populationsmodellen + en restterm

SCB

# Differensestimatorn och arbetsmodellen

$$DIFF(y,\hat{m}) = \sum_{k\in U} \hat{m}(\mathbf{x}_k) + \sum_{k\in r} \frac{y_k - \hat{m}(\mathbf{x}_k)}{\pi_k^{nr}} = DIFF(y,m) + (remainder) \qquad (5)$$

- Om resttermen är negligerbar så ”ärver” differensestimatorn med arbetsmodellen egenskaperna från differensestimatorn med populationsmodellen (som är asymptotiskt väntevärdesriktig, designkonsistent och asymptotiskt normalfördelad)

SCB

# Differensestimatorn och arbetsmodellen

$$DIFF(y,\hat{m}) = \sum_{k\in U} \hat{m}(\mathbf{x}_k) + \sum_{k\in r} \frac{y_k - \hat{m}(\mathbf{x}_k)}{\pi_k^{nr}} = DIFF(y,m) + (remainder) \qquad (5)$$

- I uppsatsen har jag inte visat huruvida resttermen är negligerbar eftersom det är tekniskt väldigt svårt och ett område i sig. Men det går bl.a. att göra med Taylor approximeringar och glattfunktioner (smoothnes conditions)

- Bafetta m.fl. har tillämpat sig av k-NN metoden i differensestimatorn tidigare och inte heller kunnat bevisa att resttermen är negligerbar men dom argumenterar, intuitivt, att den borde vara det

SCB

# Differensestimatorn och arbetsmodellen

$$DIFF(y,\hat{m}) = \sum_U \hat{m}(\mathbf{x}_k) + \sum_r \frac{y_k - \hat{m}(\mathbf{x}_k)}{\pi_k^{nr*}} \tag{11}$$

$$\hat{V}(DIFF(y,\hat{m})) = \sum_{h=1}^{H} \sum_{k\epsilon r_h} \sum_{l\epsilon r_h} \frac{\Delta_{akl}}{\pi_{kl}^{nr*}} \frac{D_k}{\pi_{ak}} \frac{D_l}{\pi_{al}} + \sum_{h=1}^{H} \sum_{k\epsilon r_h} \sum_{l\epsilon r_h} \frac{\Delta_{kl|s_a}}{\pi_{kl|s_a}^{nr}} \breve{\breve{D}}_k \breve{\breve{D}}_l \tag{12}$$

SCB

# Inklussionssannolikheter

| Condition | $\pi_{akl}$ | $\pi^{nr}_{kl\|s_a}$ |
|---|---|---|
| $k = l$ | $\frac{n_{ah}}{N_h}$ | $\frac{m_g}{n_{ag}}$ |
| $k \neq l$<br>$k \in s_{ah}$ & $l \in s_{ah}$<br>$k \in s_g$ & $l \in s_g$ | $\frac{n_{ah}}{N_h}\frac{n_{ah}-1}{N_h-1}$ | $\frac{m_g}{n_{ag}}\frac{m_g-1}{n_{ag}-1}$ |
| $k \neq l$<br>$k \in s_{ah}$ & $l \in s_{ah'}$<br>$k \in s_g$ & $l \in s_g$ | $\frac{n_{ah}}{N_h}\frac{n_{ah'}}{N_{h'}}$ | $\frac{m_g}{n_{ag}}\frac{m_g-1}{n_{ag}-1}$ |
| $k \neq l$<br>$k \in s_{ah}$ & $l \in s_{ah}$<br>$k \in s_g$ & $l \in s_{g'}$ | $\frac{n_{ah}}{N_h}\frac{n_{ah}-1}{N_h-1}$ | $\frac{m_g}{n_{ag}}\frac{m_{g'}}{n_{ag'}}$ |
| $k \neq l$<br>$k \in s_{ah}$ & $l \in s_{ah'}$<br>$k \in s_g$ & $l \in s_{g'}$ | $\frac{n_{ah}}{N_h}\frac{n_{ah'}}{N_{h'}}$ | $\frac{m_g}{n_{ag}}\frac{m_{g'}}{n_{ag'}}$ |

[1] The condition column means that if that condition is met then the inclusion probability is calculated as described for that row

# Den fullständiga estimatorn

$$\hat{t}_y = \sum_{i=1}^{4} (w_i) DIFF(y, \hat{m}_i) \qquad (13)$$

$$\hat{V}(\hat{t}_y) = \sum_{i=1}^{4} (w_i)^2 \hat{V}(DIFF(y, \hat{m}_i)) \qquad (14)$$

- W är satt proportionerligt mot antal svarande i urvalen
- Variansen är snarlik bara att när vi bryter ur den blir det vikten i kvadrat

SCB

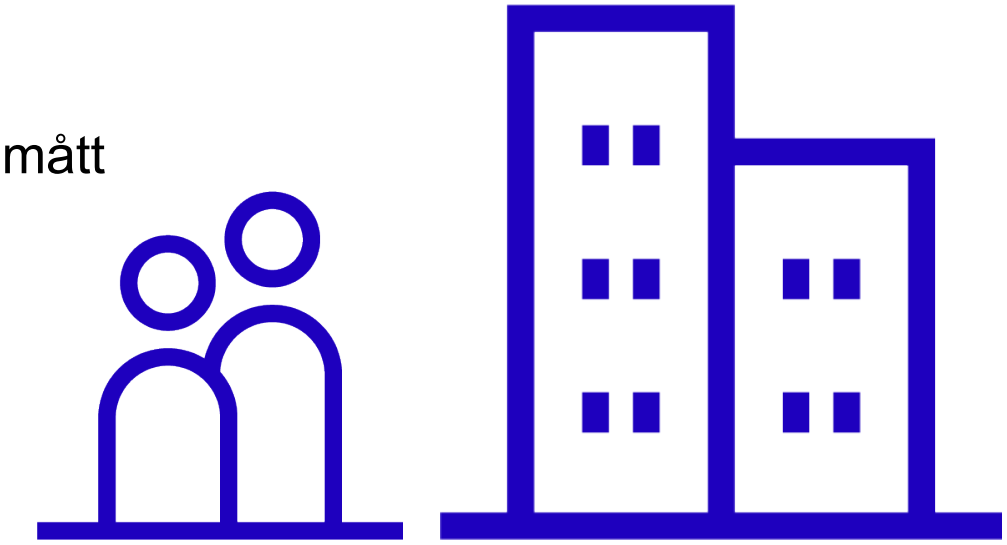
# k-NN metoden

- k-NN metoden går ut på att identifiera andra observationer som ligger väldigt nära den observation man vill förutspå

$$\hat{m}(\mathbf{x}_k) = \frac{1}{K} \sum_{l \in L_k} y_k \qquad (7)$$

- Grannskapet, $L_k$, bestäms utifrån ett avståndsmått

$$\sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$



# k-NN metoden
# Utvecklad med designvikten

- k-NN metoden går ut på att identifiera andra observationer som ligger väldigt nära den observation man vill förutspå

$$\hat{m}(\mathbf{x}_k) = \frac{\sum_{l \in L_k} \frac{y_k * w_k}{\pi_k}}{\sum_{l \in L_k} \frac{w_k}{\pi_k}} \qquad (8)$$

- Grannskapet, $L_k$, bestäms utifrån ett avståndsmått

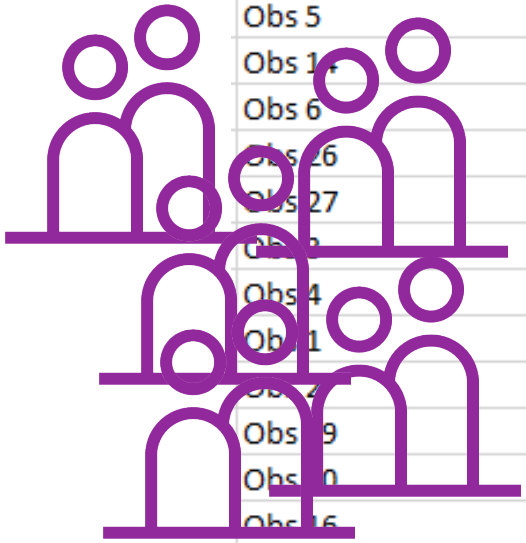
$$\sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

SCB

# k-NN metoden
# Hur funkar den?

| | Ålder | Man |
|---|---|---|
| Dag | 32 | 1 |
| | | |
| Inkomst? | | |
| k=1 | 400000 | |
| k=5 | 377500 | |
| k=10 | 344000 | |
| k=20 | 321615 | |

SCB



| Observationer från ett urval | Ålder | Man | Inkomst | Distans |
|---|---|---|---|---|
| Obs 18 | 33 | 1 | 400000 | 1 |
| Obs 9 | 30 | 0 | 160000 | 5 |
| Obs 19 | 35 | 0 | 450000 | 10 |
| Obs 20 | 38 | 1 | 500000 | 36 |
| Obs 7 | 25 | 1 | 270000 | 49 |
| Obs 17 | 25 | 1 | 150000 | 49 |
| Obs 21 | 39 | 1 | 510000 | 49 |
| Obs 8 | 25 | 0 | 300000 | 50 |
| Obs 15 | 25 | 0 | 300000 | 50 |
| Obs 12 | 40 | 1 | 400000 | 64 |
| Obs 11 | 40 | 0 | 1000000 | 65 |
| Obs 22 | 40 | 0 | 510000 | 65 |
| Obs 28 | 22 | 0 | 150000 | 101 |
| Obs 5 | 20 | 1 | 250000 | 144 |
| Obs 1 | 20 | 1 | 100000 | 144 |
| Obs 6 | 20 | 0 | 230000 | 145 |
| Obs 26 | 20 | 0 | 150000 | 145 |
| Obs 27 | 20 | 0 | 200000 | 145 |
| Obs [illegible] | 18 | 0 | 202000 | 197 |
| Obs 4 | 18 | 0 | 200300 | 197 |
| Obs 1 | 15 | 1 | 200000 | 289 |
| Obs [illegible] | 15 | 1 | 150000 | 289 |
| Obs 29 | 50 | 0 | 400000 | 325 |
| Obs 30 | 55 | 1 | 400000 | 529 |
| Obs 16 | 60 | 1 | 200000 | 784 |
| Obs 10 | 60 | 0 | 200000 | 785 |
| Obs 24 | 70 | 1 | 150000 | 1444 |
| Obs 25 | 75 | 0 | 150000 | 1850 |
| Obs 13 | 80 | 0 | 70000 | 2305 |
| Obs 23 | 80 | 0 | 60000 | 2305 |

# Applicera k-NN modellen på populationen

- Det finns alltså 66*2 kombinationer för populationen 15 – 80 år efter kön
- Man väljer en kombination och predikterar värdet för inkomst med hjälp av urvalet och sen multiplicerar med antalet
- Det gör man för samtliga och på så vis får man en skattning för populationen, plus justeringstermen

| Ålder | män | kvinnor |
|---|---|---|
| 15 år | 58915 | 55462 |
| 16 år | 57928 | 54240 |
| 17 år | 57108 | 51997 |
| 18 år | 58716 | 52341 |
| 19 år | 61135 | 51671 |
| 20 år | 59115 | 51958 |
| . | . | . |
| . | . | . |
| . | . | . |
| . | . | . |
| . | . | . |
| . | . | . |
| . | . | . |
| . | . | . |
| . | . | . |
| . | . | . |
| . | . | . |
| . | . | . |
| 75 år | 47508 | 50787 |
| 76 år | 42615 | 45668 |
| 77 år | 36201 | 39805 |
| 78 år | 33125 | 37004 |
| 79 år | 31488 | 36418 |
| 80 år | 28459 | 34250 |

# Skattning av bortfallsskevheten

$$\hat{RB}(\hat{\theta}_r) = (\hat{\theta}_r - \hat{\theta}_s)/\hat{\theta}_s$$

- Helst hade jag velat haft en proxyvariabel för fiske men någon sådan finns inte så inkomst användes vilket innebär att användningen av resultatet blir begränsat i den riktigt fritidsfiskeundersökningen men har fungerat bra för en uppsats

- Egentligen inget konstigt. Vi gör en skattning för inkomst när vi använder oss av hela urvalet och en skattning endast för de svarande. Och på den vägen får vi en skattning för bortfallsskevheten som vi sen kan sätta i relation till urvalet för att få det i relativa termer.

SCB

# Utvärdering av modellen

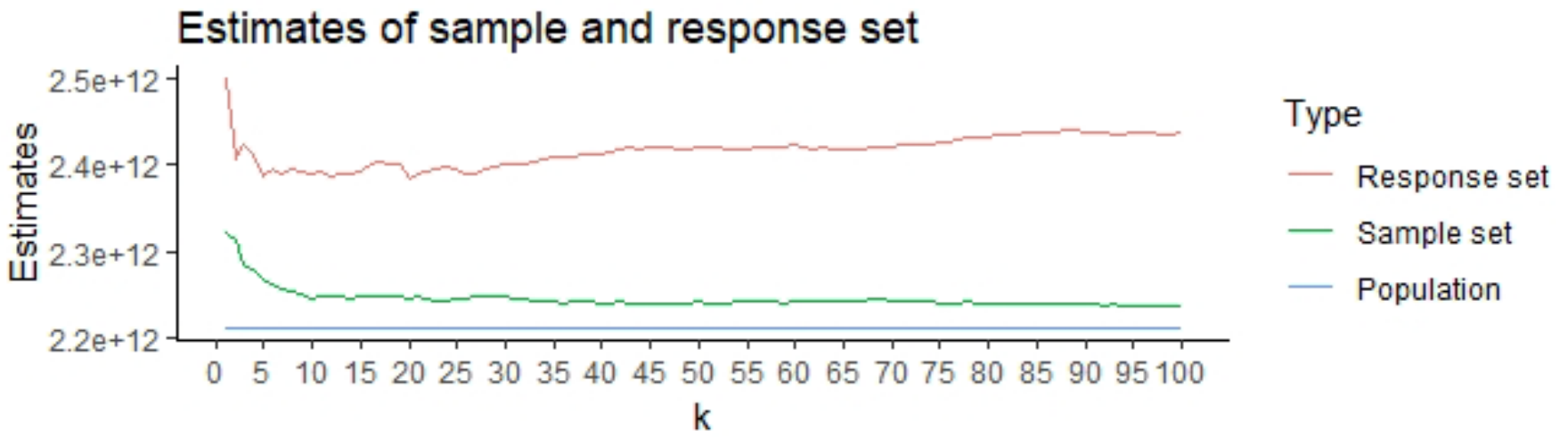
- För att hitta optimala k använde jag mig bl.a. av cross-validation (CV) indelat i segment (eller folds)
- Vilket egentligen är medelvärdet av MSE för varje segment

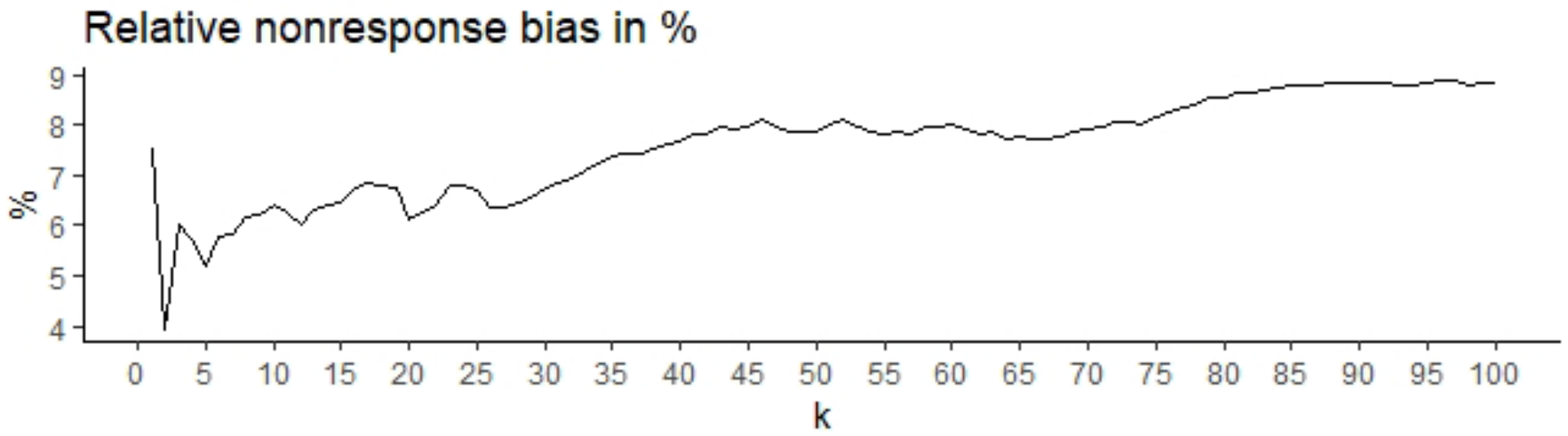
$$CV_{(m)} = \frac{1}{m}\sum_{i=1}^{m} MSE_i \qquad (9)$$

SCB

# Resultat

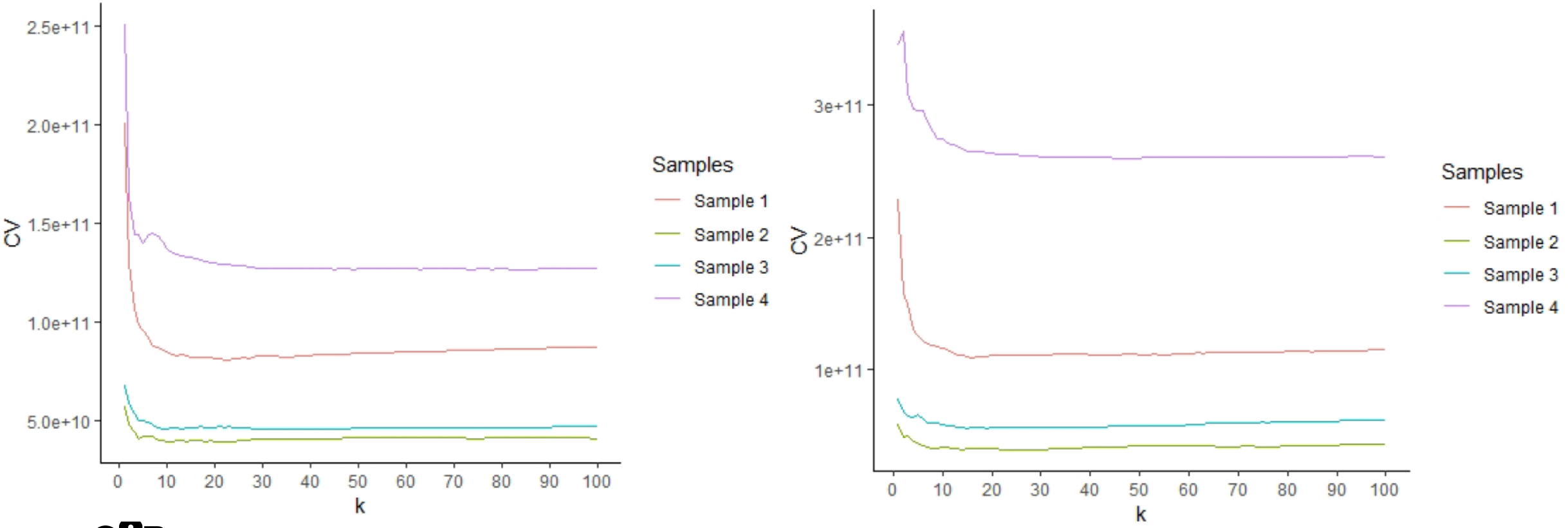
# Oviktad k-NN estimator



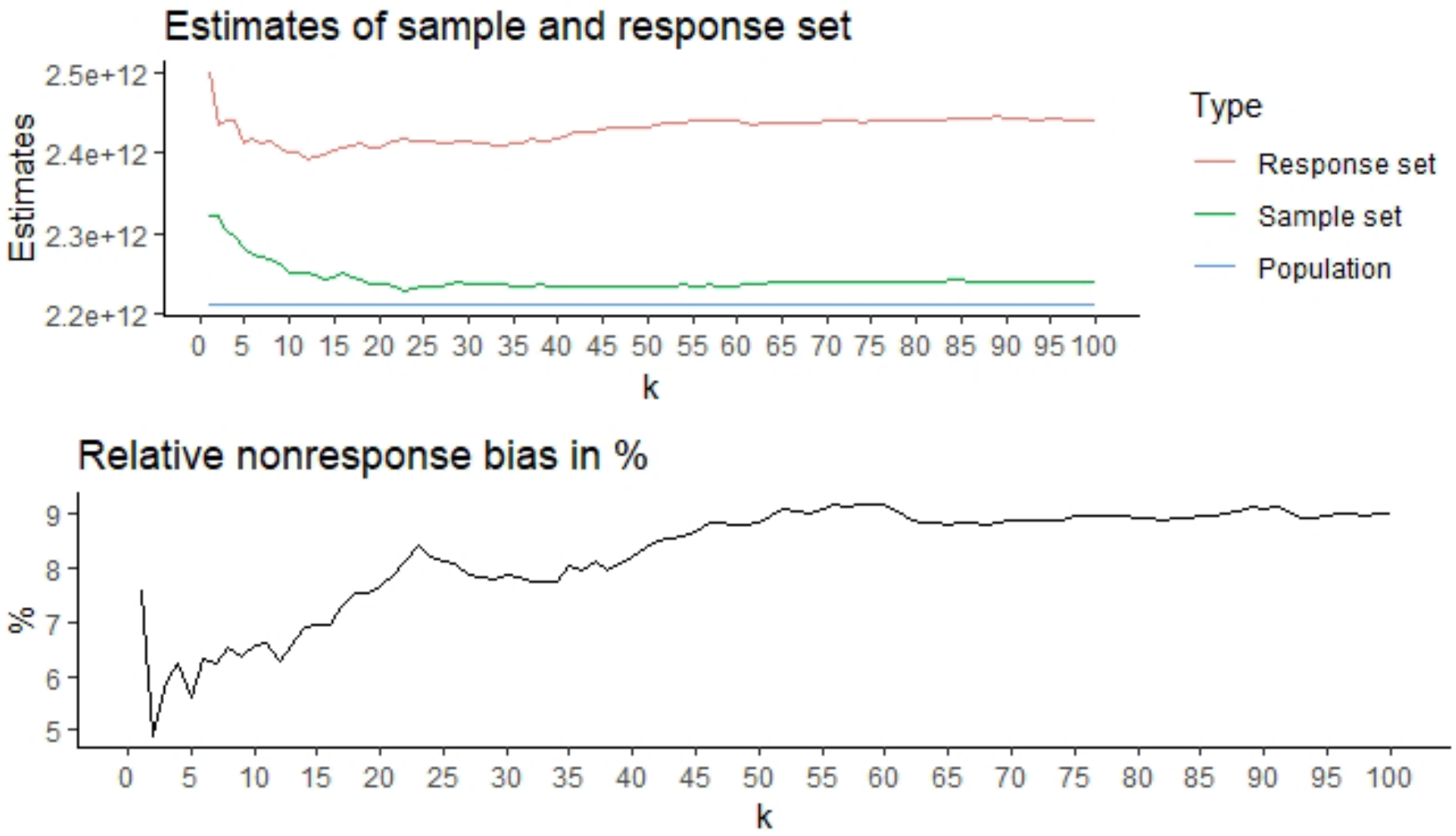

SCB

# Iviktad k-NN estimator

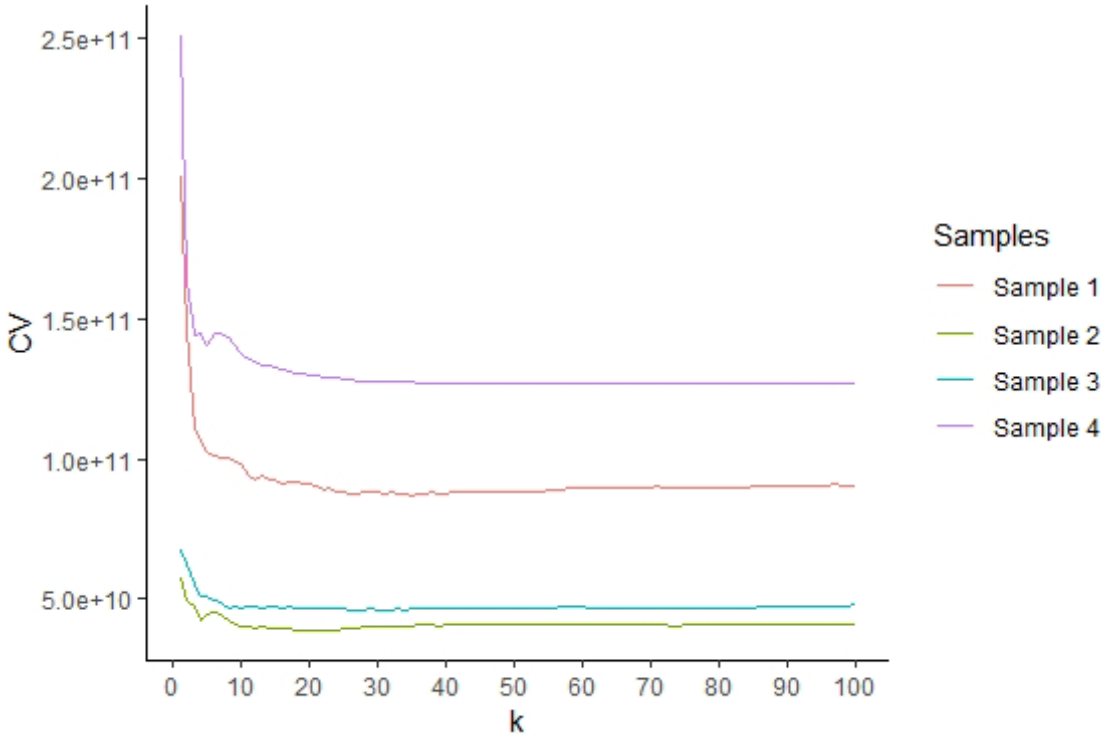


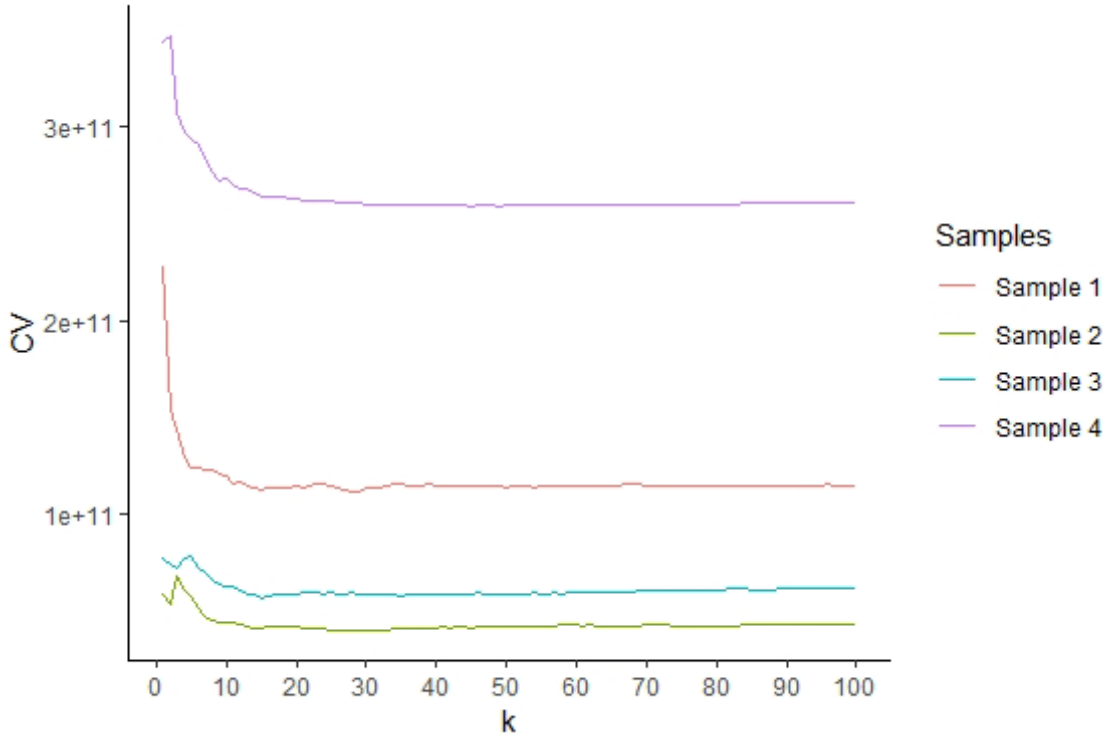

SCB

# Designviktad k-NN estimator



# Designviktad k-NN estimator

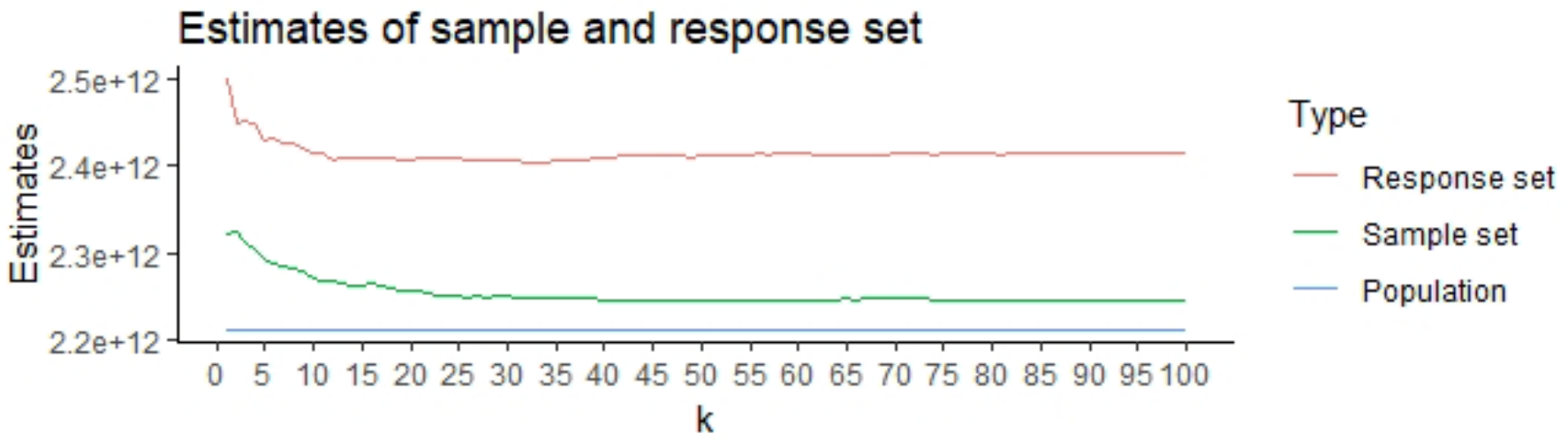


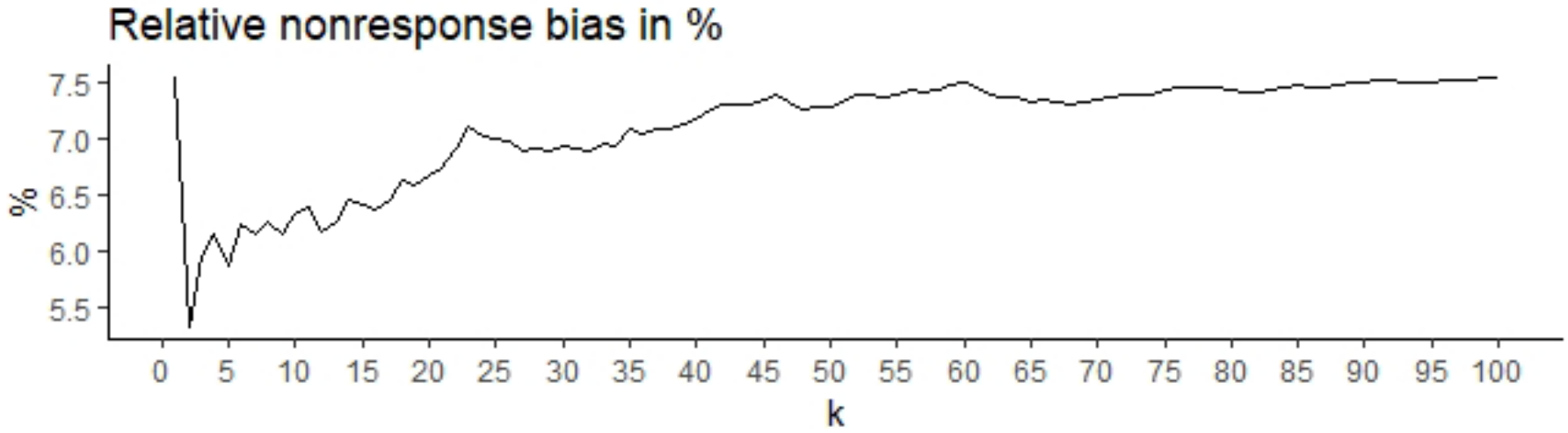




SCB

# Distans och designviktad k-NN

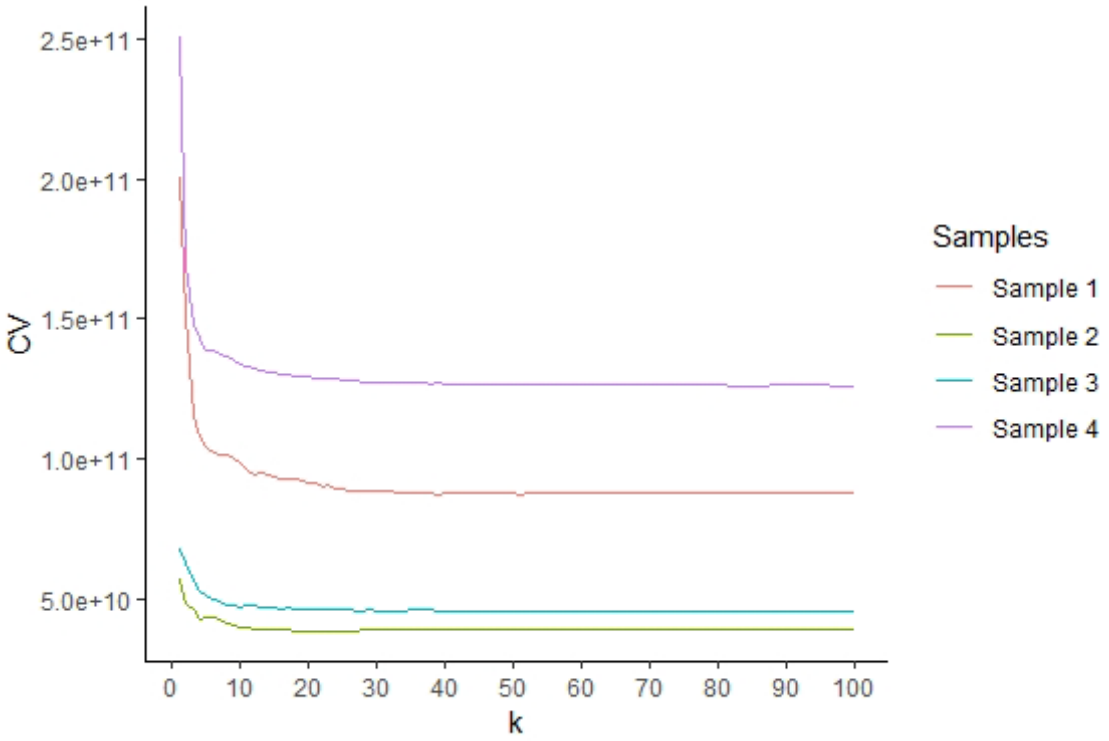
$$w_k = \frac{1}{\sum_{i=1}^{n}(q_{ik}-p_{ik})^2}$$

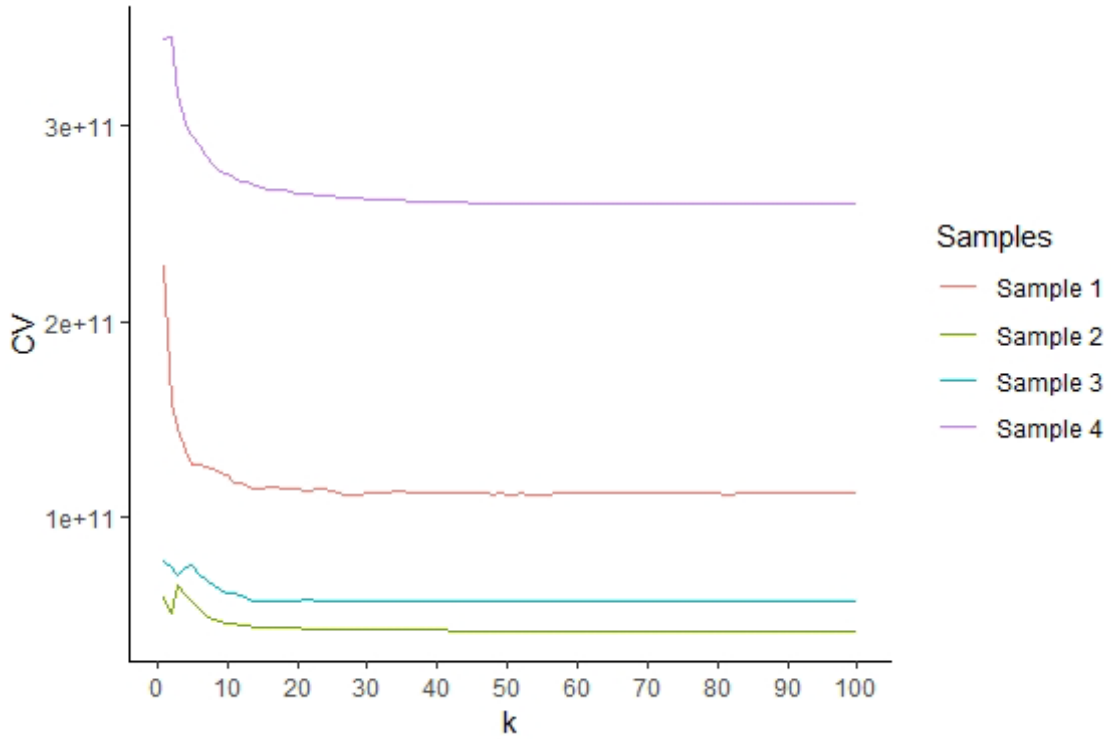



SCB

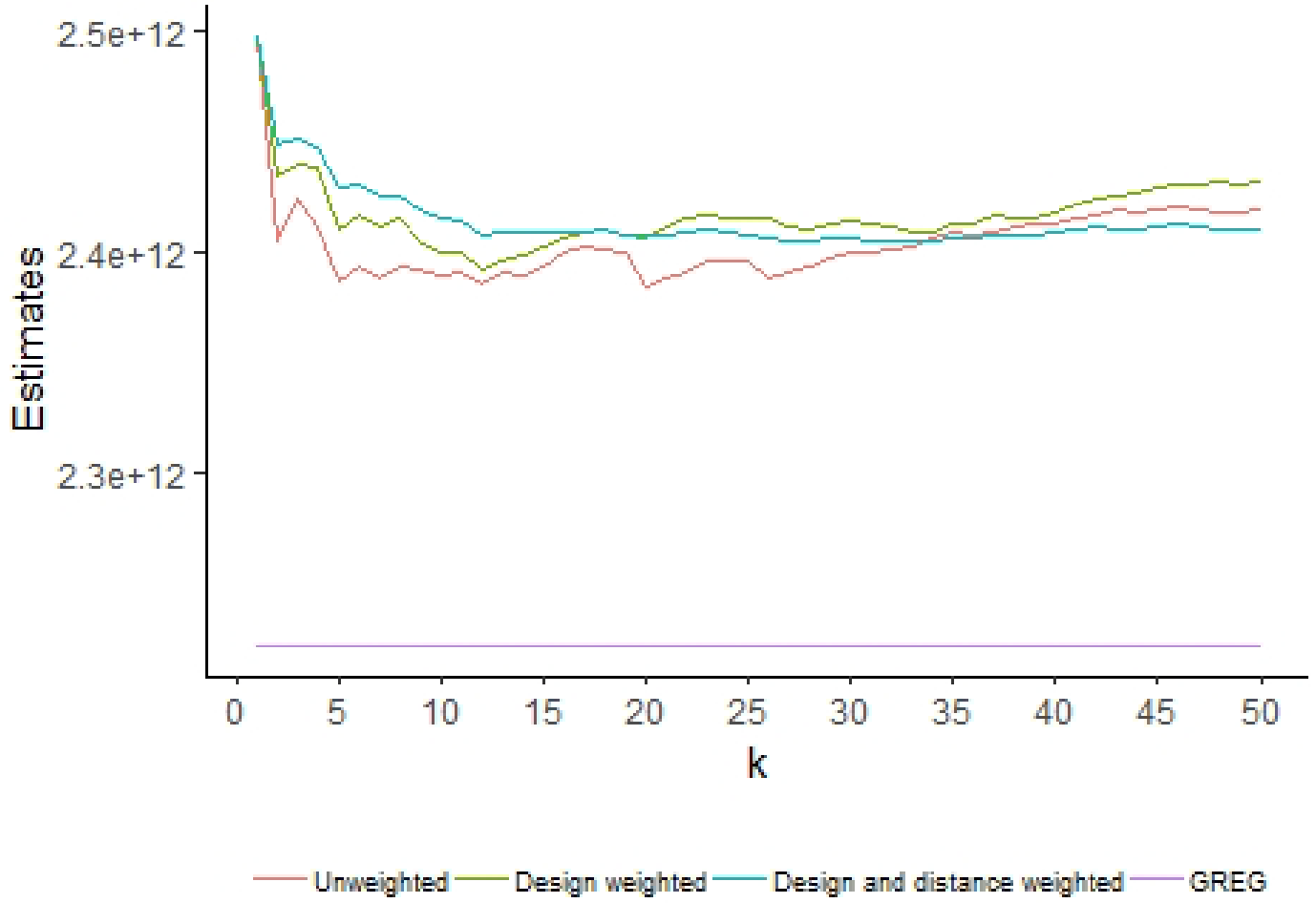
# Distans och designviktad k-NN



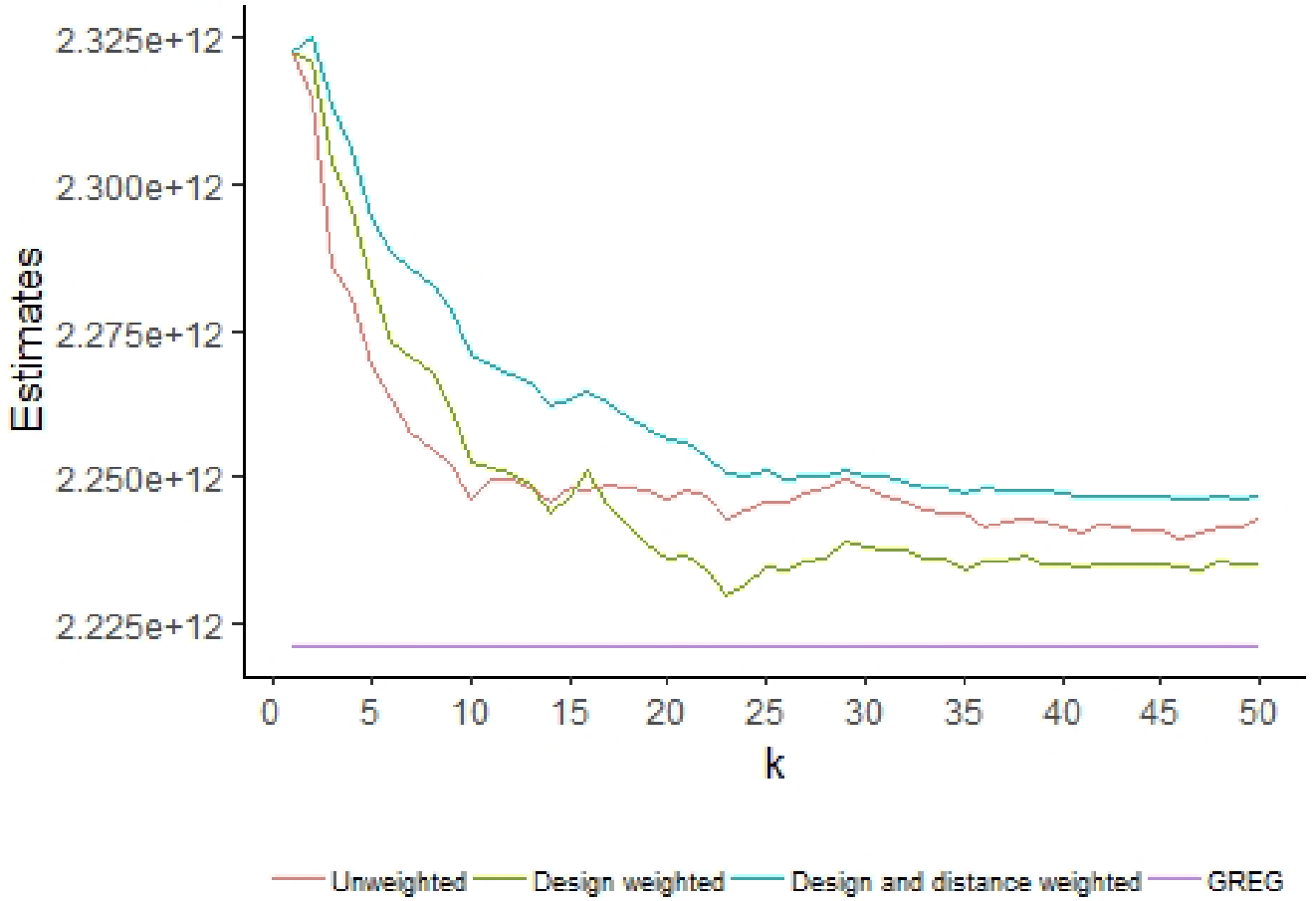




SCB

# Skillnad mellan estimatorerna Svarsmängden



# Skillnad mellan estimatorerna
# Hela urvalet



# Jämförelsetabell över estimatorerna

| Estimator | Auxiliary vector | Sample set | Response set | Relative Bias |
|---|---|---|---|---|
| Population (true) | - | 2.211 | - | - |
| HT | - | **2.213** ± 0.096 | **2.452** ± 0.171 | 10.8% |
| GREG estimator | Gender Age Bcountry Education Region Metropol | **2.221** ± 0.069 | **2.275** ± 0.125 | 2.5% |
| k-NN estimator (K=23) Unweighted | Gender Age Bcountry Education Region Metropol | **2.243** ± 0.068 | **2.395** ± 0.130 | 6.8% |
| k-NN estimator (K=23) Design weighted | Gender Age Bcountry Education Region Metropol | **2.230** ± 0.066 | **2.417** ± 0.123 | 8.4% |
| k-NN estimator (K=23) Distance and design weighted | Gender Age Bcountry Education Region Metropol | **2.250** ± 0.056 | **2.410** ± 0.107 | 7.1% |

1) All estimates are written in $10^{12}$
2) The intervall of the standard errors estimate is a 95 % confidence interval

# Slutsats?

- Det är inte alltid man kommer fram till något direkt användbart. Det blev ju inte bättre med k-NN… Men man lär sig väldigt mycket på vägen.
- k-NN var inte bättre än GREG för att minska bortfallsskevheten för fritidsfiskeundersökningen. Men den var bättre än HT-estimatorn. Sett till den hjälpinformation jag använde mig av
- Flera författare har beskrivit GREG som robust vilket mitt arbete påvisar (inte bevisar)
- k-NN är svårare att använda sig av. Det krävs en ordentlig analys för att veta att man skapar en optimal modell

SCB

# Slutsats?

- k-NN med enbart **kategoriska variabler** översätts till poststratifieringens gruppmedelvärdesmodell om antalet grannar (k) är lika med antalet observationer i varje grupp ($n_g$) av kombinationer
  - Om k är mindre kan man se prediktionerna som en skattning från ett slumpmässigt urval från varje grupp

- Hjälpinformationen och modellen är viktiga. **Stort fokus** bör läggas på att hitta relevant hjälpinformation

SCB