



Stockholms
universitet

Mätfel i åldersbestämning av fjällröding i
regressionsanalys av miljöföroreningar: en
jämförelse av errors-in-variables-modeller

Lucas Fröderberg

Kandidatuppsats 2026:5
Matematisk statistik
Maj 2026

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mätfel i åldersbestämning av fjällröding i regressionsanalys av miljöföroreningar: en jämförelse av errors-in-variables-modeller

Lucas Fröderberg*

Maj 2026

Sammanfattning

Detta examensarbete undersöker hur mätfel i åldersbestämning av fjällröding påverkar skattningen av sambandet mellan fiskens ålder och logaritmerad kvicksilverhalt. Två dataset från Naturhistoriska riksmuseet användes i analysen, ett med upprepade åldersbestämningar för att skatta mätfelets storlek och ett separat dataset med kvicksilverhalt och uppskattad ålder för regressionsanalysen. Den naiva regressionsmodellen jämfördes med två metoder för mätfelskorrigering, method of moments och SIMEX, och osäkerheten i skattningarna analyserades med parametrisk bootstrap.

Resultaten visade att den naiva modellen gav lägre lutningsskattningar än de mätfelskorrigerade metoderna. Skillnaden mellan method of moments och SIMEX var liten, medan antagandet om mätfelets storlek hade större betydelse för de korrigerade skattningarna. Arbetet visade att mätfel i ålder kan påverka skattningen av sambandet mellan ålder och kvicksilverhalt, men att antaganden som görs spelar en stor roll och bör beaktas i denna typ av analys.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: lucasfroderberg1@outlook.com. Handedare: Martin Sköld.

Förord

Denna rapport i matematisk statistik på 15 hp har genomförts i samarbete med Stockholms universitet och Naturhistoriska riksmuseet. Jag vill rikta ett stort tack till min handledare Martin Sköld för värdefulla råd och insikter under arbetets gång.

Jag vill även uppmärksamma användningen av AI-baserade språkmodeller, såsom Chat-GPT och Claude. Dessa har främst använts för att effektivisera kodning i R och LaTeX samt för språklig återkoppling och korrekturläsning.

Innehåll

1	Inledning	6
1.1	Bakgrund	6
1.2	Syfte	6
1.3	Disposition	6
2	Datamaterial	7
2.1	Åldersbestämning med replikat (agedata)	7
2.2	Kviksilverdata (hgdata)	7
2.3	Sambandet mellan dataseten	7
3	Teori och metod	8
3.1	Regression med mätfel i ålder	8
3.1.1	Enkel linjär regression	8
3.1.2	Klassiskt additivt mätfel	8
3.1.3	Attenuering och reliability ratio	9
3.1.4	Method of moments	9
3.1.5	SIMEX	10
3.2	Skattning av mätfelets komponenter	11
3.2.1	Linjär blandad modell och varianskomponenter	11
3.2.2	Relevanta varianskomponenter	11
3.3	Bootstrap för utvärdering av metoder	12
3.3.1	Bootstrap	12
3.3.2	Parametrisk bootstrap i mätfelsmodeller	12
4	Tillämpning och resultat	14
4.1	Naiv regressionsmodell för hgdata	14
4.2	Mätfelsvarians och reliability ratio	15
4.3	Korrigeringsmetoder	16
4.4	Osäkerhetsanalys	17
5	Diskussion	19
5.1	Tolkning	19
5.2	Begränsningar och antaganden	19
5.3	Överförbarhet mellan agedata och hgdata	20
6	Slutsats	20
	Referenser	21
	Appendix	22

1 Inledning

Mätningar av gifthalter i fisk är ett etablerat tillvägagångssätt inom miljöövervakning och naturvetenskaplig forskning. Kvicksilverhalt i fisk antas påverkas av fiskens ålder, eftersom fiskar som levt längre har ackumulerat kvicksilver över en längre tid. När detta samband analyseras statistiskt uppstår dock ett problem. Fiskens ålder är inte en direkt observerbar variabel, utan måste uppskattas genom åldersbestämning. Eftersom denna bestämning kan vara osäker innebär det att den förklarande variabeln i regressionsanalysen kan vara felmätt.

1.1 Bakgrund

Inom svensk miljöövervakning används fisk bland annat för att följa metaller och andra miljögifter i sjöar och marina områden [1]. Vid en statistisk analys av detta sammanhang är kvicksilverhalt en naturlig responsvariabel, medan ålder är en naturlig kovariat eftersom den kan betraktas som en proxy för exponering över tid. Beroende på art kan fiskars ålder uppskattas genom avläsning av exempelvis otoliter eller fjäll [2]. För att uppnå så hög precision som möjligt bör den metod användas som är bäst lämpad för den aktuella arten. För fjällröding, som studeras i detta arbete, har ytundersökta otoliter visat sig ge hög precision vid åldersbestämning [3].

Osäkerheten i åldersbestämningen kan bland annat bero på metodval eller på vem som utfört avläsningen. Om mätfelet i ålder ignoreras i regressionsanalysen blir skattningen av sambandet mellan ålder och gifthalt snedvriden. Detta problem brukar kallas *errors-in-variables*.

1.2 Syfte

Syftet med detta arbete är att undersöka hur mätfel i fiskars åldersbestämning påverkar skattningen av sambandet mellan ålder och logaritmerad kvicksilverhalt i fjällröding. Arbetet jämför en naiv regressionsmodell med två metoder för mätfelskorrigering, MoM och SIMEX, samt analyserar osäkerheten i de korrigerade skattningarna med parametrisk bootstrap.

1.3 Disposition

I kapitel 2 beskrivs de dataset som används i arbetet samt sambandet mellan dem. Kapitel 3 presenterar den teori och metoder som ligger till grund för analysen. I kapitel 4 tillämpas metoderna på de två dataseten, och resultat samt osäkerhetsanalys redovisas. I kapitel 5 diskuteras resultat, antaganden och begränsningar. Slutligen sammanfattas arbetets viktigaste slutsatser i kapitel 6.

2 Datamaterial

I arbetet används två dataset med olika roller. Det första används för att skatta osäkerheten i åldersbestämningen, medan det andra används i regressionsanalysen av sambandet mellan kvicksilverhalt och ålder.

2.1 Åldersbestämning med replikat (*agedata*)

Datasetet *agedata* innehåller upprepade åldersbestämningar av samma fisk utförda av flera olika oberoende läsare. För varje fisk finns mellan två och fyra tillgängliga avläsningar. Om åldersavläsningarna i *agedata* först sammanfattas till ett medelvärde per fisk varierade åldrarna mellan 2.25 och 7.00 år, med medelvärdet 4.33 och variansen 0.75. Eftersom alla fiskar inte har lika många observationer är datasetet obalanserat. Detta är viktigt i den fortsatta analysen, eftersom olika fiskar därmed bidrar med olika mycket information till skattningen av mätfelet.

Tabell 1: Fördelning av antal åldersbestämningar per fisk i *agedata*.

Antal mätningar	2	3	4
Antal fiskar	1	59	118

Tabell 1 visar att majoriteten av fiskarna har tre eller fyra åldersbestämningar. Datasetet ger därmed underlag för att studera variation både mellan fiskar och mellan läsare.

2.2 Kvicksilverdata (*hgdata*)

Datasetet *hgdata* innehåller observationer av kvicksilverhalt och uppskattad ålder för totalt 212 fiskar. För fisk i betecknas kvicksilverhalten med y_i och den observerade, uppskattade åldern med w_i . I regressionsanalysen används den logaritmerade kvicksilverhalten som responsvariabel och ålder som förklarande variabel. När kvicksilverhalt diskuteras i samband med regressionsmodellerna avses därför fortsättningsvis den logaritmerade kvicksilverhalten. Eftersom den sanna åldern inte observeras direkt, utan endast genom den uppskattade åldern w_i , är det sambandet mellan $\log(y_i)$ och w_i som kan analyseras direkt i data. De observerade åldrarna i *hgdata* varierade mellan 2 och 9 år och har medelvärdet 5.08 samt variansen 1.67.

2.3 Sambandet mellan dataseten

Båda dataseten innehåller fjällröding men inte exakt samma fiskar. Överföringen mellan dataseten bygger därför på antagandet att det mätfel som skattats från *agedata* är relevant även för de uppskattade åldrarna i *hgdata*. Dataseten skiljde sig något i åldersspridning, men hade relativt lika åldersintervall. Detta antagande diskuteras vidare i kapitel 5.

3 Teori och metod

3.1 Regression med mätfel i ålder

3.1.1 Enkel linjär regression

Som utgångspunkt betraktas den enkla linjära regressionsmodellen

$$Y_i = \beta_0 + \beta_x X_i + \varepsilon_i, \quad i = 1, \dots, N, \quad (3.1)$$

där Y_i är responsvariabeln, X_i en kovariat, β_0 interceptet, β_x regressionskoefficienten och ε_i slump termen med väntevärde noll. I detta arbete motsvarar Y_i den logaritmerade kvick-silverhalten och X_i fiskens sanna ålder, som inte observeras direkt. Modellen kan också uttryckas genom väntevärdessfunktionen

$$\mu(X_i) = E(Y_i | X_i) = \beta_0 + \beta_x X_i.$$

I den klassiska linjära regressionsmodellen behandlas kovariaten som känd utan mätfel, vilket i matrisform motsvarar att designmatrisen betraktas som given [5]. Om kovariaten i stället observeras med fel påverkas skattningen av sambandet mellan respons och kovariat, vilket motiverar mätfelsmodellen som introduceras i nästa avsnitt.

3.1.2 Klassiskt additivt mätfel

Klassiskt additivt mätfel är en typ av mätfel inom errors-in-variables-modeller som beskrivs naturligt i [7] genom att först introducera en regressionsmodell, exempelvis en enkel linjär regression med endast en kovariat som i ekvation (3.1). Antag nu att det inte är möjligt att direkt observera X_i , utan att man i stället observerar

$$W_i = X_i + U_i, \quad (3.2)$$

där U_i betecknar mätfelet (i fortsättningen av kapitlet utelämnas i-indexering för att förenkla notation). I den klassiska additiva mätfelsmodellen antas väntevärdet för mätfelet vara noll $E(U) = 0$ och konstant mätfelsvarians $\text{Var}(U) = \sigma_u^2$ samt att U är oberoende av både X och ε . Problemet uppstår då man vill undersöka sambandet mellan (Y, X) men naivt anpassar en regressionsmodell på (Y, W) , vilket motsvarar

$$Y = \beta_0 + \beta_x W + (\varepsilon - \beta_x U). \quad (3.3)$$

Det följer att denna modell inte längre har feltermen ε , utan den nya feltermen $(\varepsilon - \beta_x U)$. Skillnaden mellan dessa två regressionsmodeller är central, eftersom W i den naiva modellen är en observerad stokastisk variabel som är korrelerad med den sammansatta feltermen $(\varepsilon - \beta_x U)$. Detta kan illustreras med räknelagarna för kovarianser

$$\begin{aligned} \text{Cov}(W, \varepsilon - \beta_x U) &= \text{Cov}(X + U, \varepsilon - \beta_x U) \\ &= \text{Cov}(X, \varepsilon) - \beta_x \text{Cov}(X, U) + \text{Cov}(U, \varepsilon) - \beta_x \text{Var}(U) \\ &= -\beta_x \text{Var}(U) \neq 0. \end{aligned}$$

På grund av att den observerade kovariaten är korrelerad med feltermen i den naiva modellen, leder det till att parameterskattningen blir snedvriden. I Sundbergs kompendium [5, s.130] beskrivs problemet som att modellen inte längre kan betraktas som en vanlig regressionsmodell. För att kunna erhålla konsistenta parameterskattningar krävs därför ytterligare information om mätfelet, exempelvis dess varians σ_u^2 .

3.1.3 Attenuering och reliability ratio

I enkel linjär regression med klassiskt additivt mätfel kan biasen beskrivas explicit. Betrakta återigen modell (3.1) tillsammans med observationsmodellen (3.2), där W är den observerade felmätta kovariaten. Om man utför en naiv regression av Y på W , är den naiva regressionskoefficienten given av

$$\beta_x^* = \frac{\text{Cov}(W, Y)}{\text{Var}(W)}. \quad (3.4)$$

Eftersom U enligt antagandena är oberoende av både X och ε , kan täljaren respektive nämnaren skrivas som

$$\begin{aligned} \text{Cov}(W, Y) &= \text{Cov}(X + U, \beta_0 + \beta_x X + \varepsilon) = \beta_x \text{Var}(X), \\ \text{Var}(W) &= \text{Var}(X) + \text{Var}(U). \end{aligned}$$

Insättning i (3.4) ger

$$\beta_x^* = \frac{\beta_x \text{Var}(X)}{\text{Var}(X) + \text{Var}(U)} = \lambda \beta_x, \quad (3.5)$$

där

$$\lambda = \frac{\text{Var}(X)}{\text{Var}(X) + \text{Var}(U)} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}. \quad (3.6)$$

Storheten λ kallas i [6] för *reliability ratio* och uppfyller $0 < \lambda < 1$ i errors-in-variables-modeller. Den utgör därmed ett mått på hur starkt den naiva lutningen attenueras mot noll.

3.1.4 Method of moments

I enkel linjär regression, och om reliability ratio vore känd, så skulle en konsistent skattning av β_x kunna erhållas genom att dividera den naiva lutningsskattningen $\hat{\beta}_x^*$ med λ . Detta följer av sambandet (3.5) och i [7] kallas denna estimator för method of moments-estimatorn (MoM). Men eftersom λ i praktiken oftast inte är känd så får man skatta den. Skattningen erhålls genom att substituera σ_x^2 med $\sigma_w^2 - \sigma_u^2$ i ekvation (3.6), vilket ger den konsistenta skattningen av reliability ratio

$$\hat{\lambda} = \frac{\hat{\sigma}_w^2 - \hat{\sigma}_u^2}{\hat{\sigma}_w^2}. \quad (3.7)$$

Den praktiska MoM-estimatorn ges då av

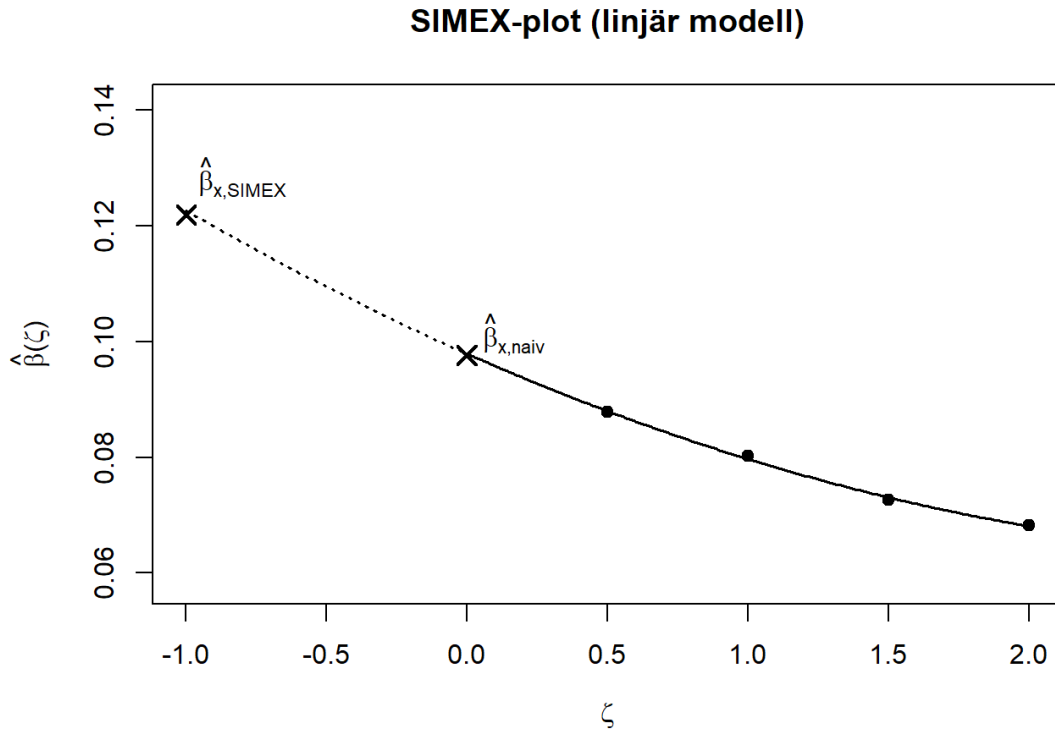
$$\hat{\beta}_{x,mm} = \frac{\hat{\beta}_x^*}{\hat{\lambda}}. \quad (3.8)$$

3.1.5 SIMEX

SIMEX (*Simulation Extrapolation* [8]) är en metod för att korrigera bias orsakad av mätfel genom att först simulera ytterligare mätfel och därefter extrapolera tillbaka till situationen utan mätfel (se närmare [7]). Idén är att undersöka hur lutningsskattningen förändras när den observerade kovariaten förorenas med successivt större mängder mätfel. I enkel linjär regression kan metoden motiveras heuristiskt genom att den naiva lutningsskattningen under ytterligare mätfel med nivå ζ har väntevärdessfunktionen

$$E(\hat{\beta}_x(\zeta)) = G(\zeta) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1 + \zeta) \sigma_u^2}, \quad \zeta \geq 0. \quad (3.9)$$

Här beskriver ζ nivån av ytterligare simulerat mätfel, och $G(\zeta)$ anger hur den förväntade lutningsskattningen förändras med denna mätfelsnivå. För varje $\zeta \geq 0$ kan pseudo-data genereras genom att addera extra mätfel till den observerade kovariaten. I praktiken upprepas simuleringen flera gånger för varje värde på ζ , där lutningsskattningarna sammanfattas genom medelvärdesbildning innan extrapoleringen utförs. Vid $\zeta = 0$ erhålls den vanliga attenuerade lutningen, eftersom $G(\zeta)$ då sammanfaller med uttrycket i (3.5). Punkten $\zeta = -1$ motsvarar frånvaro av mätfel, eftersom $(1 + \zeta) \sigma_u^2 = 0$ i (3.9). Extrapoleringen i SIMEX görs därför till denna punkt. Figur 1 illustrerar principen bakom SIMEX, där den skattade lutningen beräknas för fyra värden på ζ och därefter extrapoleras till $\zeta = -1$.



Figur 1: Illustration av SIMEX. På x-axeln anges nivån av ytterligare simulerat mätfel ζ och på y-axeln den skattade lutningen. Punkterna motsvarar de genomsnittliga skattningarna för olika värden på ζ , medan den streckade kurvan visar extrapolationsfunktionen.

3.2 Skattning av mätfelets komponenter

3.2.1 Linjär blandad modell och varianskomponenter

En linjär blandad modell används när observationer påverkas av flera olika variationskällor, där vissa effekter kan behandlas som slumpmässiga [5, s. 167–172]. För läsare j :s mätning av åldern hos fisk i kan en additiv modell med två slumpmässiga effekter skrivas som

$$A_{ij} = \mu + b_i + r_j + e_{ij}. \quad (3.10)$$

I detta arbete betecknar A_{ij} den observerade åldersbestämningen, μ ett gemensamt intercept, b_i den slumpmässiga effekten för fisk i och r_j den slumpmässiga effekten för läsare j . Residualtermen e_{ij} beskriver den återstående oförklarade variationen på observationsnivå. Om dessa effekter behandlas som slumpmässiga antas de vanligen ha väntevärde noll och varianser

$$b_i \sim N(0, \sigma_b^2), \quad r_j \sim N(0, \sigma_r^2), \quad e_{ij} \sim N(0, \sigma_e^2).$$

Varianskomponenterna σ_b^2 , σ_r^2 och σ_e^2 beskriver då variation mellan fiskar, variation mellan läsare samt återstående variation på observationsnivå.

3.2.2 Relevanta varianskomponenter

Vilka av varianskomponenterna som är relevanta för mätfel i en senare regressionsmodell beror på hur observationerna uppkommit. Om samma läsare används för samtliga observationer kan läsareffekten betraktas som en gemensam systematisk förskjutning. En sådan förskjutning påverkar främst regressionens intercept och inte lutningskoefficienten under antagandet att läsarbiasen är densamma för alla fiskar. Däremot finns fortfarande den återstående variationen på observationsnivå kvar som mätfel. I detta fall är det därför residualvariationen som är relevant för mätfelsvariansen i regressionsmodellen. Om flera läsare används och läsareffekten därmed varierar mellan observationer, bidrar även denna variation till mätfelet och bör inkluderas i mätfelsvariansen σ_u^2 . För att underlätta läsningen sammanfattas de viktigaste variansbeteckningarna i tabell 2.

Tabell 2: Sammanfattning av variansnotation som används i arbetet.

Notation	Tolkning
σ_x^2	Varians för den sanna, ej observerade åldern
σ_w^2	Varians för den observerade uppskattade åldern i <i>hgdata</i>
σ_u^2	Mätfelsvarians i regressionsmodellen
σ_b^2	Variation mellan fiskar i den blandade modellen
σ_r^2	Variation mellan läsare i den blandade modellen
σ_e^2	Residualvariation på observationsnivå i den blandade modellen
σ_ε^2	Residualvariens i regressionsmodellen för log-kvicksilverhalt

3.3 Bootstrap för utvärdering av metoder

För de korrigerade estimatorerna är strukturen av osäkerhet mer komplicerad än i vanlig linjär regression, eftersom skattningarna beror både på den naiva modellen och mätfelsvariansen. Eftersom SIMEX bygger på simulering och extrapolering är det svårt att ta fram explicita uttryck för bias och varians. I detta arbete används därför bootstrap för att approximera estimatorernas fördelning och skapa jämförbara osäkerhetsmått för både MoM och SIMEX.

3.3.1 Bootstrap

Bootstrap introducerades av [9] som en metod för att approximera fördelningen av en estimator genom simulering. Det fundamentala är att den okända datafördelningen F ersätts med en skattad fördelning $\hat{F}$. Därefter genereras bootstrapstickprov från $\hat{F}$ där estimatoren beräknas på nytt för varje iteration.

Låt $\hat{\theta} = T(x)$ beteckna punktestimatet beräknat från originaldatan x . För $r = 1, \dots, R$ genereras bootstrapstickprov x_r^* från $\hat{F}$ med motsvarande bootstrapreplikat som

$$\hat{\theta}_r^* = T(x_r^*), \quad r = 1, \dots, R.$$

Bootstrap-bias ger ett mått på eventuell systematisk förskjutning hos estimatoren under den skattade modellen och beräknas som $\hat{b} = \bar{\theta}^* - \hat{\theta}$, där

$$\bar{\theta}^* = \frac{1}{R} \sum_{r=1}^R \hat{\theta}_r^*.$$

Bootstrap-variansen $\hat{v}$ skattas med

$$\hat{v} = \frac{1}{R-1} \sum_{r=1}^R (\hat{\theta}_r^* - \bar{\theta}^*)^2,$$

där bootstrap-standardfelet definieras som $\widehat{SE}_{boot} = \sqrt{\hat{v}}$. Med tillgång till skattade bootstrap-bias $\hat{b}$ och varians $\hat{v}$ kan det sammanvägda måttet på systematisk förskjutning och variation bootstrap-medelkvadratfelet beräknas enligt

$$\widehat{MSE}_{boot} = \hat{b}^2 + \hat{v}. \quad (3.11)$$

Ett percentilbaserat konfidensintervall för θ ges av de empiriska $\alpha/2$ och $1-\alpha/2$ -kvantilerna av bootstrapreplikaten

$$\left(\hat{\theta}_{(\alpha/2)}^*, \hat{\theta}_{(1-\alpha/2)}^* \right).$$

3.3.2 Parametrisk bootstrap i mätfelsmodeller

I detta arbete används parametrisk bootstrap [10], vilket innebär att bootstrapstickproven inte erhålls genom återsampling av den empiriska datan, utan genereras genom simulering från en skattad parametrisk modell. I mätfelsmodeller där den sanna kovariaten inte observeras direkt kan bootstrapstickprov genereras genom att först simulera den latent

kovariaten, där variansen skattas utifrån sambandet i ekvation (3.2), och därefter simulera mätfel och residualer. Ett bootstrapstickprov kan då genereras enligt

$$X^* \sim N(\hat{\mu}_w, \hat{\sigma}_w^2 - \hat{\sigma}_u^2), \quad U^* \sim N(0, \hat{\sigma}_u^2), \quad \varepsilon^* \sim N(0, \hat{\sigma}_\varepsilon^2).$$

Den observerade kovariaten och responsen ges då av

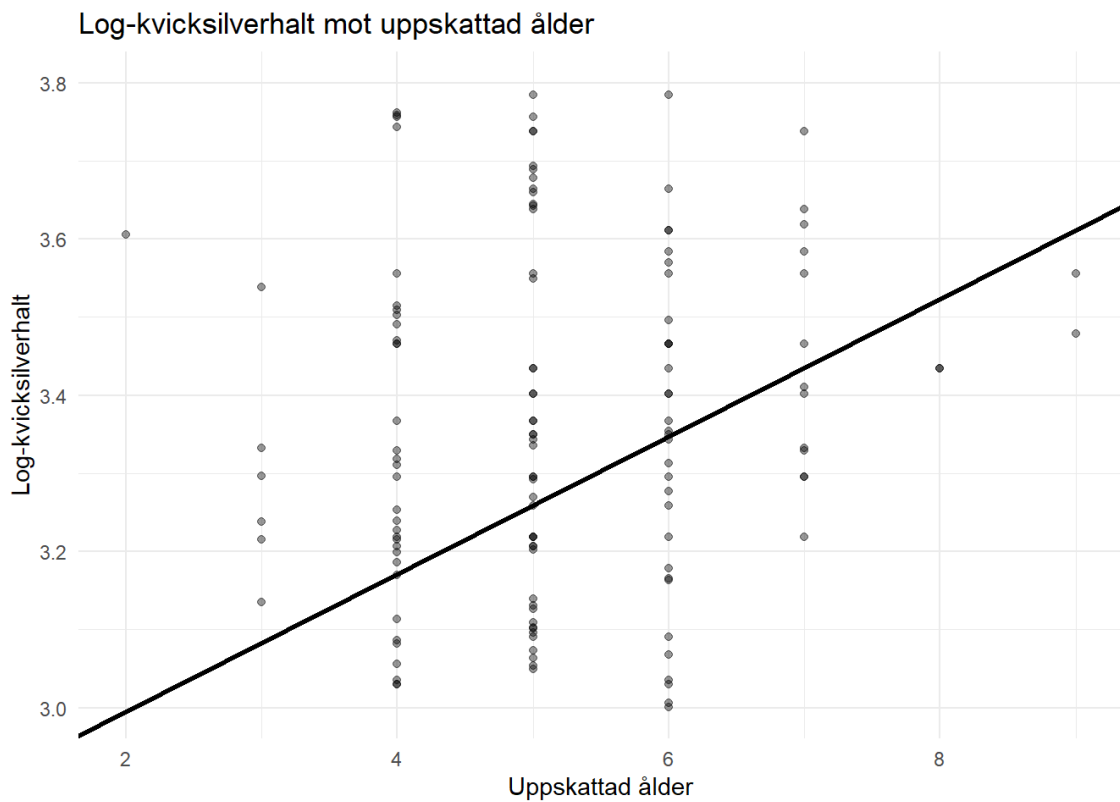
$$W^* = X^* + U^*, \quad Y^* = \hat{\beta}_0 + \hat{\beta}_x X^* + \varepsilon^*. \quad (3.12)$$

Samma metod skattas därefter på nytt i varje bootstrapstickprov, och bootstrapreplikaten används för att beräkna bias, standardfel, MSE och konfidensintervall enligt avsnitt 3.3.1. Bootstraposäkerheten ska därmed tolkas som osäkerhet under den antagna parametriska mätfelsmodellen, snarare än som en helt modellfri osäkerhetsuppskattning.

4 Tillämpning och resultat

4.1 Naiv regressionsmodell för hgdata

För att undersöka sambandet mellan kvicksilverhalten och uppskattade åldrar anpassades en naiv regressionsmodell för datasetet *hgdata* enligt ekvation (3.3). Detta genomfördes i R med funktionen `lm` och resultatet visas i figur 2.



Figur 2: Figuren visar log-kvicksilverhalten modellerad som en funktion av uppskattade åldrar på fjällröding.

Modellen gav skattningarna $\hat{\beta}_0 = 2.819$, $\hat{\beta}_x^* = 0.0881$ och $\widehat{\text{Var}}(W) = 1.67$ (se appendix för detaljer).

4.2 Mätfelsvarians och reliability ratio

För att skatta mätfelet i åldersbestämning av fjällröding anpassades den linjära blandade modellen i ekvation (3.10) på datasetet *agedata*, som innehåller upprepade åldersbestämningar. Modellen gjorde det därmed möjligt att särskilja variation mellan fiskar, läsare och residualvariation. Analysen genomfördes med funktionen `lmer` från paketet `lme4` [12] och resultaten presenteras i tabell 3.

Tabell 3: Skattade varianskomponenter från den linjära blandade modellen för *agedata*.

Komponent	Variansskattning	Tolkning
Fisk (σ_b^2)	0.6428	Variation mellan fiskar
Läsare (σ_r^2)	0.1216	Variation mellan läsare
Residual (σ_e^2)	0.3330	Återstående variation

Den största variationen återfanns mellan fiskar, men eftersom det inte är säkert att den variationen är relevant för *hgdata* så är det läsare- och residualvariation som är av intresse i den fortsatta analysen.

Dessa varianskomponenter användes därefter för att skatta reliability ratio genom ekvation (3.7). Enligt resonemanget i avsnitt 3.2.2 beror valet av mätfelsvarians på hur observationerna i *hgdata* har uppkommit. Därför redovisas de två möjliga fallen i tabell 4, där det visade sig att scenario 2 gav större mätfelsvarians och därmed lägre reliability ratio än scenario 1.

Tabell 4: Skattade mätfelsvarianser och reliability ratios, där σ_w^2 skattats från *hgdata*.

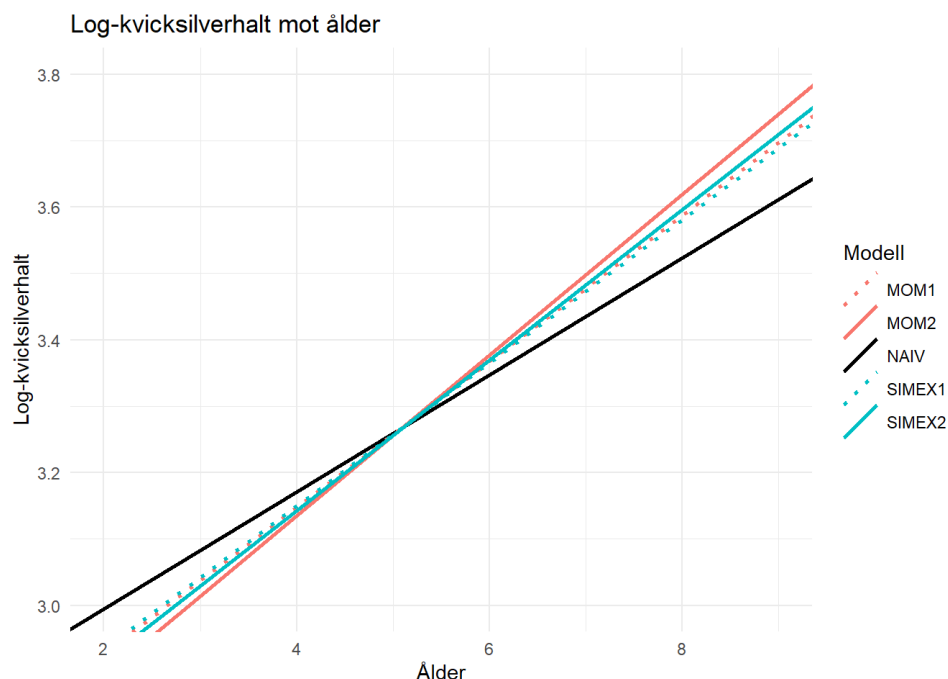
Scenario	Mätfelsvarians	σ_u	$\hat{\lambda}$
1	$\sigma_u^2 = \sigma_e^2$	0.577	0.801
2	$\sigma_u^2 = \sigma_e^2 + \sigma_r^2$	0.674	0.728

4.3 Korrigeringsmetoder

För att korrigera den attenuerade lutningskoefficienten $\hat{\beta}_x^*$ tillämpades två korrigeringsmetoder. MoM tillämpades med de skattade reliability ratios i tabell 4 enligt ekvation (3.8). SIMEX tillämpades på den naiva modellen i R med funktionen `simex` från paketet `simex` [11], där de skattade standardavvikelserna från tabell 4 användes som indata. För varje mätfelsnivå genomfördes 100 simuleringar med standardinställningarna, vilket motsvarar nivåerna $\zeta \in \{0.5, 1, 1.5, 2\}$. Resultaten redovisas i tabell 5 och figur 3.

Tabell 5: Skattade regressionskoefficienter för *hgdata* med naiva modellen, MoM och SIMEX för de två mätfelsstandardavvikelserna.

Metod	σ_u	$\hat{\beta}_0$	$\hat{\beta}_x$
Naiv	—	2.819	0.0881
SIMEX	0.577	2.720	0.1077
MoM	0.577	2.708	0.1100
SIMEX	0.674	2.690	0.1133
MoM	0.674	2.652	0.1210



Figur 3: Jämförelse mellan den naiva regressionslinjen och de mätfelskorrigerade regressionslinjerna för MoM och SIMEX i scenario 1 och 2. Linjerna visas över åldersintervallet i *hgdata*.

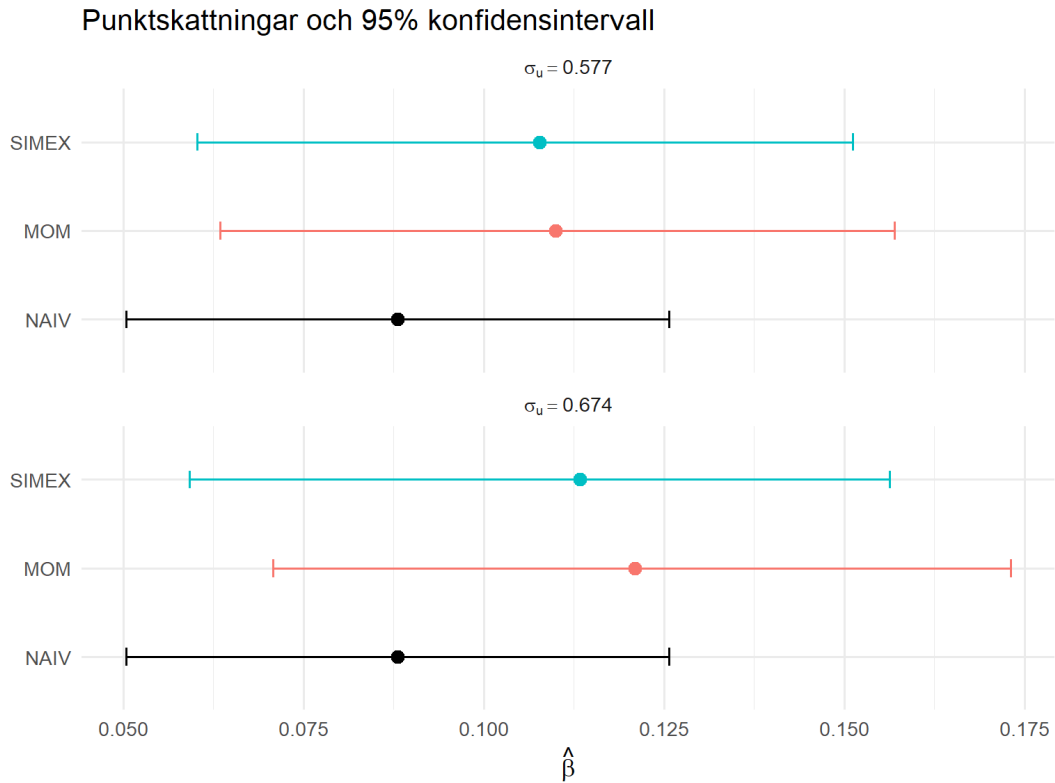
Den naiva modellen gav den lägsta lutningskoefficienten. För ett givet mätfelsscenario gav MoM något högre skattning av $\hat{\beta}_x$ än SIMEX. Lutningsskattningarna i scenario 2 var högre än i scenario 1.

4.4 Osäkerhetsanalys

Osäkerheten i de korrigerade estimatorerna skattades med parametrisk bootstrap enligt avsnitt 3.3.2. Bootstrapproceduren genomfördes separat för både MoM och SIMEX under de två skattade mätfelsscenarierna. För varje metod och scenario genererades 1000 bootstrapdataset med samma stickprovsstorlek som *hgdata*. För respektive metod användes de tidigare metods specifika skattningarna av intercept, lutningskoefficient och residualspridning (se tabell 5 och appendix B). Resultaten redovisas i tabell 6 och figur 4.

Tabell 6: Parametrisk bootstrap för MoM och SIMEX vid båda mätfelsscenarierna. Tabellen visar punktskattning $\hat{\beta}_x$, bootstrap-bias, bootstrap-standardfel samt 95% bootstrap-konfidensintervall.

Metod	σ_u	$\hat{\beta}_x$	Bootstrap-bias	Bootstrap-SE	Nedre 95% KI	Övre 95% KI
MoM	0.577	0.110	-0.000896	0.0235	0.0635	0.157
SIMEX	0.577	0.108	-0.00322	0.0236	0.0602	0.151
MoM	0.674	0.121	0.000566	0.0272	0.0708	0.173
SIMEX	0.674	0.113	-0.00697	0.0250	0.0592	0.156



Figur 4: Figuren visar punktskattningarna $\hat{\beta}_x$ och 95% bootstrap-konfidensintervall för korrigerings estimatorerna. För jämförelse visas även punktskattning och konfidensintervall för den naiva modellen.

Bootstrap-bias var liten för både MoM och SIMEX, men i absolutbelopp något mindre för MoM. För $\sigma_u = 0.577$ var bootstrap-standardfelen relativt lika mellan metoderna. För $\sigma_u = 0.674$ ökade bootstrap-standardfelet för båda metoderna, men ökningen var tydligare för MoM än för SIMEX. De percentilbaserade 95%-konfidensintervallen överlappade tydligt, medan punktskattningarna ökade när mätfelsstandardavvikelsen ökade. Figur 4 illustrerar bootstrapresultaten grafiskt och bekräftar de tydligt överlappande intervallen för de korregerade estimatorerna.

Det bootstrapbaserade medelkvadratfelet beräknades enligt (3.11). Tabell 7 visar att värdet var lägst för MoM när $\sigma_u = 0.577$, medan det var lägst för SIMEX när $\sigma_u = 0.674$.

Tabell 7: Bootstrapbaserat medelkvadratfel för MoM- och SIMEX-estimatorerna.

Metod	σ_u	$\widehat{MSE}_{boot}$
MoM	0.577	0.000553
SIMEX	0.577	0.000567
SIMEX	0.674	0.000674
MoM	0.674	0.000740

5 Diskussion

5.1 Tolkning

Sammantaget visade resultaten att båda korrigeringsmetoderna gav högre skattningar av lutningskoefficienten $\hat{\beta}_x$ än den naiva modellen. Detta talar för att mätfel i den observerade åldern ledde till attenuering i den naiva analysen. Skillnaden mellan MoM och SIMEX var liten, medan antagandet om mätfelets storlek hade tydligare betydelse för korrigeringarnas storlek.

Osäkerhetsanalysen visade samtidigt att skillnaden mellan metoderna var liten i förhållande till osäkerheten i skattningarna, vilket framgick av de tydligt överlappande bootstrapkonfidensintervallen i figur 4. Trots detta visade de korrigerade estimatorerna ett stabilt mönster, där punktskattningarna ökade vid större antaget mätfel. De bootstrapbaserade medelkvadratfelen gav ingen entydig rangordning mellan metoderna. För $\sigma_u = 0.577$ gav MoM det lägsta värdet, medan SIMEX gav det lägsta värdet för $\sigma_u = 0.674$. Någon enskild metod framstod därmed inte som bäst i båda mätfelsscenarierna. Resultaten tyder på att metodskillnaden var tydligare vid större antaget mätfel, även om skillnaden mellan MoM och SIMEX överlag var liten.

5.2 Begränsningar och antaganden

Reliability ratio har varit central genom arbetet, och därför är det viktigt att påpeka att det inte bara är mätfelsvariansen som påverkar den, utan även åldersvariationen σ_w^2 . De uppskattade åldrarna i datasetet *hgdata* låg inom ett relativt begränsat intervall (se figur 9 i appendix), särskilt i relation till att fjällröding kan leva upp till 20-40 år [4]. Hade åldrarna i stället uppvisat större variation hade reliability ratio skattats högre, och därmed gett lägre korrigeringar. Detta resonemang bygger dock på antagandet att mätfelsvariansen är konstant över olika åldrar, vilket sannolikt är en förenkling av verkligheten.

Ett annat antagande i arbetet är att mätfelet följer en klassisk additiv mätfelsmodell. I [7] beskrivs denna struktur som att den observerade kovariaten avviker från den sanna genom en additiv mätfelskomponent. Eftersom observerade åldrar i *hgdata* anges i heltal kan denna modell dock vara en förenkling. Om en läsares bedömning exempelvis ligger mellan två heltal kan den observerade åldern behöva avrundas, vilket innebär att en del av avvikelsen mellan sann och observerad ålder kan bero på avrundning och inte enbart på mätfel. Om sådana effekter förekommer kan den verkliga mätfelsstrukturen därför vara mer komplex än vad den klassiska additiva modell som antas i arbetet förutsätter.

Även osäkerhetsanalysen bygger på modellantaganden. Den parametriska bootstrappen genererade nya observationer från den skattade latent mätfelsmodellen och skattar därmed variationen hos estimatorerna under denna modell. Bootstrapintervallen bör därför tolkas som villkorade på den antagna mätfelsstrukturen och de valda mätfelsscenarierna.

Valet mellan de två mätfelsscenarierna är också en viktig osäkerhetskälla i analysen. Enligt tillgänglig information framstod det som troligt att åldersbestämningarna i *hgdata* hade utförts av en och samma läsare. I så fall är scenario 1, där endast residualvariationen bidrar till mätfelet, mer rimligt än scenario 2 för detta dataset. Scenario 2 kan därför ses

som en illustration av hur resultaten påverkas om även variation mellan läsare tillåts bidra till mätfelet, alltså som en slags känslighetsanalys. Detta skulle i sin tur innebära att den relevanta mätfelskorrigeringen sannolikt var den lägre av de två redovisade.

5.3 Överförbarhet mellan *agedata* och *hgdata*

En central förutsättning i arbetet är att den mätfelsvarians som skattats från *agedata* är relevant även för de uppskattade åldrarna i *hgdata*. Detta motsvarar ett antagande om att den del av mätfelsmodellen som används i analysen är transportabel mellan de två dataseten. Som Carroll påpekar i [7, s. 29–30] är sådana antaganden inte något som verifieras direkt, utan måste genomgå en rimlighetsbedömning. Detta antagande får ett visst stöd av att de två dataseten låg inom relativt liknande åldersintervall, även om deras åldersspridning inte var identisk. Om läsarens eller läsarnas skicklighet i åldersbestämning skiljer sig mellan *hgdata* och *agedata* blir förstås överförbarhetsantagandet svagare. Detta skulle i sin tur kunna leda till felaktiga skattningar av reliability ratio och därmed de korrigerade lutningsskattningarna. En tänkbar förbättring av analysen hade varit att ha tillgång till upprepade åldersbestämningar för samma fiskar som också ingick i *hgdata*. Det hade minskat behovet av att överföra mätfelsvariansen mellan de två dataseten.

6 Slutsats

Syftet med arbetet var att undersöka hur mätfel i åldersbestämning av fjällröding påverkar skattningen av sambandet mellan fiskens ålder och kvicksilverhalt, där kvicksilverhalten analyserades på en logaritmerad skala. För detta skattades mätfelets storlek med hjälp av ett dataset med upprepade åldersbestämningar, varefter denna skattning användes i regressionsanalysen av ett separat dataset med kvicksilverhalt och uppskattad ålder.

Resultaten visade att den naiva regressionsmodellen gav en lägre lutningsskattning än MoM och SIMEX. Större antagen mätfelsvarians gav lägre reliability ratio och därmed starkare mätfelskorrigering. MoM gav genomgående något högre skattningar än SIMEX, men skillnaderna mellan metoderna var små i förhållande till osäkerheten i skattningarna.

Bootstrapanalysen visade att ingen av metoderna framstod som entydigt bäst i båda mätfelsscenarierna. Resultaten tyder därför på att antagandet om mätfelets storlek och struktur hade större betydelse för de korrigerade skattningarna än valet mellan MoM och SIMEX.

Sammantaget visar arbetet att mätfel i ålder kan påverka skattningen av sambandet i denna typ av regressionsanalys och därför bör beaktas, särskilt när hög precision eftersträvas. Samtidigt bygger resultaten på flera antaganden om mätfelsscenarier och överförbarhet mellan dataset, vilket bör beaktas vid tolkningen.

Referenser

- [1] Havs- och vattenmyndigheten, *Miljöövervakningens metoder och övervakningsmanualer inom programområde Kust och hav*, uppdaterad 2023-05-17.
- [2] NOAA Fisheries, *Age and Growth*, uppdaterad 2025-07-28.
- [3] Timothy T. Baker and L. Saree Timmons, *Precision of Ages Estimated from Five Bony Structures of Arctic Char (*Salvelinus alpinus*) from the Wood River System, Alaska*, 1991.
- [4] Fisheries and Oceans Canada, *Arctic Char (*Salvelinus alpinus*)*, uppdaterad 2025-07-04.
- [5] Rolf Sundberg, *Lineära Statistiska Modeller kompendium*, 2025.
- [6] Wayne A. Fuller, *Measurement Error Models*, 1987.
- [7] R. Carroll et al., *Measurement Error in Nonlinear Models: A Modern Perspective*, 2006.
- [8] Cook and Stefanski, *Simulation-Extrapolation Estimation in Parametric Measurement Error Models*, 1994.
- [9] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*, 1979.
- [10] A. C. Davison, *Bootstrap Methods and their Application*, 2021.
- [11] Wolfgang Lederer and Heidi Seibold, *simex: SIMEX- and MCSIMEX-Algorithm for Measurement Error Models*, R package version 1.8.
- [12] Douglas Bates, Martin Mächler, Ben Bolker and Steve Walker, *Fitting Linear Mixed-Effects Models Using lme4*, 2015.

Appendix

A

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: age_estimate ~ 1 + (1 | id) + (1 | reader)
## Data: agedata
##
## REML criterion at convergence: 1519.3
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.8607 -0.5613 -0.0230  0.5869  3.2065
##
## Random effects:
## Groups   Name      Variance Std.Dev.
## id      (Intercept) 0.6428   0.8018
## reader  (Intercept) 0.1216   0.3487
## Residual                0.3330   0.5771
## Number of obs: 651, groups: id, 178; reader, 5
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)    4.412      0.169    26.1
```

Figur 5: Utskrift av linjär blandad modell på agedata.

```
##
## Call:
## lm(formula = log_value ~ age, data = data, x = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74506 -0.26063 -0.03002  0.21481  1.12114
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.81936    0.10024  28.126 < 2e-16 ***
## age         0.08807     0.01911   4.609 7.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3587 on 210 degrees of freedom
## Multiple R-squared:  0.09185, Adjusted R-squared:  0.08753
## F-statistic: 21.24 on 1 and 210 DF, p-value: 7.028e-06
```

Figur 6: Utskrift av naiv modell på hgdata.

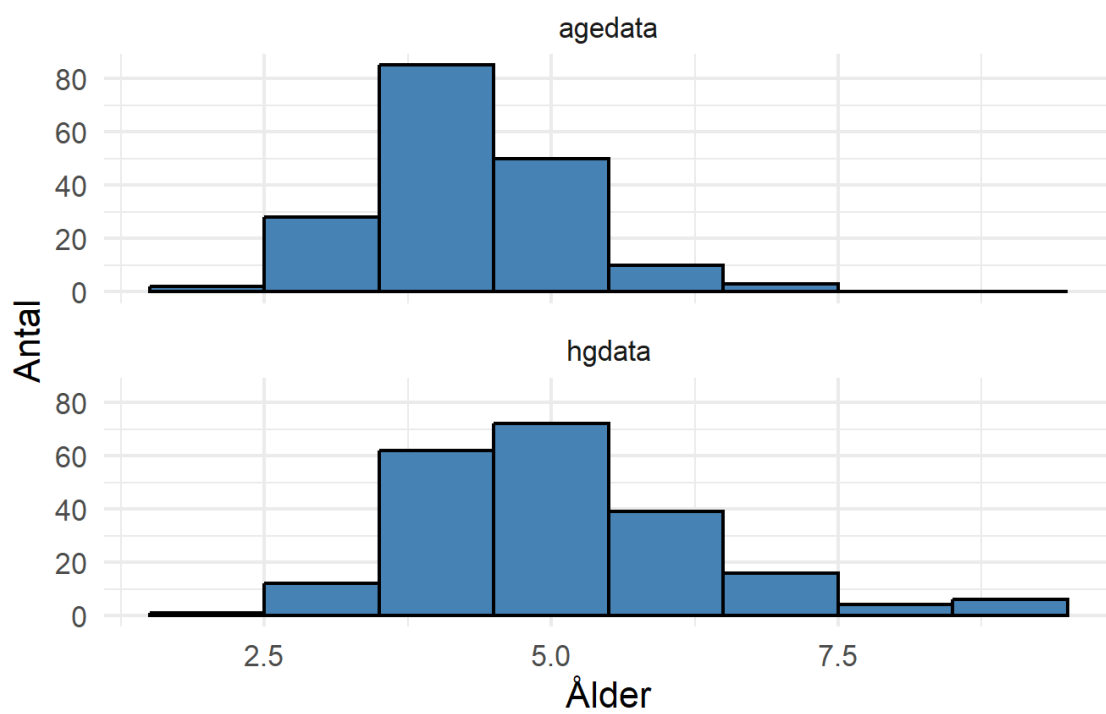
```
## Call:
## simex(model = naive_model, SIMEXvariable = "age", measurement.error = sqrt(sigmaone),
##       B = 100)
##
## Naive model:
## lm(formula = log_value ~ age, data = data, x = TRUE)
##
## Simex variable :
##              age
## Measurement error : 0.5770984
##
## Number of iterations: 100
##
## Residuals:
##      Min.      1st Qu.      Median      Mean      3rd Qu.      Max.
## -0.7634200 -0.2437167 -0.0382952 -0.0003533  0.2067367  1.1618208
##
## Coefficients:
##
## Asymptotic variance:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.71963    0.11622   23.401 < 2e-16 ***
## age          0.10775    0.02082    5.175 5.31e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Jackknife variance:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.7196    0.1208   22.510 < 2e-16 ***
## age          0.1077    0.0232    4.644 6.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figur 7: Utskrift av SIMEX på hgdata med $\hat{\sigma}_{(1)}$.

```
## Call:
## simex(model = naive_model, SIMEXvariable = "age", measurement.error = sqrt(sigmamultiple),
##       B = 100)
##
## Naive model:
## lm(formula = log_value ~ age, data = data, x = TRUE)
##
## Simex variable :
##              age
## Measurement error : 0.6742663
##
## Number of iterations: 100
##
## Residuals:
##      Min.      1st Qu.      Median      Mean      3rd Qu.      Max.
## -0.7676807 -0.2367878 -0.0341637  0.0005058  0.2081208  1.1743446
##
## Coefficients:
##
## Asymptotic variance:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.69033    0.12492   21.537 < 2e-16 ***
## age          0.11334    0.02248    5.041 9.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Jackknife variance:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.69033    0.12038   22.348 < 2e-16 ***
## age          0.11334    0.02312    4.903 1.89e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figur 8: Utskrift av SIMEX på hgdata med $\hat{\sigma}_{(2)}$.

Jämförelse av åldersfördelning i agedata och hgdata



Figur 9: Fördelningen av åldrar i *agedata* vs *hgdata*.

B

Residualspridningen som användes i den parametriska bootstrappen skattades utifrån sambandet mellan residualen mot den observerade kovariaten och residualvariationen i den latenta modellen. Utgångspunkten var den naiva modellen, tidigare definierad som

$$Y = \beta_0 + \beta_x W + (\varepsilon - \beta_x U).$$

Om vi definierar residualen mot den observerade kovariaten W som

$$R = Y - \beta_0 - \beta_x W,$$

följer att

$$R = \varepsilon - \beta_x U.$$

Under antagandet att U och ε är oberoende och har båda väntevärde noll fås

$$\text{Var}(R) = \text{Var}(\varepsilon - \beta_x U) = \sigma_\varepsilon^2 + \beta_x^2 \sigma_u^2,$$

vilket efter omarrangering ger

$$\sigma_\varepsilon^2 = \text{Var}(R) - \beta_x^2 \sigma_u^2.$$

Denna beräkning användes för både MoM och SIMEX. För respektive metod beräknades residualerna som

$$\hat{R}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_x W_i,$$

där $\hat{\beta}_0$ och $\hat{\beta}_x$ är metodens korrigerade intercept och lutningsskattning. Därefter skattades residualvariansen i den latent modellen enligt

$$\hat{\sigma}_\varepsilon^2 = \widehat{\text{Var}}(\hat{R}) - \hat{\beta}_x^2 \hat{\sigma}_u^2.$$