Analyseinstitut for Forskning

Effekten af vægtning og korrektion for partielt bortfald i modeller for regionale og andre forskelle i danske virksomheders forskning og udvikling

The Danish Institute for Studies in
Research and Research Policy
Finlandsgade 4
DK-8200 Aarhus N

Working papers 2001/9
Analyseinstitut for Forskning

Effekten af vægtning og korrektion for partielt bortfald i modeller for regionale og andre forskelle i danske virksomheders forskning og udvikling

Peter S. Mortensen

December 2001

Indholdsfortegnelse

1. Indledning

Gennem årtier har statistikere været uenige om, hvorvidt og hvordan der skal korrigeres for effekten af, at datagrundlaget ved statistiske modelleringer af fx økonometrisk tilsnit er stikprøve-baseret. Allerede ved estimation[1] af total- og middelværdier, andele og frekvenser for en endelig population på basis af stikprøveinformation skelnes mellem to tilgange:

- Det rene randomiseringssynspunkt, hvor værdierne af den målte variabel *Y* opfattes som konstanter, sådan at variationen i estimatoren for f.eks. $\overline{Y}$ kun skyldes stikprøvedesignet (Hansen et.al, 1983).
- Det rene modelbaserede synspunkt , hvor *Y* i sig selv opfattes som en stokastisk variabel, som der allerede ved udvælgelsen må opstilles en model for, baseret på en række forudsætninger (fx Royal,1992).

Når et datasæt udvalgt efter randomiseringsprincippet dernæst ønskes anvendt til afprøvninger af modeller af økonometrisk eller anden art vha. multivariate statistiske analysemetoder, skilles vandene også. I den rene modeltilgang opfattes stokastikken i responsvariablen alene at være forbundet med modelformuleringen, dvs. det forhold at en model ikke kan tage hensyn til alle faktorer, til den præcise funktionssammenhæng og til de tilfældige påvirkninger, mens datasættets observationer opfattes som ligeværdige repræsentanter for den superpopulation, hvis sammenhænge der søges modelleret med udgangspunkt i datamaterialet. Der ses altså bort fra stikprøve-designet.

I den rene stikprøvetilgang opfattes stokastikken stadig kun at være forbundet med det forhold, at ikke hele populationen indgår i stikprøven. Det betyder, at parametrene i modeller estimeres med hensyntagen til stikprøveplan (varierende udtrækssandsynlighed, stratifikations- og klyngeeffekt) og at parameterestimaternes varianser også beregnes under samme hensyn samt med hensyntagen til delpopulationernes størrelse. Endelig bygger konfidensintervaller og tests på approksimation til normalfordelingen via den centrale grænseværdi sætning.

De to tilgange giver ikke nødvendigvis de samme estimater for modelparametrene, herunder heller ikke de samme varianser på estimaterne. Begge tilgange kan kritiseres:

- stikprøvetilgangen for ikke at medtage modelstokastikken. Som minimum kan superpopulationsbegrebet introduceres ved at undlade at tage hensyn til delpopulationernes størrelse.
- modeltilgangen for ikke at tage hensyn til stikprøvedesignet i de tilfælde, hvor designmatricen korrelerer med responsmatricen.

[1] Estimaton skal forstås bredt, dvs. både estimation af punktestimater og deres varianser samt opstilling af konfidensintervaller for parametrene.

Et yderligere problem ved et stikprøvebaseret datasæt er, at der typisk er en række tomværdier i datasættet og at svarprocenten er et pænt stykke under 100%. Ved også at se bort fra disse forhold i den rene modelbaserede tilgang øges chancerne for skævvridning af estimaterne og underestimation af deres varianser, idet de to typer bortfald sjældent skyldes tilfældigheder, men ofte korrelerer med en række af variablene i datasættet. Korrektioner for tomværdier må derfor bygge på (empiriske) antagelser om bortfaldets karakter, dvs. være modelbaseret.

I dette paper undersøges effekten ved at korrigere datagrundlaget for modeller, der er estimeret ud fra en ren modelbaseret tilgang, med elementer fra stikprøvetilgangen. Der korrigeres nemlig udtrækssandsynligheden vha. vægtning af de enkelte observationer og der anvendes erstatningsværdier for tomværdierne i datasættet. Som empirisk materiale er valgt modeller for regionale og andre forskelle i danske virksomheders forskning og udvikling, udarbejdet af Analyseinstitut for Forskning, se Smith(2000) og Broberg(2001). Dette materiale præsenteres først (kapitel 2), hvorefter effekten af de to modifikationer afdækkes i kapitel 3 og 4.

2. Det empiriske grundlag for vurdering af effekten

I dette kapitel gennemgås kort det datamateriale og de modeller, som indgik i analyserne vedr. den regionale effekt på virksomheders forskning og udvikling. For yderligere detaljer, se Broberg (2001).

2.1. Datamaterialet

Datamaterialet til afprøvning af betydningen af virksomheders beliggenhed for deres FoU-adfærd er fremkommet ved at koordinere og anvende oplysninger fra den danske forskningsstatistik 1997 og Købmandsstandens Oplysningsbureau (generelle data om danske virksomheders økonomiske forhold). Den grundlæggende enhed er virksomheder forstået som juridiske enheder. Den udvalgte stikprøve omfattede 4338 virksomheder, men efter objektbortfald udgøres datamaterialet af 3435 virksomheder, dog med en del tomværdier fra begge kilder.

Som responsvariable benyttes:

- **forskningstilbøjeligheden**, dvs. sandsynligheden for at en virksomhed er engageret i forskning og udviklingsarbejde.
- **Forskningsintensiteten**, dvs. summen af virksomhedens FoU-årsværk i forhold til samtlige ansatte i virksomheden.

Disse responsvariable søges forklaret vha. determinanter, der angiver virksomhedens beliggenhed i forhold til et byområde og samtidig tager højde for en række andre karakteristika ved virksomheden. De supplerende determinanter er valgt på basis af tidligere analyser af sammenhænge mellem forskning og virksomhedskarakteristika, se f.eks. Dilling-Hansen m.fl. (1998).

Regional beliggenhed

For at kunne vurdere om virksomhedens beliggenhed har betydning for dens FoU-adfærd er det nødvendigt at afgøre, hvordan beliggenheden skal inddrages som determinant. Malecki(1983)'s by-hierarki hypotese er valgt som udgangspunkt, dvs. at den fysiske afstand mellem virksomheder og bycentre har betydning for deres FoU-adfærd, sådan at kortere afstand medfører øget FoU-tilbøjelighed og -intensitet. Alle danske kommuner er derfor blevet klassificeret i følgende fire urbaniseringsområder på basis af en inddeling fra AKF[2]:

Bycenterområder*:* Kommuner med mere end 40.000 beskæftigede og en pendling-intensitet (indkommende pendlere i forhold til udrejsende pendlere) over 2.
Andre byområder: Andre kommuner med mere end 10.000 beskæftigede og nabo til en bycenterkommune inden for 40 km fra bycenterområdets midtpunkt.
Landområder tæt på bycentre: Landkommuner i en afstand på mindre end 20 km fra en bycenterkommune eller kommuner med en afstand på mindre end 15 km fra en kommune under klassifikationen *Andre byområder*.
Land- og perifere områder: Andre landkommuner end dem i 3.

Virksomhedens størrelse

Virksomhedens størrelse er en af de determinanter, der normalt regnes for at være af stor betydning for virksomheders FoU-adfærd. Som mål for virksomhedsstørrelse anvendes antallet af ansatte i virksomheden. Der forventes en degressiv sammenhæng, så derfor anvendes logaritmen til antal ansatte i modellerne.

Markedskoncentration

Ifølge Schmooklers(1966) har bl.a. markedskoncentrationen betydning for virksomhedernes FoU-adfærd. På basis af de indsamlede data har det været muligt at beregne et salgskoncentrations-indeks af Herfindahl-typen[3], der bruges som et mål for markedskoncentrationen. Det forventes, at virksomheder på markeder med enten meget lav eller stor markedskoncentration engagerer sig mindre i FoU end virksomheder på markeder derimellem, så der anvendes en kvadratisk funktions-sammenhæng, dvs. omvendt U-formet.

[2] Amternes og Kommunernes Forskningsinstitut. Deres opdeling i 6 områder er her slået sammen til 4.

[3] Beregning pr. branche af summen af virksomhedernes kvadrerede omsætningsandele: $H = \sum_i \left(\frac{Oms_i}{Brancheoms} \right)^2$

Rentabilitet

På markeder med stærk konkurrence er der et klart incitament for virksomhederne til øget FoU-aktivitet for at kunne differentiere produkterne. Samtidig er der en tendens til, at virksomheder på markeder med stærk priskonkurrence har lav indtjening. Rentabiliteten er derfor søgt medtaget som determinant.

Ejerforhold (uafhængig vs. koncern)

Der kan argumenteres for, at koncerner råder over flere ressourcer og mere kapital samt har de nødvendige forretningsmæssige kontakter og lettere adgang til viden og finansiering. Derfor medtages ejerforholdet som en indikatorvariabel.

Virksomhedens alder

Ældre virksomheder forventes at have bedre organisatoriske og finansielle muligheder for at opnå fordele ved risikofyldte FoU-investeringer, dog kun med en degressiv stigningstakt. Omvendt kan nyetablerede virksomheder have et stort potentiale for vækst, og dermed et stærkt incitament for at investere i FoU. Denne U-formede funktionssammenhæng kan beskrives ved en kvadratisk form.

Finansiel solvens

Da FoU-investeringer er mere risikofyldte og langsigtede sammenlignet med investeringer i fysisk kapital, forventes finansielt svagere virksomheder at være mere tilbageholdende med at igangsætte nye FoU-aktiviteter, også fordi det kan være svært at rejse den nødvendige kapital til FoU-investeringer. Af denne grund er virksomhedens finansielle solvens (egenkapital i procent af de samlede aktiver) medtaget i analyserne.

Hovedbranchetype (fremstilling vs. servicevirksomhed)

Tidligere undersøgelser viser, at fremstillingsvirksomheder er mere FoU-aktive end service-virksomheder. Derfor er der medtaget en indikatorvariabel for hovedbranchetype.

2.2. Modellerne

Der er opstillet modeller for FoU-tilbøjeligheden og FoU–intensiteten. Disse to respons-variable beskrives i dette tilfælde bedst ved en logistisk regressionsmodel og ved en tobit-model, jf. nedenstående korte beskrivelser.

Virksomhedernes tilbøjelighed til at engagere sig i FoU

En virksomheds tilbøjelighed til at engagere sig i FoU kan udtrykkes som sandsynligheden for, at den pågældende virksomhed er engageret i FoU. Denne sandsynlighed, $L(w)$, kan skrives som

(1) $$L(w) = \frac{e^w}{1+e^w}$$

hvor

(2) $$w = x'\beta + u$$

w kaldes logit og er logaritmen til odds'ene for at virksomheden har FoU. x er en søjlevektor, der indeholder de forklarende variable. β er ligeledes en søjlevektor, der indeholder de forklarende variables parametre. Endelig er u søjlevektoren for fejlleddene, der antages at have en middelværdi på 0.

Logitmodellen er således en lineær sandsynlighedsmodel, hvor den afhængige målte variabel antager værdien 1 for virksomheder med FoU-udgifter og 0 ellers. w er en latent variabel, der angiver den forventede sandsynlighed (i logistisk form) givet determinanterne. Logit-funktions-formen opnås ved at forudsætte, at fejlledenes fordeling er logistisk.

FoU-intensiteten

FoU-intensiteten måles som summen af FoU-årsværk i forhold til alle ansatte i virksomheden. Der kan anvendes en tobit-model til at analysere determinanternes indflydelse på FoU-intensiteten, se Amemiya(1984). Her opfattes FoU-intensiteten, y^* som en censureret variabel ved at y^*-værdien er sat lig med nul for virksomheder uden forskning. Der analyseres på en nyskabt variabel, y, hvis fordeling er en blanding af en diskret og en kontinueret fordeling:

(3) $y = 0$ hvis $y^* \le 0$

og

(4) $y = y^*$ hvis $y^* > 0$

hvor $y_i^* = \beta_0 + \sum \beta_j x_{ij} + u_i$ estimeres vha. en censureret regressionsmodel.

3. Korrektion for variabel udtrækssandsynlighed

3.1. Teoretiske overvejelser

Ved surveys af virksomheder vil det normalt være statistisk mest effektivt at udvælge respondenterne med udtrækssandsynligheder, der afhænger af virksomhedsstørrelsen. Det skyldes, at der typisk er mange små, nogle mellemstore og ganske få helt store virksomheder, dvs. en meget skæv fordeling. En metode er at tildele virksomhederne en udtrækschance, svarende til deres størrelse, målt som omsætningen eller antal ansatte. En sådan fremgangsmåde bevirker i Danmark, at der skal foretages en totaltælling af de største virksomheder, mens kun en ganske lille andel af de mindste virksomheder udvælges.

Ved beregning af estimater for hele populationen af virksomheder, hvor analyseenheden er virksomheden – f.eks. de samlede FoU-udgifter i danske virksomheder – bliver det nødvendigt at korrigere for forskellene i udtrækssandsynlighed for at sikre forventningsrette estimater. Det gøres ved at vægte hver enhed i stikprøven med den reciprokke værdi af dens udtrækssandsynlighed, se f.eks. Mortensen(1992). Derved fremkommer et estimat på hele populationen (τ):

(5) $$\tau = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{z_i}$$

hvor z_i er det i'te udvalgte elements udtrækssandsynlighed.

Ved multivariate analyser vil lignende skævheder i estimaterne opstå, hvis de variable, som vægtene afhænger af, korrelerer med responsvariablen(e) og ikke er medtaget i modellen på netop den funktionelle måde, som sammenhængen følger, se Pfeffer-mann(1993).

I de fleste surveys er der en vis andel af respondenterne, der ikke får besvaret spørgsmålene (objektbortfaldet). For at sikre forventningsretheden må de anvendte vægte korrigeres for dette forhold, medmindre der anvendes andre metoder til korrektion for objektbortfaldet, f.eks. kalibrering eller estimation af tomværdier, se afsnit 4.

Effekten af vægtningen er som nævnt, at forventningsretheden sikres på de beregnede estimater. Estimaternes usikkerhed, udtrykt ved f.eks. variansen, bliver dog forøget pga. vægtningen. En reduktion kan eventuelt opnås ved at tage hensyn til stratifikationseffekten, hvis den estimerede parameter eller responsvariabel korrelerer positivt med vægttildelingskriterierne, f.eks. antallet af ansatte i virksomhederne og hovedbranche. Det vil dog ikke blive gjort her.

3.2. Estimationer med og uden vægte

I undersøgelsen vedr. regionale og andre faktorers indflydelse på virksomheders forskning skal den rene vægteffekt undersøges, dels i en logistisk model til forklaring af om en virksomhed bedriver forskning og udvikling, se Smith(2000) og dels i en tobit-model til forklaring af omfanget af forsknings- og udviklingsarbejdet, se Broberg (2001).

3.2.1. Virksomheders forskningstilbøjelighed

I basismodellen søger Smith(2000) at forklare forskningstilbøjeligheden vha. de i afsnit 2 nævnte variable, idet indflydelsen fra virksomhedsstørrelse gøres degressiv vha. en logaritmisk transformation, mens indflydelsen fra alder og markedskoncentration gøres kvadratisk. I tabel 3.1 sammenlignes resultaterne af denne model, når der estimeres med og uden vægtning. De overordnede modelmål (log likelihood, LR og konkordans) viser som forventet, at der er større varians på estimaterne i modellen med vægte. På de enkelte parametre i modellen ses en række større og mindre forskelle, der slår igen på de kritiske signifikansniveauer (p-værdier[4]). Mest markant er her forskellen for indikatoren for *Landområder tæt på bycentre*, for log(alder) samt de insignifikante determinanter.

Slutmodellen i Smith(2000) er reduceret med afkastet, solvensen og markedskoncentrationen. I tabel 3.1 sammenlignes estimaterne for denne endelige model, både med og uden vægte. Overordnet ses de samme forskelle som i udgangsmodellen; blandt de enkelte parametre er aldersvariablene dog klart mere insignifikante og indikatoren for *Landområder tæt på bycentre* er stadig insignifikant i den vægtede estimation.

Imidlertid kan den vægtede model reduceres yderligere. Alderen er repræsenteret ved to parametre, så på grund af den derved indbyggede multikollinearitet må disse parametre ses under et. Ved en undersøgelse af modeller med hhv. kun alder, kun kvadreret alder samt uden alders-variable afsløres det, at det er muligt helt at udelade alderen (metode: betingede χ^2-tests vha. LR-målet). Det gælder i øvrigt også i den uvægtede model.

Desuden er ikke alle regionale indikatorer signifikante og de bliver derfor omkodet. I en senere uvægtet modelestimation prøver Broberg(2001) at sammenlægge *Andre byområder* og *Land-områder tæt på bycentre*, mens *Land- og perifere områder* ikke er signifikant og derfor lægges sammen med basis, *Bycentre*. I den vægtede estimation kan den samme omkodning nok accepteres, men en anden mulighed er at beholde *Andre byområder* som eneste selvstændige

[4] P-værdierne er knyttet til hypotesen om, at variablen ingen indflydelse har på forsknings-tilbøjeligheden, dvs. H_0: $\beta_i=0$.

indikator. Den vægtede estimation af parametrene bliver altså også en delvis afkræftelse af byhierarki-hypotesen.

Tabel 3.1: Logistiske modeller for danske virksomheders FoU-tilbøjelighed, 1997.

	Basis	model	Slutmodel -Smith(2000)		Slutmodel
Parametre	Uvægtet	Vægtet	Uvægtet	Vægtet	Vægtet
Skæringspunkt	-0,7731 (0,2849)	-1,6161 (0,0119)	-0,7230 (0,3032)	-1,7540 (0,0048)	-2,1259 (0,0000)
Størrelse (Log af antal beskæftigede)	0,2747 (0,0000)	0,3079 (0,0000)	0,2710 (0,0000)	0,3008 (0,0000)	0,3112 (0,0000)
Uafhængig virksomhed	0,3538 (0,0073)	0,2929 (0,0283)	0,3601 (0,0060)	0,3122 (0,0185)	0,3010 (0,0225)
Fremstillings-virksomhed	0,8026 (0,0000)	0,6654 (0,0000)	0,8237 (0,0000)	0,6708 (0,0000)	0,7089 (0,0000)
Alder (Log af virksom-hedernes alder)	-0,7264 (0,1375)	-0,4291 (0,3159)	-0,7126 (0,1356)	-0,3410 (0,4108)	
Alder kvadreret	0,1130 (0,1780)	0,0849 (0,2488)	0,1110 (0,1750)	0,0709 (0,3188)	
Rentabilitet (Profit/ aktionærers egenkapital	-0,2862 (0,4973)	0,2850 (0,5577)			
Finansiel Solvens	-0,0217 (0,8107)	-0,1018 (0,4482)			
Markedskoncentration (Herfindahl index)	0,4898 (0,5690)	-0,3440 (0,7002)			
Markedskoncentration kvadreret	-0,1911 (0,8441)	0,7322 (0,4580)			
Andre byområder	-0,2987 (0,0500)	-0,4210 (0,0092)	-0,3179 (0,0357)	-0,4536 (0,0048)	
Landområder tæt på bycentre	-0,3790 (0,0433)	-0,1519 (0,4177)	-0,3819 (0,0398)	-0,1834 (0,3232)	
Land- og perifere områder	0,1980 (0,2893)	0,1866 (0,2884)	0,1794 (0,3325)	0,1676 (0,3366)	
Andre byområder og landområder tæt på by					-0,4126 (0,0012)
Log likelihood	1652,87	1560,10	1672,80	1574,22	1578,47
Konkordans	68,7%	68,0%	68,4%	67,8%	68,1%
Antal observationer	1296	1296	1307	1307	1307

Bemærk: Tallene i parentes er p-værdierne, se fodnote 4.

Konklusionen fra dette eksempel bliver, at der ved vægtningen fås estimater med større usikkerhed og at både parametrene og deres varianser ændrer sig. De signifikante determinanter i modellen ændrer sig derfor: alder udgår og regionalbeskrivelsen kan ændres.

3.2.2. Virksomheders forskningsintensitet

Modellerne vedrørende forskningsintensiteten tager som nævnt udgangspunkt i et revideret datasæt, men det er de samme variable som ved modelleringen af forskningstilbøjeligheden, der a priori vælges til at forklare intensiteten, se Broberg (2001). I tabel 3.2 ses estimationsresultaterne med og uden vægtning. I basismodellen er de vægtede estimationer også her behæftet med større usikkerhed; således er parameterestimaternes standardafvigelser 5-10% højere.

Udover mindre forskelle i estimaterne og deres p-værdier for en række af variablene, fås et par markante forskelle mellem estimaterne i basismodellen. Det drejer sig om virksomhedsstørrelse, der kun er en signifikant positiv parameter i den vægtede model og om markedskoncentrationen, der kun er signifikant i den uvægtede model. Desuden er der forskel i niveauet for to af de regionale indikatorer, nemlig *Land- og perifere områder* samt *Andre byområder*.

En sammenligning mellem slutmodellen i Broberg(2001) og den tilsvarende vægtede model viser, at både alder og markedskoncentration kun er signifikante i den uvægtede model, da alle fire parametre kun er cirka halvt så store i den vægtede model. Omvendt kan virksomhedsstørrelsen indgå i den vægtede model, når det suppleres med det kvadratiske led. Desuden er sammenlægningen af *Andre* byområder og *Land- og perifere* områder ikke det bedste valg i den vægtede model, men giver dog stadig en signifikant indikator.

En selvstændig modellering af den vægtede model inkluderer derfor virksomhedsstørrelse inklusive det kvadratiske led, mens regions-indikatoren kan fastholdes, selv om en anden mulighed er kun at fastholde *Andre byområder*. Mht. alders repræsentation er der ikke basis for at fastholde det kvadratiske led, ligesom markedskoncentrationsindekset (uden kvadratisk led) med en p-værdi=0,29 vanskeligt kan fastholdes. Modellen er vist i sidste kolonne af tabel 3.2. Det kan altså konkluderes, at vægtningen i dette tilfælde har betydelig indflydelse på, hvilke parametre der indgår i slutmodellen.

Tabel 3.2: Tobit-modeller for FoU-intensiteten, 1997

	Basis model		Slutmodel – Broberg(2001)		Slutmodel
	Uvægtet	Vægtet	Uvægtet	Vægtet	Vægtet
Skæringspunkt	0,1120 (0,1334)	-0,0253 (0,7225)	0,0734 (0,3024)	-0,0314 (0,6338)	-0,0295 (0,4852)
Størrelse (Log af antal beskæftigede)	0,0030 (0,5691)	0,0130 (0,0324)			-0,0422 (0,0113)
Kvadreret størrelse (Log af antal beskæftigede)					0,0084 (0,0002)
Uafhængig virksomhed	-0,0390 (0,0099)	-0,0294 (0,0681)	-0,0695 (0,0000)	-0,0629 (0,0000)	-0,0429 (0,0124)
Fremstillingsvirksomhed	0,0496 (0,0011)	0,0652 (0,0002)	0,0498 (0,0013)	0,0718 (0,0000)	0,0831 (0,0000)
Alder (Log af virksomhedernes alder)	-0,1279 (0,0071)	-0,0953 (0,0306)	-0,0985 (0,0305)	-0,0638 (0,1228)	-0,0304 (0,0039)
Alder kvadreret	0,0180 (0,0222)	0,0140 (0,0541)	0,0111 (0,1359)	0,0069 (0,2994)	
Finansiel Solvens	-0,0049 (0,5949)	-0,0157 (0,2586)			
Markedskoncentration (Herfindahl index)	0,2695 (0,0041)	0,1149 (0,2757)	0,3374 (0,0008)	0,1228 (0,2531)	
Markedskoncentration kvadreret	-0,2501 (0,0146)	-0,0794 (0,4823)	-0,3178 (0,0045)	-0,0928 (0,4300)	
Rentabilitet (Profit/ aktionærers egenkapital	-0,0409 (0,3159)	0,0352 (0,4348)			
Andre byområder	-0,0486 (0,0042)	-0,0609 (0,0014)			
Landområder tæt på bycentre	-0,0238 (0,2364)	-0,0240 (0,2534)			
Land- og perifere områder	-0,0367 (0,0778)	-0,0140 (0,5245)			
Andre byområder og Land- og perifere områder			-0,0517 (0,0006)	-0,0435 (0,0049)	-0,0441 (0,0047)
Log likelihood	568,34	662,02	951,41	922,11	896,41
Antal observationer	1285	1285	1588	1588	1577

Bemærk: Tallene i parentes er p-værdierne, se fodnote 4.

4. Korrektion for partielt bortfald

4.1. Teoretiske overvejelser

Nederst i tabel 3.1 og 3.2 er angivet stikprøvegrundlaget for hver model. Det ses heraf, at antallet varierer og at antallet er betydeligt mindre end det oprindelige grundlag på 3435 virksomheder. Årsagen er dels, at en række respondenter har undladt at udfylde et eller flere spørgsmål i indberetningsskemaet vedrørende erhvervslivets forskning 1997 og dels, at der i de anvendte virksomhedsregistre mangler en eller flere af de ønskede basis-oplysninger om en række af virksomhederne. Begge mangler kan betegnes *det partielle bortfald* i relation til de estimerede modeller.

Ved afsnit 3's vægtning af virksomhederne er der ikke taget hensyn til det partielle bortfald. Det betyder, at hvis det partielle bortfald korrelerer med de strata, som der er fastsat vægte efter, vil der stadig være en skævhed i de gennemførte modelberegninger. Denne skævhed kunne der korrigeres for ved at udregne særskilte vægtsæt for hver kombination af variable, der afprøves.

Et bedre alternativ er at bringe respondenterne med tomværdier i en eller flere af variablene ind i beregningerne ved at estimere erstatningsværdier for deres tomværdier, kaldet *imputation* på engelsk. En sådan estimation vil samtidig have den fordel, at der ved tildeling af erstatningsværdier kan tages hensyn til flere variable end dem, der ligger til grund for vægtene. Samtidig er det muligt at fastholde stokastikken i den estimerede værdi. Ved erstatningen af tomværdierne sker der en reduktion i estimaternes varianser på grund af det større stikprøvegrundlag, men denne reduktion kan dog sjældent dække den forøgelse af variansen, som estimationen af tomværdierne medfører, se fx Särndal(1992).

Der findes en række metoder til estimation af tomværdier. Det er ikke dette papers opgave at gennemgå dem alle, men her skal gives en kort beskrivelse af de typer, der er anvendt her, se eventuelt Kovar;Whitridge(1995)

- **Forventet værdi**, givet et bestemt sæt af værdier for andre variable i datasættet. Metoden gør brug af en multipel regressionsmodel med de variable, som korrelerer mest med den variabel, der skal estimeres tomværdier for. Basis for estimationen er alle respondenter uden tomværdier. Ved ordinale og nominale variable må *den mest sandsynlige værdi* anvendes som estimat, f.eks. på basis af en diskriminantanalyse.

Den forventede værdi-metode kan tilføres stokastik ved at korrigere den forventede værdi med et tilfældigt element, beregnet fra den anvendte model. Yderligere kan en sådan tilgang sofistikeres ved at generere flere erstatningsværdier, dvs. multipel imputation, se Rubin(1987).

- **Samme værdi** som en anden tilfældig respondent, der er placeret i samme sæt af udfald for udvalgte variable[5] (eng: hot deck). Gennem kombinationen af udfald på de udvalgte variable udpeger en given observation den gruppe af respondenter, hvor der skal trækkes en tilfældig erstatningsværdi fra.
- **Værdi fra en anden periode** for den samme respondent (eng: Last Value Carried Forward). Som betegnelsen angiver, anvendes der normalt erstatningsværdier fra en tidligere periode, men det kan også være fra en fremtidig periode, jf. eksemplet nedenfor.

4.2. Estimationer med estimerede værdier for partielt bortfald

Effekten af at erstatte tomværdierne med estimerede værdier skal illustreres for de to økonometriske modeller, tobitmodellen og den logistiske model, for forskningsintensitet og forskningstilbøjelighed. I tobitmodellen korrigeres kun for tomværdierne i det datasæt, der blev anvendt i afsnit 3.2.2, mens der i den logistiske model yderligere suppleres med værdier fra en anden periode/kilde for at få alle 3435 virksomheder med.

Først skal det dog dokumenteres, at der er korrelation mellem vægtgrupperne og det partielle bortfald, forårsaget af manglende registeroplysninger. Det ses af tabel 4.1, hvor gruppen af store virksomheder, der alle har deltaget i undersøgelsen, udgør 52,5% iblandt de modtagne besvarelser, men hele 63,6% blandt dem uden partielt bortfald fra registeroplysningerne, kaldet basisstikprøven.

Tabel 4.1: Fordeling af stikprøven på vægtgrupper med og uden partielt bortfald

Vægtgruppe	**Blandt besvarelserne**	**Blandt basisstikprøven**
Totaltælling	0,525	0,636
Stikprøve – industri	0,203	0,157
Stikprøve – serviceerhverv	0,272	0,207
I alt	1,000 (n=3435)	1,000 (n=1610)

[5] Intervalskalerede variable må først transformeres til ordinalskala

4.2.1. Estimation, begrænset til basisstikprøven

Basisstikprøven udgøres i det første eksempel af de virksomheder, som der er blevet fundet registeroplysninger på, da datasættet blev dannet for et par år siden. Der er imidlertid yderligere tomværdier blandt disse n=1610 respondenter. Af tabel 3.2 fremgår det, at der er mellem 1285 og 1588 virksomheder med i de viste modeller.

Tomværdierne er estimeret ved følgende procedure[6], se også bilag 1.a:

1. Variablene med tomværdier er sorteret efter antallet af tomværdier, sådan at variablen med færrest tomværdier i de fleste tilfælde estimeres først.
2. De intervalskalerede variable er estimeret som den forventede værdi i en multipel regression med de variable, der korrelerer mest med variablen som prediktorer og med den mest korrelerende ordinal/nominalskalerede variabel som grupperings-variabel.
3. De ordinal/nominalskalerede variable er estimeret med en tilfældig værdi (hot deck) fra den kombination af udfald, som de mest korrelerende variable danner.
4. Specielt med hensyn til de to økonomiske variable, forrentning og solvens, gælder, at de korrelerer meget lavt med alle øvrige variable (max r = 0,08). Estimationen af de mange tomværdier er derfor ekstra usikker, men uden betydning for den efterfølgende modellering pga. de to variables klare insignifikans.

I bilag 1.b er gennemsnit og standardafvigelser vist for de variable, der har fået tildelt erstatnings-værdier. Det ses, at bortset fra solvens-variablen og delvis også forrentnings-variablen er der tale om meget små ændringer.

Det nye datasæt med n=1610 observationer med værdier for alle variable er blevet anvendt til at reestimere tobitmodellen for forskningsintensiteten. Det er gjort ved at reestimere basis-modellen fra tabel 3.2 og dernæst modellere den bedst mulige model, se tabel 4.2. Estimaterne for basismodellen afviger på de samme punkter som den vægtede model i forhold til den uvægtede. Der er dog endnu en væsentlig afvigelse, nemlig at koefficienterne til alder og kvadreret alder har skiftet fortegn.

Fortegnsskiftet for alder holder også i slutmodellen. Desuden gælder det som i den vægtede model, at sammenlægningen af *Andre byområder* og *Land- og perifere områder* ikke er det bedste valg, men dog giver en signifikant indikator. Alt i alt er der således stadig betydelige forskelle mellem slutmodellen, som fremkommer på basis af det uvægtede datasæt og slutmodellen fra det vægtede, estimerede datasæt: Virksomhedens størrelse indgår nu, alder optræder med modsat fortegn og markedskoncentrationen er ikke signifikant.

[6] Programpakken *SOLAS for missing data analysis 3.0* er anvendt til beregningerne.

Tabel 4.2: Tobit-modeller for FoU-intensiteten, 1997

	Basismodel		Slutmodeller		
	Uvægtet	Estimeret	Uvægtet	Vægtet	Estimeret
Skæringspunkt	0,1120 (0,1334)	-0,3677 (0,0000)	0,0734 (0,3024)	-0,0295 (0,4852)	-0,2996 (0,0002)
Størrelse (Log af antal beskæftigede)	0,0030 (0,5691)	0,0101 (0,1017)		-0,0422 (0,0113)	-0,0488 (0,0052)
Kvadreret størrelse (Log antal beskæftigede)				0,0084 (0,0002)	0,0080 (0,0003)
Uafhængig virksomhed	-0,0390 (0,0099)	-0,0373 (0,0359)	-0,0695 (0,0000)	-0,0429 (0,0124)	-0,0400 (0,0229)
Fremstillings-virksomhed	0,0496 (0,0011)	0,0567 (0,0013)	0,0498 (0,0013)	0,0831 (0,0000)	0,0778 (0,0000)
Alder (Log af virksomhedsalder)	-0,1279 (0,0071)	0,1149 (0,0319)	-0,0985 (0,0305)	-0,0304 (0,0039)	0,1265 (0,0199)
Alder kvadreret	0,0180 (0,0222)	-0,0165 (0,0664)	0,0111 (0,1359)		-0,0183 (0,0444)
Finansiel Solvens	-0,0049 (0,5949)	-0,0150 (0,1838)			
Markedskoncentration (Herfindahl index)	0,2695 (0,0041)	0,1055 (0,3380)	0,3374 (0,0008)		
Markedskoncentration kvadreret	-0,2501 (0,0146)	-0,0739 (0,5400)	-0,3178 (0,0045)		
Rentabilitet (Profit/egenkapital)	-0,0409 (0,3159)	0,0416 (0,4324)			
Andre byområder	-0,0486 (0,0042)	-0,0668 (0,0008)			
Landområder tæt på bycentre	-0,0238 (0,2364)	-0,0135 (0,5385)			
Land- og perifere områder	-0,0367 (0,0778)	-0,0271 (0,2394)			
Andre byområder + Land/perifere områder			-0,0517 (0,0006)	-0,0441 (0,0047)	-0,0432 (0,0063)
Log likelihood	568,34	930,89	951,41	896,41	924,67
Antal virksomheder	1285	1610	1588	1577	1610

Bemærk: Tallene i parentes er p-værdierne, se fodnote 4.

4.2.2. Brug af andre kilder

I det andet eksempel vedrørende forskningstilbøjeligheden søges de manglende stam-oplysninger for alle virksomheder blandt de 3435, der mangler sådanne. Virksomheds-databasen NewBiz er anvendt til at søge oplysninger om virksomhedens størrelse (antal ansatte), alder (via stiftelsesår), selskabsform (A/S) og uafhængighed (koncerntilknytning). Desuden er de regnskabsbaserede variable (resultat, balance og egenkapital) hentet frem fra periode –3, der typisk er regnskabsår 1997, men kan svinge med et enkelt år til hver side. Det lykkes ved denne søgning at finde så mange oplysninger, at de resterende tomværdier blandt de 3435 virksomheder relativt set er på niveau med eller lavere end i det lille datasæt på 1610 virksomheder i det første eksempel.

De resterende tomværdier er efterfølgende estimeret på samme måde som i det første eksempel, dvs. med hhv. ”forventet værdi-metoden” og ”hot deck-metoden”. Oversigten i bilag 2 viser dels antal tomværdier og dels de variable, der er anvendt ved estimation af hver variabels tomværdier. Også her viser bilag 2.b, at der er meget begrænsede forskelle i gennemsnit og standardafvigelse før og efter estimationen.

De logistiske modeller for FoU-intensitet er dernæst blevet estimeret vha. det nye store datasæt, stadig hensyntagen til vægtningen af virksomhederne. Først er basismodellen med alle a priori-variablene estimeret, se tabel 4.3. Sammenlignes dette resultat med den uvægtede model på basis af den lille stikprøve, fås kun en beskeden stigning i LR-testoren, men et fald i konkordans-målet; der er altså mere residualvariation mellem de nye virksomheder, der er kommet til i den store stikprøve. Der ses desuden to klare forskelle ved de signifikante determinanter, nemlig at alder nu er signifikant og at *Land- og perifere områder* har skiftet fortegn til det forventede.

Slutmodellen for den store stikprøve er vist i tabel 4-3’s højre kolonne. Her er alle determinanter signifikante og ingen andre af datasættets variable kan tilføre modellen større forklaringsgrad. Ved sammenligning med denne slutmodel og de to slutmodeller for den lille stikprøve ses nogle væsentlige forskelle:

Antal ansatte: Kun den kvadrerede værdi af log(ansatte) er en signifikant determinant, mens det i de to andre modeller er log(ansatte) direkte. Andre funktionsformer for antal ansatte er afprøvet, men har ikke givet bedre resultat.
Alderen: I de to andre slutmodeller er alder reelt ikke signifikant, men det bliver den her – med negativt fortegn. Der er dog ikke noget kvadratisk led, så hele a priori hypotesen holder ikke.
Markedskoncentrationen: Heller ikke markedskoncentrationen er reelt signifikant i slutmodellerne fra den lille stikprøve. Det bliver den her, men dog ikke med det kvadrerede led. A priori hypotesen holder derfor ikke fuldt ud.

Tabel 4.3: Logistiske modeller for FoU-intensiteten, 1997

	Basismodel		Slutmodeller		
	Uvægtet	Estimeret	Uvægtet	Vægtet	Estimeret
Skæringspunkt	-0,7731 (0,2849)	-1,8221 (0,0000)	-0,7230 (0,3032)	-2,1259 (0,0000)	-1,5889 (0,0000)
Størrelse (Log af antal beskæftigede)	0,2747 (0,0000)	0,2884 (0,0000)	0,2710 (0,0000)	0,3112 (0,0000)	
Kvadreret størrelse (Log antal beskæftigede)					0,0372 (0,0000)
Uafhængig virksomhed	0,3538 (0,0073)	0,2447 (0,0296)	0,3601 (0,0060)	0,3010 (0,0225)	0,2369 (0,0297)
Fremstillings-virksomhed	0,8026 (0,0000)	0,4867 (0,0000)	0,8237 (0,0000)	0,7089 (0,0000)	0,5424 (0,0000)
Alder (Log af virksomhedsalder)	-0,7264 (0,1375)	-0,4120 (0,0098)	-0,7126 (0,1356)		-0,2871 (0,0001)
Alder kvadreret	0,1130 (0,1780)	0,0209 (0,2762)	0,1110 (0,1750)		
Rentabilitet (Profit/egenkapital)	-0,2862 (0,4973)	-0,0167 (0,9418)			
Finansiel solvens	-0,0217 (0,8107)	-0,0013 (0,9886)			
Markedskoncentration (Herfindahl index)	0,4898 (0,5690)	0,5459 (0,3997)			0,7945 (0,0001)
Markedskoncentration kvadreret	-0,1911 (0,8441)	-0,3097 (0,6679)			
Andre byområder	-0,2987 (0,0500)	-0,4632 (0,0001)	-0,3179 (0,0357)		
Landområder tæt på bycentre	-0,3790 (0,0433)	-0,2819 (0,0311)	-0,3819 (0,0398)		
Land- og perifere områder	0,1980 (0,2893)	-0,1919 (0,1336)	0,1794 (0,3325)		
Andre byområder + Land/perifere områder				-0,4126 (0,0012)	
Bycenterområder					0,3154 (0,0006)
Log likelihood	1652,9	3190,7	1672,8	1578,5	3184,9
Konkordans	68,7%	63,7%	68,4%	68,1%	63,9%
Antal virksomheder	1296	3435	1307	1307	3435

Bemærk: Tallene i parentes er p-værdierne, se fodnote 4.

De regionale indikatorer: Det uventede positive fortegn i det lille datasæts model for *Land- og perifere områder* var ikke signifikant forskellig fra basis (*Bycenter-områder*[7]). I det store datasæt er fortegnet negativt og samtidig er der forskellen mellem de tre indikatorer ikke signifikant; disse er derfor slået sammen og danner en ny basis, således at der opstår en ny signifikant og positiv indikator for virksomheder, som er placeret i bycenterområderne. Dette slutresultat kan delvist bekræfte by-hierarki hypotesen.

5. Konklusion

Eksemplet i dette paper viste, at estimaterne i økonometriske modeller på baggrund af stikprøvebaserede data kan være følsomme over for måden, som stikprøvegrundlaget inddrages på. Ændringerne i estimaterne ved ændring af behandlingen af stikprøve-grundlaget bevirkede, at der skete ændringer i de statistiske valg af determinanter i modellerne og der forekom endog fortegnsskift for et par af determinanterne. Disse resultater giver endnu en gang anledning til at anbefale, at datasæt, der er indsamlet på stikprøvebasis, må bearbejdes – anvendelse af vægte og estimation af tomværdier – før datasættet kan indgå i videre analyser. Det skyldes, at der ofte er sammenhænge mellem på den ene side stikprøvedesignet og bortfaldet og på den anden side model-sammenhængene – og disse sammenhænge kan der kun vanskeligt tages fuld højde for ved at inddrage dem i selve modellen.

[7] Bycenterområder består af områderne omkring København, Århus, Ålborg, Odense og Esbjerg.

Referenceliste

Amemiya (1984). Tobit Models: A Survey. *Journal of Econometrics* (24), 3-61.

Broberg,A.L. (2001). *En empirisk undersøgelse af regionale forskelle I virksomheders Forskning og udviklingsaktiviteter.* WP2001/3. Analyseinstitut for Forskning.

Dilling-Hansen, M., T. Eriksson, E.S. Madsen & V. Smith (1998). *Kan den økonomiske teori forklare omfanget af forskning og udvikling I danske virksomheder?*. Rapport fra Analyseinstitut for Forskning 1998/6.

Erhvervslivets forskning og udviklingsarbejde, Forskningsstatistik 1997. *Analyseinstitut for Forskning.*

Hansen,M.H; Madow, W.G and Tepping,B.J (1983): *An Evaluation of Model-dependent and Probability-sampling Inferences in Sample Surveys.* Journal of American Statististician Association, 78,776-93.

Kovar,J.G and Whitridge,P.J.(1995): *Imputation of Business Survey Data* in Cox,B.G et.al (Eds): Business Survey Methods.

Malecki E.J. (1980). Corporate organization of R and D and the location of technological activities. *Regional Studies*, 14, 219-34.

Mortensen, P.S. (1992): *Videregående stikprøveteori*. H6, Institut for Informationsbehandling, Handels-højskolen i Århus

Mortensen, P.S. (2001): *Effekten af vægtning og korrektion for partielt bortfald i modeller for regionale og andre forskelle i danske virksomheders forskning og udvikling.* WP2001/9. Analyseinstitut for Forskning.

Pfeffermann,D. (1993): *The Role of Sampling Weights when Modelling Survey Data*. International Statistical Review,61,2,317-37.

Royall, R.M (1992): Robustness and Optimal Design under Prediction Models for Finite Populations. Survey Methodology,18,179-85.

Rubin,D.B (1987): *Multiple Imputation for Nonresponse in Surveys*. Wiley,N.Y.

Särndal, C. (1992): Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used. Survey Methodology,18,2,241-52.

Schmookler J. (1966). *Invention and Economic Growth.* Harvard University Press, Cambridge, MA.

Smith,V; Broberg,A.L. & Overgard,J. (2000): *Regional Influence on R&D Behaviour Evidence from Danish Firms*. WP2000/3, Analyseinstitut for Forskning.

Bilag 1.a. Estimation af det partielle bortfald

(Til tobitmodeller vedr. forskningsintensitet)

(Basisstikprøven på n=1610)

Variabel	Antal tomværdier	Variabel-type	Prediktorer	Grupperingsvariabel
Indexh97	8	Interval	Besk97[1] D-indepe Pfou	D_fremst
Age	14	Interval	Size FoU_aar D_fremst	D_indepe
Size	12	Interval	Age[2] D_as D_fremst FoU_aar	D_indepe
Sizegrp	12	Ordinal	Size	
D_as[3]	19	Ordinal 0/1	-	D_indepe Sizegrp[4] D-fremst
Pfou[3]	16	Ordinal 0/1	-	Sizegrp[4] D_fremst D_indepe D_as
Rent	307	Interval	D_indepe D-fremst Age Pfou	D_as
Solvency	308	Interval	Rent D_fremst Size Age	D_as

Ingen overlap mellem tomværdierne i prediktoren og variablen

2 overlap mellem tomværdierne i prediktoren og variablen

Nominal variabel, så hot deck-metoden er anvendt

Size inddelt i 4 lige store grupper

Bilag 1.b. Effekten af estimationen

Målt ved gennemsnit og standardafvigelse

(Basisstikprøven på n=1610)

Variabel	Antal tomværdier	Gns. før	Gns. efter	Std.afv. før	Std.afv. efter
Indexh97	8	0.233	0.233	0.217	0.216
Size	21	3.963	3.936	1.627	1.646
Age	14	2.902	2.901	0.765	0.763
D_as	19	0.720	0.720	-	-
Pfou	16	0.431	0.429	-	-
Rent[5]	302	0.044	0.057	0.184	0.170
Solvency[6]	303	0.259	0.374	1.008	0.941

En outlier er ændret fra 1250 til 2.

En outlier er ændret fra –127 til –25.

Bilag 2.a. Estimation af det partielle bortfald

(Til logistiske modeller vedr. forskningsintensitet)

(basisstikprøven på n=3435)

Variabel	Antal tomværdier	Variabel-type	Prediktorer	Grupperingsvariabel
Indexh97	20	Interval	Size D_indepe Pfou	D_fremst
Age	27	Interval	Size D_indepe D_fremst	
Size	25	Interval	Age D_as D_indepe Pfou	D_fremst
Sizegrp[8]	25	Ordinal	Size	
D_indepe[7]	28	Ordinal		Sizegrp D-as D-fremst Pfou
Rent	223	Interval	Age Size Solvency	D_fremst
Solvency	171	Interval	Rent Age Size	D_fremst

Nominal variabel, så hot deck-metoden er anvendt

Size inddelt i 4 lige store grupper

Bilag 2.b. Effekten af estimationen

Målt ved gennemsnit og standardafvigelse

(basisstikprøven på n=3435)

Variabel	Antal tomværdier	Gns. før	Gns. efter	Std.afv. før	Std.afv. efter
Indexh97	20	0.223	0.223	0.209	0.208
Size	25	3.706	3.724	1.604	1.629
Age	27	2.788	2.830	0.807	0.933
D_indepe	28	0.482	0.484	-	-
Rent[9]	223	0.048	0.048	0.229	0.222
Solvency[9]	171	0.265	0.269	0.841	0.821

Se fodnote 5 og 6 i bilag 1.b.