En kommentar

AV SVEN BERG

Inledning

Jag har blivit ombedd av redaktionen att kommentera uppsatsen av Ersson och Wörlund (i fortsättningen E&W).

Jag vill först gärna säga att jag finner det värdefullt att forskare skaffar sig intim kännedom om sina data och möjliga analysmetoder genom ett så ambitiöst arbetsprogram, som det E&W genomför och här redovisar till gagn för kollegor, som arbetar inom samma område. E&W har som statsvetare den fördelen att slippa vissa hämningar, när det gäller det statistiska arbetet. Som statistiker tränas man att arbeta efter en mall som stundom kan kännas som en tvångströja: antaganden skall motiveras, modeller skrivas ut explicit, vissa ting är inte tillåtna, etc.

Jag har under de senaste åren i ett projekt tillsammans med Bo Bjurulf arbetat mycket med valstatistik och analys av väljares beteende och har naturligtvis härvid haft anledning att fundera över svårigheter med ekologisk inferens (facklitteraturens "ecological fallacy"). Jag väljer att redovisa några synpunkter på ekologisk inferens, som jag finner viktiga, och därtill några erfarenheter jag gjort, vilket jag hoppas skall komplettera E&W's framställning. Jag ämnar alltså inte i detalj granska E&W's uppsats här, men en passant kommer jag att hänvisa till denna, då så faller sig naturligt.

Aggregerade data alltid nödlösning?

Data erhållna vid intervjuer avser individer och borde vara de som bäst lämpar sig för studier av individers beteende. Intervjuundersökningar är emellertid kostnadskrävande och samhällsvetare blir i stor utsträckning hänvisade till den officiella statistikens aggregerade data. Detta är kanske inte alltid till nackdel?

Italien har ett stort antal (ca 50-talet) opinionsinstitut, men man har stora svårigheter att få korrekta svar vid partisympatiundersökningar. Särskilt anhängare till ytterlighetspartierna är obenägna att lämna svar och italienska psefologer är i stor utsträckning hänvisade till aggregerade data i sina undersökningar (jfr t ex Natale och Borcio 1987).

SURVEYDATA:	AGGREGERADE DATA:
Opinionsundersökning med svarsfel, intervjuarfel och helt eller partiellt bortfall	Valkretsdata med partiandelar bestämda med godtycklig precision

Uppställningen kontrasterar tillspetsat två typer av data. Naturligtvis är det inte alltid så att ekologiska data är tillförlitligare än individdata. Datakvalitetsaspekter kan emellertid leda till att man faktiskt föredrar att arbeta med aggregerade data.

Det kan också vara så att man av teoretiska skäl föredrar att arbeta med aggregerade data. Det är inte alldeles säkert att individdata medger inferens om enskilda individers beteende. En situation, där aggregerade data fungerar bättre, beskrivs i en uppsats av Kramer (1981). Jag återger här i något modifierad form Kramers exempel.

Vi har specificerat följande modell att gälla på individnivå:

$$y_i = \alpha_i + \beta x_i,\ i=1,\ldots,N_g. \qquad (1)$$

Vi tänker oss att modellens y_i, responsvariabeln, är en popularitetsnotering (säg en siffra mellan 0 och 100) för det makthavande partiet, vilket uppgetts av individ tillhörig grupp g. Modellens parametrar har följande innebörd: α_i är en lojalitetsnivå för individ i och partiet; β är en reaktionsparameter: individens respons på någon ekonomisk förändring – x_i – för vilken partiet kan

göras ansvarigt. Reaktionsparametern antas således vara gemensam för alla individer.

Summerar (aggregerar) vi över individer i gruppen g, får vi följande relation för gruppmedeltalen.

$$\bar{y}_g = \bar{\alpha}_g + \beta\bar{x}_g,\ g=1,\ldots,K \qquad (2)$$

Vi antar nu att de genomsnittliga lojaliteterna, $\bar{\alpha}_g$, är konstant över de K grupperna och att vi är intresserade av att skatta reaktionsparametern β. Det blir då av intresse att jämföra de vanliga minsta-kvadratskattningarna (OLS) baserade på respektive individ- och gruppdata. Vi antar att vi förfogar över en x-variabel, $\acute{x}$, som är en ersättare ("proxy") för den korrekta, svårmätta x. Vi kan då skriva skattningarna b av reaktionsparametern β på följande sätt (vi bortser här från slumpfelen i skattningarna):

$$b^{ind} = \text{Kov}\,(\alpha,\acute{x})/\text{Var}(\acute{x}) + \beta\,\text{Kov}(x,\acute{x})/\text{Var}(\acute{x})$$
$$= (\text{faktor} \neq 0) + \left(\begin{matrix}\text{korrekt}\\ \text{värde}\end{matrix}\right)(\text{faktor} \neq 1) \qquad (3)$$

respektive

$$b^{grupp} = \beta\ \text{Kov}\,(x_g,\acute{x}_g)$$
$$= \left(\begin{matrix}\text{korrekt}\\ \text{värde}\end{matrix}\right)(\text{faktor} \neq 1) \qquad (4)$$

Som framgår av representationerna (3) och (4) är båda skattningarna behäftade med systematiskt fel – bias. Om förklaringsvariabel (x) och proxy ($\acute{x}$) är perfekt korrelerade, försvinner den biasskapande faktorn i gruppestimatorn (4). Första snedvridande faktorn i individbaserade skattningen (3) kan emellertid kvarstå.

Man har i detta exempel god anledning förmoda att det är skattningen baserad på individdata (3), som hårdast kommer att drabbas av systematisk snedvridning. Den första termen i högra ledet av (3) kan t o m leda till att vi får fel tecken på den skattade reaktionsparametern. Kramer ser här en möjlig förklaring till vissa egendomliga resultat, som publicerats i litteraturen. Man har till tidsseriedata över s k popularitetsfunktioner anpassat regressionsekvationer med inflation och/eller arbetslöshet som förklaringsvariabler och fått resultat, som väckt viss undran och diskussion (Whiteley, 1984, är en relativt sen referens).

Kramers modellexempel är intressant och visar att man i vissa situationer bättre kommer åt individuellt beteende via aggregerade data. Langbein och Lichtman (1978) ger en lättillgänglig redovisning av de olika fall, där användning av aggregerade data fungerar vid inferens, respektive ger upphov till aggregeringsbias.

Ekologisk regression

Det E&W skriver om Goodmans teknik för ekologisk regression skulle jag vilja komplettera på ett par punkter.

Låt oss säga att vi är intresserade av att rekonstruera regionala flödesmatriser för att beskriva hur väljare byter parti mellan två val eller förblir lojala.[1]

I en viss grupp (kommun, valdistrikt etc) och för ett visst parti gäller följande "bokföringsrelation":

$$\text{Socialdemokraternas röstetal 1985} = \left\{\begin{matrix}\text{Röste-}\\ \text{tal 82}\end{matrix}\right\} \times \begin{matrix}\text{Lojalitets-}\\ \text{parameter}\end{matrix}$$
$$+ \Sigma \left\{\begin{matrix}\text{Röstetal}\\ \text{övriga 82}\end{matrix}\right\} \times \begin{matrix}\text{Bytesbe-}\\ \text{nägenhet}\end{matrix}$$

Utskrivet i regressionsekvationsform för de grupper-aggregat vi studerar, har vi

$$Y_s^{85} = P_{s\text{-}s}X_s^{82} + P_{vpk\text{-}s}X_{vpk}^{82} + \ldots + P_{m\text{-}s}X_m^{82} + P_{\ddot{o}\text{-}s}X_{ovr}^{82}$$
$$+ \text{felterm} \qquad (6)$$

med 5 riksdagspartier plus "Övriga" och en felterm tillagd, eftersom samma uppsättning p-värden givetvis inte exakt kan beskriva övergångarna.

Vi noterar till en början följande:

(a) en tillräckligt kraftig aggregeringsbias ger utslag i form av inadmissibla ("icke tillåtna") skattningar av bytesbenägenheter, modellens p-värden,[2]

men

(b) direkt erhållna, admissibla skattningar eller skattningar som korrigerats kan också de vara behäftade med systematiskt fel!

Vad som krävs är en validering mot oberoende data, t ex surveydata.

Ett besvärligt statistiskt problem med Goodmans ansats är multikolinearitet, som drabbar oss både då vi försöker rekonstruera flödesmatriser och då vi som E&W studerar klassröstande. Problemet är här att X-variablerna i (6) – röstandelarna vid första valet – med nödvändighet är korrelerade, eftersom de utgör delar av samma helhet. (Problemet försvinner då vi studerar ett

tvåpartisystem.) Multikolineariteten åstadkommer att skattningarna av parametrarna i flödesmatrisen, dvs $p_{s\text{-}s}$, ..., $p_{\ddot{o}\text{-}s}$ i ekvation (6) blir ytterst osäkert bestämda och mycket väl av denna anledning kan hamna utanför det tillåtna intervallet 0–1 (0–100 %).

S k ridgeregression är en statistisk teknik, som utvecklats för att ta hand om detta problem, eller i varje fall mildra effekterna därav. Personligen föredrar jag denna teknik framför s k constrained regression, som E&W anför som ett hjälpmedel att undvika inadmissibla skattningar. Jag ser det som otillfredsställande att, så att säga med våld, tvinga in p-skattningarna i intervallet 0–1.

Exempel: Det kan vara av intresse med en numerisk illustration till det sagda.

Vi studerar flöden av röstande mellan partierna vid kommunalvalen 1982 och 1985 med aggregeringsnivå kommuner. Goodmans metod kräver en gruppering av de regionala/administrativa enheterna för att åstadkomma homogenitet i bytesbenägenheter och partilojaliteter. Försöksvis används här principen: "kommuner med socialistisk majoritet men med socialdemokratisk röstandel < 50/%".

Det resultat som erhållits med Goodman-ansats enligt ekvation (6) redovisas i Tabell 1 i form av en matris med p-skattningar.

Av de erhållna p-skattningarna är det endast de i huvuddiagonalen – lojalitetsparametrarna – som är statistiskt signifikanta. Bytesbenägenheterna är bestämda med utomordentligt stor osäkerhet i samtliga fall. Bilden är generellt den att lojaliteterna skattas för högt, medan bytesbenägenheterna i regel blir underskattade.

I tabell 1 har inom parentes angivits motsvarande skattningar för de 100 kommunerna, erhållna med Johnstons metod (se Berg, 1987). Dessa senare skattningar verkar rimligare och ger säkert en riktigare bild av väljarnas lojaliteter och bytesbenägenheter.

Vi möter i exemplet säkerligen både multikolinearitet och specifikationsproblem: de 100 socialistiska kommunerna är inte homogena med avseende på bytesbenägenheter. En direkt, "naiv" tillämpning av Goodmans ansats ger, som i exemplet, ofta upphov till orimliga skattningar. Att gruppera för homogenitet kräver tillgång till data för ett stort antal ekologiska enheter: våra kommuner räcker uppenbarligen inte till.

En annan utväg är att i relationen (6) införa ytterligare X-variabler. Utförda studier visar att lojaliteter och bytesbenägenheter varierar lokalt med partiets styrka på platsen och vid byte har också mottagande partis styrka sin betydelse (se Johnston, 1985 och Berg, 1987). Parametern $p_{s\text{-}s}$,

Tabell 1. Kommunalvalen 1982/1985, Goodmanskattade lojaliteter och bytesbenägenheter. Kommuner med socialistisk majoritet med S-andel < 50 % (N=100). För jämförelse inom parentes Johnstonskattade flödesparametrar.

	VPK 85	S 85	C 85	FP 85	M 85	ÖVR 85	BOE 85
VPK 82	0.955	–0.167	0.019	–0.019	–0.072	0.098	0.187
	(0.673)	(0.130)	–*	(0.019)	(0.017)	(0.134)	(0.028)
S 82	0.008	0.920	–0.044	0.012	0.011	0.017	0.078
	(0.031)	(0.826)	(0.010)	(0.046)	(0.016)	(0.021)	(0.052)
C 82	–0.008	0.290	0.961	–0.033	0.022	–0.054	0.074
	(0.008)	(0.044)	(0.578)	(0.176)	(0.075)	(0.044)	(0.077)
Fp 82	0.053	0.037	0.001	1.102	–0.139	–0.054	0.006
	–	(0.056)	(0.021)	(0.650)	(0.154)	(0.017)	(0.102)
M 82	–0.024	–0.053	–0.071	0.197	0.910	0.064	–0.024
	(0.006)	(0.041)	(0.037)	(0.118)	(0.718)	(0.019)	(0.061)
Övr 82	0.025	–0.091	0.084	–0.061	–0.009	0.936	0.115
	(0.012)	(0.055)	(0.042)	(0.132)	(0.092)	(0.618)	(0.049)
BOE 82	0.004	0.164	0.015	0.125	0.097	–0.013	0.608
	–	(0.120)	(0.037)	(0.070)	(0.103)	(0.036)	(0.634)
$100R^2_{adj}$	93	87	95	78	93	75	62

* Estimering ej utförd
** Blankröster etc

Jag är tacksam för hjälp med beräkningarna av Liz Jergle vid Statistiska institutionen, Lunds universitet.

som mäter den socialdemokratiska partitroheten, bör alltså ses som en funktion av socialdemokraternas (relativa) styrka i området och av andra, relevanta kontextuella variabler. Linearisera vi sedan $p_{s\text{-}s}$ och övriga parametrar i relationen (6), får vi ett tillskott av X-variabler med tillhörande regressionskoefficienter. Vi kan då försöka på nytt med en anpassning av den sålunda modifierade modellen till data.[3] Förhoppningsvis får vi då ett bättre resultat än det som redovisas i tabell 1.

Ett problem med den senare ansatsen är, att den enkelhet och tolkbarhet parametrarna har i den ursprungliga relationen (6) går förlorad i processen. En klar fara är också, att man hamnar i ett mekaniskt anpassande av regressionsekvationer, som saknar ämnesmässig förankring och där sambandet mellan individbeteende och gruppdata är oklart.

Rekonstruktion av cellfrekvenser

Vi är intresserade av att studera klassröstande och vi arbetar med en viss regional/administrativ indelning. För argumentering behövs några beteckningar. Vi inför en indikatorvariabel:

$$y^i_{SI} = \begin{cases} 1 & \text{om individ i röstar socialdemokratiskt och är industriarbetare} \\ 0 & \text{annars.} \end{cases}$$

Uppgiften y^i_{SI} registreras inte, eftersom vi respekterar valhemligheten.

Inte heller summan $\Sigma_i\, y^i_{SI} = Y_{SI}$ – totala antalet socialdemokratiskt röstande industriarbetare i området – är tillgänglig. Uppgifter av denna typ kan bli tillgängliga via surveyer, men det gäller då i regel urval, som är representativa för samtliga regioner. Det vore av stort intresse för forskningen att få tillgång till regional utförda surveyer om väljarbeteende. Tillgänglig via officiell statistik är:

$Y_{S.}$ och $Y_{.I}$, totala antalet socialdemokratiska röster, respektive totala antalet industriarbetare i regionen.

Då vi studerar klassröstande är vi särskilt intresserade av tal av typen:

$Q_{S/I} = Y_{SI}/Y_{.I}$, proportionen socialdemokratiskt röstande bland industriarbetare i området.

Talen $Q_{S/I}$ kan tolkas som benägenheten bland industriarbetare att rösta socialdemokratiskt, eller som den konditionella sannolikheten att en industriarbetare röstar socialdemokratiskt (vi bortser här ifrån variationer över tiden i denna benägenhet).

En uppgift man då står inför, är att med hjälp av tillgänglig statistik försöka rekonstruera tal av typen $Q_{S/I}$ för partier, yrkeskategorier och lokala enheter. En väg att göra detta är via regressionsteknik på sätt som redovisas i E&W's uppsats. En annan väg anvisas av Johnston i en serie arbeten (t ex monografin, 1985). Jag skall strax återkomma till Johnstons ansats.

Låt oss först konstatera att talen-benägenheterna $Q_{S/I}$ inte utgör individrelationer och inte kan förklara varför industriarbetare röstar socialdemokratiskt. Ambitionen då vi rekonstruerar benägenheterna är att beskriva, inte att förklara individuellt beteende. Man finner typiskt att talen $Q_{S/I}$ varierar över tid och rum. *En annan uppgift,* mer ambitiös än den förra, är då att med någon statistisk modell, säg regressionsekvation, försöka förklara dessa variationer i termer av några statsvetenskapligt meningsfulla faktorer.

Jag tycker att E&W i sin uppsats glider över från den ena uppgiften ovan till den andra. Den enklaste regressionsmodellen som redovisas i arbetet, och som inte fungerar så väl, är ett försök till rekonstruktion av Q-talen. De mer komplicerade modeller, som sedan anpassas, är hybrider, med parametrar som är svårtolkade i ämnesmässiga termer. Jag tror det är viktigt att hålla isär de två uppgifter som identifierats ovan.

Då man arbetar med anpassning av regressionsmodeller bör man lägga märke till att beloppet förklarad varians är ett rent statistiskt mått. Genom att lägga till någon variabel (och parameter) till modellen ökas förklaringsgraden, men det är därmed inte sagt att den anpassade modellen *ämnesmässigt* ger mer. Det är lätt gjort att man "köper" bättre anpassning till priset av mer svårtolkade parametrar. Jag tycker inte E&W riktigt lever upp till den anspråksnivå man har i titeln på uppsatsen, där begreppet modellspecifikation ingår. Även om uppsatsens huvudsyfte är att pröva olika aggregeringsnivåer, hade man gärna sett utförligare diskussion i termer av individ- och gruppprelationer.

Johnstons maximum-entropimetod

Jag skall här kortfattat försöka förklara innebörden av denna metod, som utgör ett i vissa situationer värdefullt alternativ till regressionsmetoden.

		Region:									
Social-		1		2		3		Totalt/riket	
grupp:		I	II	I	II	I	II	I	II
Parti:	A	?	A_1	?	A_2	?	A_3	Y_{AI}	A
	B								
		G_1	N_1	G_2	N_2	G_3	N_3		N

Vi har ovan vad statistiker kallar en 3-dimensionell kontingenstabell:

- partier (A, B) representerar rader,
- socialgrupper (I, II) kolumner och
- regioner (1, 2, 3) lager

i tabellen.

Marginalerna är i tillämpningar typiskt kända: A's röster i region 2, antalet i socialgrupp II i region 1, osv. Man kan också via surveydata förfoga över hyggligt tillförlitliga uppgifter för samtliga regioner, här kallat "riket", och markerat med Y_{AI} i tabellen. Däremot är naturligtvis övriga cellfrekvenser okända, liksom i E&W's fall, och här markerat med ?-tecken i tabellen.

Med hjälp av de data vi nu har tillgängliga kan man med statistisk teknik rekonstruera cellfrekvenserna. I vanliga fall ställs statistikern inför uppgiften att jämföra *faktiskt oberoende* cellfrekvenser med på något sätt genererade cellfrekvenser. Dessa senare är då genererade under någon hypotes, som innebär radikalt förenklad parameterstruktur. I vår speciella tillämpning saknas ju cellfrekvenserna, men ingenting hindrar oss från att utnyttja skattningsteknik, som hämtats från analysen av kontingenstabeller, för att rekonstruera de saknade cellfrekvenserna. I den av Johnston lanserade metoden – av honom kallad maximum-entropimetoden – antas 3:e ordningens samspelseffekter vara av försumbar storleksordning, dvs samspelet parti – socialgrupp – region antas vara försumbart. Under detta antagande och med tillgängliga data kan man rekonstruera cellfrekvenserna regionsvis. Man skapar därigenom en grund för att studera regionala variationer i klassröstande.[4]

Man noterar att Johnstons teknik kan betecknas som "kvasiekologisk" i det avseendet att vissa surveydata måste finnas tillgängliga. Detta innebär att metoden inte kan utnyttjas vid historiska studier i partibyten eller klassröstande. 1944 års svenska val, som ingår i E&W's studie, medger således inte skattning av t ex jordbrukares benägenhet att rösta socialdemokratiskt. För 1979 års val disponerar vi surveydata och metoden har tillämpats för rekonstruktion av partibytesmatriser (Berg, 1987).

Jag vill gärna avslutningsvis nämna Risbjerg Thomsens (1987) ansats, som är "äkta" ekologisk och som bygger på en modellformulering, där individrelation explicit kopplas till motsvarande grupprelation. Liksom vid Johnstons metod är syftet att rekonstruera flödesbenägenheter.[5]

Referenser

Berg, S (1987). Spatial influence on voter transitions in Swedish elections, an application of Johnston's maximum entropy method. To appear in: *Electoral Studies.*

Kramer, G H (1981). The ecological fallacy revisited: aggregate- versus individual level findings on economics and elections, and sociotropic voting. Social Science Working Paper 424, California Institute of Technology.

Langbein, L I & Lichtman, A J (1978). *Ecological Inference.* Sage University Paper 10.

Natale, P Borcio, R (1987). Voting dynamics in parliamentary and regional elections in Italy 1983–1985: a political turning point. Paper presented at ECPR Workshop, Amsterdam.

Risbjerg Thomsen, S (1987). *Danish elections 1920–79.* A logit approach to ecological analysis and inference. Politica, Århus.

Whiteley, P 1984). Inflation, unemployment and government popularity – dynamic models for the United States, Britain and West Germany. *Electoral Studies 3.* 3–24.

Noter

[1] E&W studerar klassröstande och skall egentligen också rekonstruera matriser för flöden mellan klasser och partier, alltså i princip samma problem.

[2] E&W använder kanske något olyckligt termen "giltig" för admissibel skattning.

[3] De på detta sätt modifierade modellerna tenderar emellertid att bli mer komplicerade än den multivariata modell E&W anpassar.

[4] För en tillämpning av tekniken på svenska data hänvisas till Berg (1987).

[5] Se också recension av Risbjerg Thomsens avhandling i detta nr.