Om utnyttjande av urvalsdesignen vid regressionsanalys av surveydata - en kort introduktion

Lennart Nordberg

SCB
R&D Report
Statistics Sweden
Research-Methods-Development
1989:8

Från trycket	April 1989
Producent	Statistiska centralbyrån, Utvecklingsavdelningen
Ansvarig utgivare	Staffan Wahlström
Förfrågningar	Lennart Nordberg, tel. 019-17 60 12

ISSN 0283-8680
Printed in Sweden
Garnisonstryckeriet, Stockholm 1989

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1989:8. Om utnyttjade av urvalsdesignen vid regressionsanalys av surveydata – en kort introduktion / Lennart Nordberg.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1989-X101OP8908

OM UTNYTTJANDE AV URVALSDESIGNEN VID REGRESSIONSANALYS AV SURVEYDATA - EN KORT INTRODUKTION

AV

LENNART NORDBERG

1 Inledning

När man vill studera samband mellan två eller flera variabler med hjälp av regressionsanalys brukar man utgå från ett antal standardantaganden som (i fallet två variabler) kan se ut på följande sätt:

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad , \; i=1,2,\ldots,n$$

där

- $x_1, x_2, \ldots, x_n$ är kända tal
- $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ är okorrelerade stokastiska feltermer
- $E(\varepsilon_i)=0$, $Var(\varepsilon_i)=\sigma^2>0 \quad i=1,2,\ldots,n.$

(1.1)

Det är oklart hur modell (1.1) skall uppfattas när man har surveydata. Vilken slumpmekanism är det som genererar feltermerna? På vilket sätt kommer urvalsdesignen in i bilden?

Duger den standardlösning som bygger på antagandena i (1.1) och som finns implementerad i många standardprogrampaket eller bör man väga observationerna med hänsyn till inklusionssannolikheterna efter mönster från surveysampling-teorin?[1]

Den första ansatsen innebär att parametrarna skattas på vanligt sätt, dvs enligt följande formler:

$$\hat{\beta} = \frac{\Sigma\, y_i (x_i - \bar{x})}{\Sigma (x_i - \bar{x})^2},$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},$$

$$\hat{\sigma}^2 = \frac{1}{n-2} \Sigma\, (y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

(1.2)

---

1) Det brukar finnas möjlighet att göra en slags vägd regressionsanalys i standardprogrampaketen. Man väger då observationerna för att ta hänsyn till att feltermerna ε_i har olika varians. Observera att den sortens vägning har en helt annan innebörd och ger bl a helt annorlunda variansformler, än vägning med hänsyn till inklusionssannolikheter.

medan ansatsen med vägning naturligen leder till formlerna:

$$\tilde{\beta} = \frac{\sum \frac{1}{\pi_i} y_i (x_i - \tilde{\tilde{x}})}{\sum \frac{1}{\pi_i} (x_i - \tilde{\tilde{x}})^2}$$

$$\tilde{\alpha} = \tilde{\tilde{y}} - \tilde{\beta}\tilde{\tilde{x}} \,, \qquad (1.3)$$

$$\tilde{\sigma}^2 = \sum \frac{1}{\pi_i} (y_i - \tilde{\alpha} - \tilde{\beta} x_i)^2 / \Sigma(1/\pi_i),$$

där $\tilde{\tilde{x}} = \frac{\Sigma(x_i/\pi_i)}{\Sigma(1/\pi_i)}$,

$$\tilde{\tilde{y}} = \frac{\Sigma(y_i/\pi_i)}{\Sigma(1/\pi_i)} \,,$$

och $\pi_1, \pi_2, \ldots, \pi_n$ är inklusionssannolikheterna för respektive observation i urvalet.

Frågan om vilken av lösningarna (1.2) och (1.3) som är att föredra är litet för komplicerad för att man skall kunna ge ett omedelbart, entydigt svar. Avsikten med denna uppsats är att belysa problematiken med hjälp av exempel och att ge rekommendationer om hur man kan göra i några olika situationer.

Uppsatsen är avsedd som en introduktion till min rapport "Generalized linear modeling of sample survey data", i fortsättningen benämnd [1] som ger en mer precis och mer generell framställning av regressionsanalys av surveydata. Framställningen där bygger på att man uppfattar modellen som en tvåstegsmodell. Först genereras värden $y_1, \ldots y_N$ för hela populationen enligt (1.1) med n utbytt mot N. Därefter dras ett stickprov om n objekt från den ändliga populationen och man observerar de givna (x,y)-värdena för objekten i stickprovet. I specifikationen av (1.1) uppfattas x-värdena som kända tal trots att de rimligen borde uppfattas om stokastiska. Det kan man göra utan allvarliga konsekvenser eftersom analysen görs betingat av x.

2 Exempel

En population om 5 000 element genererades utifrån 5 000 oberoende observationer vardera av tre normalfördelade variabler x_1, x_2 och ε. Därefter genererades y utifrån följande samband.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad (2.1)$$

där

- x_1 och x_2 är N(0,1) med inbördes korrelation ca 0,3
- slumpfelstermen ε är N(0,σ) och okorrelerad med övriga termer
- modellparametrarna bestämts till

$\beta_0 = 5.0, \quad \beta_1 = 1.0, \quad \beta_2 = 0.5, \quad \sigma = 0.2$

Därefter drogs, ur den nyss konstruerade populationen, ett stratifierat urval om 500 element på följande sätt:

Populationen delades in i tre strata där:

element i tillhör stratum 1 om $x_{2i} < 0$

element i tillhör stratum 2 om $0 \leq x_{2i} < 2$ (2.2)

element i tillhör stratum 3 om $2 \leq x_{2i}$.

Strata kom därigenom att bestå av 2 499, 2 400 respektive 101 element. Inom respektive stratum drogs ett OSU om 39, 360 respektive 101 element. Dessa siffror valdes för att (förutom att summera sig till 500) åstadkomma starkt varierande inklusionssannolikheter mellan strata; ca 1,5 procent, 15 procent respektive 100 procent för stratum 1, 2 respektive 3.

Parametrarna i modell (2.1) skattades utifrån x- och y-värdena för de 500 element som ingick i urvalet. Ovägda skattningar, $\hat{\beta}$ och $\hat{\sigma}$ (där man inte tar hänsyn till de varierande inklusionssannolikheterna), togs fram såväl som, med hänsyn till inklusionssannolikheterna, vägda skattningar $\tilde{\beta}$ och $\tilde{\sigma}$.

Dessutom skattades β och σ utifrån populationens samtliga 5 000 element. Dessa skattningar betecknas $\hat{\beta}_{pop}$ och $\hat{\sigma}_{pop}$.

Resultatet framgår av följande tabell:

Parameter	Skattning		
	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}_{pop}$
β_0	4.99	4.99	5.00
β_1	0.98	1.00	1.00
β_2	0.51	0.50	0.50
	$\hat{\sigma}$	$\tilde{\sigma}$	$\hat{\sigma}_{pop}$
σ	0.19	0.19	0.20

Tabell 2.1 Resultat av parameterskattningar i modell (2.1)

Som framgår av tabell 2.1 är de ovägda och vägda skattningarna inbördes mycket lika och dessutom mycket nära de sanna parametervärdena, trots att inklusionssannolikheterna varierar så mycket. Förklaringen till att de ovägda skattningarna fungerar bra här ligger i följande resultat, som bevisas i en mer precis och mer generell version i [1].

Låt

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad i=1,2,\ldots,N \tag{2.3}$$

a) Antag att följande villkor är uppfyllda:

$$\left.\begin{aligned} &\frac{1}{\sqrt{n}} \sum_{i=1}^{N} (y_i - \mu_i)\pi_i \longrightarrow 0, \text{ när } n \text{ och } N \longrightarrow \infty, \\ &\frac{1}{\sqrt{n}} \sum_{i=1}^{N} (y_i - \mu_i)x_{1i}\pi_i \longrightarrow 0, \text{ när } n \text{ och } N \longrightarrow \infty, \\ &\frac{1}{\sqrt{n}} \sum_{i=1}^{N} (y_i - \mu_i)x_{2i}\pi_i \longrightarrow 0, \text{ när } n \text{ och } N \longrightarrow \infty, \end{aligned}\right\} \tag{2.4}$$

Då är (under ytterligare några allmänna regularitetsvillkor) den ovägda skattningen $\hat{\beta}$ designkonsistent i den meningen att $|\hat{\beta}-\hat{\beta}_{pop}|$ konvergerar mot noll i sannolikhet när n och N växer mot ∞. Villkoren i (2.4) är uppfyllda om residualerna $(y_i-\mu_i)$ är okorrelerade med π_i och med produkttermerna $\pi_i x_{1i}$ och $\pi_i x_{2i}$.

Inklusionssannolikheten π kan enligt (2.2) uttryckas som en funktion av x_2. Eftersom residualen $y-\mu$ enligt (2.1) är proportionell mot ε som i sin tur är okorrelerad med x_2 och (i stort sett) okorrelerad med π så kan man se att den ovägda skattningen $\hat{\beta}$ bör fungera i detta fall.

I praktiken har man naturligtvis inte tillgång till en så fullständig information om den "sanna" modellen som ovanstående resonemang förutsätter. Man behöver ett mer operationellt sätt att avgöra om konsistensvillkoren är (i rimlig utsträckning) uppfyllda.

En möjlighet är att göra på följande sätt: Skatta först parametrarna i (2.1) ovägt. Utvidga därefter modellen genom att bland x-variablerna inkludera π (sedd som en variabel) och produkttermerna πx_1 och πx_2. Notera att ekvationerna (2.4) är normalekvationerna, baserade på samtliga observationer i populationen, för de parametrar som svarar mot π, πx_1 respektive πx_2 i den utvidgade modellen.

Konsistensvillkoren för ovägd β-estimation är därmed, åtminstone approximativt, uppfyllda för den utvidgade modellen. Testa - med ett vanligt F-test - om den utvidgade modellen ger signifikant bättre anpassning till y än den ursprungliga modellen (2.1). Om så inte är fallet bör villkoren (2.4) vara någorlunda väl uppfyllda och därmed den ovägda skattningen i (2.1) rättfärdigad.

Om den utvidgade modellen däremot ger signifikant bättre anpassning än den ursprungliga så innebär det att π innehåller information som skulle kunna användas för att förbättra den ursprungliga modellen. Försök att hitta en variabel som är högt korrelerad med π och som det är saklogiskt rimligt att inkludera i modellbyggandet. Om detta inte är möjligt så bör man övergå till vägd estimation.

Vi illustrerar förfarandet genom att studera den ofullständiga modellen:

$$y = \beta_0' + \beta_1' x_1 + \varepsilon' \text{ , där } \varepsilon' \text{ är } N(0,\sigma'). \qquad (2.5)$$

Fortfarande förutsätts att (2.1) är den sanna modellen. Man kan tänka sig en situation där man inte är medveten om att x_2 är en betydelsefull variabel för att beskriva variationen i y.

Följande tabell visar skattningar av de tre parametrarna i (2.5) dels ovägt, dels vägt utifrån de 500 observationerna i urvalet, dels utifrån samtliga 5 000 observationerna i populationen.

Parameter	Skattning		
	$\hat{\beta}'$	$\tilde{\beta}'$	$\hat{\beta}'_{pop}$
β_0'	5.42	4.98	5.00
β_1'	1.14	1.16	1.16
	$\hat{\sigma}'$	$\tilde{\sigma}'$	$\hat{\sigma}'_{pop}$
σ'	0.50	0.50	0.51

Tabell 2.2 Skattning av parametrarna i (2.5).

Man ser att de vägda skattningarna ansluter väl till de populationsbaserade medan främst den ovägda β_0'-skattningen kraftigt avviker från motsvarande populationsbaserade skattning.

Man skulle då möjligen kunna dra den slutsatsen att den vägda skattningen bör rekommenderas i detta fall. Men det är en förhastad slutsats. Vid en jämförelse mellan tabell 2.1 och 2.2 framgår att $\sigma' >> \sigma$, dvs (2.5) ger mycket sämre anpassning till data än (2.1). Hela den y-variation som beror på variationen av x_2 är ju i modell (2.5) inbakad i slumptermen ε'.

Istället för att nöja sig med en modell med en så stor slumpterm som (2.5) bör man enligt vad som diskuterades tidigare undersöka om π bär på någon information om y som saknas i modell (2.5).

Modell (2.5) utvidgades därför med variablerna π och πx_1. Skattningen av σ i denna modell blev ca 0.35. Ett F-test av (2.5) mot den utvidgade modellen gav följande approximativa F-testvariabel:

$$F \approx \frac{(498 \cdot 0.50^2 - 496 \cdot 0,35^2)/2}{0,35^2} \approx 250,$$

som skall jämföras med det - på risknivå 5 procent - kritiska värdet som är ca 3.

Detta pekar på att π innehåller mycket information om y som inte finns i x_1. I praktiken vill man förstås inte ha med variabeln π i modellen men förfarandet riktar uppmärksamheten mot den information som finns i designen (dvs π) så att man därigenom leds att ta hänsyn till stratifieringsvariabeln x_2 i modellbyggandet. Därigenom kan man nå en väsentligt bättre modellbeskrivning än om man bara nöjer sig med att väga observationerna med hänsyn till π.

Antag nu att man istället har med x_2 i modellen men att man inte är medveten om betydelsen av x_1.

Vi studerar alltså den ofullständiga modellen:

$$y = \beta''_0 + \beta''_1 x_2 + \varepsilon'' \text{ , där } \varepsilon'' \text{ är } N(0,\sigma''). \qquad (2.6)$$

och antar fortfarande att den sanna modellen beskrivs av (2.1).

Följande tabell visar skattningarna av parametrarna i (2.6) dels ovägt, dels vägt, dels utifrån samtliga observationer i populationen.

Parameter	Skattning		
	$\hat{\beta}''$	$\tilde{\beta}''$	$\hat{\beta}''_{pop}$
β''_0	5.01	5.07	5.01
β''_2	0.86	0.84	0.82
	$\hat{\sigma}''$	$\tilde{\sigma}''$	$\hat{\sigma}''_{pop}$
σ''	0.96	0.96	0.96

Tabell 2.3 Skattning av parametrarna i (2.6).

Man kan se att de vägda och ovägda skattningarna nu är mycket lika och att de i stort sett sammanfaller med de populationsbaserade skattningarna.

Man kan också se - genom att jämföra tabellerna 2.1 och 2.3 - att $\sigma'' >> \sigma$, dvs modell (2.6) innehåller en mycket större slumpterm än modell (2.1).

Här kan man dock inte förvänta sig att man skall kunna förbättra sin modell - dvs (2.6) - genom att utnyttja π. Den väsentliga designinformationen finns redan med i modellen (och det är skälet till att de ovägda och vägda skattningarna överensstämmer här) medan den information som saknas finns i variabeln x_1. Det gäller alltså här att man - t ex utifrån saklogiska överväganden - verkligen får fatt i, och inkluderar x_1 i modellarbetet.

3 Sammanfattning

Antag att man vill skatta β-parametrarna i regressionsmodellen $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$.

Om residualen $(y-\mu)$, där $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, är okorrelerad med inklusionssannolikheten π och med produkttermerna πx_1 och πx_2 så kan den vanliga, ovägda β-skattningen användas. Den är för övrigt effektivare än den - med hänsyn till π - vägda skattningen i denna situation, se avsnitt 5 i [1].

Om $(y-\mu)$ däremot är korrelerad med π, πx_1 eller πx_2 så är den vägda skattningen bättre. Men det finns då också följande alternativ som kan vara bättre än att använda vägning. Eftersom π nu bär på sådan information om y som inte fås genom x_1 och x_2 i modellen så ligger det nära till hands att, istället för att väga med hänsyn till π, försöka förbättra modellen. Det gäller då att hitta en variabel z (eller eventuellt flera) som dels är högt korrelerad med $(y-\mu)$ och dels rent saklogiskt passar som variabel i modellen. Om residualen för den nya modellen är okorrelerad med π, πx_1, πx_2 och πz så kan parametrarna i den nya modellen skattas ovägt. Den här skisserade proceduren innebär alltså att man eliminerar den ovägda skattningens designbias samtidigt som man förbättrar regressionsmodellen.

Vi har hittills antagit att bortfall saknas. De metoder som skisserats ovan kan generaliseras på ett förhållandevis rättframt sätt så att man även kan ta hänsyn till bortfallseffekter. Detta behandlas i [1].

Referens:

[1]: L Nordberg: Generalized Linear Modeling of Sample Survey Data. SCB R&D report 1988:8.

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

R & D Reports, Statistics Sweden, are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown (beige) covers).

Reports published earlier during 1989 are:

1989:1 (grön)	**Går det att mäta produktivitetsutvecklingen för SCB? (Rune Sandström)**
1989:2 (grön)	**Slutrapporter från U-avdelningens översyn av HINK och KPI (flera författare)**
1989:3 (grön)	**A Cohort Model for Analyzing and Projecting Fertility by Birth Order (Sten Martinelle)**
1989:4 (beige)	**Abstracts I - Sammanfattningar av metodrapporter från SCB**
1989:5 (gul)	**On the use of Semantic Models for specifying Information Needs (Erik Malmborg)**
1989:6 (grön)	**On Testing for Symmetry in Business Cycles (Anders Westlund och Sven Öhlén)**
1989:7 (grön)	**Design and quality of the Swedish Family Expenditure Survey (Håkan L Lindström, Hans Lindkvist och Hans Näsholm)**

Kvarvarande BEIGE och GRÖNA exemplar av ovanstående promemorior kan rekvireras från Elisabet Klingberg, U/STM, SCB, 115 81 Stockholm, eller per telefon 08-783 41 78.

Dito GULA exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 Stockholm, eller per telefon 08-783 41 47.