Tips för analyser med regressionsmodeller på statistikdata

Resumé

Denna beskrivning tar upp några saker att tänka på när man gör analyser av samband på data från statistiska undersökningar. Beskrivningen är inriktad på regressionsanalys med praktiskt tillgängliga verktyg. Vikter och partiellt bortfall är potentiella komplikationer som det gäller att hålla reda på men som i regel har praktiska lösningar.

Beskrivningen är avsedd både som vägledning inom SCB och för spridning till kunder och andra som önskar råd för analyser.

1 Allmänt om analyser av samband

Analysbearbetningar är ofta till nytta för att påvisa och belägga samband mellan variabler i observerade data. För analyser av samband står olika metoder till buds, grundade på mer eller mindre stark teori. Verktyg för att använda olika analysmetoder är tillgängliga i välkända programvaror för persondator, såsom SAS®, Minitab®, SPSS®, STATA® och R.

Statistiska analyser kan ha olika slags syften. Man kan urskilja två huvudlinjer:

- *Explorativt*, eller *hypotesgenererande*, syfte. Där gäller det att låta observationsmaterialet tala och visa upp möjliga samband. Man är där närmast ute efter uppslag till fortsatt sökande.
- *Konfirmativt*, eller *hypotesprövande*, syfte. Där gäller det att ur observationsmaterialet dra kontrollerat säkra slutsatser om samband, eller att skatta storheter med kontrollerad säkerhet. Man är där ute efter tillförlitliga och/eller praktiskt användbara resultat.

Valet av lämplig analysmetod i olika situationer beror allmänt på vilket av dessa slags syften som är primärt i sammanhanget.

Regressionsanalys är en klass av analysmetoder som har mycket omfattande användningar och som passar bra för både explorativa och konfirmativa syften. Det gäller dock att använda metoderna på rätt sätt för det syfte man har med sin analys. Regressionsanalysen har en relativt stark teoretisk grund i statistikvetenskapen.

Dokumentet ingår i Verksamhetsstödet vid SCB:
Tillhör delprocess 6.1 (primärt), 6.5 (sekundärt).
Författare: Martin Ribe, SCB.

Denna beskrivning utgår i första hand från regressionsanalys. De allmänna principerna är dock till stor del motsvarande tillämpliga även på andra analysmetoder.

2 Problem vid analys på statistiska data

När man ska analysera data från statistiska undersökningar möter ofta vissa komplikationer. Observationsmaterialen där uppfyller inte alltid helt vad analysmetoderna förutsätter. Viktiga komplikationer är:

- *Partiellt bortfall* förekommer i statistiska undersökningar, dvs. i observationsmaterialet saknas värden på en del variabler för en del objekt. Vanliga analysmetoder klarar ofta inte detta direkt, utan förutsätter kompletta data.
- *Vikter* för objekten förekommer ofta i undersökningarna, att beaktas i beräkningen av statistikvärden. Vanliga analysmetoder kan ofta inte använda vikterna på ett korrekt sätt.
- *Beroenden mellan objekt* kan ibland uppstå genom formen för urvalsdragningen i statistiska undersökningar, t.ex. vid s.k. klusterurval. Vanliga analysmetoder förutsätter däremot ofta oberoende mellan objekten.
- *Slump och osäkerhet* betraktas på delvis olika sätt i undersökningarnas respektive analysernas metodik. Urvalsmetodiken räknar med slump i urvalsdragningen, och osäkerheten som detta resulterar i. Analysmetodiken å andra sidan räknar med slump i oförklarade avvikelser från samband, och osäkerheten som detta resulterar i.

Detta dokument behandlar praktiska sätt att ta hand om dessa komplikationer.

3 Former av regressionsanalys

För att göra sammanhanget klart ska här först rekapituleras lite om regressionsanalys allmänt. För närmare förklaringar av det grundläggande hänvisas till läroböcker i ämnet, såsom Andersson m.fl. (2007) eller Ryan (1997). Allmänna fundamentala aspekter på val av modeller och metoder för analyser förklaras informativt av t.ex. Ruist (1990) och Kennedy (2008).

Idén med regressionsanalys är att analysera hur en variabel beror av andra variabler. Indata till analysen är en mängd observationer (observationsposter) på olika objekt, med observerade värden på vissa variabler för objekten. T.ex. kan objekten vara individerna i en given grupp, och variablerna kan vara observerade egenskaper hos individerna. Observationsmaterialet kan man ha fått genom insamling i en statistisk undersökning, eller genom ett experiment på försöksobjekt, eller på annat sätt.

Avidentifierade data

Datamaterialet man analyserar är normalt *avidentifierat*, dvs. inte försett med några uppgifter om objektens identitet, såsom personnummer, namn eller löpnummer som är gemensamma med andra datakällor. Sådana uppgifter ska inte användas här och har tagits bort för bästa möjliga skydd av objektens integritet (se Svenska statistikfrämjandet 2010, avsnitt 3.1).

Variablerna i sina roller – utfallsvariabel och förklarande variabler

Variablerna har olika roller i regressionsanalysen. För det första ingår en beroende variabel, även kallad *utfallsvariabel* eller responsvariabel. För det andra ingår (en eller) flera oberoende variabler, även kallade *förklarande variabler*, eller regressorer, eller (informellt) ”*x*-variabler”. Analysen ska ge mått på hur utfallsvariabeln beror av de förklarande variablerna.

Vi betecknar den beroende variabeln med y och de förklarande variablerna med $x_1, \ldots, x_k$. Här står k för antalet förklarande variabler som ska användas i analysen.

I det följande beskrivs några olika former av regressionsanalys. I litteraturen förekommer åtskilliga varianter med i grunden likartade idéer men med olika modifieringar eller utbyggnader för att passa under olika förutsättningar. Skillnaden mellan de här beskrivna formerna har delvis att göra med vilken typ av variabel som y är.

Modeller med kvantitativ utfallsvariabel – grundform

Regressionsanalysen i sin grundform används när den beroende variabeln y är kvantitativ, alltså mått på något i siffror, t.ex. inkomst. Man ställer upp en *modell* som uttrycker ett antagande om den matematiska formen för hur y beror av de förklarande variablerna. I vanlig s.k. *linjär regressionsanalys* ser modellen ut enligt följande:

$$(1) \qquad y = b_0 + b_1 x_1 + b_2 x_2 + \ \ldots \ + b_k x_k + \varepsilon$$

Här är $b_0, b_1, \ldots, b_k$ vad som kallas regressionskoefficienter. De är modellens parametrar som analyskörningen skattar utifrån de observerade värdena på variablerna för objekten i materialet.

Termen ε kallas residual eller restterm (ibland felterm), och den fångar upp de variationer i y som inte kan förklaras av *x*-variablerna. Den termen betraktas i den matematiska modellen som en slumpvariabel (stokastisk variabel). Man antar att denna slumpvariabel är oberoende mellan objekten och har samma sannolikhetsfördelning för alla objekt, med väntevärdet noll och ändlig varians. Ibland antas även att residualens sannolikhetsfördelning är en normalfördelning.

Observera att i regressionsmodellen (1) uppfattar man de förklarande variablerna $x_1, \ldots, x_k$ som att de inte beror av slumpen utan har fixa

värden. (De ses som ”icke-stokastiska”; jfr Cramér 1946, Sect. 37.1.) Utfallsvariabeln y däremot beror av slumpen, genom att den beror av den slumpmässiga resttermen. I modellen ses alltså utfallsvariabeln som en funktion av dels de förklarande variablerna, dels slumpen. Formeln (1) som uttrycker regressionsmodellen kallas även regressionsekvation.

De skattade värdena på regressionskoefficienterna b_0 , b_1 ,…, b_k kan ses som analysens huvudsakliga resultat. De är mått på samband mellan variablerna. Analyskörningen ger även osäkerhetstal (och därmed konfidensintervall) för de skattade regressionskoefficienterna. En fördel med regressionsanalys som metod är att regressionskoefficienterna har en naturlig tolkning.

Tolkning av resultaten. Regressionskoefficienten b_1 i modellen (1) betyder (enkelt uttryckt): När x_1 ökas med 1 enhet, så ökas y med b_1 enheter.

Det man får ut här är ett mått på hur starkt utfallsvariabeln y beror av enbart variabeln x_1 för sig. Då tänker man sig att de övriga förklarande variablerna hålls konstanta (dvs. ”allt annat lika”). Exempel på tillämpningar av detta är utvärderingar, där man vill mäta hur stor effekt på y som fås genom en åtgärd på x_1 , rensat från effekter av annat som kan ha hänt eller ändrats.

Predikterade utfall. Regressionsanalysen ger också en annan form av resultat. För en observation kan man räkna ut vad högerledet i (1) blir med de skattade regressionskoefficienterna och med residualen satt till noll. Då får man det värde på utfallsvariabeln *y* som modellen predikterar, eller förutsäger teoretiskt, för observationen i fråga. Man ser att residualen kan uppfattas som avvikelsen mellan det observerade och det predikterade värdet på *y* .

Typer av förklarande variabler i modellerna

De förklarande variablerna x_1 ,…, x_k i regressionsmodellen kan vara av olika typer: antingen kvantitativa variabler eller kategorivariabler. Detta gäller både modellen (1) och de modeller som beskrivs i följande avsnitt.

Kvantitativa variabler är mått i siffror på något, t.ex. inkomst, medan kategorivariabler anger tillhörighet till en kategori, t.ex. kön eller yrke. Kategorivariabler kallas även ”kvalitativa variabler”. Kategorivariabler uttrycker man gärna i form av s.k. dummyvariabler med bara två möjliga värden, nämligen 0 och 1. De värdena står i typfallet för svaren nej resp. ja på en viss fråga om objektet.

Ibland har man att välja om en variabel ska behandlas som kvantitativ eller som kategorivariabel. *Ålder* är en sådan variabel, som mycket ofta ingår som förklarande variabel i analyser där objekten är individer.

Att behandla åldern som kvantitativ kan ofta innebära ett allför specifikt antagande om hur åldern inverkar. Lämpligt är därför kanske i regel att ta åldern som kategorivariabel. Man delar då in åldrarna i lämpliga intervall och inför en dummyvariabel för varje intervall, utom ett av dem, vilket kallas referenskategori. Regressionskoefficienten för varje åldersintervalls dummyvariabel mäter då effekten på *y* av att tillhöra det åldersintervallet, jämfört med att tillhöra referenskategorin. Idén med dummyvariabler för intervall är naturligtvis användbar även för andra variabler än ålder.

Modeller med transformerade variabler

Ofta förekommande varianter på linjär regressionsanalys är att man transformerar variablerna på ett eller annat sätt (jfr Emerson & Stoto 1983). Vanligt är t.ex. att man logaritmerar y. Det förutsätts då att y bara kan anta värden större än 0. Modellen får då detta utseende:

$$\text{(2)} \qquad \ln y = b_0 + b_1 x_1 + b_2 x_2 + \ ... \ + b_k x_k + \varepsilon$$

Här betecknar ln naturliga logaritmen. Transformationen genom logaritmeringen är ett icke-linjärt inslag i modellen. Ändå handlar det väsentligen om linjär regressionsanalys här, för när transformationen väl är gjord kör man på med vanlig linjär modell. Detta är praktiskt en stor fördel och gör analysen relativt väl hanterlig. Modellen (2) kallas ”loglinjär”. Samma modell kan alternativt skrivas som en multiplikativ modell med samma innebörd, nämligen:

$$\text{(2a)} \qquad y = B_0 \times B_1^{x_1} \times ... \times B_k^{x_k} \times e^{\varepsilon}$$

Formlerna (2) och (2a) säger samma sak, genom att vi väljer beteckningarna så att t.ex. $B_1 = e^{b_1}$.

Tolkning av resultaten. Regressionskoefficienten b_1 i modellen (2) betyder (enkelt uttryckt): När x_1 ökas med 1 enhet, så multipliceras y med talet e^{b_1}. Här är $e \approx 2{,}718$ basen för de naturliga logaritmerna. Annorlunda uttryckt: När x_1 ökas med 1 enhet, så ökas y med $(e^{b_1} - 1) \times 100$ procent, eller grovt räknat $b_1 \times 100$ procent (om b_1 inte ligger långt från 0).

Genuint icke-linjära modeller
Det kan förekomma mera genuint icke-linjära regressionsmodeller. Ett fiktivt (eller möjligen verkligt?) exempel är följande modell:

$$\text{(3)} \qquad y = b_0 + e^{b_1 x_1 + c} + b_2 x_2 + \varepsilon$$

I motsats till modellen (2) så kan denna icke-linjära modell inte lätt reduceras till vanlig linjär regressionsanalys av formen (1). Sådana genuint icke-linjära modeller går att hantera men kan innebära vissa

tekniska problem, såsom risk för falska skattningar och beroende av gissade startvärden för s.k. itereringar i skattningsalgoritmen. Tolkningen av resultaten kan också vara problematisk. Det kan därför finnas skäl att vara lite försiktig med mera obeprövade modellformer.

Modeller med ja/nej-värd (dikotom) utfallsvariabel – logistisk regressionsanalys (logitanalys)

Ofta kan utfallsvariabeln vara av det slaget att den bara kan anta två värden, som kan betecknas med ja resp. nej, eller 1 resp. 0. Den kan då ange om objektet har eller inte har en viss egenskap. Ett exempel är när objekten är individer och utfallsvariabeln *y* anger om individen har en viss sjukdom. De förklarande variablerna i analysen kan då vara riskfaktorer för denna sjukdom, tillsammans med andra variabler som kan inverka.

Då är det ofta lämpligt att ta till s.k. logistisk regressionsanalys, eller logitanalys. Där ser modellen ut enligt följande:

$$(4) \qquad \begin{cases} y \sim \text{Binomial}(1, p) \\ \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \end{cases}$$

Denna modell säger att värdet på *y* för varje objekt kan tänkas genererat genom en lottning. I lottningen är det sannolikheten p för utfallet $y = 1$, och sannolikheten $1 - p$ för utfallet $y = 0$. Sannolikheten p beror av de förklarande variablerna $x_1, \ldots, x_k$ enligt nedre raden i formlerna (4).

Som i modell (1) är här $b_0, b_1, \ldots, b_k$ regressionskoefficienter som analyskörningen skattar, tillsammans med osäkerhetstal. I motsats till modell (1) har modell (4) inte någon restterm (residual) ε. Det beror på att slumpvariationen mellan objekt, som resttermen uttrycker i modell (1), kommer in på annat sätt i modell (4), när y tänks lottat.

Tolkning av resultaten. Regressionskoefficienten b_1 i modellen (4) betyder att talet e^{b_1} kan tolkas som en s.k. oddskvot, eller kvot mellan oddstal. Närmare bestämt: När x_1 ökas med 1 enhet, så multipliceras oddstalet för utfallet $y = 1$ med talet e^{b_1}. (För förklaring, se Ribe 1997, 1999.)

Modeller med (antal) händelser som utfallsvariabel – Poissonregression

Utfallsvariabeln kan även vara av det slaget att den räknar antal händelser, t.ex. olycksfall eller dödsfall, under en period. De förklarande variablerna i analysen kan då vara riskfaktorer för händelserna i fråga, ofta tillsammans med andra variabler som kan inverka.

Då kan det vara lämpligt med s.k. Poissonregression. Där ser modellen ut enligt följande:

$$(5) \qquad \begin{cases} y \sim \text{Poisson}(\lambda) \\ \ln \lambda = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \end{cases}$$

Modellen säger att händelserna inträffar spontant och oberoende av varandra, slumpmässigt med väntevärdet λ för antalet händelser. Det förväntade antalet λ beror av de förklarande variablerna $x_1, \ldots, x_k$ enligt nedre raden i formlerna (5). Som förut är här $b_0, b_1, \ldots, b_k$ regressionskoefficienter som analyskörningen skattar, tillsammans med osäkerhetstal.

En variant på metoden är intensitetsregression. Där beaktar man även tidpunkterna för händelserna och är intresserad av hur de förklarande variablerna inverkar på händelsetakten, eller intensiteten, i antalet händelser per tidsenhet. En s.k. ”proportional-hazard”-modell är modifierad i förhållande till formen (5) så att λ står för en intensitet som är en funktion av tiden, och även b_0 är en funktion av tiden; en vanligt förekommande form kallas *Cox-modellen* (se t.ex. Lee m.fl. 2006, Sect. 10.1). Ansatsen används för överlevnadsanalys, där det t.ex. kan handla om dödsintensiteters beroende av riskfaktorer såsom sjukdomar eller miljöfaktorer, rensat från inverkan av t.ex. åldern.

I likhet med modellen (2) så är modellen (5) av loglinjär form, men den tänkta slumpmekanismen är av olika slag. I modellen (5) kan utfallsvariabeln *y* bli 0, men det kan den inte i modellen (2). Modifieringar av Poissonregressionsmodellen (5) har utvecklats för situationer där antalet händelser varierar starkare än i Poissonfördelningen; se artikeln Overdispersion (2011).

Tolkning av resultaten. Regressionskoefficienten b_1 i modellen (5) betyder att talet e^{b_1} kan tolkas som en s.k. relativ risk eller (i vissa fall) ”hazard ratio”, hasardkvot. Närmare bestämt: När x_1 ökas med 1 enhet, så multipliceras sannolikheten för händelsen i fråga med talet e^{b_1}. Alternativt, vid intensitetsregression: När x_1 ökas med 1 enhet, så multipliceras händelsetakten per tidsenhet med talet e^{b_1}.

Modeller med avgång som utfallsvariabel
– komplementär loglogmodell

I vissa tillämpningar kan utfallsvariabeln vara en ja/nej-variabel som anger avgång ur en grupp under en uppföljningsperiod. Gruppen antas vara sluten, dvs. utan nytillskott under perioden och utan möjlighet till återinträde för de avgångna. Ett exempel är en grupp nyutbildade där utfallsvariabeln anger om personen får sitt första arbete under perioden. De förklarande variablerna är där faktorer som kan ha betydelse för chansen att få ett arbete.

Då kan det vara lämpligt med en komplementär loglogmodell, eller cloglog-modell (jfr Bender & Benner 2000; Andersson & Hagsten 2011). Den kan skrivas enligt följande:

$$(6) \qquad \begin{cases} y \sim \text{Binomial}(1, p) \\ \ln(-\ln(1-p)) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \end{cases}$$

Liksom i logitmodellen (4) har man här en utfallsvariabel av ja/nej-typ. Skillnaden i modellen mellan (4) och (6) har att göra med hur man vill tolka resultaten. Benämningen på metoden kommer av att vänsterledet i nedre formeln är en s.k. komplementär loglogtransformation av p, genom att $1 - p$ är komplementära sannolikheten till p.

Tolkning av resultaten. Regressionskoefficienten b_1 i modellen (6) betyder att talet e^{b_1} kan tolkas som en relativ risk eller (exaktare) ”rate ratio”, ”hazard ratio”, hasardkvot. Närmare bestämt: När x_1 ökas med 1 enhet, så multipliceras den förväntade avgångstakten per tidsenhet med talet e^{b_1}

Kort om andra varianter på regressionsanalys

Som nämnts finns det i litteraturen en stor mångfald av ytterligare varianter på regressionsanalys, utöver dem som nu behandlats. Här ska kort kompletteras med några få valda begrepp som ofta förekommer i litteraturen. Detta avsnitt syftar enbart till en översiktlig partiell orientering.

Generalized Linear Model (GLM) är benämning på ett ganska generellt metodologiskt ramverk, där man i en enhetlig men något teoretisk formelapparat kan fånga olika former av regressionsmodeller och även vissa justeringar i modellantagandena; se t.ex. Lee m.fl. (2006). Till generaliseringarna där hör s.k. flernivåmodeller eller Hierarkiska Linjära Modeller (HLM), som syftar till analys av samband på flera aggregeringsnivåer samtidigt. *Generalized Estimating Equations (GEE)* är också en variant här. Något förvillande är distinktionen mellan de två snarlika uttrycken *Generalized Linear Model* och *General Linear Model* (utan ”-ized”, men också förkortat GLM). De står för likartade men inte identiska begrepp (jfr dessa uppslagsord i Wikipedia). – Alternativ med *Bayesiansk* regressionsanalys behandlas även i metodlitteraturen (t.ex. Walter & Augustin 2009).

Strukturekvationsmodeller (SEM) är modeller med nätverksstrukturer av regressionssamband mellan både observerade variabler och s.k. latenta, icke observerbara, variabler; se t.ex. Bollen (1989). Modellerna kan rensa för mätvariabilitet och även innehålla flernivåstrukturer. *LISREL*® (Linear Structural Relations) är en programvara för SEM med multivariata normalfördelningar; de observerrade variablerna kan vara antingen kvantitativa eller kategorivariabler. *Faktoranalys* kan ses som ett specialfall av SEM. *PLS* (Partial Least Squares Regression) är en annan metod med nätstrukturer av samband men något annan teori bakom, etablerad för bl.a. kundstudier (Fornell & Bookstein 1982). *Latent Class Analysis* (LCA) är en ansats för SEM med kategorivariabler som latenta variabler.

Robust regressionsanalys används i princip för samma slags syften som vanlig regressionsanalys med kvantitativ utfallsvariabel, men skattningsmetoden för regressionskoefficienterna är mindre känslig än den vanliga för störningar från

exceptionellt utstickande värden (outliers). *Kvantilregression* är en sådan robust regressionsmetod. *Tobitanalys* är regressionsanalys med kvantitativ utfallsvariabel som är begränsad till ett visst intervall (se Kennedy 2008, Sect. 17.2; Tobin 1958).

Multinomial Logit Models (MNL) har utfallsvariabler med få men fler än två möjliga värden. Metoden beror i sin grundform av ett modellantagande som kan vara problematiskt, det s.k. Axiom of Independence of Irrelevant Alternatives (IIA). Har utfallsalternativen en naturlig hierarkisk struktur kan dock denna ofta modelleras så att antagandet inte behövs, genom mera avancerade *Mixed Multinomial Logit Models* (MMNL) med särskild programvara (McFadden 2001; Hensher & Greene 2001). – *Probitanalys* är mycket snarlik logitanalys men saknar resultatens tolkbarhet som oddskvoter.

4 Modellantaganden och modellanpassning

Modellens innebörd

När man ställer upp en regressionsmodell så gör man vissa antaganden om verkligheten. Det är antaganden om på vilket sätt utfallsvariabeln y beror av de förklarande variablerna $x_1, \ldots, x_k$. Exempelvis kan antagandet vara det som uttrycks av formeln (1) ovan. Gemensamt för de olika modellerna som beskrevs ovan är att man implicit gör dessa två antaganden:

(i) Utfallet är oberoende mellan observationerna.
(ii) Utfallets sannolikhetsfördelning följer den uppställda modellen.

Exempelvis för modellen (1) innebär antagandena (i) och (ii) detta:

(i)' Residualerna ε är oberoende mellan observationerna.
(ii)' Sannolikhetsfördelningen för residualerna ε är lika för alla objekt, eller åtminstone har alla väntevärdet 0 och samma varians.

En konsekvens av antagandet (ii) är vidare att utfallsvariabeln förväntas bero av de förklarande variablerna på ett för alla objekten likartat sätt.

I praktiken måste man räkna med att dessa antaganden aldrig är exakt uppfyllda. Modellens tillämplighet hänger på om den ändå kan ses som en tillräckligt god approximation av verkligheten. (”All models are wrong, but some are useful”, G.E.P. Box.) När man ska välja och bedöma modellen är det bra att tänka på vad man vet om dels hur den modellerade verkligheten fungerar, dels hur modellen fungerar.

Diagnostika

Analyskörningarna med analysprogramvaror genererar en rikhaltig mängd s.k. diagnostika, som är processdata om hur analysen fungerar på datamaterialet i fråga. Dessa diagnostika innehåller nyttig information och kan varna för problem i modellens tillämplighet på observationsmaterialet. De är dock inte speciellt avpassade för komplikationerna till följd av urvalsdesign och liknande.

Regressionsdiagnostika handlar bland annat om dessa aspekter:

(i) Hur nära modellen passar till det verkliga utfallet.
(ii) Vilken påverkan enskilda observationer har – gäller s.k. outliers.
(iii) Hur likartade variabler kan störa varandras effekter – s.k. multikollinearitet.
(iv) Om residualerna beter sig i strid mot modellen, genom att deras observerade varianser varierar beroende på *x*-variablerna (s.k. heteroskedasticitet), eller genom att de är beroende sinsemellan.

För förklaringar hänvisas till förslagsvis Ryan (1997), Atkinson & Riani (2000), Welsley & Kuh (1980), och även Kennedy (2008) och SAS (2008). Här ska bara påminnas om ett par särskilt centrala grepp.

Förklaringsgraden, ofta betecknad R^2 eller R-kvadrat och även benämnd determinationskoefficienten, är ett mått på hur nära modellen följer de observerade utfallen, i t.ex. modellen (1). Detta mått är ett tal mellan 0 och 1, och det anger kvoten mellan de predikterade och de observerade *y*-värdenas varianser. Ett värde nära 1 betyder att modellen nära fullständigt förklarar variationerna i y , medan ett värde nära 0 betyder att modellen nästan inte alls förklarar något av variationerna i y .

Värdena på R-kvadrat fås i körningarna i två varianter, nämligen ojusterad och justerad (*adjusted*), där den senare är lägre. Det är i regel lämpligt att i första hand titta på det justerade talet. Justeringen innebär kort uttryckt att talen görs lite bättre jämförbara mellan olika antal observationer och olika antal variabler.

Man ska inte direkt se R-kvadrat som ett mått på analysens ”kvalitet”. Det kan t.ex. ibland vara mest relevant att göra separata analyser på delar av ett material, även om R-kvadrat då blir lägre än i en analys på hela materialet samtidigt. (T.ex. sett över många decennier kan det lätt uppstå starka nonsensartade samband, som mellan storkpopulation och nativitet, genom att mycket hinner bli radikalt annorlunda under en så lång tid.)

R-kvadrat har inte någon direkt mening för t.ex. logitmodellen (4), även om vissa mått med analog tolkning har föreslagits där (jfr Ryan 1997; SAS 2008).

Residualernas observerade värden för observationerna kan vara nyttiga att titta på, gärna i form av plottdiagram (spridningsdiagram); se t.ex. Goodall (1983). Residualerna bör ligga någorlunda väl samlade symmetriskt kring nollnivån. Om man får ett annat mönster så är det att se som varningstecken, att modellen kanske inte passar verkligheten så bra och analysen kan vara missvisande.

5 Sätt att hantera analysproblem på statistiska data

Partiellt bortfall del I – saknade värden på utfallsvariabeln

Vanlig regressionsanalys förutsätter egentligen att alla observationer har värden på alla variablerna i modellen, alltså på både den beroende variabeln och de förklarande variablerna. Detta passar inte så bra med tanke på att vi i praktiken oftast har partiellt bortfall i statistiska observationsmaterial. Det vill säga, i regel gäller att inte alla observationer i ett material har värden på alla variablerna i modellen.

Man behöver alltså något sätt att hantera det partiella bortfallet. Här är det skäl att skilja mellan olika slags variabler.

Objekt som saknar värde på utfallsvariabeln är det troligen oftast relevant att behandla som objektbortfall och normalt utesluta ur analysen. Detsamma gäller objekt som saknar värde på den eventuella förklarande variabel som är primärt mest intressant. När man tolkar resultaten av analysen får man reservera sig för en möjlig snedvridande selektionseffekt genom dessa uteslutningar. En mera avancerad och svårare alternativ ansats att hantera objektbortfallet i regressionsanalys är att modellera bortfallsorsakerna i själva analysmodellen, enligt Heckman (1979); jfr Kennedy (2008, Sect. 17.3).

Partiellt bortfall del II – saknade värden på förklarande variabler

För saknade värden på förklarande variabler i allmänhet kan olika ansatser tänkas. Potentiellt möjliga alternativ är följande, sammanfattat kommenterade:

(i) Utesluta alla ofullständiga observationer
- *Kan snedvrida, undvik helst*
- *Tänkbart vid ringa partiellt bortfall*

(ii) Imputera med t.ex. medelvärde
- *Blir lätt missvisande, undvik som standardgrepp*

(iii) Imputera med typvärde
- *Relevant när ett enstaka värde dominerar starkt*
- *Kan förutsätta viss eftertanke i läsning av resultaten*

(iv) Sätta indikatorer (dummyvariabler) för saknade värden
- *Informativt, transparent, användbart*
- *Kan förutsätta viss eftertanke i läsning av resultaten*

(v) Imputera modellbaserat
- *Avancerad ansats, kan i princip fungera utmärkt*
- *Extra modellberoende; svårgenomskådade förutsättningar*

(vi) Använda analysberäkning som delvis tål partiellt bortfall
- *Kan vara tänkbar utväg vid stora modeller*
- *Svårgenomskådade förutsättningar; robusthetsproblem*

Imputera som nämns här betyder att ersätta ett saknat variabelvärde med ett antaget värde, tillfälligt för analyskörningen enligt någon regel.

De nämnda alternativen beskrivs något närmare i de följande delavsnitten.

(i) Alternativet: Utesluta alla ofullständiga observationer
Ett slags nollalternativ är att köra analysprogrammet direkt på det givna materialet, utan att göra något åt saknade variabelvärden. Vad som händer om man gör så är normalt att analysen utesluter alla ofullständiga observationer, och den körs alltså på enbart de objekt som har värden på alla variablerna i modellen.

Det alternativet kan i regel vara mindre lämpligt. Det kan nämligen lätt bli en snedvridande selektering som gör resultaten missvisande, och dessutom blir det större slumposäkerhet genom det minskade underlaget. Om modellen har fler än mycket få förklarande variabler, så kan det lätt bli en avsevärd andel av observationerna som saknar värde på åtminstone någon variabel. Alla de observationerna skulle behöva kasseras helt ur analysen, med den nämnda ansatsen.

Detta alternativ kan dock eventuellt vara tänkbart om partiellt bortfall är sällsynt i materialet, så att det bara är undantagsvis som någon observation inte har värden på alla variabler. Men även då kan det vara skäl att vara försiktig och tänka efter, om det kan vara något särskilt med de ofullständiga observationerna och i så fall vad, och om de principiellt sett är relevanta eller inte för den rätta bilden.

(ii) Alternativet: Imputera med t.ex. medelvärde
Att imputera, dvs. ersätta, saknade värden med medelvärdet på variabeln kan möjligen tyckas ligga till hands, men allmänt är det skäl att varna för detta alternativ. Medelvärdet behöver nämligen inte vara relevant för de objekt som saknar värde på variabeln, och imputering med medelvärdet kan då störa analysen mer eller mindre starkt och göra den missvisande. Undantag kan naturligtvis tänkas där specifika förhållanden råder.

(iii) Alternativet: Imputera med typvärde
Att imputera (ersätta) saknade värden med typvärdet kan däremot vara vettigt när typvärdet är starkt dominerande, så att en relativ stor andel av objekten ligger på typvärdet för variabeln.

Ett viktigt specialfall är när variabeln är en dummyvariabel som indikerar en mindre vanlig egenskap, t.ex. att en individ har en viss ovanlig sjukdom. Då innebär detta alternativ att objekt utan uppgift om egenskapen betraktas som om de inte har egenskapen.

Denna utväg har fördelen att den är ganska transparent, i att det är lätt att fatta vad som är gjort. När man läser resultaten får man dock tänka på att variabeln har getts en annan innebörd än den ursprungliga. Variabeln visar nu inte om objektet har egenskapen i fråga, utan om objektet är känt för att ha egenskapen.

(iv) Alternativet: Sätta indikatorer (dummyvariabler) för saknade värden
Ett ofta användbart alternativ är att behandla saknat värde som en egen svarskategori. För en given förklarande variabel inför man då en kompletterande dummyvariabel, som anger när värde saknas på den givna variabeln.

Om den givna variabeln är en kategorivariabel och i sin tur består av en eller flera dummyvariabler för svarskategorier, så är gången denna: När värde saknas så sätts värdet till 0 (dvs. ”tillhör inte kategorin”) på de givna dummyvariablerna och till 1 på den bortfallsindikerande dummyvariabeln.

Greppet är lämpligt även för kvantitativa variabler. För objekt som saknar värde på den givna kvantitativa variabeln så sätts denna till ett konstant värde (i princip godtyckligt valt men gemensamt), samtidigt som den bortfallsindikerande dummyvariabeln sätts till 1.

Detta är ett transparent och rättframt angreppssätt. Man visar fram det partiella bortfallet som det är och hur det inverkar. Samtidigt förutsätts en viss medvetenhet hos dem som tar del av resultaten.

(v) Alternativet: Imputera modellbaserat
Finns det tillgång till bra välutvecklade modeller för att imputera saknade värden så kan de naturligtvis i princip utnyttjas. Man kör då analysen med imputerade värden insatta i stället för observerade värden, på objekt där sådana saknas på respektive variabel.

Detta bör i princip kunna fungera bra, men samtidigt blir analysen väsentligt mera komplex. Det gör att det kan bli mera problematiskt att hålla reda på om förutsättningarna och modellantagandena är tillräckligt väl uppfyllda. Ansatsen förutsätter därför strängt taget kvalificerad metodinsikt för att ge fullt tillförlitliga resultat.

Exempelvis skulle det kunna hända att imputeringen i sin tur bygger på en regressionsmodell, som kan ha variabler gemensamma med den analys man vill göra. De inblandade modellerna kan då riskera att störa varandra på ett sätt som kan vara svårt att genomskåda.

Ett annat problem är att imputerade värden naturligen kan få mindre variabilitet än observerade. Det kan ge risk för att osäkerhetstal blir för optimistiskt små och att signifikanstester får för optimistiskt låga risknivåer (p-värden). En modern teknik för att möta detta problem är s.k. multipel imputering; se t.ex. Heeringa m.fl. (2010, Ch. 11). Denna teknik gör dock också situationen mera komplex, om den används för underlag till sambandsanalys. Vidare har den kanske ännu inte i större omfattning avpassats till och demonstrerats på data från statistiska undersökningar.

(vi) Alternativet: Använda analysberäkning som delvis tål partiellt bortfall
I stället för att imputera saknade värden kan man ibland ha möjlighet att modifiera själva analysberäkningen, så att den blir mindre krävande på observationernas fullständighet.

Exempelvis vid skattning av modellen (1) behöver man inte nödvändigtvis beräkna regressionskoefficienterna direkt från mikrodata, dvs. data för de enskilda observationerna. En alternativ beräkningsgång är att först beräkna aggregerade tal i form av kovariansmatrisen för modellens variabler, utifrån mikrodata, och sedan därifrån beräkna regressionskoefficienterna.

Detta öppnar följande möjlighet. Kovarianserna (och varianserna) i kovariansmatrisen kan var för sig beräknas utifrån observerade värden på enbart de två variabler (eller den enda variabel) som talet avser. Beräkningen av kovariansmatrisen skulle därför i princip kunna göras med utnyttjande även av objekt som inte har observerade värden på alla variablerna. Sedan skulle man beräkna regressionskoefficienterna från kovariansmatrisen på samma sätt som nyss antyddes.

Förfarandet skulle ha fördelen att utnyttja även ofullständiga observationer, utan att någon imputering behöver göras. Men förfarandet är knappast oproblematiskt, för risk kan finnas att kovariansmatrisen inte får de rätta egenskaperna, när den styckevis grundas på varierande delar av underlaget. Möjligen kan någon kompletterande justeringsmetod behövas.

Idén att modifiera själva analysberäkningarna för att bättre tåla partiellt bortfall skulle allmänt behöva utredas med kvalificerad metodkompetens, för det slags datamaterial som det är fråga om, innan den tillämpas där.

Hänsyn till urvalsdesign del I – viktning

I datamaterial från statistiska undersökningar kan urvalsdesignen göra att förutsättningarna för en regressionsmodell inte uppfylls utan vidare. Särskilt är frågan hur man ska göra med vikter.

I material från statistiska undersökningar är observationerna ofta försedda med vikter som varierar mellan observationerna och som används i statistikberäkningen. Vikterna är till för att kompensera för sådant som att urvalet kan vara draget med olika urvalssannolikheter för olika objekt, och att svarsbortfallet kan variera mellan grupper. (För förklaringar se t.ex. S.Lohr 2010; Särndal m.fl., 1992.)

Regressionsanalys och statistikprogramvaror kan i regel inte utan vidare använda vikterna på rätt sätt. För att komma rätt med regressionsanalysen på materialen i fråga behöver man ett lämpligt sätt att förhålla sig till vikterna. Möjliga alternativ är följande, sammanfattat kommenterade:

(i) Analysera oviktat
- *Ofta väl användbart, teoretiskt försvarligt*
- *Rättvisande under vissa naturliga förutsättningar*
- *Inte konsistent med viktade estimatorer (skattningar)*
- *Bör hålla slumposäkerheten relativt låg*
- *Någon viktberäkning stör inte analysen, som hålls ”ren”*

(ii) Vikta enkelt i analysen
- *Rättvisande punktskattningar av regressionskoefficienterna*
- *Men ger inte användbara osäkerhetstal eller tester*
- *Konsistent med deskriptiv statistik (viktade estimatorer)*
- *Viktningen kan ev. dra upp slumposäkerheten*
- *Viktberäkningen kan ev. störa analysen*

(iii) Kombinera oviktad och enkelt viktad analys
- *Viktat för punktskattning och oviktat justerat för osäkerhetstal*
- *Kan ibland vara tänkbart under viss försiktighet*
- *Konsistent med deskriptiv statistik (viktade estimatorer)*
- *Problematiskt med underlag för justeringen*
- *Viktningen kan ev. dra upp slumposäkerheten*
- *Viktberäkningen kan ev. störa analysen*

(iv) Analysera med metod som beaktar urvalsdesign
- *Avancerad ansats, teoretiskt ”rätt” i viss mening*
- *Beaktar både vägning och andra designaspekter, t.ex. kluster*
- *Begränsad tillgång till verktyg*
- *Konsistent med deskriptiv statistik (viktade estimatorer)*
- *Viktningen kan ev. dra upp slumposäkerheten*
- *Viktberäkningen kan ev. störa analysen*

(v) Utvärdera genom parallellkörning av alternativa ansatser
- *Lämpligt för metodstudier*
- *Inte så lämpligt i ”skarpt läge” för resultat som ska användas*
- *Tolkningen kan vara problematisk*

Dessa alternativ beskrivs något närmare i de följande delavsnitten.

(i) Alternativet: Analysera oviktat
Ofta kan det faktiskt vara lämpligt att köra analysen oviktat (ovägt), dvs. utan viktning av observationerna. Detta trots att materialet är försett med vikter som är avsedda att användas när man räknar statistik på det.

Ett gott teoretiskt stöd för att köra oviktat finns i den klassiska Gauss-Markovs sats (se t.ex. Wilks, 1962, Sect. 10.3). Denna sats medför bland annat att regressionskoefficienterna skattas väntevärdesriktigt vid oviktad analys, under förutsättningen att modellen beskriver verkligheten korrekt. Denna förutsättning är i princip aldrig exakt uppfylld, men villkoret kan göras svagare och ändå lämna slutsatsen väsentligen giltig.

Nordberg (1989) visar nämligen att skattningarna blir väntevärdesriktiga även under vissa svagare förutsättningar. Dessa är väsentligen att residualerna ska uppfylla en form av oberoende gentemot designvariablerna (vanligen stratifieringsvariablerna) som vikterna beror av. Resultatet är dock tämligen generellt och gäller även på ett motsvarande sätt för modellerna av formerna (4)–(6) som inte uttryckligen innehåller residualer.

I praktiken betyder detta att analysen normalt bör kunna köras oviktat utan att det gör den mindre korrekt, ifall man har med designvariablerna som förklarande variabler i modellen. Och det har man kanske oftast anledning till ändå, av saklogiska skäl.

En väsentlig fördel med oviktad analys är att analysen får en ren form och inte svårgenomskådat störs av metoder för viktberäkning. Att köra oviktat kan också väntas ha viss fördel i fråga om precisionen, genom att då inte variationer i vikter bidrar till att dra upp osäkerheten.

Något som kan vara en nackdel är att resultaten av oviktad analys inte blir helt konsistenta med statistik som är beräknad med hjälp av vikterna. Logiskt borde t.ex. medelvärdena av de observerade och de predikterade *y*-värdena i modell (1) stämma överens, men det kan de inte väntas göra om det ena medelvärdet är räknat viktat och det andra oviktat.

(ii) Alternativet: Vikta enkelt i analysen
Programvaror för analys har ofta en möjlighet att beakta vikter på observationerna, men i en annan mening än vikter i statistiska undersökningar.

Denna viktningsmöjlighet är bara delvis användbar med vikter på observationer från statistiska undersökningar. Den bör i princip ge rätt punktskattningar på regressionskoefficienterna. Däremot ger den ingen användbar information om osäkerhetstal och signifikanstester, utan de siffror som körningen ger om detta är helt fel.

Att köra viktat på detta enkla sätt kan alltså vara gångbart om man bara är ute efter punktskattningar, inte efter osäkerhetstal eller tester. Variationer i vikterna kan dock tendera att ge något sämre precision än oviktad analys. Å andra sidan kan det vara en fördel att resultaten vid viktad analys blir bättre konsistenta med annan statistik som är räknad med vikterna.

Är man intresserad bara av punktskattningarna är alltså både oviktad och enkelt viktad analys möjliga alternativ. Vilket som är lämpligast av dessa två alternativ beror på hur de nämnda för- och nackdelarna ställer sig för det aktuella syftet.

(iii) Alternativet: Kombinera oviktad och enkelt viktad analys
En möjlighet kan vara att kombinera enkelt viktad och oviktad analys. Man kan då köra viktad analys för punktskattningar av regressionskoefficienterna, och oviktad för dessa skattningars osäkerhetstal.

Man måste då på något sätt justera osäkerhetstalen så att de blir rättvisande. Detta behövs för att osäkerhetstalen är räknade oviktat men ska användas för skattningar som är räknade viktat. Om osäkerhetstalen inte behöver uppfylla högre krav på tillförlitlighet utan mest är till för orientering kan möjligen denna justering göras något schablonmässigt. En ansats kan vara att multiplicera osäkerhetstalet med en faktor (”designeffekt”), troligen något över 1, som är grundad på tidigare erfarenhet från specifikt liknande situationer.

(iv) Alternativet: Analysera med metod som beaktar urvalsdesign
Numera finns dock även teori och programvara för att göra analyser ”surveyanpassat”, med korrekt beaktande av vikter på observationer från statistiska undersökningar. Förklaringar ges av Lohr (2010, Ch. 11); Chambers & Skinner (2003); och Heeringa m.fl. (2010). Exempel på tillgängliga verktyg av detta slag är procedurerna SURVEYREG, SURVEYLOGISTIC och SURVEYPHREG i programsystemet SAS/STAT; se SAS (2008).

Denna ansats är naturligtvis allmänt sett tilltalande. Den bygger på moderna verktyg för att göra analysen på ett sätt som kan sägas vara rätt för data från statistiska undersökningar. Detta genom att vikterna beaktas på det sätt som de är avsedda för. Analysresultaten kan då också bli väl konsistenta med andra statistikresultat från undersökningarna i fråga, om man där har beaktat vikterna på det avsedda sättet. Den här typen av verktyg kan beakta inte bara vikterna utan även andra egenskaper hos data från statistiska undersökningar, såsom beroenden mellan observationer genom klusterurval.

Praktiska begränsningar ligger i att sådana surveyanpassade analysverktyg inte finns utvecklade för ett lika rikhaltigt spektrum av varianter på analysmetoder och verktygsmiljöer som är fallet med andra analysverktyg.

Det är också skäl att tänka igenom syfte och förutsättningar ordentligt, innan man väljer att använda sådana surveyanpassade verktyg för en analysuppgift. I det fördolda kan situationen bli ganska komplex, genom att bl.a. metoden att räkna fram vikterna kommer in som en del av analysen. Analysen blir mindre ren, mindre transparent än när man kör oviktat.

Om delvis samma variabler används i viktberäkningen och i analysen, så kan det finnas risk att vikterna påverkar analysresultaten på ett svårgenomskådat sätt. De surveyanpassade analysmetoderna förutsätter också att vissa antaganden är uppfyllda, och de gör knappast att antagandena i själva regressionsmodellen blir mindre kritiska.

Det är därför skäl att noga ta del av undersökningens dokumentation och sätta sig in i hur vikterna är beräknade, innan man använder dem i analysen. Ansatsen förutsätter strängt taget kvalificerad metodinsikt för att ge fullt tillförlitliga resultat.

Hur de beräknade och faktiska osäkerhetstalen kan påverkas är knappast helt oproblematiskt. I princip ska osäkerhetstalen beräknas korrekt med de surveyanpassade verktygen, men om detta fungerar fullt ut kan också hänga på om viktberäkningen i undersökningen svarar mot verktygens förutsättningar.

(v) Alternativet: Utvärdera genom parallellkörning av alternativa ansatser
I metodstudiesyfte kan man naturligtvis göra känslighetsanalyser genom att jämföra olika analysansatsers utfall på observationsmaterial av intresse. Man kan t.ex. jämföra regressionskoefficienterna och deras osäkerhetstal mellan olika skattningar, såsom dels oviktat, dels med korrekt beaktande av vikterna genom surveyanpassade analysverktyg.

När man planerar en sådan metodstudie är det lämpligt att tänka igenom vad man ska titta efter och hur tänkbara utfall kan tolkas. När två jämförda metoder ger i stort sett nära samstämmiga resultat så kan det sägas ge stöd för båda metodernas tillförlitlighet. Värre är det i det motsatta fallet, att metoderna ger märkbart skiljaktiga resultat. Då kan man normalt inte ta för givet vilket av resultaten som är mera rätt, utan man får då i nästa steg gräva vidare och försöka hitta evidens för vad som kan ligga bakom skillnaden.

För diagnostik kan det vidare vara bra att titta på hur starkt vikterna varierar mellan objekten. Om vikterna varierar relativt måttligt så bör man knappast vänta sig någon större skillnad i utfall mellan oviktad och viktad analys. Om det ändå blir en stor skillnad där trots ”lugna” vikter, så kan det vara skäl att se efter om t.ex. problem med utstickande värden i datamaterialet kan ligga bakom.

Känslighetsanalyser med jämförelser av metoder lämpar sig i första hand bara för särskilda metodstudier, och inte gärna för eventuellt metodbyte under produktion i ”skarpt läge”. I produktionen ska man normalt ha utformat metoderna innan och sedan hålla fast vid dem, så att det inte kan finnas risk för godtycke där det aktuella utfallet kan påverka metodvalet.

Hänsyn till urvalsdesign del II – beroende mellan observationer

Designen i statistiska undersökningar kan medföra även andra avvikelser än de redan behandlade, från förutsättningarna för vanliga regressionsmodeller. Förutom att vikterna kan variera så kan det även förekomma beroende mellan observationer i data från statistiska undersökningar. De

vanliga regressionsmodellerna förutsätter däremot, som tidigare sagts, att observationerna är sinsemellan oberoende.

Beroende mellan observationer kan till stor del undgås i data från svenska statistiska undersökningar, där man ofta kan dra urvalen ur bra ramar i form av register.

I undersökningar med klusterurval blir det ett beroende mellan observationerna, genom att urvalssteget med dragningen av klustren berör alla observationerna i respektive kluster. Ett exempel är en undersökning på elever i skolor, där man först drar ett urval av skolor och sedan drar elever i de dragna skolorna.

Sådana situationer kan hanteras på lite olika sätt, beroende på syftet med analysen. Ett sätt är att använda surveyanpassade verktyg som redan har nämnts. Ett annat sätt är att använda en analysmodell för hierarkiska data (flernivåmodell), med ett verktyg som kan skatta en sådan modell. Det senare kan vara lämpligt om man i analysen är intresserad av att särskilt kunna mäta effekter på klusternivå. Här ska inte gås närmare in på detta.

Faror med designvariabler

Problem med antagandena bakom regressionsmodellen (jfr avsnitt 4 ovan) kan vidare uppstå genom valet av designvariabler. Om utfallsvariabeln *y* skulle finnas med bland designvariablerna (stratifieringsvariablerna) i undersökningen så kan modellantagandena störas så att analysen blir missvisande. Då skulle nämligen sannolikhetsfördelningen för *y* kunna variera i strid med modellen och få ett konstlat metodberoende, så att analysresultaten inte kan antas spegla verkligheten. Samma sak kan gälla om utfallsvariabeln skulle vara direkt relaterad till någon designvariabel i undersökningen.

Det nämnda problemet bör inte kunna uppstå om designvariablerna finns med som förklarande variabler i analysmodellen. Förmodligen finns kanske i regel också naturliga saklogiska relevansskäl att ändå ha med variablerna i fråga som förklarande i modellen.

I vissa typer av studier kan man dock behöva arbeta med analyser som har en designvariabel som utfallsvariabel, och inte förklarande variabel, såsom s.k. fall-kontroll-studier. På vissa villkor kan regressionsmodeller vara användbara även i sådana situationer, men vissa korrigeringar kan behövas (se t.ex. Schlesselman 1982, Sect. 8.2; Kennedy 2008, Ch. 17).

Modell- och urvalsperspektiv

Begreppsmässigt kan man säga att slump och osäkerhet egentligen kommer in på fundamentalt olika sätt i regressionsmodeller jämfört med data från urvalsundersökningar (jfr Lohr 2010,Sect. 11.4):

- I regressionsmodellen ses utfallsvariabelns värde som slumpmässigt, givet de oberoende variablernas värden vilka ses som fixa. (Jfr förklaringarna efter formel (1).)
- I urvalsmaterialet ses urvalsdragningen som slumpmässig, medan variabelvärdena ses som fixa för populationens objekt.

Denna distinktion är det bra att vara medveten om. När man arbetar med analysmodeller behöver man i regel osäkerhetsmått som innefattar osäkerhet i själva modellen. Sådan osäkerhet råder även när analysen görs på hela populationen och inte bara på ett urval. Det innebär ett annat synsätt än det brukliga för deskriptiv statistik på urval, där det redovisas mått på den osäkerhet som kommer sig av urvalsdragningen och inte har någon motsvarighet i totalräknad statistik.

6 Referenser

Not: Listan över referenser får inte ses som någon generell rekommendation av metoderna i dem utan enbart som möjligheter för läsaren att få förklaringar enligt hänvisningarna i texten.

F. Andersson & E. Hagsten (2011), Konjunkturberoende i inflödet till och utflödet från högre studier, Arbetsmarknads- och utbildningsstatistik, Bakgrundsfakta 2011:7, SCB.

G. Andersson, U. Jorner & A. Ågren (2007), Regressions- och Tidsserieanalys, Tredje Upplagan, Lund: Studentlitteratur.

A. Atkinson & M. Riani (2000), Robust Diagnostic Regression Analysis, New York: Springer.

D.A. Belsley, E. Kuh & R.E. Welsch (1980), Regression Diagnostics, New York: Wiley.

R. Bender & A. Benner (2000), Calculating ordinal regression models in SAS and S-Plus, Biometrical Journal 42, 677-699.

K.A. Bollen (1989), Structural Equations with Latent Variables, New York: Wiley.

R.L. Chambers & C.J. Skinner (2003), Analysis of Survey Data, Hoboken: Wiley.

H. Cramér (1946), Mathematical Methods of Statistics, Princeton: Princeton University Press.

J.D. Emerson & M.A. Stoto (1983), Transforming data, in Understanding Robust and Exploratory Data Analysis, Ed. by D.C. Hoaglin et al., New York: Wiley (reprinted 2000), Ch. 4, pp. 97-128.

C. Fornell & F.L. Bookstein (1982), Two structural equation models: LISREL and PLS applied to consumer exit-voice theory, Journal of Marketing research, 19, 440-452.

C. Goodall (1983), Examining residuals, in Understanding Robust and Exploratory Data Analysis, Ed. by D.C. Hoaglin et al., New York: Wiley (reprinted 2000), Ch. 7, pp. 211-246.

J.J. Heckman (1979), Sample selection bias as a specification error, Econometrica, 47, 153-161.

S.G. Heeringa, B.T. West & P.A. Berglund (2010), Applied Survey Data Analysis, Boca Raton: Taylor and Francis.

D.A. Hensher & W.H. Greene (2001), The Mixed Logit Model: The state of practice and warnings for the unwary, http://people.stern.nyu.edu/wgreene/MixedLogitSOP.pdf

P. Kennedy (2008), A Guide to Econometrics, Sixth Edition, Malden: Blackwell.

Y. Lee, J.A. Nelder & Y. Pawitan (2006), Generalized Linear Models with Random Effects, Boca Raton: Chapman & Hall/CRC.

S. Lohr (2010), Sampling: Design and Analysis, Boston: Brooks/Cole.

D. McFadden, (2001), Disaggregate behavioral travel demand's RUM side, in Hensher, D.A. (ed.) Travel Behaviour Research: The Leading Edge, Pergamon Press, Oxford, pp.17-64.

L. Nordberg (1989), Generalized linear modeling of sample survey data, Journal of Official Statistics, 5, 223-239.

Overdispersion (2011), www.wikipedia.org , English.

M. Ribe (1997), Analysmodeller på olika vis, Statistikskolan [ur VälfärdsBulletinen nr 1, 1997], SCB, www.scb.se

M. Ribe (1999), Oddskvoter berättar, Statistikskolan [ur VälfärdsBulletinen nr 4, 1999], SCB, www.scb.se

E. Ruist (1990), Modellbygge för Empirisk Analys, Lund: Studentlitteratur.

T.P. Ryan (1997), Modern Regression Methods, New York: Wiley.

J.J. Schlesselman (1982), Case-Control Studies, New York: Oxford University Press.

C.-E. Särndal, B. Swensson & J. Wretman (1992), Model Assisted Survey Sampling, New York: Springer.

SAS (2008), SAS/STAT® 9.2 User's Guide, Cary NC: SAS Institute Inc.

Svenska statistikfrämjandet (2010), Svenska statistikfrämjandets etiska kod för statistiker och statistisk verksamhet, http://statistikframjandet.se/wp-content/uploads/2010/12/etisk_kod_final.pdf

J. Tobin (1958), Estimation of relationships for limited dependent variables, Econometrica, 26, 24-36.

G. Walter & T. Augustin (2009), Bayesian Linear Regression – Different conjugate models and their (in)sensitivity to prior-data conflict, Technical Report 069, Department of Statistics, University of Munich, http://epub.ub.uni-muenchen.de/11050/1/tr069.pdf

S.S. Wilks (1962), Mathematical Statistics, New York: Wiley.

Not: Heckman, McFadden och Tobin erhöll Priset i Ekonomi till Nobels minne.