Innehållsförteckning

1 Inledning

Vid analys av surveydata görs en distinktion mellan deskriptiva och analytiska urvalsundersökningar, här benämnt som survey. Kortfattat kan deskriptiva surveys sägas syfta till att besvara frågor om hur en begränsad population ser ut genom skattning av målparametrar såsom medelvärden och andelar, och analytiska syfta surveys till att besvara frågan varför det ser ut som det gör i en ”oändlig” population genom skattning av målparametrar såsom regressionkoefficienter.

Då logistisk regression är en metod som kan används för att undersöka sambandet mellan en beroende kategorisk variabel och en eller flera oberoende variabler, med målet att prediktera, klassificera eller att fastställa associationer, är metoden således tillämplig på analytiska survey. Logistisk regression är en populär metod eftersom den ej innehåller något antagande om normalfördelning. Den är också flexibel eftersom de oberoende variablerna kan vara såväl diskreta som kontinuerliga. Metoden används ofta i kliniska och epidemiologiska studier, såsom populationsbaserade hälsoundersökningar, där data kan ha inhämtats från en survey.

Korn och Graubard (1999) beskriver ett antal faktorer vid denna typ av studier som är mer vanliga med surveydata än vid icke-surveydata. Urvalsstorlekarna är ofta väldigt stora då data härrör från observationer snarare än experiment. Bortfall är också vanligt, vilket alltid måste hanteras. Ofta används också förfinade urval i syfte att representera komplexa underliggande populationsstrukturer. Ett exempel är klusterurval, vilket dock kan ge upphov till korrelation mellan observationer och härigenom riskerar att underskatta varianser.

Stratifierade urval är också vanligt, till exempel då intresse finns för att kunna undersöka olika delpopulationer. Populationen delas då upp i olika strata, varigenom sannolikheten att komma med i urvalet blir lika inom ett stratum men kan skilja sig mellan strata. Om samband mellan variabler skiljer sig mellan strata skulle detta dock kunna påverka slutsatserna av en logistisk regression. Ofta antas att stratifiering inte har någon påverkan, även om det är möjligt att ta hänsyn till den. Inom survey finns designbaserat och modellbaserat synsätt, vilka i sina rena former har olika syn på hur stratifieringen ska tas hänsyn till. Valet av synsätt kan få konsekvenser för slutsatserna, även om synsätten i praktiken kan ses som komplement.

1.1 Syfte

Syftet är att ge en allmän beskrivning av binär logistisk regression samt att beskriva tillämpning av metoden på stratifierade surveydata.

1.2 Metod

En litteraturstudie genomförs inom aktuella områden. Vidare illustreras binär logistisk regression med studien Liv & Hälsa (2000).

1.3 Avgränsning

I denna uppsats avgränsas till binär logistisk regression. Övriga avgränsningar ges i texten.

1.4 Disposition

Kapitel 2 beskriver surveydata i allmänhet och stratifiering i synnerhet. I kapitel 3 introduceras binär logistisk regression i allmänhet, samt de förutsättningar som gäller vid analys utifrån modell- respektive designbaserat synsätt då urvalet är stratifierat. En illustration ges sedan i kapitel 4 av logistisk regression utifrån de två synsätten. I kapitel 5 sammanfattas och dras slutsatser, samt ges synpunkter på framtida forskning.

2 Survey och analys av surveydata

I detta kapitel ges en genomgång av survey och analys av surveydata. I avsnitt 2.1 ges en introduktion till och syfte med survey, liksom vissa begrepp. Avsnitt 2.2 fokuserar sedan på definition av populationen och målparametrar. I avsnitt 2.3 berörs urvalsmekanismer, designvikt och hantering av bortfall. Inferens från surveydata tas sedan upp i avsnitt 2.4, och i 2.5 presenteras estimation. En introduktion till algebra med matriser och vektorer ges i bilaga 1. I uppsatsen används i regel skalär algebra (gemener), men matrisalgebra (versaler) förekommer vid vissa multivariata modeller, se bilaga 1 för en introduktion till matrisalgebra.

2.1 Introduktion till och syfte med survey

Biemer och Lyberg (2003) presenterar sju krav som Dahlenius (1985) anser måste vara uppfyllda för att en studie ska kallas för en survey.

1. *A survey concerns a set of objects comprising a population.*
2. *The population under study has one or more measurable properties.*
3. *The goal of the project is to describe the population by one or more parameters defined in terms of the measurable properties.*
4. *To get observational access to the population, a frame is needed (i.e., an operational representation of the population units, such as a list of all objects in the population under study or a map of a geographical area.)*
5. *A sample of objects is selected from the frame in accordance with a sampling design that specifies a probability mechanism and a sample size.*
6. *Observations are made on the sample in accordance with a measurement process (i.e., a measurement method and a prescription as to its use).*
7. *Based on the measurements, an estimation process is applied to compute estimates of the parameters when making inference from the sample to the population.*

Från tabell 1.1 i Biemer och Lyberg (2003), sidan 4, vilken är baserad på Dahlenius (1985).

Surveyprocessen kan således sägas bestå av att bestämma syfte och målpopulation, samt välja datainsamlingsmetod och urvalsdesign. Urvalsdesignen beskriver här urvalsramen, den använda metoden för urvalsdragning vilken även benämns som urvalsmekanism, samt urvalsstorleken. En urvalsmekanism med planerad slump kallas sannolikhetsurval. De objekt som ingår i en population kallas för enheter.

Målet med en survey är därefter att kunna dra inferens om målpopulationen. Inferens beskrivs av Särndal (1985, sidan 50) som "*a statement made about an unknown population quantity, in terms not of fully certainty but of probability.*". För att minska osäkerhet i inferens bör skattning av målparametrarna ske med effektiva estimatorer utan bias. Kravet på avsaknad av bias ersätts ibland med det svagare kravet konsistens, eftersom kravet kan bli för strikt vid icke-linjära estimatorer såsom varianser. Konsistens innebär avsaknad av bias i stora urval, och brukar betraktas som ett nödvändigt villkor för en estimator. Ett vanligt kriterium är därefter att välja den estimator som har minsta kvadrerade felet (MSE; eng. *mean squared error*). MSE kombinerar egenskaperna effektivitet och bias genom att addera varians och kvadrerad bias för en estimator: $MSE\left(\hat{\theta}\right) = E\left(\hat{\theta} - \theta\right)^2 = Var\left(\hat{\theta}\right) + \left[B\left(\hat{\theta}\right)\right]^2$.

Ett delmål med en survey är att mäta värdena på undersökningsvariablerna för alla enheter som ingår i urvalet. Ibland används begreppet observationsprocess för den process varigenom värdena på undersökningsvariablerna transformeras till ett dataset. Detta inkluderar således urvalsmekanismen såväl som bortfalls- och mätprocessen.

I denna uppsats avgränsas till surveys där varje enhet i urvalet endast observeras vid ett tillfälle, en så kallad tvärsnittsstudie. Denna typ av studie syftar normalt till att beskriva undersökningsvariablerna inom populationen, eller att studerar associationer mellan undersökningsvariablerna, och benämns därför som deskriptiva respektive analytiska survey. Skinner, Holt och Smith (1989) framhåller J. Neyman som den som lagt grunden till teorin för deskriptiva survey, och P. F. Lazarsfield till analytiska survey. På sidan 1 beskriver de vidare syftet med deskriptiva survey som *"directed at estimation of summary measures of the population, such as means and frequencies."*, medan Deming (1950, sidan 249), citerad i Kalton (2002), beskriver syftet med analytisk survey som *"directed at the underlying causes that have made the frequencies of various classes of the population what they are, and will govern the frequencies of these classes in time to come"*. Analytiska survey kan därför sägas syfta till att gå bakom de beskrivande måtten för att söka förklara samband. Kortfattat kan således ett deskriptivt syfte sägas vara att besvara frågan *hur många?* och ett analytiskt syfte att besvara frågan *varför?*.

Kendall and Lazarsfield (1950) beskriver kausala samband mellan händelserna A och B som att *"A causes B if (a) A precedes B in time, (b) variations in A have corresponding variations in B, (c) other variables fail to account for the association between A and B*". Det är således inte möjligt i en tvärsnittsstudie att påvisa huruvida associationer är kausala eller ej då mätning endast sker vid ett tillfälle. Associationer kan dock användas som hypotetiska samband mellan orsak och verkan. Variabler tänkta som orsaks- och verkan benämns normalt som oberoende och beroende variabler. Diskussionen om huruvida det över huvud taget är möjligt att påvisa kausalitet i en studie lämnas därvid.

Surveys används inom flera olika områden. Inom hälsoområdet kan survey användas för att kartlägga sjukdomar och hälsorelaterade variabler i stora populationer. Exemplen i denna uppsats berör huvudsakligen denna typ av surveys. Detta motsätter naturligtvis inte möjligheten att generalisera slutsatser även till andra områden.

Korn och Graubard (1999) framhåller ett flertal fördelar med hälsosurveys i jämförelse med andra typer av observationsdata. Ofta kan målpopulationen lättare definieras än i typiska epidemiologiska och kliniska studier. Härigenom uppstår mindre problem med vem frågorna berör, samt att vissa typer av bias kan minimeras. Trots den normalt sett stora variabiliteten i hälsodata är det ändå ofta möjligt att fastställa vetenskapligt meningsfulla effekter, tack vare att stora urval ofta används. Även om undersökningarna kan vara mycket resurskrävande, innebär god dokumentationen dock att det ofta är enkelt även för utomstående att sätta sig in i och använda materialet i efterhand. Givet ett stort antal variabler är det också möjligt att undersöka en uppsjö av associationer. Dessa behöver dock inte vara beroende, utan kan uppkomma på grund av en annan variabel som är relaterad till både den beroende och oberoende variabeln, en så kallad confounder. Stora stickprov möjliggör att ta hänsyn till confounders. Med stora stickprov kan dock även ett litet samband bli signifikant även om det inte är intressant ur vetenskaplig synvinkel. Rekommendationen de ger för att undvika feltolkningar är därför att noga undersöka sambanden genom till exempel konfidensintervall. Interaktion innebär att ett samband mellan den beroende och en oberoende variabel kan se olika ut beroende av nivån på en annan variabel. Även detta kan behöva prövas i en modell.

2.2 Population och parameter

Enligt Chambers och Skinner (2003) härrör målparametern normalt vid deskriptiva survey till en ändlig population, och inferens görs därefter om denna ändliga populationsparameter. En ändlig population U består av N enheter indexerade med t, $U = \{1,...,t,...,N\}$, där storleken på

N antas vara känd. Dessa är vanligtvis individer som en uppsättning undersökningsvariabler mäts på. Undersökningsvariablerna består här av beroende variabler *Y* samt oberoende variabler *X*. För enkelhet i notationen används tills vidare *Y* som en sammanfattande vektor för både *Y* och *X*, medan skiljs sedan åt med början i slutet av avsnitt 2.4.3.

Då avgränsningen här är till tvärsnittsstudier kan *U* antas vara konstant, till skillnad från exempelvis longitudinella studier där vilka enheter som ingår i populationen kan förändras över tiden. Ett urval *s* av storleken *n* är då en delmängd av *U*. Om vektorn θ av storleken $k \times 1$ innehåller målparametrarna, kan således ett estimat baserat på urvalet skrivas som $\hat{\theta}(s)$.

Vektorn av värdena som ingår i *Y* för enhet *t* skrivs som y_t. Matrisen för alla populationsvärden betecknas då med y_U, och raderna i denna skrivs som $y_1,...,y_N$. Observationsprocessen i en survey syftar då till att mäta värdena på *Y* för alla enheter *n* som ingår i *s*, för att sedan transformera dessa till ett dataset.

De variabler som är tillgängliga i urvalsramen och är kända för alla enheter kallas för hjälpvariabler. Dessa betecknas med *Z*, och är ofta regionala, demografiska eller socioekonomiska variabler. Vektorn av värden för enhet *t* betecknas då som z_t och matrisen för alla populationsvärden $z_1,...,z_N$ betecknas som z_U.

Enhet	Variabler		
t=1, 2, …, N	X	Y	Z
1	x_1	y_1	z_1
2	x_2	y_2	z_2
…	…	…	…
t	x_t	y_t	z_t
N	x_N	y_N	z_N

Figur 1.1 Vektorvärden för populationsvariablerna (med X och Y skiljda)

Vid deskriptiva survey består målparametern θ av tämligen enkla funktioner $f(y_1,...,y_N,z_1,...,z_N)$ av populationsvärdena, och refereras därför till som ändliga populationsparameter. En utförlig beskrivning av inferens vid deskriptiva survey ges till exempel av Cochran (1977).

Eftersom analytiska survey har fokus på hur bakomliggande processer ger upphov till olika samband snarare än den begränsade populationens utseende, bör målparametern θ här enligt Chambers och Skinner (2003) definieras i förhållande till en obegränsad population. Vanligtvis antas därför en superpopulationsmodell, vilken styrs av ett begränsat antal parametrar i θ. Fokus är därför att dra inferens om denna superpopulationsparameter.

Målparametern θ kan till exempel vara koefficienter i en regressionmodell. Om värdena på *Y* ses som en realisering av en slumpmatris Y_U vars fördelning styrs av parametervektorn θ, så kan en superpopulationsmodell då uttryckas som $f(y_U;\theta)$. Denna är då en täthetsfunktion som anger fördelningen för undersökningsvariablerna *Y* styrda av parametern θ. Denna fördelning brukar även benämnas ξ-fördelningen.

Enligt Kalton (2002) skulle ett urval från ett superpopulationsperspektiv kunna betraktas i två steg. I det första steget dras först ett urval från en oändlig population som får utgöra den begränsade populationen som faktiskt kan observeras. Ett urval dras sedan från denna begränsade population. Den ändliga populationens värden $y_1,...,y_N$ kan således ses som en slumpmässig realisering av en modell som genererat den ändliga populationen, där modellen är just superpopulationsmodellen. Detta möjliggör också användandet av deskriptiva estimatorer för att dra inferens om en superpopulation.

Regressionparametrar kan skattas som sammansättningar av olika populationstotaler, se till exempel Särndal, Swensson och Wretman (1993) för en närmare beskrivning. En regressionsparameter baserad på en ändlig population kallas då för en censusparameter, och benämns som θ_U . Givet en modell som är rätt specificerad så går det i många fall att visa att θ_U kommer att befinna sig nära superpopulationsparametern θ. Det är även alltid möjligt att tolka θ_U då denna härrör till den ändliga populationen. Censusparametern kan således sägas utgöra en länk mellan de estimatorer som används inom deskriptiva och analytiska survey.

Enligt Chambers och Skinner (2003) är det dock sällan rimligt att en modell stämmer exakt. Givet att θ_U befinner sig nära θ så att modellen är tolkningsbar kan dock modellen betraktas som robust. Med robusthet avses i allmänhet att dragna slutsatser är okänsliga mot förändringar i antagandena, se Särndal (1985). Vid analytiska survey brukar den numeriska skillnaden mellan populations- och superpopulationsparametrar i regel också minska vid stora populationer enligt Skinner m fl (1989).

2.3 Urvalsmekanism, designvikt och bortfall

Den metod som används för att dra ett urval från en population *U* kallas för urvalsmekanism. En urvalsmekanism med planerad slump kallas för ett sannolikhetsurval. Vid sannolikhetsurval kommer urvalsmekanismen således att vara den funktion som tilldelar vardera av alla möjliga stickprov en känd sannolikhet $p(s)$. Denna benämns därför även som p-fördelningen.

För urvalsmekanismen måste gälla att $\sum_{alla\ s} p(s) = 1$, samt att varje element $t = 1,...,N$ som ingår i *U* har en positiv och känd inklusionssannolikhet π_t att komma med vid urvalsdragningen. Genom att representera urvalet med en slumpvektor I_U kan de realiserade värdena betraktas som en bernoullifördelad inklusionsindikatorvariabel, $i_t = \begin{cases} 1 \; om \; t \in s \\ 0 \; om \; t \notin s \end{cases}$, där $t = 1,...,N$. De *N* värdena $i_1,...,i_N$ utgör således elementen i vektorn i_U , vilken har storleken $N \times 1$. Urvalsmekanismen kan därför uttryckas som en funktion $f(i_U)$ av i_U , vilken anger sannolikheten för att erhålla något av de 2^N möjliga urvalen från populationen, se Chambers och Skinner (2003).

Den grundläggande typen av urval inom survey är ett obundet slumpmässigt urval utan återläggning (OSU). Här har varje *s* av den bestämda storleken *n* samma sannolikhet att dras,

medan alla *s* med $n_s \neq n$ ges sannolikheten noll, $p(s) = \begin{cases} 1/\binom{N}{n} \; om \; n_s = n \\ 0 \; annars \end{cases}$. Urvalsdragning sker här genom att *n* element dras från *N*. Inklusionssannolikheten beräknas sedan för varje element *t* som ingår i *n* som $\pi_t = \frac{n}{N}$, och inversen av inklusionssannolikheten kallas för elementets designvikt, $\frac{1}{\pi_t} = w_t$. Designvikten kan således liknas vid det antal enheter i populationen som varje vald enhet representerar. Andelen dragna element av möjliga benämns även som urvalsfraktionen $f = \frac{n}{N}$. Baserad på att varje element $t = 1,...,N$ som ingår i *U* har en positiv och känd inklusionssannolikhet π_t med OSU, är en estimator av en populationstotal *T* som är fri från bias den välkända Horvitz-Thompson estimatorn $\hat{T} = \sum_{t=1}^{N} i_t \frac{y_t}{\pi_t} = \sum_{t=1}^{n} \frac{y_t}{\pi_t} = \sum_{t=1}^{n} w_t y_t$, se Lohr (1999).

Några andra exempel på urval än OSU är bernoulliurval, systematiskt urval, poissonurval, proportionella urval, och klusterurval. Dessa beskrivs dock inte här, utan hänvisning görs till Särndal m fl (1992). Här beskrivs däremot stratifierat urval. Stratifiering sker genom att populationens samtliga element delas in i *H* olika delpopulationer ($U_1,...,U_h,...,U_H$) baserat på en eller flera hjälpvariabler. Varje element ska härigenom ingå i ett och endast ett stratum. De hjälpvariabler *Z* som används för att definiera strata kallas designvariabler.

Vid stratifiering dras ofta ett OSU inom respektive stratum oberoende av varandra, även om det är möjligt att använda olika metoder för urvalsdragning i olika stratum. Härigenom erhålls ett stratifierat OSU (STOSU). Antalet element i stratum *h* betecknas då som N_h, och urvalsstorleken i varje stratum som n_h. Varje element inom samma stratum $h = 1,...,H$ kommer då att ha samma inklusionssannolikhet och designvikt $\pi_t = \frac{1}{w_{ht}} = \frac{n_h}{N_h}$, även om de kan skilja mellan strata. När alla enheter i hela urvalet har samma inklusionssannolikhet benämns urvalet som självviktat. Ett OSU är således alltid självviktat, liksom ett proportionellt STOSU där för varje stratum $h = 1,...,H$ gäller att $w_{ht} = \frac{N_h}{N} = \frac{n_h}{n}$.

Syftet med att använda ett STOSU istället för ett OSU kan vara att vilja tillförsäkra sig observationer från vissa delpopulationer eller att erhålla tillräcklig precision inom dessa genom att välja tillräckligt stora n_h. Vidare kan bortfalls- och mätproblem skilja mellan olika strata, eller rent administrativa geografiska indelningar motivera en specifik indelning i strata. En annan viktig anledning att använda STOSU är variansreduktion. Korn och Graubard (1999) beskriver detta som att om enheterna inom respektive stratum är homogena blir variansen inom varje stratum liten. Den poolade variansen mellan strata kommer då att vara mindre än variansen från ett OSU med motsvarande urvalsstorlek *n*.

Variansreduktionen kan påverkas på två sätt. Först kan tillgänglig information användas för att välja strata så att de blir homogena. Denna information kan även användas för att minska variansen genom poststratifiering, se till exempel Lohr (1999) för en beskrivning av

poststratifiering. Det andra sättet utgår från att välja n_h så att den poolade stratumvariansen minimeras. Detta förutsätter kunskap om N_h samt varianserna S_h^2 inom strata. Givet kunskap om dessa samt om kostnaden för stratifieringen beskriver Cochran (1977) hur en optimal stratiferingsplan kan utformas**.**

Vid jämförelse mellan en vald design och ett OSU med samma urvalsstorlek kan effektivitet hos en parameterskattning $\hat{\theta}$ mätas genom designeffekten, $deff = \frac{\text{var}_{vald\ design}(\hat{\theta})}{\text{var}_{OSU}(\hat{\theta})}$. Skinner m fl (1989) argumenterar dock för att denna endast är lämplig för valet av design under designstadiet. I analysstadiet då en design redan är realiserad argumenterar de därför för att istället studera effekten av en design utifrån möjliga variansestimatorer, vilket de benämner som misspecificeringseffekt (meff), se Skinner m fl (1989) för en utförligare beskrivning.

Bortfall är ett vanligt problem med surveydata, och måste alltid hanteras på något sätt. Skillnad görs normalt mellan enhets-/svarsbortfall och partiellt bortfall. Enhet *t* klassas normalt som enhetsbortfall då y_t ej är observerad, även om z_t är tillgänglig. Detta är till exempel fallet då en person avstått från att besvara en enkät, men där registerdata finns tillgängliga. Partiellt bortfall föreligger då ett eller flera värden i vektorn y_t saknas för enhet *t*, till exempel då en person avstått från eller missat att fylla i en eller flera frågor i en enkät.

I denna uppsats hanteras enhetsbortfallet genom att anta att det uppkommit slumpmässigt (MCAR, eng. *missing completely at random*) inom urvalet. Vid hänsyn till stratifiering antas att urvalet är MCAR inom respektive stratum. Detta innebär således att de tillgängliga enheterna inom ett stratum antas vara ett slumpmässigt urval av de dragna enheterna inom stratumet. Eftersom z_U antas känd kan detta även uttryckas som enhetsbortfallet är (MAR, eng. *missing at random*).

Partiellt bortfall hanteras här genom att utesluta enheten helt och därefter betrakta även denna enhet som en del av enhetsbortfallet. Svarsandel r_h i stratum *h* beräknas då som antalet tillgängliga av antalet dragna enheter n_h i stratumet. En enhets designvikt kan därefter justeras med hänsyn till r_h som $w_{ht} = \frac{N_h}{r_h n_h}$. För en mer heltäckande beskrivning av justering för bortfall, se Little och Rubin (2002).

2.4 Inferens

Vid inferens om målparametern θ i populationen *U* bör skattning av θ ske med effektiva och konsistenta estimatorer. Vid survey baseras skattningen på de observerade värdena av y_U från undersökningsvariablerna Y_U, men inferens rör populationen som helhet. Givet ett observerat urval i_U från I_U krävs därför att relationen mellan de observerade $(y_U \mid i_U = 1)$ och icke-observerade enheterna $(y_U \mid i_U = 0)$ kan specificeras. Om (Y_U, I_U) betraktas som en slumpmatris, innebär detta i praktiken att ge en beskrivning av de möjliga utfallen från den simultana fördelningen (y_U, i_U).

Som ovan noterats anger urvalsmekanismen $f(i_U)$ sannolikheten för de möjliga utfallen i_U på I_U, samt $f(y_U;\theta)$ fördelningen för undersökningsvariablerna *Y* styrda av parametern θ. En beskrivning av de möjliga utfallen för den simultana fördelningen (y_U, i_U) skulle då innebära en specificering av den simultana funktionen $f(y_U;\theta)f(i_U)$. För att detta ska vara möjligt krävs dock att specificera hur $f(y_U;\theta)$ och $f(i_U)$ kan antas oberoende. I annat fall är risken att estimatorer inte blir konsistenta och effektiva, och inferensen därigenom felaktig.

I avsnitt 2.4.1 presenteras två huvudsakliga synsätt på hur oberoendet mellan $f(y_U;\theta)$ och $f(i_U)$ kan betraktas. Avsnitt 2.4.2 fördjupar sedan diskussion av vad som krävs för att specificera $f(y_U;\theta)f(i_U)$, och i 2.4.3 diskuteras valet mellan synsätten 2.4.1.

2.4.1 Modell- och designbaserat synsätt

De två synsätten för inferens med surveydata brukar benämnas som design- respektive modellbaserat synsätt. Enligt Särndal (1985) omfatta begreppet synsätt valet och genomförandet av en urvalsdesign, valet av estimator inklusive valet av variansestimator, samt inferens om en eller flera parametrar. Designbaserat synsätt beskrivs bland annat i Cochran (1977), och ett modellbaserat synsätt intas av bland annat Valliant, Dorfman och Royall (2000).

Enligt Smith (1994) var debattklimatet under en längre tid hård mellan företrädare för de två synsätten. Under de senaste årtiondena har dock i högre grad präglats av konsensus, där de båda synsätten mer kommit att betraktas som kompletterande varandra. Smith framhåller bland annat hur stratifiering förordas utifrån designbaserat synsätt, och beträffande modellbaserat synsätt skriver han att

> "*Most workers in this area have come to the conclusion that the most useful class of models with some guarantee of robustness over a wide range of alternatives is the class of stratification models.*" Smith (1994, sidan 9)

Även Skinner m fl (1989) betonar att synsätten framförallt bör betraktas som kompletterande varandra. Därför är gränsen mellan dem inte uppenbar. Som exempel kan nämnas det designbaserade modellassisterade synsättet som beskrivs av Särndal m fl (1992). Detta är dock i huvudsak ett designbaserat synsätt. Kalton (2002) tar upp att det kan vara möjligt att vikta mellan design- och en modellbaserade estimat. Detta görs av Little (1991) utifrån bayesiansk synsätt, och är således främst är att betrakta som ett modellbaserat synsätt. Little (2003) vidareutvecklar även denna idé i ett pågående arbete. För att undersöka skillnaderna mellan synsätten görs dock i denna uppsats en uppdelning mellan design- och modellbaserat synsätt vid klassisk inferens baserat på asymptotisk teori, där stora urval innebär att centrala gränsvärdessatsen är tillämplig. Särndal (1985) använder ett citat av Smith (1978) för att påvisa den grundläggande förutsättningen för inferens utifrån respektive synsätt.

> "*For survey analysis we can distinguish two principal contenders. These are (i) inferences based on the p-distribution generated by the randomisation in the design and (ii) inferences based on the ξ-distribution, a hypothetical distribution of errors associated with a stochastic model which is assumed to underlie the data.*". Smith (1978) citerad i Särndal (1985), sidan 52.

Enligt ovan kan p-fördelningen respektive ξ-fördelningen uttryckas med funktionerna $f(i_U)$ respektive $f(y_U;\theta)$. Utifrån ett designbaserat synsätt antas enligt Lohr (1999) att värdena y_U

på undersökningsvariablerna är okända men fixa värden. Således kan $f(y_U;\theta)$ betraktas som en okänd konstant och därför ej användas för att beskriva relationen mellan de observerade enheterna och populationen som helhet. Däremot antas att urvalsmekanismen $f(i_U)$ är känd. Relationen mellan enheterna i urvalet och enheterna utanför urvalet kan således beskrivas som att de senare hade kunnat väljas om ett annat i_U erhållits vid urvalsdragningen. Sannolikheten att observera ett visst utfall på en variabel kommer därför endast att bero av sannolikheten för de möjliga urvalen av enheter, vilket bestäms av $f(i_U)$.

Urvalsmekanismen $f(i_U)$ fungerar således som den kända slumpvariabel vilken avgör vilka enheter som ska ingå i urvalet. En förutsättning för designbaserad analys är således att sannolikhetsurval använts, så att $f(i_U)$ är känd för alla möjliga värden på i_U. För att korrekt representera urvalet krävs därför att hänsyn tas till inklusionssannolikheterna π_t. Då $f(y_U;\theta)$ antas men konstant behövs därför inga antaganden om undersökningsvariablerna *Y*:s fördelning i θ. Inget motsäger dock att funktionen faktiskt beskriver hur data genererats. Ett designbaserat synsätt kan därför sägas utgöra en icke-parametrisk väg till inferens.

Utifrån ett modellbaserat synsätt antas istället att urvalsmekanismen $f(i_U)$ är okänd men fix, och därför kan betraktas som en konstant. Däremot betraktas undersökningsvariablerna *Y* som slumpvariabler vilka genererats från en superpopulationsmodell med parametern θ. Härigenom kommer fördelningen för undersökningsvariablerna $f(y_U;\theta)$ att utgöra en länk mellan de enheter som återfinns i urvalet och enheterna utanför urvalet, där antagandet är att modellen kan förklara samtliga dessa. Eftersom $f(y_U;\theta)$ innebär ett antagande om θ så kan modellbaserad inferens sägas vara parametrisk. Detta är ett starkare antagande än vid designbaserat synsätt, eftersom inferens endast kommer att vara giltig givet att den antagna modellen är korrekt specificerad. Modellbaserat synsätt förutsätter således inte sannolikhetsurval. Här antas istället att designinformation kan tas hänsyn till vid specificering av den aktuella modellen. Detta motsäger dock inte användandet av sannolikhetsurval.

Klassisk inferens är möjlig oavsett vilket av de två synsätten som intas. Utifrån modellbaserat synsätt är det även möjligt att använda bayesiansk inferens genom att en priorfördelning $p(Y)=\int p(Y\mid\theta)\,p(\theta)d(\theta)$ specificeras för populationsvärdena, se Little (2003). Modell- och designbaserat synsätt leder dock enligt Lohr (1999) inte sällan till samma resultat även om innebörden av tolkningarna skiljer sig åt.

Ett designbaserat konfidensintervall för ett populationsmedelvärde utifrån klassisk inferens tolkas här som att om konfidensintervall med konfidensgraden α bildas för ett medelvärde i alla möjliga oberoende slumpmässiga urval (OSU) av storleken *n* från en begränsad population av storleken *N*, så ska $1-\alpha$ av dessa täcka det sanna populationsparametervärdet. Tolkningen utgår således från urvalen i_U vilka erhållits genom repeterade dragningar från slumpvektorn I_U. Ett modellbaserad konfidensintervall tolkas däremot givet den valda modellen $f(y_U;\theta)$, där den övre och den undre gränsen i konfidensintervallet ses som slumpvariabler. Även här kan en repetitiv tolkning göras genom att anta att värdena för en population kan genereras från $f(y_U;\theta)$ i princip i all oändlighet, samtidigt som konfidensintervall med konfidensgraden α bildas för varje urval. Av alla möjliga urval som

kan genereras av modellen, så kommer den förväntade andelen vilka täcker det sanna populationsmedelvärdet då att vara $1-\alpha$.

2.4.2 Ignorable och informativ design

Oavsett vilket av synsätten som antas kommer giltigheten av slutsatserna vara avhängig av om $f(y_U;\theta)$ och $f(i_U)$ är oberoende eller ej. Detta antagande benämns som att urvalsdesignen är ignorable. Binder och Roberts (2001) ger en generell definition av begreppet:

> *"Essentially, a sample is said to be ignorable for a variable of interest if the inference based on all the known information, including the sample design information, is equivalent to the inference based on the same information, excluding the outcomes of the random variables corresponding to whether each unit is in the sample."* Binder och Roberts (2001) sidan 2.

De beskriver också en regressionsmodell där hänsyn tagits till designen:

> *"Here, there is no information contained in the model about the sample design beyond what is explicitly specified in the model. We note that the definition of an ignorable sample"…"allows for design variables, such as stratum identifiers in the case of stratified sample, to be part of the model specification. Therefore, if all the relevant features of the sample design are correctly incorporated into the model, the design is ignorable. If, on the other hand, there are features of the design which would make the regression model invalid for at least some observations, the design may be non-ignorable."*. Binder och Roberts (2001) sidan 2.

Närliggande är begreppet ickeinformativ, vilket inte berör inferens utan variablerna. En design är ickeinformativ om det erhållna urvalet har en sannolikhetsfördelning som överensstämmer med den valda modellens. I annat fall är designen informativ. En design som är ickeinformativ kommer alltid att vara ignorable, men däremot behöver en ignorable design inte vara ickeinformativ. Således räcker det att visa att en design är ickeinformativ för att den också ska vara ignorable.

Avgörande vid designbaserat synsätt är huruvida det genomförda urvalet faktiskt överensstämmer med det avsedda. En korrekt genomförd urvalsdesign utan bortfall eller andra urvalsfel, en korrekt estimator innehållande designvikter, samt ett tillräckligt stort urval, leder här till giltig inferens. Antagandet att designen är ignorable kan således uppfyllas genom att inkludera designvikterna. Vid modellbaserat synsätt krävs dock att inferens inte påverkas av någon information i $f(i_U)$. I annat fall krävs att hänsyn tas till denna information genom modellering av $f(y_U;\theta)$, vilket kan vara komplicerat.

Chambers och Skinner (2003) använder en fall-kontroll-studie som exempel på en informativ design. Utfallet av en dikotom variabel y_t anger här huruvida enhet *t* tillhör gruppen fall eller kontroll. Dessa två grupper kan således sägas bilda varsitt stratum, och urvalet görs normalt som ett OSU inom respektive stratum. Givet att fall- och kontrollgruppen faktiskt skiljer sig åt med avseende på undersökningsvariablerna, så kommer urvalsmekanismen $f(i_U)$ att bero av värdena på y_U och således vara informativ. För att kunna dra inferens om den simultana fördelningen för (y_U, i_U) krävs dock att designen är ickeinformativ. Detta kan uppnås genom att $f(i_U)$ betingas på värdet av y_U, så att den simultana fördelningen kan skrivas som $f(i_U \mid Y_U = y_U) f(y_U;\theta)$. För att modellen ska gälla för samtliga enheter i populationen krävs

därför utöver att specificera modellen $f(y_U;\theta)$ för Y_U även att modellera urvalsdesignen $f(i_U \mid Y_U)$ så att den gäller för alternativa utfall för Y_U annat än bara y_U.

Ett annat exempel på en informativ design som Chambers och Skinner (2003) tar upp är då designvariabler används för att definiera strata. Enligt ovan kan z_U här ses som ett utfall på slumpmatrisen Z_U. För att visa på urvalsdesignens beroende av z_U kan $f(i_U)$ då skrivas som $f(i_U \mid Z_U = z_U)$. Givet att värdena på Z_U hålls konstanta vid z_U när urvalsmekanismen specificeras, måste dessa också hållas konstanta när den simultana fördelningen för Y_U och I_U specificeras. Fördelningen för Y_U med Z_U konstant vid z_U kan därför skrivas som $f(y_U \mid Z_U = z_U;\phi)$. Anledningen att parametervektorn ϕ används istället för θ är att de betingade fördelningarna kan tänkas skilja sig från den ursprungliga $f(y_U \mid \theta)$. Betingningen innebär således att θ byts ut mot ϕ. Givet att det inte förkommer något annat direkt beroende mellan $f(i_U)$ och y_U så kan den simultana fördelningen för Y_U och I_U då uttryckas som $f(I_U \mid Z_U = z_U) f(Y_U \mid Z_U = z_U;\phi)$. Här antas alltså att Y_U och I_U blir oberoende betingat på z_U, och således att designen blir ickeinformativ.

Chambers och Skinner anser att betingning på z_U endast bör göras om olika värden på z_U faktiskt representerar relevanta delgrupper av en population där fördelningen för utfallsvariablerna Y_U skiljer sig åt mellan strata. Detta kan göras genom att z_U inkluderas som en oberoende variabel i modellen, eventuellt med interaktionseffekter. De anser dock inte att det vid bestämmandet av fördelningen för Y_U är nödvändigt att betinga på de designvariabler i z_U som är av administrativ karaktär. Detta motiveras med att det i allmänhet inte är lämpligt om urvalsmetoden är drivande i specificeringen av målparametrarna, då modellen för Y_U ändras från $f(y_U \mid \theta)$ till $f(y_U \mid Z_U = z_U;\phi)$, där θ kan skilja sig från ϕ.

2.4.3 Designhänsyn och val av synsätt

Enligt Skinner m fl (1989) är tre hänsyn till designen hos en survey aktuella vid inferens. Dessa är definitionen av målparametern θ, valet av punktestimatorn $\hat{\theta}$, samt valet av standardfelsestimatorn $s.\hat{e}.(\hat{\theta})$. Det sista är ekvivalent med valet av variansestimator $V\hat{a}r(\hat{\theta})$. Dessa punkter återkommer i diskussionen nedan, samt i avsnitt 2.5 om estimation.

Deskriptiva survey utgår vanligtvis från ett designbaserat synsätt vid inferens om ändliga populationsparametrar. Modellbaserat synsätt förekommer dock enligt Kalton (2002) vid väldigt skeva fördelningar. Vid analytiska survey bestäms målparametern ofta genom en iterativ process då en exakt modell i regel inte finns bestämd på förhand. Här är det vanligare att utgå från ett modellbaserat synsätt eftersom superpopulationsparametrar skattas. Ett designbaserat synsätt kan dock intas genom att utgå från en ändlig censusparameter θ_U. Som tidigare nämnts kommer den rent numeriska skillnaden mellan populations- och superpopulationsparametrar vid analytiska survey i regel också att minska vid stora populationer. Oavsett vilken målparameter som är aktuell benämns denna i den senare texten som θ, och antas om inget annat anges att vara väldefinierad.

Användandet av designvikter är allmänt accepterat vid deskriptiva survey, se till exempel Kish (1992). Beträffande analytiska survey går dock åsikterna mer isär enligt Pfeffermann

(1993). I praktiken består skillnaden mellan ett designbaserat och ett modellbaserat synsätt ofta av att det tidigare tar hänsyn till designvikterna, även om det inte finns något som motsäger att designvikter även kan användas vid modellbaserat synsätt, vilket Pfeffermann ger en översikt över hur det kan göras. Syftet här är dock att åtskilja synsätten, varför här antas att designvikter endast används vid designbaserat synsätt. Som tidigare nämnts betraktas synsätten i praktiken dock främst som kompletterande varandra.

Då designbaserat synsätt använder sig av designvikterna är det härigenom lättare att erhålla konsistenta estimatorer. Enklast ses detta genom att betrakta skattning av ett medelvärde. Om urvalet är draget som ett oproportionellt STOSU där medelvärdet mellan strata skiljer sig åt, så kommer designvikterna att kompensera för denna skillnad. Detta riskerar dock en modellbaserad estimator att missa om den utgår från att urvalet är OSU. Användandet av designvikter kan således sägas skydda mot informativa urval. Om urvalet istället varit proportionellt STOSU och således självviktat, skulle estimatet dock ha varit konsistent.

Om modellen är felaktig riskerar således modellbaserade estimatorer att inte vara konsistenta. Designbaserade estimatorer är dock konsistenta oavsett om modellen är korrekt eller ej. I fallet vid en felspecificerad modell, till exempel vid utelämnade variabler eller fel funktionell form, kommer dock designbaserade estimatorer endast att vara konsistenta för den ändliga populationsparameter de faktiskt estimerar enligt modellen. Åtgärder vid en modell som är felspecificerad annat än i förhållande till designen diskuteras ej här utan den intresserade hänvisas till exempelvis Stock och Watson (2003).

Att däremot alltid ta hänsyn till designvikterna riskerar dock att vara ineffektivt. Om urvalet i exemplet ovan gjorts av administrativa skäl där variansen inom de olika strata ej skiljer sig nämnvärt åt så kommer designbaserad synsätt med designvikter leda till en överskattning av populationsvariansen för medelvärdet. Skattning av variansen utifrån modellbaserat synsätt kommer då att vara mer effektivt. Givet en ickeinformativ design kommer således användandet av designvikter att leda till en effektivitetsförlust. Ju mer designvikterna varierar desto större kommer också denna förlust att vara, se Korn och Graubard (1999).

Om däremot designen är informativ är risken stor att variansestimaten vid ett modellbaserat synsätt blir felaktiga såvida inte detta tas hänsyn till i modellen. Även om modellen är felspecificerad vid designbaserat synsätt men variansen skattas för en censusparameter där de deskriptiva måtten är korrekta, så kommer dock variansen här att skattas korrekt.

Givet att en modell är korrekt specificerad är således en avgörande fråga för valet av synsätt huruvida designen är informativ eller ej. Om så är fallet så kommer designvikterna vid ett designbaserat synsätt direkt att justera för skillnaden mellan populationen och urvalet, och härigenom garantera att skattningarna är approximativt konsistenta. Vid en ickeinformativ design riskerar däremot ett designbaserat synsätt att leda till en effektivitetsförlust.

Ett modellbaserat synsätt är att föredra om en modell är korrekt specificerad, eftersom variansskattningarna då kan antas vara mer effektiva än vid designbaserad synsätt där designvikter används. Om modellen däremot är felaktig så kan ett designbaserat synsätt vara att föredra då synsättet delvis kan sägas skydda mot en felspecificerad modell i populationen. Detta ses genom att designviktade estimat relativt sett påverkas mindre av att till exempel en eller flera oberoende variabler saknas i en modell. Valet av variabler är ju också begränsat till dem som faktiskt observerats, se Lohr (1999).

Lohr (1999) diskuterar vilket synsätt som är lämpligt vid regressionsanalys. Vid ett teoretiskt välkänt samband är modellbaserat det naturliga synsättet. Fördelarna med modellbaserat synsätt framhålls som många då det överensstämmer med samhällsvetenskapliga teorier, är konsistent med andra delar inom statistik, möjliggör hänsyn till bortfall, samt ger ett ramverk för att jämföra teorier om strukturella samband. Modellbaserat synsätt möjliggör också estimation vid små eller icke-sannolikhetsbaserade urval. Som framhållits kan dock en felspecificerad modell såsom utelämnade variabler vara ett problem vid modellbaserat synsätt, i synnerhet om dessa är relaterade till designen. Då antagandet är att en modell även antas passa samtliga observationer i populationen och inte bara dem som ingår i urvalet poängteras dock vikten av att undersöka antagandena, till exempel om de stämmer för olika delgrupper.

Med utgångspunkt från syftet med regressionsanalys ger Lohr (1999) följande råd om när designvikter bör användas eller ej, vilket här implicerar valet mellan synsätt:

- Vid beslutsunderlag baserad på officiell statistikproduktion för estimation av parametrar där inferens dras utifrån designen.
- Om designvikter används deskriptivt och syftet även är analytiskt bör man vara konsistent och även använda dem i det senare fallet.
- Ej vid icke-sannolikhetsurval eller små urval då modellbaserat synsätt bör väljas.
- Ej om tidigare teori och kunskap talar för en viss modell eftersom detta talar för ett modellbaserat synsätt.

Enligt Chambers, Dorfman och Sverchkov (2003) finns idag inget sätt att definitivt bestämma om ett urval är draget via en informativ urvalsmetod. De beskriver praxis som att enbart använda designvikter om det leder till att resultatet skiljer sig från metoder som genomförts utan designvikter. Detta sker genom att först beräkna viktade och oviktade skattningar för målparametrarna samt beräkna design- och modellbaserade varianser för de viktade estimaten. Om urvalet är ickeinformativt kommer estimaten och varianserna att vara nära varandra. Om däremot estimaten är olika kan urvalet vara informativt, och om estimaten är lika men varianserna är olika är urvalet informativt.

De beskriver olika sätt hur detta skulle kunna testas. Det första bygger på en Wald-statistiska, vilket dock kräver en jackknife simulering av en kovariansmatris. Jackknife beskrivs bland annat av Lohr (1999). Detta test synes dock enligt författarna att fungera sämre vid heteroskedasticitet. Ytterligare ett test baserat på en multivariat normalfördelning synes ha samma bekymmer. Ett tredje test de föreslår baseras på antagandet om oberoende mellan inklusionsindikatorvariabeln I samt variabeln Y betingat på X. Testet är dock beroende av urvalsstorleken samt antalet grupperingar som görs av X. Dessa diskuteras därför inte vidare här utan den intresserade hänvisas till Chambers m fl (2003).

2.5 Estimation

Såsom beskrivits ovan är eftersträvansvärda egenskaper hos en estimator att den är fri från bias samt är effektiv. Som omnämnts brukar kravet på avsaknad av bias ersättas med det svagare kravet att estimatorn ska vara konsistent. Detta gäller till exempel vid icke-linjära estimatorer såsom varianser eller regressionskoefficienter vanliga vid analytiska survey.

Givet konsistens bör valet av estimator därefter utgå från effektivitet, med målet att minimera MSE. En distinktion mellan modell- och designbaserad konsistens, varians, och MSE, görs utifrån att måtten baseras på p-fördelningen respektive ξ-fördelningen. Som kommenterats ovan leder dock modell- och designbaserat synsätt med klassisk inferens baserat på centrala

gränsvärdessatsen ofta till samma resultat även om innebörden skiljer sig åt. I båda fallen är en estimator konsistent om differensen mellan det sanna värdet och det förväntade värdet av samplingfördelningen för estimatorn går mot noll när urvalsstorleken går mot oändligheten. Samplingfördelningen kan dock se olika ut i de båda fallen. För en vidare diskussion hänvisas till Skinner m fl (1989).

Oavsett vilket synsätt som intas är det ofta möjligt att explicit beräkna estimatorer för deskriptiva punktestimat och variansestimat. Detsamma gäller linjära punktestimat såsom enkla linjära regressionskoefficienter. Dessa skattas normalt medelst minsta-kvadrat-metoden (OLS; eng. *ordinary least squares*), men även maximum likelihood (ML) estimation är möjlig. I bilaga 2 ges en kort introduktion till ML.

2.5.1 Replikering och linjärisering

Icke-linjära estimat såsom varianser kan dock vara mer problematiska att skatta. Ett alternativ är då att approximera estimaten med hjälp av replikeringsmetoder eller linjärisering. Replikering förekommer i ett flertal olika varianter såsom random groups, balanced repeated replication (BRR), jackknife, och bootstrap. Samtliga baseras dock på en grundläggande princip där det tillgängliga urvalet hanteras som en ny population varifrån nya urval kan dras. Därefter beräknas de estimat man är intresserad utifrån de nya urvalen, och variabilitet i dessa nya estimat används därefter för att beräkna det eftersökta variansestimatet. Replikeringstekniker beskrivs av bland annat Lohr (1999).

Linjärisering sker i regel genom expansion av Taylorserier, utgåendes från Taylors formel, se Persson och Böiers (2001). Taylors formel säger att funktionen $f(x)$ i en omgivning av punkten $x = a$ kan approximeras genom att utveckling av en talföljd enligt formeln $f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + ... + \frac{f^{(n)}(a)}{n!}(x-a)^n + \frac{f^{(n+1)}(b)}{(n+1)!}(x-a)^{(n+1)}$, där *b* är ett tal mellan *a* och *x*. Högerledet benämns här som Taylorpolynomet av ordning *n*, där den sista termen är en restterm vilken anger felet i approximationen av $f(x)$. Ju större *n* som väljs desto mindre blir denna term och desto närmare kommer approximation det sanna funktionsvärdet $f(x)$.

Eftersom en varians kan uttryckas som $Var[f(x)] = E[f^2(x)] - E^2[f(x)]$ så kan variansen för funktionen $f(x)$ härigenom approximeras med Taylorpolynomet av första ordningen genom att *a* byts ut mot $E(x)$ i Taylors formel så att följande erhålls:
$Var[f(x)] \approx f(E(x)) + f'[E(x)][x - E(x)] = f'[E(x)]^2 Var[x - E(x)]$.

Givet att $Var[x - E(x)] = Var[x]$ är känd samt att första derivatan av $f(x)$ existerar kan således variansen för $f(x)$ härigenom approximeras. På samma sätt kan funktioner av flera variabler expanderas för att erhålla variansapproximationer, förutsatt att funktionen är deriverbar så att partiella derivator existerar, se Lee, Forthofer och Lorimor (1989).

Principen som används vid linjärisering är att uttrycka den parameter θ som önskas skattas som en funktion av enklare parametrar med kända variansuttryck, till exempel $\theta = f(T_1,...,T_k)$ där $j = 1,...,k$ är populationstotalerna T_j med kända variansuttryck $Var(T_j)$. Därefter används skattningarna $\hat{T}_j$ i Taylorpolynomet av första ordningen för att erhålla en approximation $\hat{\theta}$, varefter variansuttrycken $Var(\hat{T}_j)$ utnyttjas för beräkning av $V\hat{a}r(\hat{\theta})$.

Enligt Lohr (1999) så är denna approximation i regel god då resttermen i normala fall är relativt liten. Fördelar som framhålls är att teorin bakom Taylorserier är välutvecklad, samt givet att partiella derivator existerar så ger metoden alltid ett variansestimat. Nackdelen är dock att beräkningarna kan bli röriga eftersom flera variansuttryck av olika typ kan krävas. Metoden kan också vara svår att tillämpa vid komplexa funktioner med vikter. Exempelvis kan ej heller median eller andra kvartiler uttryckas som funktioner av populationstotalerna T_j.

Skinner m fl (1989) beskriver ett problem som kan uppstå vid estimation av kovariansmatriser vid analytiska survey, orsakat av små urvalsstorlekar eller många skattade parametrar. Om antalet frihetsgrader är färre än dimensionen av kovariansmatrisen, så kommer estimatorn av denna att bli singulär. Således finns ingen entydig lösning till ekvationssystemet. Om antalet frihetsgrader endast är något fler än dimensionen av matrisen riskerar estimaten också att ha en negativ bias och vara instabila. Det skulle dock kunna lösas genom att modellera variansen genom utjämning eller genom linjära samband mellan delestimaten och dess varianser. För en beskrivning av detta hänvisas dock till Wolter (1985).

2.5.2 Exempel på estimatorer

I exemplen nedan presenteras punkt- och variansestimat utifrån design- och modellbaserat synsätt. Dessa estimat kan användas för hypotestest eller för att bilda konfidensintervall. Exemplen nedan berör skattning av ett populationsmedelvärde, samt regressionskoefficienter β vid enkel och multipel linjär regression med tillhörande variansestimat. I kapitel 3 byggs dessa på till att omfatta regressionskoefficienter vid logistisk regression.

Utgångspunkten är att ett STOSU av storleken *n* med *H* stratum genomförts. Vid designbaserat synsätt antas här att stratifiering kan vara informativ vilket tas hänsyn till genom designvikter. Vidare antas att urvalet är draget från en begränsad population, vilket innebär att en korrektionsfaktor beräknad som ett minus urvalsfraktionen används vid variansestimation. Utifrån modellbaserade synsätt antas däremot att stratifieringen är ickeinformativ och därför kan bortses ifrån. Då populationen här antas vara dragen från en superpopulation där *N* är oändligt stort, kommer korrektionsfaktor här att vara lika med ett varför denna kan bortses ifrån vid variansestimation, se Korn och Graubard (1999).

Som estimator för ett populationsmedelvärde för variabel *Y* används vid modellbaserat synsätt urvalsmedelvärdet $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$ vilket är samma estimator som används vid självviktade urval med ett designbaserat synsätt. I praktiken innebär detta alltså att urvalet här betraktas som det vore draget genom ett OSU. Variansen för populationsmedelvärdet skattas då som $\hat{Var}(\bar{y}) = \frac{s^2}{n}$, där $s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(y_t - \bar{y})^2$. Utifrån ett designbaserat synsätt är motsvarande estimator $\bar{y} = \frac{1}{N}\sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H}\sum_{t=1}^{n_h} w_{ht} y_{ht}$. Variansen för populationsmedelvärdet skattas då som $\hat{Var}(\bar{y}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right)\frac{s_h^2}{n_h}$ där $s_h^2 = \frac{1}{n_h - 1}\sum_{t=1}^{n_h}(y_{ht} - \bar{y}_h)^2$ är urvalsvariansen i stratum *h*.

Vid linjär regression kan observationerna utifrån ett modellbaserat synsätt modelleras som $y_t = \beta_0 + \beta_1 x_t + e_t$, givet normala regression antaganden, se till exempel Stock och Watson (2003). OLS-estimat kan då erhållas som $\hat{\beta}_1 = \dfrac{\sum_{t=1}^{n}(y_t - \bar{y})(x_t - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$ och $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Vid antagandet att variansen (σ_e^2) för e_t är homoskedastisk kan denna skattas som $s_e^2 = \dfrac{1}{n-2}\sum_{t=1}^{n}\left[y_t - \left(\hat{\beta}_0 + \hat{\beta}_1 x_t\right)\right]^2$, och variansen för $\hat{\beta}_1$ som $\hat{Var}(\hat{\beta}_1) = \dfrac{s_e^2}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$.

Vid multipel linjär regression skrivs modellen normalt som $y_t = \beta_0 + \beta_1 x_{t,1} + ... + \beta_{k-1} x_{t,k-1} + e_t$, vilket med matrisnotation kan uttrycks som $y = X\beta + e$. OLS-estimatet för parametorvektorn β kan då skrivas som $\hat{\beta} = (X'X)^{-1}X'y$. Antagande homoskedastisk varians ges kovariansestimatorn för $\hat{\beta}$ av $\hat{Cov}(\hat{\beta}) = s_e^2(X'X)^{-1}X'y$, där $s_e^2 = \dfrac{1}{n-k}\sum_{t=1}^{n} r_t^2$ och $r_t = y_t - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{t,1} + ... + \hat{\beta}_{k-1} x_{t,k-1}\right)$ är residualerna.

Ett designbaserat synsätt vid analytiska survey kräver dock att målparametrarna definieras i förhållande till en superpopulation trots att utgångspunkten här är en ändlig population. Detta görs då genom att anta att målparametrarna ses som ändliga censusparametrar vilka närmar sig superpopulationsparametrarna, såsom beskrivits i avsnitt 2.1. Därför används här samma parameterbeteckningar som vid modellbaserat synsätt.

Vid enkel linjär regression erhålls OLS-estimatet som $\hat{\beta}_1 = \dfrac{\sum_{h=1}^{H}\sum_{t=1}^{n_h} w_{ht}(y_{ht} - \bar{y}_h)(x_{ht} - \bar{x}_h)}{\sum_{h=1}^{H}\sum_{t=1}^{n_h} w_{ht}(x_{ht} - \bar{x}_h)^2}$, där $\bar{x}_h = \dfrac{1}{n_h}\sum_{t=1}^{n_h} x_{ht}$ och $\bar{y}_h = \dfrac{1}{n_h}\sum_{t=1}^{n_h} y_{ht}$ är medelvärden inom strata. Vid multipel linjär regression skrivs den viktade OLS-estimatorn för β som $\hat{\beta} = (X'WX)^{-1}X'Wy$ där W är en matris av storleken $n \times n$ med vikterna w_{ht} i diagonalen.

Kovariansmatrisen skrivs då som $\hat{Cov}(\hat{\beta}) = \dfrac{n}{n-1}(X'WX)^{-1}s(X'WX)^{-1}$, där $s = \sum_{t=1}^{n_h}\left(1 - \dfrac{n_h}{N_h}\right)s_h$ är den stratum-poolade urvalsvariansen och s_h är urvalsvariansen i stratum *h*. Kovariansmatrisen kan skattas som ovan, men enklast är att skatta medelst linjärisering. Härigenom släpps även kravet på att feltermerna ska vara homoskedastiska, se Korn och Graubard (1999).

När antagandet om homoskedastiska feltermer ej är uppfyllt, är ett alternativ till OLS att använda estimationsmetoden maximum likelihood (ML), se beskrivning i bilaga 2. Givet att

stratifieringen är ignorable så kommer skattning medelst ML utifrån ett modellbaserat synsätt inte att skilja sig från beskrivningen i bilaga 2.

Om man däremot vill använda sig av ett designbaserat synsätt men fortfarande vill dra inferens om superpopulationsparametern, så kan man använda sig av vad som Skinner m fl (1989) benämner som pseudo-ML (PML). I princip går PML ut på att låta designvikterna förstora upp urvalet så att det motsvarar hela populationen. Härigenom kommer likelihoodfunktionen att omfatta en fiktiv population, trots att denna endast är baserad på de tillgängliga observationerna. Detta skiljer sig från ML som förutsätter att likelihoodfunktionen baseras på observerationer, och är således anledningen till att begreppet pseudo används.

Målparametern kan vid PML definieras på flera sätt, men enligt Skinner m fl (1989) ger alla lösningen till score-ekvationerna, se bilaga 2. Eftersom målparametern här är censusparametern är som tidigare noterats skattningarna i princip av deskriptiv karaktär, se Chambers och Skinner (2003). Binder (2003) visar dock att PML estimatorer kan rättfärdigas asymptotiskt även vid analytiska survey. Binder (1983) visar också hur linjärisering vid komplexa surveys kan användas även när θ är lösningen till ekvationen $f(\theta, T_1, ..., T_k) = 0$ men θ inte kan uttryckas som en explicit funktion av populationstotalerna $T_1, ..., T_k$, jämför med avsnitt 2.5.1. Detta kan till exempel vara fallet när PML används vid logistisk regression.

3 Binär logistisk regression

I detta kapitel presenteras logistisk regression. Beskrivningen avgränsar sig till binär logistisk regression. För en beskrivning av multinominal logistisk regression, se Hosmer och Lemeshow (2000). I första hand beskrivs de test vid logistisk regression som finns tillgängliga i SAS version 9.1.

Två områden som endast diskuteras perifert här är plottar i allmänhet, liksom residualgranskning. Anledningen till detta är dels av utrymmesskäl, dels att dessa ej är tillgängliga i någon större omfattning i programvaran SAS 9.1 som används i kapitel 4 för att illustrera hur hänsyn kan tas till surveydata. Vikten av dessa delar kan dock inte nog understrykas.

Beskrivningen i avsnitt 3.1-3.5 utgår från ett modellbaserat synsätt där surveydata ej tas hänsyn till. Inledningsvis i avsnitt 3.1 presenteras generaliserade linjära modeller följt av förutsättningar och antaganden för logistisk regression. Därefter presenteras logistisk regression samt jämförs med enkel linjär regression och diskriminantanalys. Avsnitt 3.2 beskriver skattning och tolkning av koefficienter, medan avsnitt 3.3 beskriver grundläggande principer för modellbygge. I avsnitt 3.4 beskrivs test för kompletta modeller och enskilda koefficienter. Vidare beskrivs test för hur bra modellen skattar utfallet i den beroende variabeln i avsnitt 3.5. Utifrån ett modell- respektive designbaserat synsätt diskuteras sedan i avsnitt 3.6 hur hänsyn kan tas vid tillämpning på stratifierade surveydata.

3.1 Förutsättningar, antaganden och modellbeskrivning

Logistisk regressionsanalys går främst ut på att beskriva sambandet mellan en dikotom beroende variabel och en eller flera oberoende variabler. Syftet kan också vara att klassificera objekt (individer) i någon av två grupper, eller att skatta sannolikheten för ett utfall (prediktion).

Logistisk regression ingår i en grupp statistisk modeller benämnda generaliserade linjära modeller (GLM). Såsom namnet antyder är detta en sammanfattande metod för att hantera modeller som är linjära i parametrarna. De statistiska modeller som ingår i GLM har ett antal liknande egenskaper vilket gör det möjligt att kategorisera dessa modeller i en gemensam grupp. McCullagh och Nelder (1989) ger exempel på modeller som ingår i GLM, linjär regression, variansanalys, logit- och probitmodeller, loglinjära modeller, multinominala responsmodeller och vissa modeller för överlevnadsdata.

McCullagh och Nelder (1989) beskriver hur modellerna i GLM baseras på tre komponenter. Den första är en slumpmässig komponent vilken innehåller de beroende variabler *Y* där observationerna antas vara oberoende och likafördelade med konstant varians. Den andra är en systematisk komponent innehållande de oberoende variablerna i *X*, vilka antas ge en linjär skattning. Den tredje komponenten utgörs av länken mellan den slumpmässiga och den systematiska komponenten, länkfunktionen *g(.)*. I tabell 3.1 visas hur komponenterna är sammansatta för några exempel av modeller inom GLM.

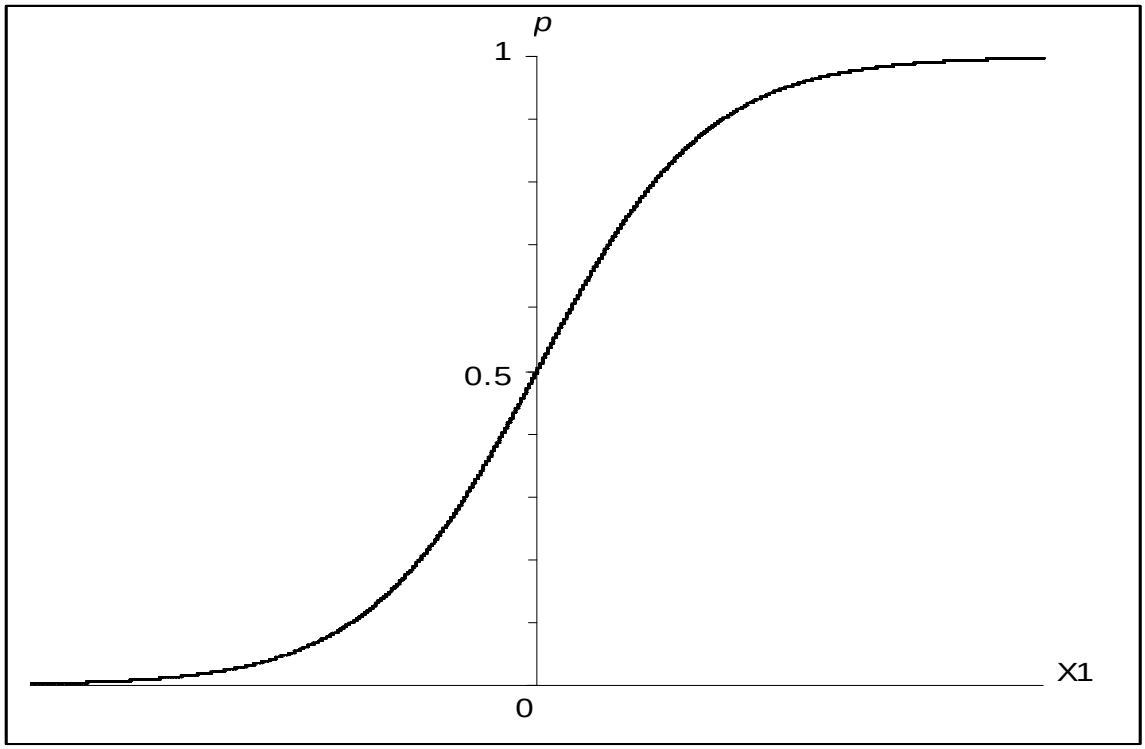
Tabell 3.1 Exempel på statistiska modeller inom GLM

Fördelning	**Modell**	**Länkfunktion**	**Väntevärde**
Normal	Identitet	$X\beta = \mu$	$\mu = X\beta$
Exponential	Invers	$X\beta = \mu^{-1}$	$\mu = (X\beta)^{-1}$
Gamma			
Poisson	Log	$X\beta = \ln(\mu)$	$\mu = \exp(X\beta)$
Binomial	Logit	$X\beta = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(X\beta)}{1-\exp(X\beta)}$
Multinominal			

Logit i GLM är densamma som logistisk regression. Den fortsatta beskrivningen av logistisk regression baseras på den begreppsflora som används av Hosmer och Lemeshow (2000) snarare än den GLM-terminologi som används av McCullagh och Nelder (1989).

Sharma (1996) framhåller att logistisk regression har många likheter med diskriminantanalys. Vid diskriminantanalys antas att de oberoende variablerna är kontinuerliga och följer en multivariat normalfördelning. Detta antagande kommer ej att vara uppfyllt om de oberoende variablerna är diskreta. En lösning är då att använda logistisk regression, vilket ej antar en multivariat normalfördelning. Sharma (1996) visar också att logistisk regression med enbart en kategorisk oberoende variabel kan reduceras till analys av en korstabell.

Hosmer och Lemeshow (2000) gör en pedagogisk poäng i att jämföra linjär och logistisk regression med en oberoende variabel, där de visar på de grundläggande skillnaderna i antaganden. Vid linjär regression antas att förhållandet är linjärt och kan uttryckas som $E(Y \mid x) = \beta_0 + \beta_1 x$, se identitetsmodellen i GLM, tabell 3.1. Ekvationen förutsätter att $E(Y \mid x)$ kan anta alla värden på den reella tallinjen, $-\infty < E(Y \mid x) < +\infty$. Vid en dikotom beroende variabel är dock $E(Y \mid x)$ begränsad till ett sannolikhetsintervall för det binära utfallet, $0 \le E(Y \mid x) \le 1$. Förändringen i $E(Y \mid x)$ då x förändras en enhet är skarpt avtagande då $E(Y \mid x)$ närmar sig 0 eller 1, vilket ger upphov till en S- kurva. Denna benämns även som den logistiska kurvan, se figur 3.1.



Figur 3.1 Logistiska kurvan

Då väntevärdet kommer att vara en sannolikhetsfunktion beroende av *x* kan detta här skrivas som $p(x) = E(Y \mid x)$. Denna sannolikhetsfunktion kallas även för den logistiska fördelningsfunktionen. Förhållandet till parametrarna kan här skrivas som:

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

Linjär regression är per definition linjär i sina parametrar. För att detta även ska gälla vid logistisk regression krävs att en länkfunktion används. Vid logistisk regression kallas denna länkfunktion för logit, se tabell 3.1. Logit beräknas som det logaritmerade oddset för sannolikhetsfördelningen genom transformationsfunktionen $g(x)$. Denna kommer att kunna anta alla värden på den reella tallinjen, $-\infty < g(x) < +\infty$, samt är linjär i sina parametrar:

$$g(x) = \ln\left[\frac{p(x)}{1 - p(x)}\right] = \beta_0 + \beta_1 x.$$

Vidare förutsätts vid linjär regression att en observation i den beroende variabeln kan beskrivas som $y_t = E(Y_t \mid x_t) + \varepsilon_t$, där ε_t är en felterm. Ofta antas att ε_t är normalfördelad med medelvärdet noll och konstant varians, $\varepsilon_t \sim N(0, \sigma^2)$. Av detta följer att den betingade fördelningen är normalfördelad med medelvärdet μ och konstant varians σ^2, $E(Y_t \mid x_t) \sim N(\mu, \sigma^2)$. Då *Y* är en dikotom variabel stämmer dock inte detta antagande, eftersom ε_t i $y_t = p(x_t) + \varepsilon_t$ endast kan anta två värden. Detta eftersom om $y_t = 1$ med sannolikheten $p(x_t)$ så är $\varepsilon_t = 1 - p(x_t)$, medan om $y_t = 0$ med sannolikheten $1 - p(x_t)$ så är $\varepsilon = -p(x)$. Härigenom kommer feltermen vid logistisk regression att följa en betingad

binomialfördelning, $\varepsilon_t \mid x_t \sim (0, p(x_t)[1-p(x_t)])$, vilken av Collet (1991) benämns som en ”*shifted binomial distribution*”.

Givet förutsättningar med transformationen genom logit samt fördelningen för feltermerna kommer principerna och tillvägagångssätten för linjär och logistisk regressionsanalys i stort att vara desamma.

Vid *k* oberoende variabler utökas analysen till multipel logistisk regression, vilket är den generella modellen för logistisk regressionsanalys. Transformationsfunktionen för den multipla regressionsmodellen kan då skrivas som $g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$, och den logistiska sannolikhetsfunktionen skrivas som $p(x) = \frac{e^{g(x)}}{1+e^{g(x)}}$.

Om oberoende variabler återfinns på nominalskala bör dessa kodas som dummyvariabler. I annat fall kan skattningen av koefficienterna bli svårtolkade och medföra felaktiga samband i modellen. En modell innehållande en nominal variabel med *k* kategorier ger en uppsättning av *k-1* dummyvarialber: $g(x) = \beta_0 + \sum_{j=1}^{k-1} \beta_j Dummy_j$. Kategori *k* kommer här att vara referenskategorin. Hardy (1993) framhåller tre riktlinjer för valet av referenskategori. Först ska referenskategorin vara väl definierad. Om det är osäkert huruvida kategorierna kan ordnas, bör detta prövas genom att testa variabeln som kontinuerlig, i syfte att underlätta tolkningen. Vidare bör referenskategorin innehålla tillräckligt många observationer för att betraktas som stabil.

3.2 Skattning och tolkning av koefficienter

Om feltermerna ε vid linjär regression antas ha konstant varians kan koefficienterna skattas med OLS. Eftersom ε vid logistisk regression däremot följer en betingad binomialfördelning, krävs att någon annan skattningsmetod används, företrädesvis ML, se bilaga 2 för en övergripande beskrivning. ML-metoden syftar till hitta de koefficienter som maximerar likelihoodfunktionen. Detta görs genom att logaritmera och derivera denna med avseende på respektive koefficient. De härigenom erhållna score sätts sedan lika med noll, varefter ML-estimaten för koefficienterna kan lösas ut.

Likelihoodfunktionen består här av produkten av likelihooden för varje enskild observation *t*. För de *n* observationerna kan denna vid binär logistisk regression där y endast antar värdena 0 och 1 skrivas som $L(\beta) = \prod_{t=1}^{n} p(x_t)^{y_t} (1-p(x_t))^{1-y_t}$. Den logaritmerade likelihoodfunktionen erhålls därefter som $\sum_{t=1}^{n} \{y_t \ln[p(x_t)] + (1-y_t)\ln[1-(p(x_t)]\}$. Vid skattning deriveras således denna med avseende på de *k* koefficienterna för att erhålla *k* score. Vidare erhålls score-ekvationen för interceptet som $\sum_{t=1}^{n} [y_t - p(x_t)] = 0$ och för de *k-1* övriga koefficienterna som $\sum_{t=1}^{n} x_t [y_t - p(x_t)] = 0$.

Kovariansmatrisen av dimensionen *k* erhålls därefter som inversen av informationsmatrisen,

$I^{-1}(\hat{\beta})$. Informationsmatrisen skattas här som $\hat{I}(\beta) = X'VX$ där *X* är observerade data och *V* är en diagonalmatris med $\hat{p}(x_t)(1-\hat{p}(x_t))$ för de *t=1,…,n* observationerna,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ & \vdots & & \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix} \quad \mathbf{V} = \begin{bmatrix} \hat{p}(x_1)(1-\hat{p}(x_1)) & 0 & \cdots & 0 \\ 0 & \hat{p}(x_2)(1-\hat{p}(x_2)) & \cdots & 0 \\ 0 & 0 & \vdots & 0 \\ 0 & 0 & \cdots & \hat{p}(x_n)(1-\hat{p}(x_n)) \end{bmatrix}.$$

Vid tolkning av de erhållna koefficienterna är det viktigt att dessa är relevanta för den frågeställning som motiverar studien. Hosmer och Lemeshow (2000) menar att tolkningen involverar två steg. Det första är att fastställa relationen mellan den beroende variabeln och en eller fler oberoende variabler, och således att finna en rimlig tolkning för riktningen på sambanden. Nästa steg innebär att tolka storleken på sambandet mellan de oberoende och den beroende variabeln.

Ett centralt begrepp här är odds vilket anger den relativa sannolikheten för en händelse i två grupper. Givet en händelse $x = 1$ i två grupper $y = 1$ och $y = 0$ så kan oddset för $y = 1$ beräknas som kvoten mellan sannolikheten $p(y = 1 \mid x = 1)$ att händelsen inträffar i grupp $y = 1$ och sannolikheten $p(x = 1 \mid y = 0)$ att händelsen inträffar i grupp $y = 0$. Således beräknas oddset som $\frac{p(y = 1 \mid x = 1)}{p(y = 0 \mid x = 1)}$.

Kvoten mellan två odds kallas i regel för oddskvot (OR=Odds Ratio), och anger det relativa oddset för två sannolikheter. Om två odds beräknats för samma grupper $y = 1$ och $y = 0$ men för olika sannolikheter $p(x = 1)$ respektive $p(x = 0)$ så kan OR tolkas som den relativa risken för händelsen $x = 1$ i förhållande till händelsen $x = 0$ mellan grupperna:

$$OR = \frac{p(y = 1 \mid x = 1)/(p(y = 0 \mid x = 1))}{p(y = 1 \mid x = 0)/(p(y = 0 \mid x = 0))}.$$

Vid linjär regression är de skattade parametrarna ofta enkla att direkt tolka genom det linjära sambandet. Då logit används som länkfunktion vid logistisk regression är detta emellertid inte möjligt. Som ovan beskrivits kan dock transformationsfunktionen vid en variabel uttryckas som den naturliga logaritmen av oddset för den logistiska regressionsmodellen.

Om *X* är en dikotom variabel som antar värdena 0 och 1, så kan en logit differens därmed beräknas som

$$g(1) - g(0) = \ln\left[\frac{p(1)}{1-p(1)}\right] - \ln\left[\frac{p(0)}{1-p(0)}\right] = \ln\left(\left[\frac{p(1)}{1-p(1)}\right] \Big/ \left[\frac{p(0)}{1-p(0)}\right]\right).$$

Denna differens kommer således att vara logaritmen av oddskvoten för $p(x = 1)$ i förhållande till $p(x = 0)$ för grupperna $y = 1$ och $y = 0$. Uttryckt med parametrar kan detta skrivas som $g(1) - g(0) = [\beta_0 + \beta_1] - \beta_0 = \beta_1$. Således är den logaritmerade oddskvoten lika med den skattade koefficienten. Oddskvoten kan härigenom erhållas genom att beräkna exponenten av koefficienten, $OR = \exp(\beta_1)$.

Generellt vid skattning av OR för en oberoende variabel som kan anta fler än två värden gäller att om $x = a$ och $x = b$ så erhålls den logaritmerade skattningen av OR som: $\ln[\hat{OR}(a,b) = \hat{g}(x = a) - \hat{g}(x = b) = (\hat{\beta}_0 + \hat{\beta}_1 xa) - (\hat{\beta}_0 + \hat{\beta}_1 xb) = \hat{\beta}_1 x(a - b)$. Således kommer värdet på den oberoende variabeln att ha betydelse vid tolkningen av OR.

Enligt Hosmer och Lemeshow (2000) är relationen mellan koefficienten och oddskvoten en viktig anledning till att logistisk regression är ett kraftfullt analysverktyg. De framhåller dock att Skattningen av OR tenderar att ha en skev fördelning eftersom $0 < OR < +\infty$ med en nollpunkt som motsvaras av 1. De framhåller dock att mycket stora stickprov tenderar att göra skattningen av OR normalfördelad, även om dessa i praktiken behöver vara så stora att de saknar verklighetsanknytning.

Emellertid så kommer fördelningen för β_1 att vara approximativt normalfördelad vid mer realistiska stickprovsstorlekar. För skattning av konfidensintervall eller signifikanstest utnyttjas därför istället fördelning för koefficienten, och först därefter transformeras till OR genom att beräkna exponenten. Således kan ett 100(1-α) % konfidensintervall för OR av skattningen för β_1 beräknas som $\exp\left[\hat{\beta}_1 \pm z_{1-\alpha/2} \times \hat{SE}(\beta_1)\right]$.

3.3 Metoder för uppbyggnad av modell

Såsom vid alla analys av data är det svårt att beskriva någon universell process för uppbyggnad av en modell. Vissa punkter är dock viktiga att ta i beaktande. Den första berör kontroll av data. Här bör stickprovsstorlekar och bortfall kontrolleras. För hantering av enhets- och partiellt bortfall hänvisas till Rubin och Little (2002). Inför analysen av materialet bör sedan de hypoteser som ska testas ställas upp. Vidare bör val av variabler bestämmas. Om fler än en variabel ska användas bör en metod för variabelinkludering också bestämmas.

Processen för val av variabler bör enligt Hosmer och Lemeshow (2000) börja med en noggrann univariabel analys av de potentiella variablerna. Datamaterialet bör här kontrolleras deskriptivt med plottar och tabeller för att undersöka hur data fördelar sig, och för att upptäcka skevheter, kurtosis eller outliers. De nämner till exempel att för nominella, ordinala, samt kontinuerliga data med relativt få unika värden kan korstabeller för den beroende variabeln mot den oberoende variabeln tas fram. Sambandet kan sedan testas med till exempel Pearsons χ^2-test eller LR-test. Vidare bör univariata oddskvoter och Wald-test beräknas för samtliga variabler, se avsnitt 3.4 för beskrivning av detta.

Många oberoende variabler leder i regel till en komplex beslutssituation rörande vilka variabler som ska inkluderas i en modell. Olika skolor har här olika syn på hur variabelselektionen ska gå till. Inom exempelvis epidemiologi menar vissa att samtliga påtänkta variabler ska inkluderas i modellen. Detta kan dock leda till att modellen blir instabil samt att risken för confounding och multikolinjäritet ökar. Hosmer och Lemeshow (2002) skriver att den avgörande frågan som bör ställas är:

> *”Does the model that includes the variable in question tell us more about the outcome (or response) variable than a model that not include the variable?”*
> Hosmer och Lemeshow (2002), sidan 11.

Vidare anser de att en metod behövs för urvalet av de variabler som ska resultera i en bästa modell ur ett vetenskapligt perspektiv. Denna metod bör grunda sig på en plan för att

selektera de variablerna som ska ingå i modellen samt metoder för att fastställa modellens giltighet. Här ingår de bland annat de test som beskrivs i avsnitt 3.4.

Således prövas såväl enskilda variabler liksom modellen som helhet. Generella metoder kan användas för att ta hänsyn till båda kriterierna. I regel innebär dessa att minimera antalet variabler givet att relevanta samband bibehålls i modellen. Å ena sidan så innebär fler variabler att mer sofistikerade samband kan beskrivas, men å den andra kommer det skattade standardfelet att öka, och modellen således bli mer beroende av observerade data och således bli mindre generell.

Om fler än en oberoende variabler övervägs att inkluderas i analysen bör dessa testas för multikolinjäritet med hjälp av korstabeller och korrelationsmatriser. Vid multikolinjäritet kan variablerna till exempel aggregeras. Alternativt kan någon eller några av de kolinjära variablerna uteslutas. En tumregel som Hosmer & Lemeshow (2000) ger är att samtliga variabler med ett univariat p-värde mindre än 0,25 bör tas med som potentiella variabler till modellen, även om variabler som hamnar under detta p-värde men som anses vara av klinisk relevans för undersökningen ändå bör tas med.

En stegvis urvalsprocess kan användas för att välja en uppsättning oberoende variabler för modellen. Detta kan till exempel användas då data består av många variabler eller om de bakomliggande sambanden är oklara. Den stegvisa proceduren är då en algoritm som utifrån olika kriterier väljer ut vilka variabler som bör ingå i modellen. Denna är huvudsakligen uppdelad i tre tillvägagångssätt. Framåtinkludering med test för bakåtelimination, bakåteliminering med test för framåtinkludering, eller en kombination av dessa två. En selektering av variabler kan till exempel göras med hjälp av LR eller Wald-test, se avsnitt 3.4. Vid varje steg i processen inkluderas, alternativt exkluderas, någon variabel baserat på dessa test i förhållande till den tidigare modellen.

Vid många variabler föreslår Korn och Graubard (1999, sidan 94-95) istället att en uppdelning av variablerna görs i en primär och sekundär grupp utifrån relevans samt presenterar en tillhörande strategi för variabelinkludering. De påpekar dock att valet av strategi bör bero av undersökningens syfte. En annan idé kan vara att göra en uppdelning av datasetet i en explorativ och en konfirmativ del. Vid den explorativa analysen antas att ingen eller liten kunskap finns om de faktiska sambanden i data. Denna används därför för att ta fram en modell. Därefter kan den konfirmativa anlaysen användas för att pröva om den explorativa modellens hypotetiserade samband stämmer.

När modellen är skattad bör denna prövas genom generella modelltest. Vidare bör kontrolleras huruvida de skattade värdena överensstämmer med de observerade värdena ur ett absolut perspektiv. De senare benämner Hosmer och Lemeshow (2000) som test av Goodness-of-Fit. Hosmer och Lemeshow (2000) gör dock en distinktion mellan grundläggande test av enskilda koefficienter och modellen ur ett relativt perspektiv, samt hur väl modellerna passar data. Dock menar de att även andra mått såsom klassificering och R^2 även kan användas som mått på Goodness-of-Fit även om dessa inte mäter avståndet mellan skattade och observerade värden. Många författare är dock otydliga vid definitionen av Goodness-of-Fit vid logistisk regression. Modelltest och Goodness-of-Fit tenderar att även sammanblandas. Tabachnick och Fidell (2001) kallar till exempel LR test av modellen som helhet för Goodness-of-Fit. Collett (1991, sidan 62) lägger in en brasklapp och skriver att test som jämför skattade och observerade värdena *”is widely referred to as goodness of fit”*.

Därutöver bör residualanalys genomföras för att hitta inflytelserika observationer, samt antaganden och koefficienter prövas genom känslighetsanalys. Givet att en modell klarar de uppställda kraven, kan grundfrågeställningen därefter prövas utifrån uppställda hypoteser och tidigare kunskap. Om modellen bedöms riktig kan sedan resultat implementeras i form av slutsatser, klassificering eller prognoser.

Följande citat kan ses som en ledstjärna för att lösa problematiken kring valet av variabler vid en logistisk regressionsmodell:

> *"Successful modeling of a complex dataset is part science, part statistical methods, and part experience and common sense."* Hosmer och Lemeshow (2000), sidan 91.

3.4 Koefficient- och modelltest

Såsom beskrivits i avgränsningen kommenteras här endast de test som finns tillgängliga för kommandona PROC LOGISTIC och PROC SURVEYLOGISTIC i SAS 9.1, med undantag av de som benämns som residualer eller som baseras på skillnaden mellan observerade och skattade värden på den beroende variabeln. De test som presenteras används sedan vid illustrationen i kapitel 4. Vad som finns tillgängligt i SAS 9.1 presenteras i bilaga 4.

3.4.1 Likelihoodbaserade test

Första steget är att undersöka om modellen som helhet passar till de underliggande observationerna. En teststatistika skapas av likelihoodfunktionen, se avsnitt 3.2, genom att denna logaritmeras och multipliceras med -2. Denna benämns här som *-2logL* och följer en χ^2 -fördelning.

-2logL används även för att testa huruvida en införd variabel tillför information till modellen som helhet. Nollhypotesen skrivs då som att $\beta_j = 0$ och mothypotesen som $\beta_j \neq 0$, där *j=1,..,k*. Detta kallas då för Likelihoodkvottest (LR=likelihood ratio) och genomförs genom att beräkna statistikan $LR = -2\log L_R - (-2\log L_F) = -2\log\left(\frac{L_R}{L_F}\right)$, där L_R är likelihooden för modellen som helhet. L_F är likelihooden för en modell som endast innehåller interceptet vid test av modellen som helhet är. Alternativt är L_F likelihooden för modellen med exkluderade variabler vid test av en eller flera variabler. Teststatistikan kommer att följa en χ^2 -fördelning med *k-q* frihetsgrader, där *k* är antal skattade parametrar vid L_R och *q* är antal skattade parametrar vid L_F. Förkastas nollhypotesen kan antas att modellen baserad på L_R passar data bättre än modellen baserad på L_F. Detta benämns även som nestade modelltest.

3.4.2 Wald test

Wald test kan ses som ett komplement till LR-test. Wald jämför en eller flera koefficientskattningar med dess standardfel. Hypoteserna ställs upp på motsvarande sätt som för LR. De univariata teststatistikorna skrivs som: $Wald_j = \frac{\hat{\beta}_j}{S\hat{E}(\hat{\beta}_j)}$, där *j=1,…k,* och kommer approximativt att följa standardnormalfördelning med väntevärde 0 och varians 1.

Den multivariabla teststatistikan beräknas som $Wald = \hat{\beta}'\left[V\hat{a}r(\hat{\beta})\right]^{-1}\hat{\beta} = \hat{\beta}'(XVX)\hat{\beta}$. Denna kommer approximativt att följa en χ^2 -fördelning med *k* frihetsgrader motsvarande antalet

skattade parametrar.

3.4.3 Score-test

Till skillnad från LR och Wald vilka är baserade på ML-skattningarna av koefficienterna, så kan det univariata Score-test användas på motsvarande sätt men utan att kräva ML-skattningar. Score-testet är baserat på derivatan av loglikelihood vilket redovisas i bilag 2. Det univariata Score-testet beräknas genom att sätta $\beta_0 = \ln(n_1 / n_0)$ och $\beta_1 = 0$ för score-ekvationerna. Detta ger att $\hat{p} = n_1 / n = \bar{y}$. Från score-ekvationen ges att $\sum x(y - \bar{y})$, och den skattade variansen kan skrivas som, $\bar{y}(1-\bar{y})\sum (x - \bar{x})^2$. Teststatistikan för det univariata Score-testet (ST) beräknas som $ST = \dfrac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n} (x_i - \bar{x})^2}}$ och kommer approximativt att följa standardnormalfördelning med väntevärde 0 och varians 1. Den multivariata motsvarigheten beräknas med hjälp av score och inversen av informationsmatrisen (bilaga 2) enligt $U'(\hat{\beta}) I^{-1}(\hat{\beta}) U(\hat{\beta})$. Det multivariata Score-testet följer approximativt en χ^2 -fördelning med *k* frihetsgrader motsvarande antalet skattade parametrar. I det multivariata fallet krävs emellertid ML-skattade koefficienter.

3.5 Goodness-of-Fit

Givet att en modell och dess variabler klarat genomförda test som beskrivits i avsnitt 3.4 är målet därefter att pröva modellens förmåga att beskriva hur bra modellen passar data. Detta benämns allmänt som modellens Goodness-of-Fit. Det är dock svårt att ge någon entydig definition av begreppet då det ofta är vagt beskrivet. Här används därför indelningen som görs av Hosmer och Lemeshow (2000), utan att närmare definiera begreppet.

För att fastställa Goodness-of-fit för en modell föreslår Hosmer och Lemeshow (2000) att beräkning och tolkning först görs av helhetsmått för anpassningen. Därefter bör enskilda komponenter av de övergripande måtten granskas, vilket ofta sker grafiskt. Slutligen kan andra mått för avståndet mellan skattning och observation för den beroende variabeln prövas.

Olika test finns utvecklade för att testa helhetsanpassningen av en modell. Några exempel är Pearson χ^2 test, Deviance test, och Osius and Rojek test. Dessa mått baseras på differensen mellan de observerade och skattade värden på den beroende variabeln. Differenserna beräknas inom alla grupper som är möjliga att skapa baserad på alla möjliga kombinationer av värden på för de oberoende variablerna. Om antalet unika kombinationer av värden på de oberoende variablerna är stort i förhållande till antalet observationer är det dock inte möjligt att beräkna dessa mått, vilket är vanligt vid kvantitativa eller många diskreta variabler. Detta är till exempel fallet för illustrationen i kapitel 4. Dessa mått beskrivs därför inte närmare här, utan hänvisning görs till Hosmer och Lemeshow (2000).

3.5.1 Hosmer och Lemeshow-test

För att lösa problemet med att Pearson χ^2 och Deviance inte kan beräknas inom ovan nämnda situationer har Hosmer och Lemeshow utvecklat ett alternativt test. Nollhypotesen här är dock att modellen inte har någon bristande anpassning. Förkastas denna är detta således en indikation på att modellen har dålig anpassning.

Istället för att dela upp data efter antalet möjliga kombinationer av de oberoende variablerna delar man upp data i mellan fem och tio lika stora klasser *c* utifrån storleken av de predikterade värdena för den beroende variabeln. Antalet observationer i klass *j* där den beroende variabeln antar värdet 1 betecknas som O_{j1} och antalet observationer där den antar värdet 0 betecknas som O_{j0}, och antalet observationer i klass *j* kommer således att vara $n_j = O_{j1} + O_{j0}$.

Förväntat antal värden som är lika med 1 i klass *j* kan då beräknas som summan av de skattade sannolikheterna i denne klass $E_{j1} = \sum_{t=1}^{n_j} \hat{p}_t(x)$, och förväntat antal värden lika med 0 beräknas som $E_{j0} = \sum_{t=1}^{n_j} 1 - \hat{p}_t(x) = n_j - E_{j1}$. Teststatikan beräknas sedan som $HL = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(O_{jk} - E_{jk})^2}{E_{jk}}$. Hosmer och Lemeshow har genom simulering visat att denna statistika approximativt kommer att följa en χ^2-fördelning med *c-2* frihetsgrader.

3.5.2 Klassificeringsförmåga

Klassificering avser förmågan att kunna placera in skattade värden från en modell i korrekt grupp utifrån de observerade värdena på den beroende variabeln. Detta skiljer sig från residualanalys, vilket istället jämför de skattade värden från en modell med de faktiska observerade värdena på den beroende variabeln.

En 2x2 tabell, vilken benämns som klassificeringstabell, kan ställas upp, se tabell 3.2. Klassificeringstabellen visar andelen korrekta och andelen icke korrekta klassificeringar för den beroende variabeln. I tabellens kolumn visas de observerade värdena för skattningarna av den beroende variabeln medan i raden visas de klassificerade värdena.

Tabell 3.2 Klassificeringstabell

Klassificerade värden	Observerade värden		
	y=1	y=0	Totalt
y=1	n_{11}	n_{10}	n_{11+} n_{10}
y=0	n_{01}	n_{00}	n_{01+} n_{00}
Totalt	n_{11+} n_{01}	n_{10+} n_{00}	n_{11+} n_{10+} n_{01+} n_{00}

Klassificering av en skattning görs med ett cutoffvärde. Cutoffvärdet anger vid vilken skattad sannolikhet $\hat{p}(x)$ som observationen ska klassas som 1 respektive 0. Som standard är denna gräns satt till 0,5. Om $\hat{p}(x)$ då överstiger 0,5 klassas observationen som 1, och i annat fall som 0. Vid en perfekt modell återfinns samtliga värden i diagonalen i tabellen, och andelen korrekta klassificeringar är därmed 100 %. Ett problem är dock att cutoffvärdet är begränsat till ett värde, och inte tar hänsyn till hur nära $\hat{p}(x)$ befinner sig de faktiska värdena 1 eller 0. Därför bör en enkel klassificeringstabell inte användas som ett Goodness-of-Fit mått.

En lösning är dock att tillåta att cutoffvärdet varierar. Denna teknik härstammar egentligen från signalteorin för att visa hur en mottagare hanterar signaler då det förekommer brus, där

sannolikheten att få korrekt signal (sensitivitet) och sannolikheten att få felaktig signal (1-specificivitet) studeras inom ett intervall av cutoffvärden.

Om sensitivitet beräknas som andel korrekt klassificerade för de observerade värden y=1, $\frac{n_{11}}{n_{11}+n_{01}}$, och specificitet beräknas som andel korrekt klassificerade för de observerade värdena y=0, $\frac{n_{00}}{n_{01}+n_{00}}$, så kan en tabell ställas upp för olika värden på dessa baserat på olika cutoffvärden. Ett receiver-operating-characteristics (ROC) diagram kan sedan ritas upp med 1-specificitet på x-axeln och sensitivitet på y-axeln.

Kurvan som ses i diagrammet kallas för ROC-kurvan. Arean under kurvan kommer här att variera mellan 0 och 1, och vara ett mått på modellens förmåga att diskriminera mellan objekt för vilka *y=1* och för objekt där *y=0*, det vill säga förmågan att korrekt klassificera observationer. I tabell 3.3 ses den tumregel Hosmer och Lemeshow (2000) ger för tolkning av måttet.

Tabell 3.3 Tumregel för tolkning av arean under ROC-kurvan

Utfall	Diskrimineringsförmåga
ROC=0,5	”no discrimination”
0,7≤ROC<0,8	”acceptabele discrimination”
0,8≤ROC<0,9	”excellent discrimination”
ROC≥0,9	”outstanding discrimination”

Hosmer och Lemeshow (2000), sidan 162.

Ett antal rankkorrelationsmått mellan skattad sannolikhet och det observerade värden kan också beräknas. Dessa mått framställs genom att först beräkna antalet möjliga par av observationer $(n_{11}+n_{01})\times(n_{10}+n_{00})$. Därefter benämns paren enligt tabell 3.4.

Tabell 3.4 Benämningar vid beräkning av rangkorrelationer

Relation mellan skattade sannolikheter för par	Benämning av par	Antal par
$\hat{p}(y=1\mid x) > \hat{p}(y=0\mid x)$	Samstämmiga	n_c
$\hat{p}(y=1\mid x) < \hat{p}(y=0\mid x)$	Motstridiga	n_d
$\hat{p}(y=1\mid x) = \hat{p}(y=0\mid x)$	Tied	n_t
Totalt antal par		$n_c+n_d+n_t$

Baserat på hjälpfil i SAS 9.1.

Rangkorrelationsmåttet $C=\frac{n_c+0{,}5(n_t)}{n_c+n_d+n_t}$ som anger andelen korrekt klassificerade par då y=1, kommer vid binär logistisk regression att motsvara arean under ROC-kurvan.

Ett annat rangkorrelationsmåttet är $Somer's\ D=\frac{n_c-n_d}{n_c+n_d+n_t}$ vilket kan används för att fastställa styrkan och riktningen mellan par. Måttet kommer att variera mellan -1, inget par överensstämmer, och 1, samtliga par överenstämmer.

$Goodman - Kruskal\ Gamma = \frac{n_c - n_d}{n_c + n_d}$ beräknas på liknande sätt men bortser från antalet ties n_t och kommer därigenom generellt att vara högre än Somer's D. Måttet varierar mellan -1 (ingen association) och 1 (perfekt association).

Kendalls's Tau-a är en modifiering av Somer's D och definieras som $\frac{2(n_c - n_d)}{n(n-1)}$. Måttet tar hänsyn till den beräknade skillnaden mellan antal möjliga par av observationer och antal par av observationer med olika respons, och är vanligen mindre än Somer's D eftersom det ofta förekommer par med samma respons.

3.5.3 Ytterligare mått

Vid linjär regression brukar R^2 användas som mått på hur stor andel av variationen i den beroende variabeln som en modell förklarar. Då den beroende variabeln vid logistisk regression är dikotom kommer variansen härigenom att vara störst då antalet observationer för y=1 och y=0 är lika stora, och ju skevare en fördelning är desto mindre blir variansen. Härigenom kommer R^2-mått som används vid linjär regression inte att vara jämförbara om fördelningarna för de oberoende variablerna är olika. De R^2-mått som finns framtagna för logistisk regression ska därför endast ses som en approximation av måtten som förekommer vid linjär regression, och inte som andelen förklarad varians.

Två av de R^2-mått som föreslagits är Cox & Snell R^2_{SC} och Nagelkerke R^2_N. Måtten utgår från likelihoodberäkningar som presenterats i avsnitt 3.4.1, och tar hänsyn till urvalsstorleken *n*.

Cox & Snell beräknas som $R^2_{SC} = 1 - \left(\frac{L_F}{L_R}\right)^{\frac{2}{n}}$. Måttet kommer dock att nå sitt maximum vid $1 - (L_R)^{\frac{2}{n}} < 1$. Nagelkerke R^2_N justerar därför måttet så det når sitt max vid 1, $R^2_N = \frac{R^2_{SC}}{1 - (L_R)^{\frac{2}{n}}}$.

Akaike's information criterion (AIC) och Schwartz's criterion (SC) används ofta vid tidsserieanalys för att bestämma antalet laggar som ska inkluderas vid autoregressiva modeller, se till exempel Stock och Watson (2003). Målet är då att minimera värdet på dessa statistikor. Sharma (1996) innefattar AIC och SC som Goodness-of-Fit och exemplifierar användandet av dessa vid stegvis logistisk regression som ett fristående kriterium för att ge en stoppunkt vid inkluderade eller exkludering av variabler. Statistikorna beräknas som $AIC = -2LogL + 2k$ och $SC = -2LogL + kLog(n)$, där *k* är antalet parametrar i modellen.

Dessa två statistikor baseras således på -2LogL men med justering för de skattade parametrarna De kommer därför inte att ha någon entydig samplingfördelning, och används därför normalt självständigt för att jämföra olika modeller skattade från ett och samma dataset. Sharma (1996) påpekar också att det inte finns några specifika regler för att avgöra hur litet ett värde ska vara för att vara lågt eftersom detta beror av de aktuella data. Ju lägre värde statistikorna har desto bättre anpassad kan dock en modell anses vara.

3.6 Tillämpning vid surveydata

Beskrivningen av logistisk regression har i avsnitt 3.1-3.4 gjorts utifrån ett modellbaserat synsätt utan hänsyn till surveydata. Här presenteras dock hur hänsyn kan tas till stratifierade surveydata utifrån modell- och designbaserade synsätt som beskrivits i kapitel 2.

Målparametern vid modellbaserat synsätt antas som tidigare vara en väldefinierad superpopulationsparameter och vid designbaserat synsätt en censusparameter vilken närmar sig denna. De estimatorer som presenteras i detta avsnitt kan sägas vara en fortsättning på de exempel som gavs i avsnitt 2.5. Skillnaden är att här också diskuteras utförligare hur hänsyn kan tas till informativ stratifiering utifrån ett modellbaserat synsätt.

Lohr (1999) framhåller att det kan vara en utmaning att plotta data från komplexa survey. Beträffande stratifierade data rekommenderar hon att antingen plotta olika strata var för sig, eller att använda olika markörer för olika strata. Vidare framhåller hon att data bör plottas med och utan vikter. Här diskuteras dock inte plottar vidare utan den intresserade hänvisas till Korn och Graubard (1999).

3.6.1 Modellbaserat synsätt

Giltigheten hos en modell beror av hur huruvida den är korrekt eller ej. Som konstaterats i avsnitt 2.4.2 är en förutsättning för detta att designen är ignorable. För att detta ska vara uppfyllt räcker det att designen är ickeinformativ, det vill säga att det erhållna urvalet har en sannolikhetsfördelning som överensstämmer med den valda modellens.

Om endast den beroende variabeln *Y* i en logistisk regressionsmodell beror av designvariablerna *Z*, till exempel vid en fall-kontroll studie, har dock Prentice och Pyke (1979) enligt Lohr (1999) visat att givet att modellen stämmer så kommer interceptet att vara den enda koefficient som påverkas. Om fokus endast är att studera sambandet mellan variablerna och interceptet ej är av intresse kan därför detta bortses ifrån. Om man vill ta hänsyn till stratifieringen när antalet strata, och därigenom antalet skattade parametrar, är stort i förhållande till antalet observationer, vilket ofta är fallet vid fall-kontroll studier, kan dock krävas att använda en typ av betingad (conditional) logistisk regression. Detta beskrivs av bland annat Kleinbaum, Kupper, Muller och Nizam (1998).

Om däremot designen inte är informativ genom att de oberoende variablerna *X* är beroende av stratifieringsvariablerna *Z* bör detta tas hänsyn till i modellen. Praxis enligt Chambers och Skinner (2003) är då att inkludera stratifieringsvariablerna som oberoende variabler i modellen, se kapitel 2.4.2. Hur detta görs vid logistisk regression beskrivs av Hosmer och Lemeshow (1998) enligt följande.

Om modellen utan hänsyn är av formen $g(x_t) = \ln\left[\frac{p(x_t)}{1-p(x_t)}\right] = \beta_0 + \beta_1 x_t$ så kan en modell där en fast effekt antas för varje stratum *H* med *j=1, 2… n_j* observationer skrivas som $g(x_{hj}, z_h) = \ln\left[\frac{p(x_{hj}, z_h)}{1-p(x_{hj}, z_h)}\right] = \alpha_k z_k + \beta_0 + \beta_1 x_{hj}$. Detta innebär i praktiken ett intercept definierade av variabel *Z*. Om *Z* här är en dikotom variabel kan då modellen för referenskategorin skrivas som $g(x_{hj}, z_h = 0) = \beta_0 + \beta_1 x_{hj}$ och modellen för den andra kategorin som $g(x_{hj}, z_h = 1) = (\alpha_k + \beta_0) + \beta_1 x_{hj}$.

För att därutöver tillåta *Z* att påverka parameterkoefficienterna för *X* så att effekten ser olika ut i olika strata kan en interaktionsparameter γ_k tillföras mellan Z och *X* $g(x_{hj}, z_h) = \alpha_k z_k + z_k \gamma_k x_{hj} + \beta_0 + \beta_1 x_{hj}$. Modellen för referenskategorin kan då skrivas som $g(x_{hj}, z_h = 0) = \beta_0 + \beta_1 x_{hj}$ och för den andra kategorin som $g(x_{hj}, z_h = 1) = (\alpha_k + \beta_0) + (\gamma_k + \beta_1) x_{hj}$.

I båda fallen kommer likelihoodfunktionen att behöva utvidgas så att den beräknas över de olika strata, se Hosmer och Lemeshow (1998). Om flera variabler ingår i *Z* kan i första hand de enskilda stratifieringsvariablerna inkluderas, men det är även möjligt att inkludera interaktionseffekter mellan dessa. Som påpekats ovan under 2.4.2 kan det dock enligt Chambers och Skinner (2003) vara olämpligt att inkludera variabler som endast är av administrativ karaktär och ej är relevanta för modellen ifråga eftersom det inte är troligt att dessa kommer att vara informativa.

Estimationen av kovariansmatrisen kommer här att ske på samma sätt som tidigare, med enda skillnad att en utvidgning sker till skattningar av parameterkoefficienterna för de extra stratifieringsvariabler *Z* och eventuella interaktionstermer som tas hänsyn till i modellen. De modelltest som beskrivits i avsnitt 3.2 påverkas också endast på motsvarande sätt.

Även om utgångspunkten här är att designvikter är förbihållet ett designbaserat synsätt, så är det i praktiken även möjligt att ta hänsyn till designinformation genom att inkludera designvikterna direkt i en modell. Detta beskrivs dock inte här, utan hänvisning görs till Pfeffermann (1993) som beskriver hur detta kan göras.

3.6.2 Designbaserat synsätt

Grundprincipen för att ta hänsyn till designen vid designbaserat synsätt är som tidigare beskrivits att inkludera designvikterna i estimatorerna. Det är dock fullt möjligt att de stratifieringsvariabler och interaktionstermer som beskrivits i 3.6.1 även kan vara aktuella att inkludera i en modell även om designvikterna tagits hänsyn till. Syftet med detta är att dessa variabler antas mäta ett samband aktuellt för modellen i fråga som inte är relaterat till designen, eftersom denna tas hänsyn till genom designvikterna.

Utöver den beskrivning Pfeffermann (1993) ger för hur designvikter kan inkluderas direkt i en modell ger han även en översikt över hur designvikter kan tas hänsyn till vid estimation av censusparametrar utifrån ett designbaserat synsätt. Här begränsas till att beskriva hur skattning görs medelst PML. Designvikterna används här för att skatta likelihoodekvationer för en fiktivt totalt observerad ändlig population. Likelihoodfunktion för populationen approximeras således här med en likelihoodfunktion baserad på urvalet och designvikterna.

Detta beskrivs av Hosmer och Lemeshow (2000) vid logistisk regression enligt följande. Först sätts log-likelihoodfunktionen: $\sum_{h=1}^{H}\sum_{t=1}^{n_h} w_{ht}\left[(\ln(p_{ht}))' y_{ht} + \ln(1-p_{ht})(1-y_{ht})\right]$ upp. På vanligt sätt kan sedan funktionen differentieras med avseende på de okända regressionskoefficienterna β för att erhålla en vektor innehållande *k* score-ekvationer $(X'W)(y - p(X)) = 0$, där $p(X) = \left[p(x_{11}),..., p(x_{Hn_h})\right]^{-1}$ är en $n \times 1$ vektor. Härigenom kan PML-skattningarna för $\hat{\beta}$ sedan lösas ut. Enskilda element i score-ekvationerna bildas här

som $\tau'_{ht} = (x'_{ht} w_{ht})(y_{ht} - p(x_{ht}))$, där summan av de n_h enheterna i stratum *h* är $\tau_h = \sum_{t=1}^{n_h} \tau_{ht}$ och medelvärdet av alla strata är $\bar{\tau}_h = \frac{1}{n_h} \sum_{t=1}^{n_h} \tau_{ht}$.

En naiv skattning av kovariansmatrisen skulle här vara $\hat{Var}(\hat{\beta}) = (X'DX)^{-1}$, där $D=WV$, varav V är diagonalmatrisen av storleken $n \times n$ med $\hat{p}_t(1-\hat{p}_t)$ i diagonalen och W är viktmatrisen av storleken $n \times n$ innehållande vikterna w_{ht}. Den korrekta estimatorn kommer dock att vara $\hat{Var}(\hat{\beta}) = (X'DX)^{-1} S (X'DX)^{-1}$ där S är den poolad inomstratum-kovariansmatrisen. Som estimator för S används normalt den stratum-poolade variansestimatorn $\hat{S} = \sum_{t=1}^{n_h} \left(1 - \frac{n_h}{N_h}\right) \hat{S}_h$, där $\hat{S}_h = \frac{n_h}{n_h - 1} \sum_{t=1}^{n_h} (\tau_h - \bar{\tau}_h)(\tau_h - \bar{\tau}_h)'$ är den estimerade variansen i stratum *h*. En approximation av $\hat{Var}(\hat{\beta})$ kan sedan göras genom linjärisering eller replikering.

Roberts, Rao och Kumar (1987) beskriver hur teststatistikor baserade på χ^2 och likelihood påverkas av surveydesignen. Deras resonemang baseras på en generaliserad design effekt. Genom att vikter appliceras på log-likelihooden, $-2 \log L = -2 \sum_{t=1}^{n} w_t \log \hat{p}_t$, innebär detta att likelihoodkvottest, scoretest, samt förklaringsgrad påverkas. Beträffande pearson χ^2 kan detta enligt Rao och Thomas (2003) beräknas genom en relativ viktning, och Roberts m fl (1987) har visat när denna kommer att följa en χ^2-fördelning. Detta test finns dock ej tillgängligt i SAS 9.1, varför vi avstår från att vidare kommentera detta.

Wald test skulle kunna genomföras som beskrivits ovan. En förutsättning är dock att hela den genom designvikterna justerade kovariansmatrisen kan skattas vilket krävs för att genomföra hypotestesten. Ett problem är att Wald statistikan inte är oberoende av icke-linjära transformationer av målparametern. Ett quasi-score test har därför föreslagits baserat på linjärisering eller replikering, vilket gör statistikan oberoende av icke-linjära transformationer av målparametern, se Rao och Thomas (2003). Hosmer och Lemeshow (2000) visar också på hur ett modifierat Wald test bestående av F-statistikan $F = \frac{(n-q)}{nq} Wald$ där *q* är antalet parametrar som under nollhypotesen är lika med noll skulle kunna användas istället. Statistikan har *q* frihetsgrader i täljaren och *H-q* frihetsgrader i nämnaren. Om *q* är stort i förhållande till *n* finns dock effektivare test, se Korn och Graubard (1999).

Klassificering och rangkorrelation påverkas inte annat än indirekt genom modellen. Roberts m fl (1987) påtalar problem med residualanalys med surveydata, men ger förslag på hur denna skulle kunna genomföras. Detta finns dock inte tillgängligt i SAS 9.1 och är avgränsat ifrån i denna uppsats, varför det lämnas utan vidare kommentar. Hosmer och Lemeshow (2000) föreslår generellt att när modelltest inte finns tillgängliga för designbaserat synsätt kan motsvarande modellbaserad test beräknas, för att sedan utgå från dessa eller att söka inkorporera dem i det designbaserade synsättet.

4 Illustration – studien Liv & Hälsa (2000) i Örebro län

I detta kapitel ges ett exempel på hur hänsyn kan tas till stratifiering vid surveydata, då logistisk regression används för att studera associationer mellan inkontinens och hälsorelaterade faktorer. Data är hämtade från enkätstudie ”*Liv & Hälsa – en undersökning om hälsa, levnadsvanor och livsvillkor i Örebro*”, här refererat till som Liv & Hälsa (2000).

I avsnitt 4.1 ges först en presentation av studiedesignen samt de för regressionmodellen relevanta variablerna. Därefter presenteras i avsnitt 4.2 hur hänsyn kan tas till designen utifrån såväl modell- som designbaserad synsätt, och de erhållna resultaten diskuteras. Slutligen förs i avsnitt 4.3 en diskussion om utfallet av modelltesten, betydelsen av att ta hänsyn till stratifieringen baserat på kapitel 2, samt övriga kommenterar.

Regressionsparametrar för variabler och dess enskilda kategorier testades med Wald. Enskilda modellers signifikans som helhet testades med Score, Wald, och LR. Det senare gjordes båda för modellen som helhet, LR(modell), och för stegen mellan modellerna, LR(variabel). Av de test som genomfördes här var det endast HL som ej var tillgängligt vid båda synsätten, se bilaga 4 för en beskrivning av vad som finns tillgängligt i SAS 9.1. Ytterligare tillgängliga mått på goodness-of-fit togs också fram.

Då syftet med denna uppsats även är att beskriva logistisk regression i allmänhet återges även utfallet av de test som beskrivits i kapitel 3. Utöver vad som återges i detta kapitel återfinns i bilaga 3 en variabelbeskrivning, i bilaga 4 en överblick vad som finns tillgängligt i SAS 9.1 vid logistisk regression, samt i bilaga 5 utfallet för genomförda test.

4.1 Studiedesign och variabler

Huvudsyftet med Liv & Hälsa (2000) var att ge underlag för hur befolkningen i Örebro län mådde, och genomfördes som en postenkät som skickades ut till urvalspersonerna. Datainsamlingen genomfördes under perioden mars-maj 2000 av Statistiska centralbyrån (SCB) på uppdrag av Örebro läns landsting.

Urvalsramen skapades den 31:a januari 2000 från registret för totalbefolkningen (RTB). Rampopulationen innehöll *198 491* personer, varav *98 812* män och *99 679* kvinnor Urvalet drogs den 7:e mars 2000. Härigenom var samtliga stratumtotaler N_h för hjälpvariablerna Z kända. Totalt drogs $n_h=120$ personer från *128* olika stratum som ett STOSU uppdelat efter kön, fyra åldersgrupper, och *16* geografiska områden det vill säga totalt *15 360* personer, se tabell 4.1.

Tabell 4.1 Stratifieringsvariabler i undersökningen Liv & Hälsa (2000)

Kön	**Ålder**	**Geografiska områden**	
Män	65-79 år	Askersund	Lindesberg
Kvinnor	50-64 år	Degerfors	Ljusnarsberg
	35-49 år	Hallsberg	Nora
	18-34 år	Hällefors	Örebro , Innerstaden
		Karlskoga	Örebro, Miljonprogrammet
		Kumla	Örebro, Villaområden
		Laxå	Örebro, Ytterområden
		Lekeberg	Örebro, Övriga områden

Datainsamlingen resulterade i *9 836* svar (svarsandelen *64 %*) varav *4 557* män (*59 %*) och *5 279* kvinnor (*69 %*), men varierade mellan *41* till *101* (*36 %* till *84 %*) mellan strata. Det

partiella bortfallet var mindre än *5 %* för de flesta frågorna. Såväl enhets- som partiellt bortfall hanteras här såsom beskrivits i avsnitt 2.3. En vidare diskussion om bortfallet återfinns i avsnitt 4.3. Övertäcknings- och undertäckningsfel bedömdes vara små. Övriga felkällor kommenteras ej här.

I den slutgiltiga datafilen återfanns registerdata från RTB och Utbildningsregistret (UTB), svar på huvudenkätens frågor, samt svar på frågor i en tilläggsenkät vilken var ämnad att besvaras av personer med inkontinensproblem/urinläckage. Svarande respektive icke-svarande på denna tilläggsenkät kom efter viss rensning att utgöra den beroende variabeln (INKONT) i regressionen.

Syftet med regressionen i det här presenterade exemplet var att studera associationer mellan inkontinens (INKONT) och en uppsättning variabler från huvudenkäten inklusive registerdata bland kvinnor i åldern 18-79 år i Örebro län. En övergripande hypotes var att inkontinenta är en grupp som är mer utsatta än andra. Därför används endast *64* strata för kvinnor baserat på de fyra åldersgrupperna samt de *16* geografiska områden enligt tabell 4.1. De två hjälpvariablerna som här använts var således dummyvariabler motsvarande åldersgrupp (S_ALDER#) där #=1,2,3,4, samt dummyvariabler motsvarande geografiskt område (S_OMR#) där #=1,2,….,16. Urvalsfraktionerna inom de 64 strata $f_h = \frac{n_h}{N_h}$ varierade mellan 0,03-0,29, och svarsandelen r_h mellan *55 %* till *83 %*. Av de svarande klassades *1 571* (*30 %*) kvinnor som inkontinenta och *3 708* (*70 %*) som friska. Jämförelsevis klassades *12 %* av männen som inkontinenta.

Efter en grundlig genomgång av tillgängliga variabler, inklusive omkodningar och sammanslagningar, återstod följande variabler vilka antingen var tydligt associerade med inkontinens eller kunde motiveras av urologens analysgrupp; ålder i år (ALDER); body mass index (BMI) beräknat som kvoten mellan en persons vikt i kg och kvadraten av längden i meter (kg/m^2); har under de 3 senaste månaderna upplevt värk i rygg/höfter och/eller under de 3 senaste månaderna upplevt värk i händer, armar, ben, knän eller fötter (FYS); har under de 3 senaste månaderna upplevt sömnproblem och/eller under de 3 senaste månaderna upplevt trötthet och kraftlöshet (PSYK); skulle klara att skaffa fram 18 000 kronor på en vecka och/eller under de 3 senaste månaderna haft svårt att klara löpande utgifter (EKON); har under de 3 senaste månaderna upplevt att någon behandlat dig nedlåtande och/eller under de 3 senaste månaderna upplevt att någon gjort dig till åtlöje inför andra (SKAM); har under de 3 senaste månaderna ansett dig vara i behov av läkarvård men inte sökt sådan (VARD). Av dessa variabler är ALDER den enda kontinuerliga variabel. Resten av variablerna kan betraktas som ordinala med antingen två eller tre kategorier, se bilaga 3 för en utförligare variabelbeskrivning. Då en variabel skrivs med ett nummer inom parentes efter, avser denna den kategori som anges i bilaga 3 och ej variabeln som helhet.

En variabel som hade varit av intresse men som inte fanns tillgänglig var antal genomgångna förlossningar. Anledningen att denna fråga aldrig kom med i huvudenkäten kan vara att denna analys inte var påtänkt vid studiens genomförande. Efter justering för bortfallet kvarstod *4 609* komplett observerade enheter motsvarande *60 %* av de ursprungligen dragna. Av dess var *1 332* (*29 %*) inkontinenta. Detta utgjorde därefter det undersökta datasetet, och designvikter beräknades för dessa såsom beskrivits i avsnitt 2.3.

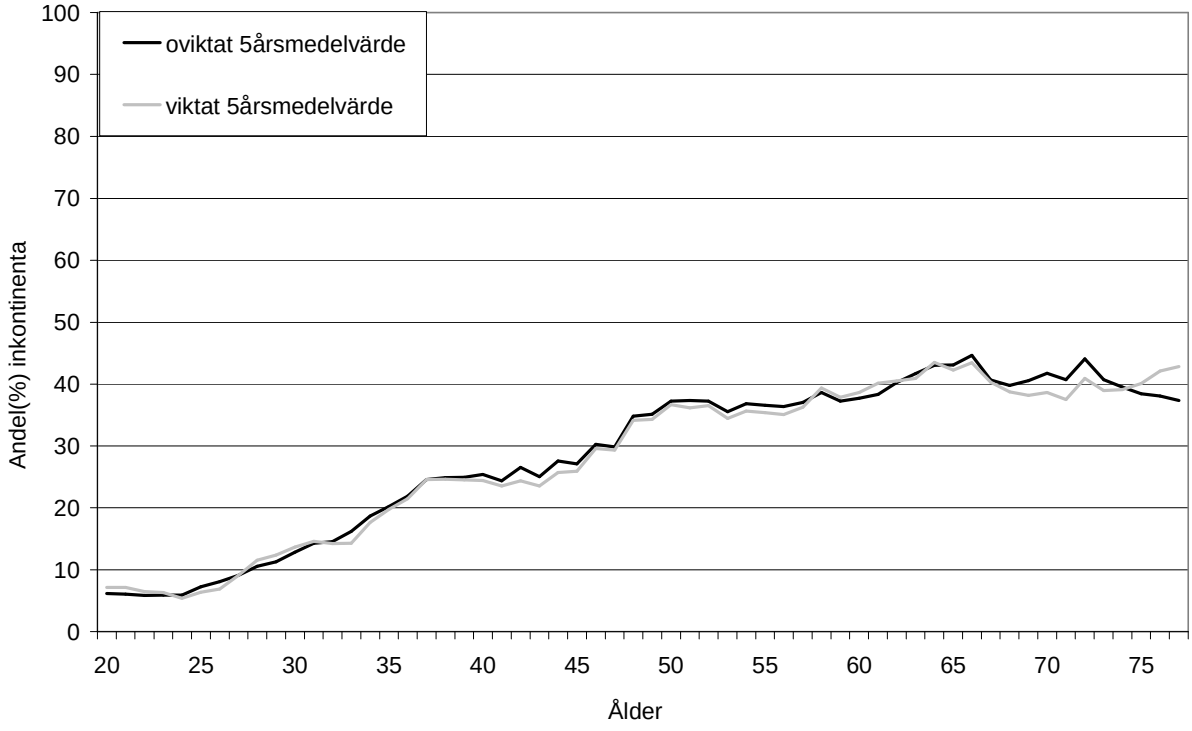
4.2 Modell, metod och resultat

Vid beräkning av den logistiska regressionsmodellen användes programvaran SAS 9.1. För modellbaserat synsätt användes kommandot PROC LOGISTIC och för designbaserat PROC SURVEYLOGISTIC. ML-estimat av regressionsparametrarna beräknades här genom Fisher scoring, se bilaga 2. I bilaga 4 återfinns en lista över tillgängliga modelltest för logistisk regression med PROC LOGISTIC och PROC SURVEYLOGISTIC i SAS 9.1, med undantag av plottar vilka ej kommenteras närmare i denna uppsats.

Utifrån ett modellbaserat (avsnitt 4.2.1) respektive designbaserat (avsnitt 4.2.2) synsätt presenteras hur variabler stegvis exkluderas baserat på högst p-värde för Wald-statistikan. Den första modellen (Mod1 respektive Des1) innehåller således samtliga variabler, och den sista (Mod9 respektive Des9) endast en.

Först undersöktes sambandet mellan variablerna med hjälp av korrelationsmått och korstabeller. Inga större skillnader sågs vid användandet av oviktade och viktade data med undantag av ALDER. Detta är också den enda kontinuerliga variabeln, varför endast denna var intressant att plotta mot de övriga. När denna plottades mot INKONT konstaterades en tydligt stigande tendens fanns upp till ca 50 års ålder, varefter kurvan blev mer flack. Denna effekt syntes något tydligare vid oviktade data, medan viktade data hade något högre värden för de allra högsta värdena på ALDER, se figur 4.1. Noterbart var att sambandet påminner om en logistisk S-kurva fram till dess att den flackar ut.

Efter prövning med ett flertal modeller innehållande olika exponenter och dummyvariabler valdes dock att modellera om ALDER genom att ålder över 50 år kodades som 50 år. Detta motiverades av att denna både var enkel och rimlig att tolka samt syntes ge liknande resultat som andra prövade kodningar och därför fanns vara robust. Härigenom erhölls en semikontinuerlig variabel vilken benämndes som ALDER_.



Figur 4.1 Andel inkontinenta över ålder för viktade och oviktade värden (glidande 5-årsmedelvärden).

4.2.1 Modellbaserat synsätt (MOD)

För att ta hänsyn till stratifieringen inkluderas här utöver de tilltänkta oberoende variablerna i *X* även stratifieringsvariablerna i *Z* som oberoende variabler. Eftersom dessa är ömsesidigt uteslutande variabler utesluts en från vardera uppsättning (S_ALDER1 och S_OMR1) vilka fick utgöra referenskategorierna. Både univariata oddskvoter samt nio olika modeller (Mod1-Mod9) där variabler successivt exkluderats presenteras i tabell 4.2. De p-värden som ges avser här Wald-test för enstaka kategorier eller variabler. De univariata oddskvoterna är signifikanta för samtliga ingående oberoende variabler med undantag av SKAM, vars oddskvoter också befann sig nära 1. Övriga variabler har enligt förväntan oddskvoter större än 1. Beträffande stratifieringsvariablerna hade S_ALDER# ett gemensamt p-värde<0,01 med höga oddskvoter (3,32-6,03), medan S_OMR# hade ett p-värde på 0,74 (0,84-1,23).

Tabell 4.2 Univariata och prövade modellers oddskvoter med PROC LOGISTIC

Oddskvoter	Univariata	Multivariata								
		Mod1	Mod2	Mod3	Mod4	Mod5	Mod6	Mod7	Mod8	Mod9
Intercept	0,40	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,02	0,02
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ALDER_	1,08	1,06	1,07	1,08	1,07	1,07	1,07	1,07	1,07	1,08
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
BMI (p-värde)	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
BMI (1)	2,99	2,05	2,06	2,06	2,06	2,11	2,12	2,14	2,42	
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
BMI (2)	1,63	1,19	1,19	1,20	1,19	1,20	1,20	1,20	1,28	
p-värde	0,00	0,03	0,03	0,02	0,02	0,02	0,02	0,02	0,00	
FYS (p-värde)	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00		
FYS (1)	2,50	1,58	1,58	1,59	1,61	1,64	1,83	2,02		
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00		
FYS (2)	1,48	1,29	1,29	1,29	1,31	1,32	1,39	1,43		
p-värde	0,00	0,02	0,02	0,02	0,01	0,01	0,00	0,00		
VARD (1)	1,63	1,37	1,38	1,37	1,41	1,45	1,54			
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00			
PSYK (p-värde)	0,00	0,00	0,00	0,00	0,00	0,00				
PSYK (1)	2,18	1,61	1,60	1,59	1,66	1,69				
p-värde	0,00	0,00	0,00	0,00	0,00	0,00				
PSYK (2)	1,54	1,48	1,47	1,47	1,51	1,53				
p-värde	0,00	0,00	0,00	0,00	0,00	0,00				
EKON (1)	1,31	1,38	1,37	1,37	1,42					
p-värde	0,00	0,00	0,00	0,00	0,00					
SKAM (p-värde)	0,70	0,00	0,00	0,00						
SKAM (1)	1,04	1,39	1,39	1,35						
p-värde	0,70	0,01	0,01	0,01						
SKAM (2)	1,06	1,28	1,28	1,24						
p-värde	0,41	0,00	0,00	0,00						
S_ALDER4 (1)	1,95		1,22							
p-värde	0,00		0,02							
S_ALDER#	3,32 - 6,03	1,34 - 1,78								
p-värde	0,00	0,07								
S_OMR#	0,84 - 1,23	0,87 - 1,26								
p-värde	0,74	0,84								

I Mod1 ingick samtliga variabler i *X*. Alla variabler hade p-värden<0,01 med oddskvoter större än 1 enligt förväntat. Enstaka kategorier hade dock högre p-värden. Noterbart är att oddskvoterna för SKAM ökat markant jämfört med de univariata oddskvoterna. Beträffande variablerna i *Z* hade S_ALDER# ett gemensamt p-värde på 0,07 och betydligt lägre oddskvoter (1,34-1,78) än de univariata. Detta berodde av att ALDER_ återfanns i modellen. S_OMR# hade däremot ett gemensamt p-värde på 0,84 och oddskvoter (0,84-1,23) vilka ungefär motsvarade de univariata.

Från Mod1 till Mod2 genomfördes en reduktion av variabler med ett p-värde>0,05. Detta innebar att samtliga kategorier ingående i *Z* reducerades med undantag av S_ALDER4, vilken är dummy för åldersgruppen 65-79. Denna är dock den variabel som har högst p-värde (0,02). Noterbart är att övriga oddskvoterna ändrats ytterst marginellt i jämförelse med Mod1. Även om förändringen av oddskvoten för ALDER_ var liten i absoluta tal bör betänkas att denna får en stor effekt då den var kontinuerlig upp till 50 år. Förändringen var naturlig då tre av dummyvariablerna i S_ALDER# reducerats. Samtliga bivariata interaktionseffekter testades, men ingen befanns ha ett p-värde<0,05.

Mod3 till och med Mod9 innebär att den variabel med högst p-värde reducerades. Noterbart var att Mod3 motsvarar den modell som skulle ha erhållits om inga stratifieringsvariabler inkluderats från början. Samtliga variabler i denna modell hade ett p-värde<0,01, och förändringar av oddskvoter var tämligen marginella jämfört med Mod2, möjligtvis med undantag för SKAM som sjunker något. Allteftersom variabler reduceras behöll samtliga ett p-värde<0,01. Oddskvoterna var i regel oförändrade eller ökade något. Detta syntes rimligt då variablerna var positivt korrelerade och i någon form kunde antas förklara utsatthet.

4.2.2 Designbaserat synsätt (DES)

Med PROC SURVEYLOGISTIC togs hänsyn till designinformation genom att i underkommandot specificera strata, populationstotaler och designvikter. På samma sätt som i avsnitt 4.2.1 prövades nio modeller (Des1-Des9) genom successiv exkludering av variabler, se tabell 4.3.

Först undersöktes samtliga univariata oddskvoter. Ingen större skillnad från 4.2.1 sågs här med undantag för S_ALDER4 och BMI (1), det vill säga dummy för 65-79 år samt kraftigt överviktiga, vilka befanns ha en något högre oddskvot.

För jämförbarhet med Mod1 inkluderades i den första modellen (Des1) utöver modellvariablerna i *X* även stratifieringsvariabler i *Z*. Liknande skillnader som sågs för Mod1 sågs även här i jämförelse med de univariata oddskvoterna. I direkt jämförelse med Mod1 sågs relativt små skillnader. Ett undantag var FYS som hade ett högre p-värde (0,01) och oddskvoterna som för båda kategorierna var lägre. Stratifieringsvariablerna S_ALDER# hade här som grupp också ett lägre p-värde (0,01) än i Mod1.

Till Mod2 reducerades sedan på samma sätt som i 4.2.1 stratifieringsvariablerna till dess att endast S_ALDER4 befanns ha ett p-värde <0,05. Samtliga variabler hade då ett p-värde<0,01. Förändringarna av oddskvoterna synes tämligen små från Mod1 med undantag av förändringen i ALDER_ vilken är samma förväntade förändring som ovan i 4.2.1. Även denna modell jämfördes med motsvarande i 4.2.1. Här sågs framförallt att oddskvoterna för FYS var lägre på än de i Mod1, och även S_ALDER4 var något större. Samtliga bivariata interaktionseffekter testades här, men ingen befanns ha ett p-värde<0,05.

Mod3 till och med Mod9 innebar på samma sätt som i 4.2.1 att den variabel med högst p-värde reducerades. Noterbart var att denna ordning skiljde sig något från i 4.2.1. Till exempel reduceras FYS först, och först därefter S_ALDER4. Även här skedde ungefär samma förändringar av oddskvoterna som i 4.2.1.

Tabell 4.3 Univariata och prövade modellers oddskvoter med PROC SURVEYLOGISTIC

Oddskvoter	Univariata	Multivariata								
		Des1	Des2	Des3	Des4	Des5	Des6	Des7	Des8	Des9
Intercept	0,39	0,01	0,01	0,01	0,01	0,01	0,01	0,01	0,02	0,02
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
ALDER_	1,08	1,05	1,07	1,08	1,08	1,08	1,07	1,07	1,07	1,08
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
BMI (p-värde)	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
BMI (1)	3,25	2,08	2,10	2,21	2,23	2,25	2,31	2,40	2,51	
p –värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
BMI (2)	1,72	1,18	1,18	1,22	1,24	1,23	1,24	1,26	1,28	
p-värde	0,00	0,05	0,05	0,02	0,01	0,01	0,01	0,01	0,00	
PSYK (p-värde)	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00		
PSYK (1)	2,18	1,65	1,64	1,78	1,76	1,85	1,89	2,12		
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00		
PSYK (2)	1,50	1,48	1,48	1,54	1,54	1,58	1,61	1,67		
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00		
VARD (1)	1,69	1,43	1,45	1,52	1,52	1,57	1,63			
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00			
EKON (1)	1,33	1,42	1,42	1,43	1,43	1,48				
p-värde	0,01	0,00	0,02	0,01	0,01	0,00				
SKAM (p-värde)	0,89	0,00	0,00	0,00	0,01					
SKAM (1)	0,99	1,38	1,37	1,38	1,31					
p-värde	0,91	0,01	0,02	0,02	0,04					
SKAM (2)	1,04	1,31	1,30	1,32	1,26					
p-värde	0,66	0,00	0,00	0,00	0,01					
S_ALDER4 (1)	2,17		1,30	1,31						
p-värde	0,00		0,00	0,01						
FYS (p-värde)	0,00	0,01	0,01							
FYS (1)	2,36	1,42	1,42							
p-värde	0,00	0,00	0,00							
FYS (2)	1,44	1,22	1,23							
p-värde	0,00	0,09	0,08							
S_ALDER#	3,23-6,36	1,62-2,09								
p-värde	0,00	0,01								
S_OMR#	0,74-1,34	0,83-1,23								
p-värde	0,17	0,64								

4.3 Diskussion om test, stratifieringshänsyn, samt kommentarer

Oavsett om MOD eller DES avses ses att samtliga likelihoodkvottest, benämnt såsom LR (modell), samt Score och Wald har p-värden<0,01 för samtliga prövade modeller, se bilaga 5. För likelihoodkvottest mellan två nestade modeller, LR (variabel), ses dock att skillnaden mellan Mod1 och Mod2 respektive Des1 och Des2, har klart högre p-värden. Detta bekräftar

de höga p-värdena för Wald avseende stratifieringsvariablerna som ses i tabell 4.2 och 4.3. Mod1 innehållande samtliga stratifieringsvariabler avvisas således tydligt.

Beträffande G-o-F ses att både Cox & Snell R^2 och Nagelkerke R^2 minskar successivt från ungefär samma nivåer, 0,12 respektive 0,17 för MOD och 0,12 respektive 0,18 för DES. Att minskningarna blir större ju mindre modellerna är stämmer väl med vad som kan förväntas då de oberoende variablernas gemensamma förklaring av den beroende variabeln delas på färre och färre variabler.

För MOD är AIC lägst för Mod2 och SC för Mod4. Mod1 har femte respektive nionde lägsta värde, vilket således talar emot denna modell som innehåller samtliga stratifieringsvariabler. För DES är AIC lägst för Des2 och SC för Des5. Des1 har fjärde respektive nionde lägsta värde, vilket således talar emot modellen innehållande samtliga stratifieringsvariabler.

Beträffande klassificeringsförmågan är nivåerna för MOD och DES i princip desamma. Förändringarna i Somer´s D, Kendall´s Tau-a och C är relativt små mellan modellerna med undantag för Mod7 till Mod9 respektive Des7 till Des9. Goodman-Kruskal Gamma ses i det närmsta oförändrad mellan modellerna. C kommer här att motsvara arean under ROC-kurvan. I jämförelse med Hosmer och Lemeshows tumregeln ger samtliga modeller ”*acceptable discrimination*” med undantag av Mod8 och Mod9 respektive Des8 och Des9, vilka ger ”*no discrimination*”. Den faktiska skillnaden synes dock liten mellan modellerna.

HL är endast möjligt att beräkna vid MOD. Statistikan hade i samtliga fall ett p-värde>0,05, vilket inte talar mot någon modell. Noterbart är att p-värdet är högst för Mod1 och Mod2 där S_ALDER4 ingår. Således syntes det som att passningen förbättras genom att inkludera denna variabel i modellen. Lägst p-värde hade Mod4, Mod7, och Mod9, vilket antyder att dessa kan passa data något sämre.

Baserat på modelltesten och G-o-F var det inte uppenbart vilken modell inom MOD respektive DES som var att föredra. Flera av testen talade dock för de större modellerna, undantaget Mod1 och Des1. Noterbart är att en okritisk användning av en signifikansnivå på 0,01 för likelihoodkvottestet, LR (variabel) skulle ha förkastat Mod2 men ej DES2. Det kan dock förmodas att en känslighetsanalys och närmare granskning skulle upptäcka och pröva detta samband ytterligare. Likaledes hade ett okritiskt användande av den tumregel som ges beträffande diskrimineringsförmågan förkastat de minsta modellerna, trots att den faktiska skillnaden syntes liten mellan de enskilda modellerna. Tumregeln att endast behålla variabler med univariata p-värden<0,25 hade tidigt sållat bort variabeln SKAM. Sammantaget talar detta för att modellerna bör granskas ytterligare, samt att med fackkunskap diskutera rimliga tolkningar. Två viktiga moment som ingår här är känslighets- och residualanalys.

Fokus i studien Liv & Hälsa (2000) var att undersöka sambanden mellan inkontinens och en uppsättning variabler. Designen kommer då att vara informativ om sambanden inom de olika strata skiljer sig åt mellan strata, samtidigt som urvalsfraktionerna skiljer sig åt. Olika urvalsfraktioner kan således förstärka respektive försvaga sambandet från olika strata, så att den totala effekten avviker från vad som är representativt för populationen som studeras.

Av de större modellerna innehåller endast Mod1 och Des1 samt Mod2 och Des2 samma uppsättning oberoende variabler. Här valdes att jämföra Mod2 och Des2 eftersom stratifieringsvariablerna har p-värden>0,05 i Mod1 och Des1. I tabell 4.4 återges skattningar av oddskvoter, parametrar och standardavvikelser, liksom kvoten mellan dessa skattningar.

Såsom konstaterats skiljde sig oddskvoterna ganska lite mellan MOD och DES, med störst skillnad för FYS och S_ALDER4. Eftersom ALDER_ är en semikontinuerlig kommer dock små skillnader i denna att innebära stora skillnader om en effekt motsvarande flera år avses upp till 50 års ålder. Exempelvis innebär en ökning av den skattade oddskvoten från 1,07 till 1,08 en ökning av den fleråriga oddskvoten mer än 1 % gånger antalet år. På grund av den av den icke-linjära transformationen blir den relativa skillnaden större om parameterskattningar studeras. Beträffande parameterskattningarnas standardavvikelser sågs att dessa alltid är större för Des2. Ökningen är mellan 8 % och 14%, omräknat till varianser mellan 16 % och 29 %.

Tabell 4:4 Skattningar av oddskvoter, parametrar och standardavvikelser för Mod2 och Des2

	Oddskvot			Parameterskattning			Standardavvikelse		
	MOD	DES	DES/MOD	MOD	DES	DES/MOD	MOD	DES	DES/MOD
Intercept	0,01	0,01	1,03	-5,20	-5,17	1,00	0,26	0,29	1,12
ALDER_	1,07	1,07	1,00	0,07	0,07	1,00	0,01	0,01	1,14
BMI (1)	2,04	2,10	1,03	0,71	0,74	1,04	0,10	0,11	1,08
BMI (2)	1,19	1,18	0,99	0,17	0,17	0,96	0,08	0,08	1,08
FYS (1)	1,58	1,42	0,90	0,46	0,35	0,77	0,10	0,12	1,10
FYS (2)	1,29	1,23	0,96	0,25	0,21	0,82	0,11	0,12	1,11
VARD (1)	1,38	1,45	1,05	0,32	0,37	1,15	0,09	0,09	1,09
PSYK (1)	1,60	1,64	1,03	0,47	0,49	1,06	0,11	0,12	1,12
PSYK (2)	1,47	1,47	1,00	0,39	0,39	1,00	0,11	0,13	1,12
EKON (1)	1,37	1,41	1,03	0,32	0,35	1,09	0,10	0,11	1,12
SKAM (1)	1,39	1,37	0,98	0,33	0,31	0,95	0,12	0,13	1,08
SKAM (2)	1,28	1,30	1,02	0,25	0,27	1,06	0,08	0,09	1,09
S_ALDER4	1,22	1,30	1,07	0,20	0,26	1,34	0,08	0,09	1,11

Att ta hänsyn till stratifieringen synes således ha relativt liten, men inte obefintlig, påverkan på skattningarna av regressionskoefficienterna, och således även på oddskvoterna. Beträffande den geografiska stratifieringen skulle denna förmodligen kunna betraktas som varande av administrativ karaktär, eftersom denna främst används för att studera delgrupper i de primära studierna. Det är därför mindre sannolikt att denna skulle vara informativ. Detta stärks också av de höga univariata p-värdena, samt höga p-värdena i Mod1 och Des1. Det synes därför rimligt att pröva om dessa kan exkluderas av effektivitetsskäl.

Ålder är intressant då denna är relaterad till den beroende variabeln, vilket ses i figur 4.1. Vid univariat analys fick också dummyvariablerna för samtliga åldersgrupper höga oddskvoter och låga p-värden. När däremot den semikontinuerliga variabeln inkluderas som en oberoende variabel hade endast dummyvariabeln 65-79 år ett lågt p-värde. Designvikterna syntes således inte att vara informativa för sambandet mellan ålder och den beroende variabeln i sig. Ett möjligt undantag är den äldsta åldersgruppen, vilket också kan vara orsaken till att p-värdet för dummyvariabeln 65-79 år blir något lägre vid DES än vid MOD. Det i sammanhanget höga p-värdet kan dock delvis bero av att kodningen för ålder kanske inte är perfekt, utan att dummyvariabeln fångar upp denna effekt.

De övriga sambanden förändrades relativt lite när hänsyn togs till designvikterna. Undantagen var regressionskoefficienterna för fysisk värk samt ovan kommenterade dummyvariabel för ålder 65-79. Detta antyder således att designen kan vara svagt informativ. Förändringarna syntes som sagt vara relativt små.

En förväntad effektivitetsförlust sågs i variansskattningarna när hänsyn togs till designvikterna. Som ses i tabell 4.4 gav Mod2 lägre variansskattningar än Des2, vilket kan antas vara en följd av att den senare tar hänsyn till designvikterna. Detta skulle kunna jämställas med en misspecificeringseffekt. Effekten syntes dock relativt liten. Anledningen härtill kan vara att spridningen i designvikter är relativt begränsad, men dock märkbar, varför effektivitetsförlusten också kan antas vara liten. Urvalet är också tämligen stort vilket innebär att ändlighetsjusteringar får liten betydelse.

Då residualanalys utifrån designbaserat synsätt inte är möjlig med SAS 9.1, men designen endast syntes svagt informativ, så torde dock en residualanalys utifrån ett modellbaserat synsätt kunna vara tillräcklig. De föreslagna testen i avsnitt 2.4.3 skulle också kunna genomföras. För att ytterligare validera en modell bör även annan typ av känslighetsanalys genomföras för att testa de underliggande antagandena. Detta beskrivs dock inte här.

Felkällor såsom bortfall bör också om möjligt undersökas närmare, och tillgängliga hjälpvariabler användas för att pröva justering för såväl partiellt som enhetsbortfall genom exempelvis poststratifiering. Om en bortfallsprocess lett till en selektion skulle även denna kunna vara informativ. Inkontinens upplevs ofta som något skamset vilket kan leda till att omfattningen underskattas, samt att individerna felklassificeras. Detta har dock förhoppningsvis begränsats av att postenkäten var anonym.

Då inkontinens även är korrelerat med andra sjukdomstillstånd finns här en uppenbar risk för selektion av svarande vid högre ålder. Risken finns därför för överskattning av antalet friska personer då dessa kan antas ha en högre svarsfrekvens. En högre andel med andra sjukdomstillstånd kan även antas var förknippat med högre dödligheten bland inkontinenta, vilket kan antas bidra till att figur 4.1 påminner om en S-kurva. Det är dock inte säkert att de skattade koefficienterna i den logistiska regressionmodell skulle påverkas av detta då sambanden fortfarande kan vara desamma bland till exempel svarande och icke-svarande.

Vad beträffar tolkningen av resultaten i själva modellerna var det tydligt hur ålder utgjorde den dominerande effekten. Detta ses bland annat på storleken av LR, Wald och Score, samt R^2 och klassificeringsmåtten, se bilaga 5. Att BMI är näst starkast i modellen synes också rimligt då denna liksom ålder är ett fysiologiskt mått. Övriga variabler kan antas mäta olika former av utsatthet, och utfallen syntes rimliga och tolkningsbara. Confounding torde föreligga, vilket kan förklara varför oddskvoterna förändras när ytterligare variabler inkluderas i modellerna. Inga interaktionseffekter befanns dock ha höga p-värden.

5 Sammanfattning och slutliga kommentarer

I detta kapitel sammanfattas och diskuteras resultatet av uppsatsen. Uppsatsens syfte var att ge en allmän beskrivning av binär logistisk regression samt att beskriva tillämpning av metoden på stratifierade surveydata. För att uppfylla syftet sammanfattades i kapitel 2 och 3 en litteraturstudie, vilken illustrerades i kapitel 4 med en tidigare genomförd hälsoenkät, Liv & Hälsa (2000).

Målet med en survey är kunna dra inferens om en målpopulation, där målparametrarna bör skattas med estimatorer som är effektiva och fria från bias eller konsistenta. Vid tillämpning på surveydata kan hänsyn behöva tas till designen om denna inte är ignorable. Detta kan vara fallet vid stratifiering, då populationen delats upp och olika urval sedan dragits från de olika delarna. Estimat kan annars innehålla bias och felaktiga varianser skattas, så att inferens utan hänsyn till inklusionssannolikheterna skiljer sig från inferens med hänsyn tagen till dessa. Närliggande är begreppet ickeinformativ. En design är ickeinformativ om sannolikhetsfördelningen för det erhållna urvalet överensstämmer med fördelningen för den valda modellen. En design som är ickeinformativ kommer alltid att vara ignorable, men däremot behöver en ignorable design inte vara ickeinformativ. Således räcker det att visa att en design är ickeinformativ för att den också ska vara ignorable.

Såväl modell- som designbaserat synsätt kan användas för att dra inferens vid surveydata. Här görs en åtskillnad mellan de två synsätten, trots att de i praktiken främst bör ses som komplement till varandra. Modellbaserat synsätt vilar på antagandet att en superpopulation genererat den population som urvalet dragits ifrån, den så kallade ξ-fördelningen. Målparametrarna härrör således till superpopulationen. Inferens vid designbaserat synsätt utgår istället från en ändlig population, där urvalsmekanismen är en funktion som tilldelar vardera av alla möjliga stickprov en känd sannolikhet, vilken benämns som p-fördelningen. Detta antagande omfattar således inte populationsparametrarna, och kan därför ses som svagare än vid modellbaserat synsätt. Ett urval från ett sannolikhetsurval kan dock genom att populationen betraktas som dragen från en superpopulation antas ha en ändlig målparametrar som närmar sig superpopulationsparametern. Denna kallas då för en censusparameter.

Designbaserat synsätt kan skydda mot felspecificering av modellen och mot informativ design. Även om den valda modellen inte är korrekt så kommer ett designbaserat synsätt ändå att ge konsistenta skattningar för den begränsade population som modellen avser, och variansskattningarna kan vara korrekta om komponenterna som ingår i skattningen av denna är korrekta. Modellbaserat synsätt riskerar dock att ge icke-konsistenta estimat med felaktiga variansskattningar om modellen är felspecificerad. Om däremot modellen är sann eller nästan sann kan modellbaserat synsätt tack vare parameterantagandena ge mer effektiva estimat än modellbaserat synsätt. Oavsett synsätt kommer estimaten att vara konsistenta med korrekta variansskattningar, även om designbaserat kan vara mindre effektiva. Här kan det alltså vara aktuellt med en avvägning mellan bias och effektivitet för att minimera MSE.

Binär logistisk regression kan användas för att beskriva sambandet mellan en dikotom beroende variabel och en eller flera kategoriska eller kontinuerliga oberoende variabler. Inom surveyanalys sker tillämpning således på analytiska survey. Utöver att undersöka associationer kan syftet även vara att prediktera eller klassificera ett utfall. Metoden används ofta inom hälsoundersökningar, där logit-transformationen och skattade koefficienter omräknade till oddskvoter ofta har relevanta tolkningar. Stratifierade urval är även vanligt förekommande.

Logistisk regression tillhör GLM och har inom denna familj flera likheter med linjär regression. Till skillnad från linjär regression antas dock ej att feltermerna är normalfördelade med konstant varians, varför skattningen av koefficienterna i logistisk regression i regel görs med maximum likelihood metoden. Logistisk regression har även många likheter med diskriminantanalys men antar ej att en multivariat normalfördelning för de oberoende variablerna föreligger. Vid endast en kategorisk oberoende variabel kan metoden reduceras till analys av en korstabell.

Ett flertal test finns utvecklade för att utvärdera logistiska regressionsmodeller. Dessa kan grovt delas upp i modelltest och test av goodness-of-fit. Många av modelltesten utgår från likelihoodfunktionen, medan goodness-of-fit i regel bygger på differensen mellan skattade och observerade värde för den beroende variabeln. Begreppet goodness-of-fit används dock inte entydigt i litteraturen. Till exempel definieras AIC, SC, och R^2 som goodness-of-fit av vissa författare trots att dessa är baserade på likelihoodfunktionen och ej på avvikelsen mellan skattade och observerade värde.

Stratifiering vid surveydata kan vara informativ vid logistisk regression om de associationer som undersöks ser olika ut i olika stratum. För att estimatorer ska vara konsistenta och effektiva krävs därför att detta tas hänsyn till. Här blir den praktiska distinktionen mellan de modell- och designbaserat synsätt tydlig. Utifrån modellbaserat synsätt krävs att hänsyn tas till om stratifieringen är informativ direkt i modellen. Ett sätt att göra detta är att inkludera stratifieringsvariablerna som oberoende variabler i modellen. Vid designbaserat synsätt används i regel designvikter för att ta hänsyn till skilda urvalssannolikheter, och såväl punkt- som variansestimat viktas. Härigenom förstoras urvalet så att en fiktiv total population erhålls. På grund av detta används psuedo maximum likelihood istället för maximum likelihood, då det senare kräver att likelihoodberäkningar baseras på faktiska och inte fiktiva observationer.

Vissa mått som kan användas vid logistiska regression finns endast tillgängliga vid modellbaserat synsätt. I SAS 9.1 finns till exempel inte Hosmer Lemeshow testet eller residualer tillgängliga vid designbaserat synsätt. En lösning som föreslagits här är att använda modellbaserat synsätt när designbaserat inte finns tillgängligt, och om möjligt modifiera detta. Det är ett område som framtida forskning kan fokusera på. Åtminstone så bör beprövade mått och test kunna integreras i programvara.

Ett annat intressant område kan vara att söka integrera design- och modellbaserat synsätt ytterligare, såsom föreslagits från bayesianskt håll. Detta är dock förmodligen mer intressant vid andra problem än de som diskuterats i denna uppsats, till exempel vid små urval eller olika typer av urvalsfel såsom bortfalls- och mätfel. En entydig teoretisk bas är också något som låter tilltalande.

Studien Liv & Hälsa (2002) som presenterades användes endast för att illustrera tillämpningen av binär logistisk regression samt hur stratifierade surveydata kan tas hänsyn till. Resultaten av studien i sig är dock inte del i uppfyllandet av uppsatsens syfte. Sammantaget var valet av modell här inte självklart. Exempelvis sågs hur en naiv tillämpning av tumregler eller beslutsregler baserade på signifikansnivåer kunde vara problematiska. Att döma av erhållna skattningar syntes stratifieringen endast att vara svagt informativ. Tolkningen av modellen var att ålder var den dominerande faktorn, men övriga variablers parametrar syntes rimliga och tolkbara. Viktigt för att slutligen avgöra om en modell är bra är dock att genomföra residual- och känslighetsanalys, något denna uppsats inte berört närmare.

Referenser

Biemer, P. P., Lyberg, L. E. (2003). Introduction to survey quality. Hoboken: John Wiley & Sons

Binder, D. A. (1983). "On the Variances of Asymptotically Normal Estimators from Complex Surveys". *International Statistical Review*, **51**, 279-292.

Binder, D. A., Roberts, G. R. (2001). “Can informative designs be ignorable?”. American statistical association. *Survey research methods section newsletter*, **12**, 1-3

Binder, D. A., Roberts, G. R. (2003). “Design-based and model-based methods for estimating model parameters”. Kapitel 3 i Chambers, R. L., Skinner, C. J., editors (2003).

Chambers, R. L., Skinner, C. J., editors (2003). Analysis of survey data. Chichester: John Wiley & Sons.

Chambers, R. L., Dorfman A. H., Sverchkov M.Y. (2003). “Nonparametric regression with complex survey data”. Kapitel 11 i Chambers, R. L., Skinner, C. J., editors (2003).

Cochran, W. G. (1977). Sampling techniques. 3e upplagan. New York: John Wiley & Sons.

Collet, D. (1991). Modelling binary data. London: Chapman & Hall.

Dahlenius, T. (1985). Elements of survey sampling. Stockholm: Swedish agency for research cooperation with developing countries.

Deming, W. E. (1950). Some theory of sampling. New York: John Wiley & Sons.

Fuller, W. A. (1984). “Least squares and related analyses for complex survey designs”. *Survey methodology*, **10**(1), 97-118.

Godambe, V.P., Thompson, M.E (1986). “Parameters of superpopulations and survey population: Their relationship and estimation.”. *International statistical review*, **54**(2), 37-59.

Hardy, Melissa A. (1993). Regression With Dummy Variables. Newbury Park, CA: Sage.

Hellström, L., Johansson P-G., Morander S., Tengstrand A. (2001). Elementär algebra. 2a upplagan. Lund: Studentlitteratur.

Hosmer, D.W., Lemeshow, S. (1998). “Logistic regression, conditional”. I Encyclopedia of biostatistics, volym 3. Chichester: John Wiley & Sons.

Hosmer, D.W., Lemeshow, S. (2000). Applied logistic regression. 2a upplagan. New York: John Wiley & Sons.

Kalton, G. (2002). “Models in the practice of survey sampling” *Journal of official statistics*, **18**(2), 129-154.

Kendall, P. L., Lazarsfield, P. F. (1950). Problems of survey analysis. I Merton, K., Lazarsfield, P. F., editors. Continuities in social research: studies in the scope and method of the 'The American soldier'. Chicago: Free press.

Kish, L. (1992). "Weighting for unequal P". *Journal of official statistics* **8**(2), 183-200.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E., Nizam, A. (1989). Applied regression analysis and multivariate methods. 3e upplagan. Pacific Grove: Brooks/Cole publishing company.

Korn, E. L., Graubard, B. I. (1999). Analysis of health surveys. New York: John Wiley & Sons.

Liv & Hälsa (2000). "Liv & Hälsa – en undersökning om hälsa, levnadsvanor och livsvillkor i Örebro". Teknisk rapport framställd av SCB för Örebro läns landsting.

Lee, E. S., Forthofer, R. N., Lorimor, R. J. (1989). Analyzing complex survey data. Sage university paper series on quantitative applications in the social sciences, 07-071. Newbury Park: Sage Pubns.

Little, R. J. A. (1991). "Inference with survey weights". *Journal of official* statistics, **7**(4), 405-424.

Little, R. J. A., Rubin, D. B. (2002). Statistical analysis with missing data. 2a upplagan. Hoboken: John Wiley & Sons.

Little, R. J. A. (2003). "To model or not to model? Competing modes of inference for finite population sampling". The university of Michigan department of biostatistics working paper series. Working paper 4. http://www.bepress.com/umichbiostat/paper4.

Lohr, S. L. (1999). Sampling: design and analysis. Pacific Grove: Brooks/Cole publishing company.

Pfeffermann, D. (1993). "The role of sampling weights when modelling survey data". *International statistical review*, **61**(2), 317-337.

Persson, A., Böiers, L-C. (2001). Analys i en variabel. 2a upplagan. Lund: Studentlitteratur.

Prentice, R.L., Pyke, R. (1979). "Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403-411.

Rao, J.N.K., Thomas, D.R. (2003). "Analysis of categorical response data from complex surveys: an appraisal and update". Kapitel 7 i Chambers, R. L., Skinner, C. J., editors (2003).

Roberts, G., Rao, J.N.K., Kumar, S. (1987). Logistic regression analysis of sample survey Data. *Biometrika*, **74**, 1-12.

Sharma, S. (1996). Applied multivariate techniques. Hoboken: John Wiley & Sons

Skinner, C. J., Holt D., Smith, T.M.F. (1989). Analysis of complex surveys. West Sussex: John Wiley & Sons.

Smith, T.M.F. (1978). “A model building approach to survey analysis”. Paper presented at the European meeting of statisticians, Oslo, August, 1978.

Smith, T.M.F. (1994). ”Sample surveys 1975-1990; an Age of Reconciliation?” (with discussion). *International statistical review*, **62**, 5-34.

Stock, J. H., Watson, M. W. (2003). Introduction to econometrics. New York: Pearson education.

Särndal, C-E. (1985). “How survey methodologists communicate”. *Journal of official statistics* **1**(1), 49-63.

Särndal, C-E., Swensson, B., Wretman, J. (1992). Model assisted survey sampling. New York: Springer-Verlag.

Tabachnick, B. G., Fidell, L. S. (2001). Using multivariate statistics. Needham Heights: Allyn & Bacon.

Valliant, R., Dorfman, A. H., Royall, R. M. (2000). Finite population sampling and inference: a prediction approach. New York: John Wiley & Sons.

Wolter, K. M. (1985). Introduction to variance estimation. New York: Springer-Verlag.

Bilaga 1 Matrisalgebra

Matriser är ett effektivt sätt att presentera omfattande ekvationssystem. Matrisalgebra skiljer sig från skalär algebra genom att en uppsättning av liknande ekvationer uttrycks med hjälp av en ekvation, vilken sedan kan användas vid beräkning av andra matrisuttryck.

En linjär regressionmodell bestående av *n* ekvationer och *k* parametrar uttrycks normalt i kortform som

$$y_t = \beta_0 + \beta_1 x_{t,1} + ... + \beta_{k-1} x_{t,k-1} + e_t \text{ där } t = 1,...,n.$$

Ekvationssystemet som helhet kan således skrivas som:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{1,1} + ... + \beta_{k-1} x_{k-1,1} + e_1 \\
y_2 &= \beta_0 + \beta_1 x_{1,2} + ... + \beta_{k-1} x_{k-1,2} + e_2 \\
&..... \\
y_n &= \beta_0 + \beta_1 x_{1,n} + ... + \beta_{k-1} x_{k-1,n} + e_n
\end{aligned}$$

Motsvarande ekvationssystem skulle i utvecklad matrisfrom kunna skrivas som (med förenklad matrisform under själva matrisuttrycket inklusive dimensioner inom klammer):

$$\begin{bmatrix} Y_1 \\ Y_2 \\ .... \\ .... \\ .... \\ Y_n \end{bmatrix} = \begin{bmatrix} 1\ X_{1,1} ... X_{k-1,1} \\ 1\ X_{1,2} ... X_{k-1,2} \\ ..... \\ ..... \\ ..... \\ 1\ X_{1,n} ... X_{k-1,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ .... \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} e_1\ e_2\ ...\ ...\ ...\ e_n \end{bmatrix}.$$

$$y_{[n\times 1]} = \quad X_{[n\times k]} \quad \beta_{[k\times 1]} + \quad e_{[1\times n]}$$

I den förenklade matrisformen anger första dimensionssiffran antal rader och den andra antalet kolumner. Såväl *y*, *X*, β och *e* är här olika matriser. Då *y* och β har kolumndimensionen 1 kallas dessa för kolumnvektorer och *e* som har raddimensionen 1 kallas för en radvektor. En matris med både rad- och kolumndimensionen 1 kallas en skalär. Normal skaläralgebra är således ett specialfall av matrisalgebra.

Givet vissa förutsättningar så kan de fyra vanliga räknesätten tillämpas på matriser. Vid matrismultiplikation krävs att kolumndimensionen i den första termen är av samma ordning som raddimensionen i den andra termen. Om matrisen ovan studeras ses att så är fallet då *X* innehåller *k* kolumner och β *k* rader, samt β innehåller 1 kolumn och *e* 1 rad.

Genom multiplikation av matriserna X och β kan således en ny matris $X\beta$ bildas. Denna matris kommer då att ha raddimension av ordning motsvarande *X* raddimension och kolumndimension av ordning motsvarande β:s kolumndimension.

Denna nya matris är således en kolumnvektor med *n* rader, där elementen i rad *n* beräknas som produktsumman av element i rad *n* från X och elementen i kolumn *k* från β. Denna beräkning upprepas för varje element i X.

$$\begin{bmatrix} 1\times\beta_0 + X_{1,1}\times\beta_2 + ... + X_{k-1,1}\times\beta_{k-1} \\ 1\times\beta_0 + X_{1,2}\times\beta_2 + ... + X_{k-1,2}\times\beta_{k-1} \\ ..... \\ ..... \\ ..... \\ 1\times\beta_0 + X_{1,n}\times\beta_2 + ... + X_{k-1,n}\times\beta_{k-1} \end{bmatrix}.$$

$$X\beta_{[n\times 1]}$$

Matrisen X kan transponeras till en ny matris genom att elementet i rad *n* och kolumn *k* istället placeras i rad *k* och kolumn *n*, vilket upprepas för varje element i X. Härigenom erhålls matrisen X' med *k* rader och *n* kolumner, vilken benämns som transponatet av X. Transponering möjliggör att X kan multipliceras med X' för att erhålla matrisen XX' med både rad- och kolumndimensionerna *k*. En sådan matris kallas för en kvadratisk matris, och motsvarar X^2 vid skalär algebra.

En kvadratisk matris *KM* där alla andra element än de som återfinns i diagonalen är 0 kallas för en diagonalmatris.

$$KM = \begin{bmatrix} 3\ 0\ ...\ 0 \\ 0\ 7\ ...\ 0 \\ ...\ ...\ ...\ ... \\ 0\ 0\ ...\ 2 \end{bmatrix}$$

En diagonalmatris med elementen 1 i diagonalen betecknas som en enhetsmatris *EM*. En enhetsmatris av storleken 1×1 är således talet 1 vid skalär algebra.

$$EM = \begin{bmatrix} 1\ 0\ ...\ 0 \\ 0\ 1\ ...\ 0 \\ ...\ ...\ ...\ ... \\ 0\ 0\ ...\ 1 \end{bmatrix}.$$

Den kvadratiska matrisen Δ sägs vara inverterbar om det gäller att $\Delta\Delta^{-1} = \Delta^{-1}\Delta = E$ där Δ^{-1} är matrisinversen av Δ. En förutsättning för att en matris ska vara inverterbar är att den är kvadratisk.

För en utförligare beskrivning av matrisalgebra, se till exempel Hellström, Johansson, Morander och Tengstrand (2001).

Bilaga 2 Maximum likelihood metoden

Denna bilaga baseras på Collet (1991). Maximum likelihood (ML) är en vanligt förkommande estimationsmetod då ML-estimatorer i många fall är konsistenta och effektiva. Dess samplingfördelningar är asymptotiskt normalfördelade, och estimaten kan vid stora urval även antas vara approximativt normalfördelade. Vid skattning av regressionsparametrar kräver ML till skillnad från OLS inte att feltermerna är likafördelade. ML-estimatorer har därför många egenskaper som gör dem användbara, till exempel vid logistisk regression.

Metoden utgår från att estimaten väljs så att om de hade varit sanna, så skulle sannolikheten maximeras för att observera de data som har observerats. Vid skattning av parameterarna β givet observerade data Y är likelihoodfunktionen $L(\beta \mid Y)$ produkten av täthetsfunktionerna $f(y_t \mid \beta)$ för de t enskilda enheterna. Likelihoodfunktionen består således av produkten av likelihooden för varje observation t, och kräver således att samtliga dessa kan beräknas.

Olika algoritmer kan användas för att beräkna ML-estimaten. Genom att först logartimera $L(\beta \mid Y)$ och därefter derivera denna med avseende på de $j = 1,2,...k$ parametrarna, $\frac{\partial \log L(\beta \mid Y)}{\partial \beta_j}$, erhålls k score vilka tillsammans skrivs $u(\beta)$. Vardera score sätts sedan lika med noll för att erhålla de $j = 1,2,...k$ score-ekvationerna $\frac{\partial \log L(\beta \mid Y)}{\partial \beta_j} = 0$ vilka kan sammanfattas som $u(\beta) = 0$.

De partiella derivatorna med avseende på de k ingående parametrarna benämns som Hessianmatrisen $H(\beta)$ av storleken $k \times k$. För varje element (j_1, j_2), där $j_1 = 1,2,...k$ och $j_2 = 1,2,...k$, beräknas således $\frac{\partial^2 \log L(\beta \mid Y)}{\partial \beta_{j_1} \partial \beta_{j_2}}$.

Matrisen av de förväntade värdena av elementen i $H(\beta)$ multiplicerad med -1 kallas för informationsmatrisen, $I(\beta)$. Elementet (j_1, j_2) skrivs här som $-E\left\{\frac{\partial^2 \log L(\beta \mid Y)}{\partial \beta_{j_1} \partial \beta_{j_2}}\right\}$. Inversen av $I(\beta)$ är då den asymptotiska kovariansmatrisen för ML-estimaten, $I^{-1}(\hat{\beta})$.

Skattningar av β genom Newton-Raphson, kan erhållas från $u(\beta) = 0$ medelst utveckling av en Taylorserie. Genom att β^* antas befinna sig nära $\hat{\beta}$ görs approximationen $u(\hat{\beta}) \approx u(\beta^*) + H(\beta^*)(\hat{\beta} - \beta^*)$. Härav följer att $\hat{\beta} \approx \beta^* - H^{-1}(\beta^*)u(\beta^*)$. Genom en iterativ process där det $(r+1)$ estimatet för β ges som $\hat{\beta}_{r+1} = \hat{\beta}_r - H^{-1}(\hat{\beta}_r)u(\hat{\beta}_r)$ erhålls ML-estimat.

Alternativt kan ML-estimat erhållas med Fisher scoring tekniken. Här ersätts $H(\beta)$ med $-I(\beta)$, och därefter erhålls estimaten iterativt genom $\hat{\beta}_{r+1} = \hat{\beta}_r + I^{-1}(\hat{\beta}_r)u(\hat{\beta}_r)$.

Bilaga 3 Variabelbeskrivning - Liv & Hälsa (2000)

Tabell B3.1 Variabelbeskrivning för Liv & Hälsa (2000)

n	Variabel*	Kategori
3277	INKONT (1)	Har ej besvarat tilläggsenkäten eller bedömts som frisk
1332	INKONT (2)	Har besvarat tilläggsenkäten och ej bedömts som frisk
4609	ALDER_	Ålder 18-50 år. Ålder 51-79 kodad som 50
635	BMI (1)	Kraftigt överviktig (BMI>30)
1476	BMI (2)	Överviktig (BMI 25-29,99)
2498	BMI (3)	Under/normalviktig (BMI<25) (referens)
2207	PSYK (1)	Har minst någon enstaka gång under de 3 senaste månaderna upplevt trötthet och kraftlöshet samt sömnproblem
1628	PSYK (2)	Har minst någon enstaka gång under de 3 senaste månaderna upplevt trötthet och kraftlöshet eller sömnproblem
774	PSYK (3)	Under de 3 senaste månaderna ej besvärats av trötthet och kraftlöshet eller av sömnproblem (referens)
2126	FYS (1)	Har minst någon enstaka gång under de 3 senaste månaderna upplevt värk i rygg,höft samt värk i händer,armar,ben,knän,fötter
1528	FYS (2)	Har minst någon enstaka gång under de 3 senaste månaderna upplevt värk i rygg,höft eller värk i händer,armar,ben,knän,fötter
955	FYS (3)	Under de 3 senaste månaderna ej besvärats av värk i rygg,höft eller värk i händer,armar,ben,knän,fötter (referens)
615	EKON (1)	Har minst någon enstaka månad under de 3 senaste månaderna haft svårt att klara löpande utgifter samt skulle inte i en oförutsedd situation klara att skaffa fram 18´ kronor på en vecka
3994	EKON (2)	Har ej minst någon enstaka månad under de 3 senaste månaderna haft svårt att klara löpande utgifter samt skulle inte i en oförutsedd situation klara att skaffa fram 18´ kronor på en vecka (referens)
520	SKAM (1)	Har minst någon gång under de 3 senaste månaderna upplevt att någon behandlat dig på ett nedlåtande sätt samt att någon gång gjort dig till åtlöje inför andra
1561	SKAM (2)	Har minst någon gång under de 3 senaste månaderna upplevt att någon behandlat dig på ett nedlåtande sätt eller att någon gång gjort dig till åtlöje inför andra
2528	SKAM (3)	Har under de 3 senaste månaderna ej upplevt att någon behandlat dig på ett nedlåtande sätt eller gjort dig till åtlöje inför andra (referens)
955	VARD (1)	Har minst någon gång under de 3 senaste månaderna ansett dig vara i behov av läkarvård men inte sökt sådan
3654	VARD (2)	Har ej under de 3 senaste månaderna ansett dig vara i behov av läkarvård men ändå inte sökt sådan (referens)

* Numret inom parentes efter ett variabelnamn anger kategori inom respektive variabel.

Bilaga 4 Tillgängligt för binär logistisk regression i SAS 9.1

Tabell B4.1 Tillgängligt i SAS 9.1 för binär logistisk regression med undantag plottar

	PROC LOGISTIC	PROC SURVEYLOGISTIC	Kommando	Beskrivet i avsnitt
Modell- och koefficienttest				
-2logL	Ja	Ja	*	3.4.1
Likelihood Ratio	Ja	Ja	*	3.4.1
Wald	Ja	Ja	*	3.4.2
Score	Ja	Ja	*	3.4.3
Wald konfidensintervall	Ja	Ja	*	3.2
LR konfidensintervall	Ja	Nej	CLPARM	
Goodness-of-fit				
Pearson chisquare	Ja	Nej	SCALE	3.5**
Deviance	Ja	Nej	SCALE	3.5**
Williams	Ja	Nej	SCALE	
Hosmer Lemeshow G o F	Ja	Nej	LACKFIT	3.5.1
Klassificeringstabell	Ja	Ja	CTABLE	3.5.2
Somer´s D	Ja	Ja	*	3.5.3
Goodman-Kruskal Gamma	Ja	Ja	*	3.5.3
Kendall´s Tau-a	Ja	Ja	*	3.5.3
C	Ja	Ja	*	3.5.3
area under ROC	Ja	Ja	OUTROC	3.5.2
Cox & Snell R2	Ja	Ja	RSQ	3.5.3
Nagelkerke R2	Ja	Ja	RSQ	3.5.3
AIC	Ja	Ja	*	3.5.3
SC	Ja	Ja	*	3.5.3
Residualer				
Influence	Ja	Nej	INFLUENCE	
hat matrix	Ja	Nej	INFLUENCE	
Pearson residual	Ja	Nej	INFLUENCE	
Deviance residual	Ja	Nej	INFLUENCE	
DFBETA	Ja	Nej	INFLUENCE	
C and CBAR (based on Cook)	Ja	Nej	INFLUENCE	
DIFDEV	Ja	Nej	INFLUENCE	
DIFCHISQ	Ja	Nej	INFLUENCE	

* Beräknas alltid. ** Endast refererat till.

Bilaga 5 Test i Liv & Hälsa (2000)

Tabell B5.1 Modelltest och goodness-of-fit för MOD med PROC LOGISTIC

	Mod1	Mod2	Mod3	Mod4	Mod5	Mod6	Mod7	Mod8	Mod9	Intercept*
# parametrar	29	12	11	9	8	6	5	3	1	
-2 LOGL	4963,41	4974,89	4980,28	4990,86	5002,23	5025,98	5051,89	5110,15	5193,42	5542,36
Modelltest										
LR (variabel)	11,48	5,38	10,58	11,38	23,74	25,91	58,26	83,27	348,94	
p-värde	0,83	0,02	0,01	0,00	0,00	0,00	0,00	0,00	0,00	
LR (modell)	578,95	567,46	562,08	551,50	540,12	516,38	490,47	432,21	348,94	
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
Wald	450,99	444,23	437,10	431,85	427,73	409,52	390,97	344,07	270,89	
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
Score	523,52	511,24	503,32	495,69	488,19	467,22	443,34	389,76	304,37	
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
G-o-F										
Cox & Snell R^2	0,12	0,12	0,11	0,11	0,11	0,11	0,10	0,09	0,07	
Nagelkerke R^2	0,17	0,17	0,16	0,16	0,16	0,15	0,14	0,13	0,10	
AIC	5023,41	5000,89	5004,28	5010,86	5020,23	5039,98	5063,89	5118,15	5197,42	5544,36
SC	5216,48	5084,56	5081,51	5075,21	5078,16	5085,03	5102,50	5143,89	5210,29	5550,79
Somer´s D	0,43	0,43	0,43	0,42	0,42	0,41	0,40	0,36	0,30	
Goodman-Kruskal Gamma	0,43	0,43	0,43	0,43	0,42	0,42	0,41	0,41	0,43	
Kendall´s Tau-a	0,18	0,18	0,18	0,17	0,17	0,17	0,16	0,15	0,13	
C	0,72	0,71	0,71	0,71	0,71	0,70	0,70	0,68	0,65	
HL statistika	2,35	3,20	7,00	10,50	6,44	5,93	10,66	4,61	5,47	
HL frihetsgrader	8	8	8	8	8	8	8	7	4	
p-värde	0,97	0,92	0,54	0,23	0,60	0,66	0,22	0,71	0,24	

** Intercept avser modell där endast ett intercept ingår.*

Tabell B5.2 Modelltest och goodness-of-fit för DES med PROC SURVEYLOGISTIC

	Des1	Des2	Des3	Des4	Des5	Des6	Des7	Des8	Des9	Intercept*
# parametrar	29	12	10	9	7	6	5	3	1	
-2 LOGL	4836,78	4850,52	4862,34	4871,55	4882,25	4896,05	4928,75	4981,40	5066,04	5449,00
Modelltest										
LR (variabel)	13,74	11,82	9,21	10,71	13,80	32,69	52,65	84,64	382,96	
p-värde	0,69	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
LR (modell)	612,22	598,48	586,66	577,45	566,75	552,95	520,25	467,60	382,96	
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
Wald	364,28	352,45	345,76	330,24	322,54	324,22	305,54	265,71	209,50	
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
Score	557,84	542,50	532,37	519,18	511,21	501,76	472,20	424,51	336,29	
p-värde	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	
G-o-F										
Cox & Snell R^2	0,12	0,12	0,12	0,12	0,12	0,11	0,11	0,10	0,08	
Nagelkerke R^2	0,18	0,18	0,17	0,17	0,17	0,16	0,15	0,14	0,12	
AIC	4896,78	4876,52	4884,34	4891,55	4898,25	4910,05	4940,75	4989,40	5070,04	5451,00
SC	5089,86	4960,18	4955,13	4955,90	4949,74	4955,10	4979,36	5015,14	5082,92	5457,44
Somer´s D	0,43	0,43	0,42	0,42	0,41	0,41	0,39	0,36	0,30	
Goodman-Kruskal Gamma	0,43	0,43	0,42	0,42	0,42	0,42	0,41	0,41	0,43	
Kendall´s Tau-a	0,18	0,18	0,17	0,17	0,17	0,17	0,16	0,15	0,13	
C	0,72	0,71	0,71	0,71	0,71	0,70	0,70	0,68	0,65	

** Intercept avser modell där endast ett intercept ingår.*