EKONOMIHÖGSKOLAN
Lunds universitet
Statistiska institutionen

Matrismodellen vs Two-part regressionsmodeller -effekter på Region Skånes resursfördelning-

Av: Jennifer Ericsson

Uppsats i Statistik
15 hp
Nivå 61-90 poäng
September 2007

Handledare: Mats Hagnell
Bihandledare: Juan Merlo

ABSTRACT

An important task for Region Skåne is to allocate resources to the health care districts. From 1999 to 2002 Region Skåne used needs-based resource allocation as a model for allocating resources. In a needs-based resource allocation individuals with the same socioeconomic and demographic characteristics are assumed to have the same level of need and are therefore allocated the same amount of resources. During the period of needs-based resource allocation a matrix model was used as a method. In the matrix model individuals were divided into cells after each combination of the socioeconomic and demographic variables. Mean costs in each cell were then calculated and summed for each health care district.

An alternative to the matrix model is to use regression analysis. However, the dependent variable health care cost is characterized by a large fraction of individuals with zero costs and few individuals with very high costs; hence health care cost has a highly skewed distribution. Health care cost is therefore assumed to have a mixed distribution; i.e. it is both discrete and continuous. Two-part models are specially developed for this type of distribution. By applying a two-part model a more precise resource allocation is assumed to be accomplished.

In this thesis the matrix model is compared to two different specifications of the two-part model but also with an ordinary multiple regression model. The focus is on how the different models affect the resource allocation. The result shows that the two-part models allocate fewer resources in total than the other models. The conclusion is that the advantage of applying a two-part model is low. This is due to the fact that the analysis of the two-part model is complicated, both theoretically and practically.

INNEHÅLLSFÖRTECKNING

1. Inledning

1.1 Bakgrund

Region Skåne har som en del i sin verksamhet att fördela resurser till sjukvården. Det finns dock flera olika metoder för att fördela resurser till sjukvården. Ett sätt är ersättning genom att resurser fördelas med en given summa per individ oavsett hur mycket vård som faktiskt utförs. Ersättningen till sjukvårdsdistrikt för varje individ skulle då kunna utgöras av den genomsnittliga kostnaden för alla individer i Skåne. Denna metod tar dock ingen hänsyn till att individer kan ha olika behov av sjukvård. Ett alternativ är istället att anta att individer med samma socioekonomiska och demografiska karakteristiker har samma behov. Resurser fördelas därmed istället efter genomsnittliga kostnader för individer med samma karakteristiker. Denna metod kallas behovsbaserad resursfördelning. Ett problem är dock att det inte är möjligt att fullt ut tillgodose individers exakta behov. Målet med den behovsbaserade resursfördelningen är således att individer med samma behov ska ges tillgång till lika vård givet sjukvårdens begränsade resurser.

Givet att en behovsbaserad resursfördelning används, behövs en modell för att fördela resurserna. I Region Skåne användes under åren 1999 till 2002 en behovsbaserad resursfördelningsmodell, där resurser fördelades med en så kallad matrismodell. I matrismodellen beräknas genomsnittliga kostnader för individer med samma kombination av karakteristiker. Den genomsnittliga kostnaden multipliceras sedan med antalet personer i varje sjukvårdsdistrikt med den givna kombinationen av karakteristiker.

Istället för att använda en matrismodell för att beräkna genomsnittliga kostnader för individer med samma karakteristiker skulle en regressionsmodell kunna användas. Sjukvårdskostnader karakteriseras dock ofta av en stor andel individer med noll kostnader och ett fåtal individer med mycket höga kostnader vilket ger en starkt skev fördelning. Dessutom förekommer ofta heteroskedastisitet, det vill säga icke-konstant residualvarians. Detta gör att en regressionsanalys av sjukvårdskostnader blir komplicerad. För att komma till rätta med problem som skev fördelning och heteroskedastisitet vid analys av sjukvårdskostnader kan en så kallad two-part modell användas. I en two-part modell sker analysen i två steg. Den första delen modellerar sannolikheten att en individ har kostnader och den andra delen storleken på dessa. I den första delen används oftast en logistisk regression medan en multipel regression med logaritmerad beroende variabel är vanligast i den andra delen. Det kan hävdas att en two-part modell stämmer bättre överrens med sjukvårdskostnaders underliggande fördelning vilket i sådana fall skulle leda till bättre skattning av kostnader.

1.2 Syfte

Att modellera behov, det vill säga använda de variabler som bäst förklarar skillnader i behov, på ett korrekt sätt är väsentligt för att en behovsbaserad resursfördelning ska fungera. Denna uppsats utgår dock från Region Skånes modellering av behov och syftar inte till att utvärdera huruvida denna kombination av variabler är korrekt eller ej. Istället behandlar denna uppsats effekter av olika modeller för en behovsbaserad resursfördelning. Syftet med denna uppsats är således att jämföra en matrismodell med two-part regressionsmodeller avseende effekter på resursfördelningen.

1.3 Avgränsningar

I denna studie används endast data från år 1999. Materialet består endast av individer mellan 45 och 64 år då det var detta material som stod till förfogande.

1.4 Disposition

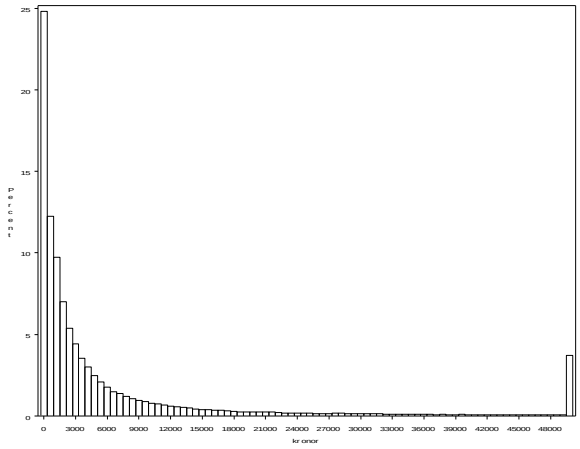
I kapitel två beskrivs populationen samt variablerna som används i de olika modellerna. Kapitel tre till fem ger en ingående beskrivning av matrismodellen, multipel linjär regression vid analys av sjukvårdskostnader samt two-part modellen samtidigt som det i respektive kapitel utvecklas modeller. De tre olika typerna av modeller jämförs i kapitel sex för att sammanfattas i en diskussion i kapitel sju.

2. Data

2.1 Variabler och population

I denna studie har Region Skånes register för resursfördelning från år 1999 använts. Detta register består av uppgifter, från Statistiska Centralbyrån, om sociodemografiska och socioekonomiska variabler på individnivå. I registret finns antalet personer boende i Region Skåne 1999-12-31, vilket ger totalt 1 153 633 individer. På grund av materialets storlek analyseras i denna uppsats endast individer mellan 45-64 år, vilket ger totalt 285 431 individer.

För varje individ finns i registret uppgifter om kostnader för sjukvård. Kostnaden för varje individ inkluderar både öppen och sluten vård, men exkluderar läkemedelskostnader (som fördelas via en separat budget). I tabell 1 finns beskrivande statistik för sjukvårdskostnader. Sjukvårdskostnader har en lång högersvans vilket histogrammet i figur 1 visar. Även toppigheten är hög vilket innebär högre sannolikhet för extrema värden än om variabeln skulle vara normalfördelad. Variabeln sjukvårdskostnader används som beroende variabel i regressionsmodellerna.



Figur 1 *Histogram över sjukvårdskostnader*

Variabel	Medelvärde	Summa	Min	Max	Skevhet	Toppighet
Sjukvårdskostnader N=285431	9320[1]	2660229234	0	3106236	17.6	633

Tabell 1 *Beskrivande statistik för variabeln sjukvårdskostnader, i kronor*

De socioekonomiska och sociodemografiska uppgifterna för varje individ som registret innehåller används som förklarande variabler. De socioekonomiska variablerna är sysselsättning, inkomst, boendetyp och utbildningsnivå och de sociodemografiska är ålder, kön och civilstånd. Från de ursprungliga uppgifterna i Statistiska Centralbyråns register har sammanslagningar i olika grupper gjorts. I tabell 2 finns indelningen av variabler som används i matrismodellen. Då variablerna utbildningsnivå och inkomstgrupp innehåller tre kategorier bildas i regressionsmodellerna två dummyvariabler för varje variabel. Tabell 3 visar variablerna som används i regressionsmodellerna.

[1] Medelkostnaden för hela Skåne år 1999 var 8979 kr (Lithman, 2001)

Variabel	Värden
Ålder	45-54, 55-64
Kön	man, kvinna
Civilstånd	gift/registrerad partner, ogift/skild/änka/änkling
Sysselsättning	sysselsatta med kontrolluppgift, personer med kontrolluppgift ej sysselsatta
Inkomstgrupp	nollinkomsttagare, under medianinkomst, över eller lika med medianinkomst
Boendetyp	småhus och jordbruksfastighet, övriga fastighetstyper
Utbildningsnivå	folkskola/grundskola/uppgift saknas, gymnasieskola, högskola/forskarutbildning

Tabell 2 *Variabler i matrismodellen*

Variabel	Typ	Värden	Andel
Ålder	Dikotom	1 = 55-64 år	55,1
		0 = 45-54 år	44,9
Kön	Dikotom	1 = Man	49,9
		0 = Kvinna	50,1
Civilstånd	Dikotom	1 = Ogift, skild, änka, änkling	35,5
		0 = Gift, registrerad partner	64,5
Sysselsättning	Dikotom	1 = Ej förvärvsarbetande	5,4
		0 = Förvärvarbetande	94,7
Inkomst1	Dikotom	1 = Under medianinkomst	29,5
		0 = Annars	70,6
Inkomst2	Dikotom	1 = Över eller lika med medianinkomst	66,6
		0 = Annars	33,4
Boendetyp	Dikotom	1 = Småhus, jordbruksfastighet	66,5
		0 = Övriga fastighetstyper	33,5
Utbildning1	Dikotom	1 = Högskola, forskarutbildning	26,4
		0 = Annars	73,7
Utbildning2	Dikotom	1 = Gymnasieutbildning	40,5
		0 = Annars	59,5
Vårdtung diagnos	Dikotom	1 = Vårdtung diagnos	6,5
		0 = Ej vårdtung diagnos	93,5

Tabell 3 *Förklarande variabler i regressionsmodellerna*

I materialet finns även uppgifter om individen har någon vårdtung diagnos. Vårdtunga diagnoser består av grupper av diagnoser som anses kostsamma för samhället. De vårdtunga diagnoserna med ICD-10 koder[2] finns i tabell 4 och följer indelningen från Behov och resurser i vården (SOU 1996:163). Två vårdtunga diagnoser, avlidna under året samt astma och kronisk obstruktiv lungsjukdom, har exkluderats på grund av att de inte fanns med i materialet. Tre diagnosgrupper har lagts till, dessa finns beskrivna i tabell 5. Andelen individer mellan 45-64 år med någon av de vårdtunga diagnoserna beskrivna i tabell 4 och 5 är 6,6 procent. Denna grupp av individer står för 47 procent av de totala kostnaderna.

Vårdtung grupp	ICD-10
Artros	M15-M19
Cancer	C00-C97
Cerebrovaskulär sjukdom	I60-I69, G45
Diabetes	E10-E14
Höftfraktur	S720-S722
Inflammatorisk ledsjukdom	M05
Ischemisk hjärtsjukdom	I20-I22, I50
Schizofreni och övriga psykoser	F00-F09, F10-F19, F20-F39, F40-F48

Tabell 4 *Vårdtunga diagnoser*

[2] ICD-10 koder är ett klassificeringssystem för sjukdomar och symptom

Diagnos	ICD-10
Dialys	Z49
Grå starr	H25
Övriga skador	S00-T98 exkl. S720-S722

Tabell 5 *Övriga diagnoser*

Fördelningen av resurser sker till de fem sjukvårdsdistrikten[3] som kommunerna i Skåne är uppdelade i. Tabell 6 visar den relativa fördelningen av individer per sjukvårdsdistrikt i materialet.

Sjukvårdsdistrikt	Frekvens	Procent
Mellersta	72 047	25.2
Nordvästra	62 772	22.0
Nordöstra	43 640	15.3
Sydvästra	83 323	29.2
Ystad-Österlen	23 649	8.3
Totalt	285 431	100

Tabell 6 *Sjukvårdsdistrikt*

Region Skånes register för resursfördelning har av Statistiska Centralbyrån godkänts för användning vid analys vid resursfördelning. Registret innehåller inga personnummer och alla analyser har utförts så att anonymiteten för individerna bevarats.

För beskrivande statistik och matrismodellen har SPSS version 14.0 använts medan alla regressionsmodeller har analyserats i SAS version 9.1, huvudsakligen i Proc REG, Proc GENMOD samt Proc LOGISTIC.

2.2 Korsvalidering

För att kunna utvärdera och jämföra regressionsmodellerna med varandra och med matrismodellen delas materialet i en skattningsdel och en valideringsdel. I skattningsdelen skattas regressionsmodellerna. De skattade koefficienterna från regressionsmodellerna används för att prediktera kostnader för individer i valideringsdelen. En matrismodell anpassas till valideringsdelen då det inte är praktiskt genomförbart, eller intressant, att utforma denna modell i två steg. Skattningsdelen består av ett stratifierat urval efter sjukvårdsdistrikt på 2,1 procent. Detta ger totalt 6000 individer vilket innebär 279 431 individer i valideringsdelen.

[3] Uppdelningen i fem sjukvårdsdistrikt försvann år 2007

3. Matrismodellen

3.1 Fördelning av resurser efter behov

Fördelning av resurser efter behov grundas på att behov indirekt mäts av olika socioekonomiska och demografiska variabler. Dessa variabler förklarar dock endast en liten andel av skillnader i vårdkostnader. Då behov modelleras finns därmed en stor risk att variabler som ytterligare förklarar vårdbehov utelämnas, men det går inte heller att bortse från att behov av sjukvård karakteriseras av en stor del slumpmässighet (Smith, Rice & Carr-Hill, 2001). Detta medför svårigheter att på ett systematiskt sätt fördela resurser efter behov.

Vid en behovsbaserad resursfördelning måste behov uppskattas. Det finns dock inget direkt mått på behov, utan snarare finns flera olika metoder att uppskatta behov med. En metod är att indirekt mäta vårdbehov genom olika demografiska och socioekonomiska variabler. Genom tillgång till individdata kan dessa demografiska och socioekonomiska variabler kopplas till vårdkostnader, varvid ett monetärt mått på behov erhålls. Det ska dock poängteras att variationer i faktorer som ålder, kön, etnicitet, civilstånd, utbildning och yrke endast förklarar några få procent av variationen mellan individer (Ljung m.fl, 2001). Om uppgifter om tidigare sjukdomar och vård inkluderas kan andelen förklarad varians höjas, men sällan till mer än 20 procent. Resterande variation beror på slumpen och andra delvis okända biologiska samt ärftliga faktorer.

Behov modelleras på olika sätt i de länder behovsbaserad resursfördelning används. I flera länder saknas data på individnivå varvid modeller baserade på aggregerade data används. De statistiska metoder som används för att fördela resurser efter behov skiljer sig åt mellan länder. Rice och Smith (1999) genomgång av ersättningssystem efter behov i 19 länder visar att både matrismodeller och regressionsmodeller används. Modellernas omfattning är varierande, vissa länder grundar sina modeller på endast ålder och kön medan andra länder har flera andra förklarande variabler med.

Resursfördelningsmodellen bygger på ett antagande om att skillnader i vårdutnyttjande mellan olika grupper avspeglar lika stora skillnader i vårdbehov. Detta är ett starkt antagande som inte tycks stämma fullt ut. Studier visar att låginkomsttagare och arbetslösa tenderar att underutnyttja den öppna vården i förhållande till sin självrapporterade ohälsa (Walander & Burström, 2005). Det finns också andra grupper som tycks ha ett lägre vårdutnyttjande. Utomnordiska invandrare har visat sig ha ett lågt utnyttjande av den psykiatriska vården, utan att det kan påvisas att de har ett lägre behov av psykiatrisk vård (Diderichsen & Varde, 1996).

Den behovsbaserade resursfördelningen tilldelar varje individ en viss summa. Det är dock inte rimligt att anta att varje individ kommer att ta exakt de resurser som tilldelats i anspråk. Resurser som fördelas efter behov bör därför ses som en förväntad kostnad, en viss variation kommer således alltid att finnas.

3.2 Matrismodellen i Region Skåne

Region Skåne bildades år 1998 och innebar att de forna länen Kristianstads län och Malmöhus län fick en gemensam organisation. Den första resursfördelningen på regionnivå gjordes år 1999, då resurser fördelades efter behov med en matrismodell. Denna modell användes fram till år 2002 då Region Skåne övergick till anslagsfinansiering. Matrismodellen används dock fortfarande som metod för kommunalekonomisk utjämning.

Behov i Region Skånes resursfördelning skattas enligt variabler beskrivna i tabell 2. Dessa variabler har vid en utvärdering med hjälp av regressionsanalys ansetts ge den högsta förklaringen av skillnader i behov.

En möjlig förklaring till skillnader i sjukvårdskostnader mellan individer är lokalt vårdutbud och närhet till sjukvård. För att reducera inverkan av vårdutbud och andra lokala påverkansfaktorer beräknas därför genomsnittskostnaderna på hela Region Skåne.

3.3 Beräkningsmetod

Befolkningen delas först upp i celler efter varje kombination av variablerna i tabell 2. I varje cell beräknas därefter en genomsnittlig kostnad. Den genomsnittliga kostnaden baseras således på individer i hela Region Skåne. Därefter beräknas antalet personer i varje cell i matrisen för varje sjukvårdsdistrikt. Antalet personer i cellen uppdelat på varje sjukvårdsdistrikt multipliceras med genomsnittskostnaden för cellen. De totala kostnaderna för varje sjukvårdsdistrikt erhålls genom att summera kostnadscellerna för respektive sjukvårdsdistrikt.

Vid resursfördelningen år 1999-2002 då matrismodellen användes gjordes en separat beräkning för vissa vårdtunga individer. Till dessa individer räknades avlidna under året och individer med diagnos cancer. För att ytterligare kontrollera för vårdtunga grupper görs i denna uppsats en separat beräkning för alla diagnoser från tabell 3 och 4. En separat beräkning görs således också för alla individer utan vårdtung diagnos.

Då matrismodellen användes i Region Skåne gjordes även en framskrivning för efterföljande år. Kostnader för befolkningen per kommun, ålder och kön multiplicerades med kvoten av befolkningen år 2000 genom befolkningen år 1999 inom dessa tre variabler. En sådan framskrivning kommer dock inte att göras i denna uppsats då syftet är att jämföra hur väl matrismodellen skattar kostnaderna i materialet jämfört med regressionsmodeller.

4 Multipel linjär regression

4.1 Analys av sjukvårdskostnader

Regressionsanalys bygger på att ett antal antaganden är uppfyllda. Särskilt ska residualerna vara oberoende och homoskedastiska. För att test av koefficienter ska vara giltiga krävs även att residualerna är normalfördelade. Vid analys av sjukvårdskostnader uppfylls dock sällan dessa antaganden. Vanligtvis uppstår nedan beskrivna problem.

Sjukvårdskostnader karakteriseras av en hög andel individer som inte uppsökt läkarvård och därmed har noll kostnader. Den kumulativa fördelningen av kostnader har därför en ”spets” vid nollpunkten. Därav följer kostnader en så kallad mixad fördelning, det vill säga en fördelning som både är diskret och kontinuerlig. En mixad fördelning definieras av att det i vissa punkter finns positiv sannolikhet samtidigt som sannolikheten för övriga värden är ett intervall (där sannolikheten i varje punkt följaktligen är lika med noll). För sjukvårdskostnader finns alltså en positiv sannolikhet vid nollpunkten, för övriga värden är fördelningen kontinuerlig. Vid regressionsanalys kan sjukvårdskostnaders mixade fördelning ignoreras och en vanlig multipel linjär regressionsmodell användas. Alternativt kan en two-part modell anpassas som grundas på den mixade fördelningen.

Sjukvårdskostnader är, enligt beskrivning ovan, sällan normalfördelade. Stora avvikelser från normalitet av den beroende variabeln, i synnerhet om fördelningen är starkt skev, medför att koefficienterna inte heller blir normalfördelade och därmed kommer t-testen ge felaktiga signifikanstest. Koefficienterna är dock asymptotiskt normalfördelade även om den beroende variabeln avviker starkt från normalitet. En beroende variabel som inte är normalfördelad ger också residualer som inte heller är normalfördelade, vilket innebär att antagandet om normalfördelade residualer ej uppfylls. Antagandet om normalfördelade residualer påverkar dock endast test och konfidensintervall för koefficienter, det vill säga koefficienterna kommer att vara Best Linear Unbiased Estimator (BLUE) även om residualerna inte är normalfördelade. För att åstadkomma en mer symmetrisk fördelning logaritmeras vanligen kostnader. De logaritmerade kostnaderna följer en normalfördelning, vilken kan definieras som fördelningen av en slumpvariabel vars logaritm är normalfördelad. Alternativt kan en generaliserad linjär modell anpassas vilket innebär att den beroende variabeln inte behöver transformeras.

Ett antagande i regressionsmodellen är att residualvariansen är konstant, det vill säga homoskedastisk. Det är troligt att detta antagande inte uppfylls då sjukvårdskostnader analyseras. Effekten av heteroskedastisitet är att OLS-skattningarna inte längre är BLUE och effektiva. Koefficienterna och prediktioner baserade på dessa är dock fortfarande utan systematiska avvikelser och konsistenta. Koefficienternas standardfel kommer dock att ha systematiska avvikelser och vara icke-konsistenta vilket innebär att hypotestest inte är giltiga. Flera studier (Blough mfl, 1999; Diehr mfl, 1999; Blough & Ramsey, 2000) har observerat att standardavvikelsen är högre för grupper av individer där medelvärdet också är högre, det vill säga sjukvårdkostnader tenderar att variera mer för höga kostnader än för låga. För att komma till rätta med detta problem kan logaritmen av kostnader användas, alternativt kan en generaliserad modell anpassas där en lämplig variansfunktion specificeras.

4.2 Multipel linjär regression - modellen

I den multipla linjära regressionsmodellen ignoreras sjukvårdskostnaders mixade fördelning. Regressionen ger en modell som är additiv, där koefficienterna tolkas som den genomsnittliga förändringen i Y då en given oberoende variabel ökar med en enhet, givet att övriga variabler är konstanta.

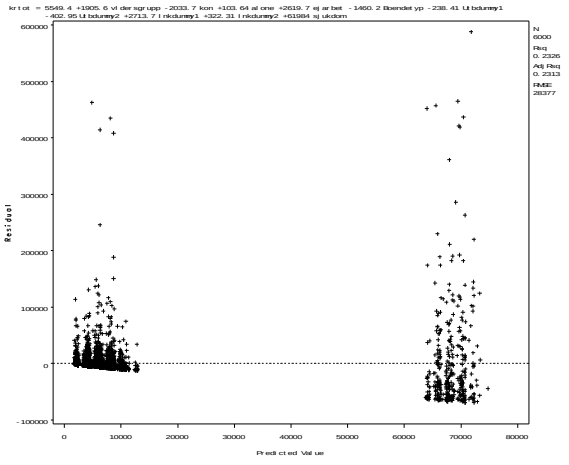
Detta ger följande regressionsmodell;

$$sjukvårdskostnad_i = \beta_0 + \beta_1 ålder_i + \beta_2 kön_i + \beta_3 civilstånd_i + \beta_4 sysselsättning_i + \beta_5 inkomst1_i$$
$$+ \beta_6 inkomst2_i + \beta_7 boendet yp_i + \beta_8 utbildning1_i + \beta_9 utbildning2_i + \beta_{10} vårdtungdiagnos_i + \varepsilon_i$$

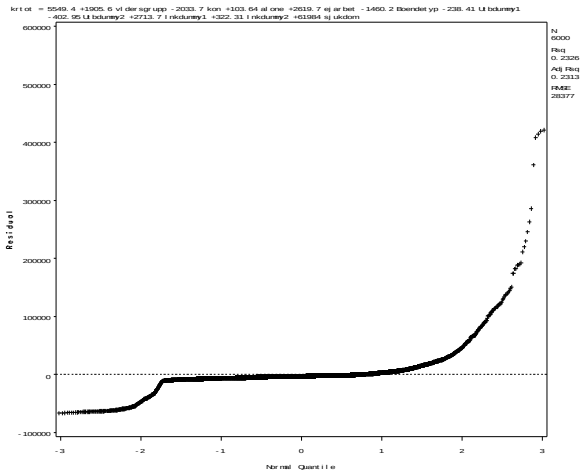
där ε_i är Nf $(0,\sigma^2)$

4.3 Validering

En residualanalys utförs i syfte att kontrollera modellens antaganden. Figur 2 där residualerna plottas mot predikterade kostnader visar att variansen ökar kraftigt för högre predikterade kostnader. Ett Breusch-Pagan test utförs därmed för att formellt kontrollera om heteroskedastisitet föreligger. Resultatet av testet visar att så är fallet. Figur 3 visar ett normal kvartil diagram där det kan utläsas att residualerna avviker från normalitet genom en lång högersvans. Antagandet om homoskedastisitet och normalitet är därmed inte uppfyllt vilket innebär att OLS-skattningarna inte är effektiva och BLUE samt att test inte är giltiga.



Figur 2 *Residualplott, vanlig multipel linjär regression*



Figur 3 *Normal-kvartil diagram*

5 Two-part modellen

Resultatet av den vanliga multipla linjära modellen visar att modellen ger en dålig anpassning till materialet. I detta avsnitt appliceras därför en two-part modell.

5.1 Two-part modellens fördelning

Two-part modellen utvecklades för att hantera den mängd nollvärden som material, från skiftande discipliner, kan medföra. Inom sjukvården användes modellen först för att modellera efterfrågan på sjukvård (Duan mfl, 1983). Därefter har modellen blivit en populär metod för att prediktera sjukvårdskostnader. Den första delen skattas med logistisk regression medan den andra delen traditionellt skattats med minsta-kvadrat metoden och logaritmerad beroende variabel. På senare år har dock generaliserade linjära modeller i större utsträckning börjat användas eftersom systematiska avvikelser vid en tillbakatransformering på så sätt undviks (Blough & Ramsey, 2000; Manning & Mullahy, 2001).

Two-part modellen består av två ekvationer. Den första delen är en logistisk regression för det dikotoma utfallet att en individ har eller inte har sjukvårdskostnader.

$$\mathbf{I}_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \tag{5.1}$$

där **X** är en vektor av kovariat, **β** en vektor av parametrar, **ε** en vektor av residualer och

$$\mathbf{I} = \begin{cases} 1 \ då \ Y > 0 \\ 0 \ då \ Y = 0 \end{cases} \tag{5.2}$$

Den andra delen utgörs av en multipel regression som kan specificeras på olika sätt. Som exempel används en vanlig multipel regression.

$$\mathbf{Y}_i|\mathrm{I}=1 = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \quad \text{där } \boldsymbol{\varepsilon}_i \text{ är Nf } (0,\sigma^2) \tag{5.3}$$

Täthetsfunktionen för two-part modellen är;

$$f_Y\left(y,\boldsymbol{\beta}\middle|x\right) = \begin{cases} P\left(I=0\middle|x\right) & då \ y=0 \\ f_Y\left(y,\beta\middle|I=1,x\right)P\left(I=1\middle|x\right) & då \ y>0 \\ 0 & då \ y<0 \end{cases} \tag{5.4}$$

Maximum likelihoodfunktionen av 5.4 kan bestämmas på följande sätt. Givet *n* observationer $I_1, x_1, y_1, I_2, x_2, y_2, \ldots, I_n, x_n, y_n$, där materialet är sorterat så att de första n_1 observationerna har $y_i = 0$ (och $I = 0$) och de återstående *n*- n_1 observationerna har $y_i > 0$ (och $I = 1$), är likelihoodfunktionen för **β**;

$$\prod_{i-1}^{n} f_Y\left(y_i;\boldsymbol{\beta}|x\right)=\prod_{i=1}^{n_1} P\left(I=0|x\right)\prod_{i=n_1+1}^{n} f_Y\left(y_i,\boldsymbol{\beta}|I=1,x\right)P\left(I=1|x\right)$$

$$=\left[P\left(I=0|x\right)\right]^{n_1}\left[P\left(I=1|x\right)\right]^{n-n_1}\times\prod_{i=n_1+1}^{n} f_Y\left(y_i,\boldsymbol{\beta}|I=1,x\right) \tag{5.5}$$

= (Likelihood för del 1) × (Likelihood för del 2)

Likelihoodfunktionen delas alltså in i två delar. Parametrarna i den första delen beror endast på parametrarna i 5.1 och parametrarna i del två beror endast på parametrarna i 5.3. Detta gör de två delarna oberoende och kan därmed separeras och maximeras var för sig. Del ett tolkas som sannolikheten att en individ har sjukvårdskostnader och del två storleken på dessa sjukvårdskostnader. Den totala kostnaden för en individ predikteras genom att först skatta sannolikheten att individen har kostnader och därefter, givet att individen har kostnader, skatta nivån på dessa. Fördelen med att använda en two-part modell är att dess fördelning stämmer bättre överrens med sjukvårdskostnaders fördelning än en vanlig modell. En two-part modell borde därmed ge mer precisa skattningar av en individs sjukvårdskostnader. De totala sjukvårdskostnaderna predikteras alltså genom att multiplicera ihop sannolikheten för del ett med predikterad kostnad i del två. Sätt $\hat{p}_i$ = skattad sannolikhet från del ett och sätt $\hat{\mu}_i$ = skattad medelkostnad från del två, givet att individen har kostnader. Detta ger;

$$\hat{y}_i = \hat{p}_i \times \hat{\mu}_i \tag{5.6}$$

den totala predikterade kostnaden för individ i.

Eftersom de två delarna är oberoende ger two-part modellen ett standardfel för varje del. Genom att kombinera standardfelen för de två delarna erhålls standardfelet för den totala skattningen. Standardfelet för skattningen av totala sjukvårdskostnader är roten ur;

$$Var\left(\hat{y}\right)=Var\left(\hat{p}\hat{\mu}\right)=\hat{p}^2Var\left(\hat{\mu}\right)+\hat{\mu}^2Var\left(\hat{p}\right) \tag{5.7}$$

5.2 Del ett

Del ett av two-part modellen skattar sannolikheten för en individ att ha kostnader. Då den beroende variabeln är binär, det vill säga antingen har en individ kostnader eller inte, används logistisk regression[4]. Låt Y vara en Bernoullifördelad variabel. Sannolikhetsfördelningen blir då

$$Y_i=\begin{cases}1\,med\ sannolikheten & p\\ 0\,med\ sannolikheten & 1-p\end{cases} \tag{5.8}$$

Sannolikheten att ha kostnader ser ut som följer;

[4] En probit modell skulle också kunna användas. Vanligtvis ger de båda modellerna samma resultat, vilket innebär att det inte spelar någon roll vilken modell som används.

$$\Pr(y>0|x)=\frac{\exp\left(\beta_0+\sum\beta_j X_{ij}\right)}{1+\exp\left(\beta_0+\sum\beta_j X_{ij}\right)} \qquad (5.9)$$

I denna modell tas alla variabler från tabell 3 med förutom variabeln vårdtung diagnos. Variabeln vårdtung diagnos utesluts då den inte kan anses vara en förklarande variabel för om en individ kostar eller ej.

5.3 Del två

I den andra delen av two-part modellen skattas kostnader, för de individer som har kostnader, med multipel linjär regression och generaliserad linjär regression.

5.3.1 Multipel linjär regression med logaritmerad beroende variabel

För att komma till rätta med problem som icke normalfördelade variabler och heteroskedastisitet kan den beroende variabeln transformeras. Flera olika transformationer är möjliga; Box-Cox transformering, kvadratrotstransformering och logaritmering. I denna uppsats logaritmeras den beroende variabeln sjukvårdskostnader på grundval av att denna transformering, vid analys av sjukvårdsdata, är den i särklass vanligaste transformeringen (Manning & Mullahy, 2001). Detta ger följande modell;

$$\ln(sjukvårdskostnad_i)=\beta_0+\beta_1 ålder_i+\beta_2 kön_i+\beta_3 civilstånd_i+\beta_4 sysselsättning_i+\beta_5 inkomst1_i$$
$$+\beta_6 inkomst2_i+\beta_7 boendeform_i+\beta_8 utbildning1_i+\beta_9 utbildning2_i+\beta_{10} vårdtungdiagnos_i+\varepsilon_i$$

där ε_i är Nf $(0,\sigma^2)$

En regressionsmodell med logaritmerad beroende variabel ger en multiplikativ modell. Koefficienterna tolkas som den relativa förändringen i Y då en given oberoende variabel ökar med en enhet, givet att övriga variabler hålls konstanta.

Då det för beslutsfattare inte är intressant att fördela logaritmerade kostnader måste modellen transformeras tillbaka till ursprunglig skala. Det är dock inte möjligt att endast exponentiera modellen och därefter prediktera kostnader eftersom $E\left(\exp(\mathbf{x\beta})\right)$ i den lognormala fördelningen ger medianen istället för det aritmetiska medelvärdet. Det förväntade värdet för en lognormal variabel är istället

$$E(y|x)=\exp\left(x\beta+0.5\sigma_\varepsilon^2\right) \text{ då } \varepsilon \text{ är } N\left(0,\sigma_\varepsilon^2\right) \qquad (5.10)$$

Detta innebär att $0.5\sigma_\varepsilon^2$ måste skattas för att undvika att systematiska avvikelser, som underskattar prediktioner, uppstår vid en transformering tillbaka till normal skala. Ett alternativ är att skatta $0.5\sigma_\varepsilon^2$ (normalteoriskattning), vilket dock förutsätter att residualerna är normalfördelade. Alternativt kan en ickeparametrisk metod kallad Duan's smearingfaktor användas, vilken förutsätter att residualerna är oberoende och identiskt fördelade (Duan, 1983). Smearingfaktorn är medelvärdet av de exponentierade residualerna

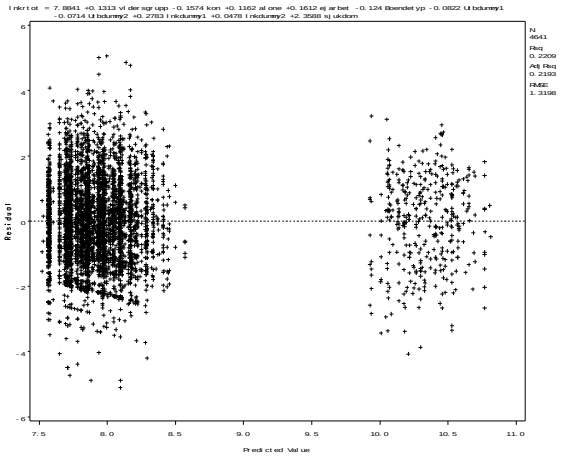
$$S = \frac{1}{N}\sum_{i=1}^{N}\exp(e_i) \qquad \text{där } e_i = \ln(y_i) - x_i\beta \tag{5.11}$$

Normalteoriskattningarna är betydligt mer effektiva än smearingfaktorn då residualerna är normalfördelade men är samtidigt känsliga för avvikelser från normalitet. Vidare kommer smearingfaktorn endast att ge icke-skeva skattningar då residualerna är homoskedastiska (Manning, 1998). Om residualerna är heteroskedastiska, det vill säga om residualvariansen beror på **xβ**, bör detta modelleras för att undvika systematiska avvikelser. Heteroskedasticitet kan medföra att en modell med en smearingfaktor under- eller överskattar skattade kostnader i vissa intervall. Uppvisar residualerna heteroskedasticitet bör därför olika smearingfaktorer, baserade på de grupper av individer för vilka residualvariansen skiljer, användas. En smearingfaktor används på grund av att det inte är troligt att residualerna är perfekt normalfördelade. Den skattade kostnaden för individ i blir då;

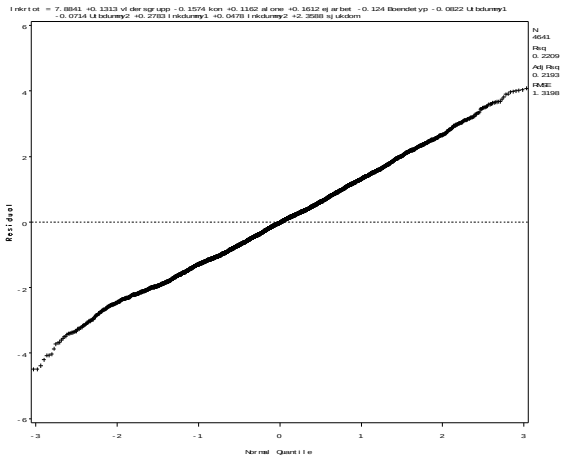
$$E[y_i|x_i] = \Pr(y_i > 0|x_i) \times E(y_i|x_i, y_i > 0) \times S \tag{5.12}$$

5.3.1.1 Validering

För att kontrollera om det finns avvikelser från modellens grundläggande antaganden utförs en residualanalys. Residualerna plottas mot predikterade värden i figur 4. Inga tecken på avvikelser från antagandet om homoskedastisitet kan avläsas, vilket även bekräftas av Breusch-Pagan testet. Normalitet kontrolleras med ett normal-kvartildiagram i figur 5, vilket visar tecken på tjocka svansar. Jämfört med den vanliga multipla linjära modellen erhålls här en regression med homoskedatiska och bättre normalfördelade residualer.



Figur 4 *Residualplott, 2-part OLS*



Figur 5 *Normal-kvartil diagram*

5.3.2 Generaliserad linjär modell (GLM)

Ett antagande i den klassiska linjära modellen är att den beroende variabeln Y_i är oberoende normalfördelad med konstant varians. En generaliserad linjär modell är en utvidgning av den klassiska linjära modellen på så sätt att Y_i kan tillhöra vilken fördelning som helst inom den exponentiella familjen (Olsson, 2002). Denna familj av fördelningar kan skrivas som

$$f(y;\theta,\phi)=\exp\left[\frac{(y\theta-b(\theta))}{a(\phi)}+c(y,\phi)\right] \tag{5.13}$$

där $a(\cdot)$, $b(\cdot)$ och $c(\cdot)$ är funktioner. Till denna familj av fördelningar tillhör bland andra normal-, Poisson-, binomial- och gammafördelningen.

Den klassiska linjära modellen är **y** = **Xβ** + **e**, där **η** = **Xβ** kallas den linjära komponenten. Det förväntade värdet i den klassiska linjära modellen är $\mu=E(y)$. I den generaliserade linjära modellen (GLM) är istället det förväntade värdet en funktion av den linjära komponenten $\eta=x\beta$, så att $g(\mu)=\eta=X\beta$. Funktionen $g(\mu)$ kallas länkfunktion och länkar alltså det förväntade värdet av Y till de förklarande variablerna $X_1 \ldots X_n$. Skillnaden mellan länkfunktionen i den klassiska linjära modellen och länkfunktionen i en GLM är att i den senare modelleras länken mellan de förklarande variablerna och det förväntade värdet av Y icke-linjärt. Länkfunktionen erhålls genom att derivera Maximum Likelihood skattningarna av parametrarna. För att underlätta beräkningar beräknas log likelihood funktionen av 5.13, så att $l(\theta,\phi;y)=\log f(y;\theta,\phi)=(y\theta-b(\theta))/a(\phi)+c(y,\phi)$. Detta ger;

$$\frac{\partial l}{\partial\theta}=[y-b'(\theta)]/a(\phi) \tag{5.14}$$

där b' betecknar förstaderivatan av b med avseende på θ. Genom att ta förväntat värde av 5.14 och sätta denna till noll erhålls länkfunktionen;

$$E\left(\frac{\partial l}{\partial\theta}\right)=E\{[y-b'(\theta)]/a(\phi)\}=0 \tag{5.15}$$

Länkfunktionen blir; $E(Y)=\mu=b'(\theta)$

Vissa länkfunktioner är naturliga för en given fördelning på så sätt att $g(\mu)=\theta$. Dessa kallas kanoniska länkar. Det finns dock ingen garanti för att de kanoniska länkarna alltid ger den bästa anpassningen till materialet. Vissa material har särskilda egenskaper så att en annan länkfunktion än den kanoniska är mest lämplig. Till exempel används ofta log-länken $g(\mu_i)=\log(\mu_i)$ med gammafördelningen vid sjukvårdskostnader.

Variansen av Y erhålls genom andraderivatan av den logaritmerade likelihoood funktionen 5.13 samt genom att utnyttja att enligt likelihood teori är $E\left(\frac{\partial^2 l}{\partial\theta^2}\right)+E\left[\left(\frac{\partial l}{\partial\theta}\right)^2\right]=0$. Andraderivatan blir;

$$\frac{\partial^2 l}{\partial\theta^2}=-b''(\theta)/a(\phi) \tag{5.16}$$

Detta ger;

$$-\frac{b''(\theta)}{a(\phi)}+\frac{Var(Y)}{a^2(\phi)}=0 \tag{5.17}$$

vilket ger $Var(Y)=a(\phi)\cdot b''(\theta)$

Parametern ϕ kallas dispersionsparametern och $b''(\theta)$ kallas variansfunktionen. Variansfunktionen skrivs ofta $V(\mu)=b''(\theta)$ och visar hur variansen beror på medelvärdet μ i

fördelningen, där μ i sin tur är en funktion av θ. Tabell 7 visar några vanliga exponentiella fördelningar och deras respektive länkfunktion och variansfunktion.

Fördelning	Kanonisk länkfunktion	Variansfunktion
Normal	η=μ	1
Poisson	η=log(μ)	μ
Gamma	η=1 / μ	μ^2
Inverse Gaussian	η=1 / μ^2	μ^3

Tabell 7 *Länk- och variansfunktioner för exponentiella fördelningar*

Vid modellering av en generaliserad linjär modell behöver alltså en fördelning, en länkfunktion och en variansfunktion specificeras. För att undersöka vilken typ av fördelning som bör anpassas studeras variansfunktionen. För de fördelningar som har en variansfunktion upphöjd till ett visst värde, λ, kan variansen generaliseras och skrivas;

$$\mathrm{Var}(y|x) = k(\mu(x\beta))^{\lambda} \tag{5.18}$$

där λ måste vara ändlig och icke-negativ. Då λ = 0 får vi en vanlig icke-linjär minsta-kvadrat skattning med konstant varians. Om λ = 1 får vi en Poisson fördelning där variansen är proportionell mot medelvärdet. Om λ = 2 erhålls gamma, homoskedastisk lognormal, Weibull och Chi-två fördelningar där standardavvikelsen är proportionell mot medelvärdet och om λ = 3 fås en inverse Gaussian fördelning.

Manning och Mullahy (2001) föreslår en utvidgning av Park's test[5] för specificering av variansfunktionen. För att utföra testet behövs residualer från en GLM eller logaritmerad OLS. Därefter utförs en regression med residualerna i kvadrat som beroende variabel och logaritmerade skattade värden som oberoende variabel. Detta ger;

$$\ln\left(y_i - \hat{y}_i\right)^2 = \lambda_0 + \lambda_1\left(\ln\left(\hat{y}_i\right)\right) + \varepsilon_i \tag{5.19}$$

där $\hat{y}_i$ är $\exp\left(x_i\hat{\beta} + 0.5\hat{\sigma}^2(x)\right)$ från OLS-modellen. Värdet på λ_1 avgör vilken GLM modell som bör användas. Regressionen ger λ_1=1,85 vilket ligger närmast en gammafördelning.

Som länkfunktion väljs en log-länk. Detta motiveras med att det är denna länkfunktion som vanligen väljs då sjukvårdskostnader modelleras (Buntin & Zaslavsky, 2004). Med denna länkfunktion ger en förändring i en förklarande variabel en multiplikativ effekt på sjukvårdskostnader. Modellen blir;

$$\log\left(\mu_i\right) = \log\left(E\left(sjukvårdskostnad_i\right)\right) = \beta_0 + \beta_1 ålder_i + \beta_2 kön_i + \beta_3 civilstånd_i + \beta_4 sysselsättning_i$$
$$+\beta_5 inkomst1_i + \beta_6 inkomst2_i + \beta_7 boendet yp_i + \beta_8 utbildning1_i + \beta_9 utbildning2_i + \beta_{10} vårdtungdiagnos_i$$

Kostnader modelleras på den ursprungliga skalan, vilket innebär att modellen inte behöver transformeras tillbaka. För den generaliserade modellen blir den skattade kostnaden för en individ;

$$E\left[y_i|x_i\right] = \Pr\left(y_i > 0|x_i\right) \times E\left(y_i|x_i, y_i > 0\right) \tag{5.20}$$

[5] Park's test användes ursprungligen för att testa heteroskedastisitet (Park, 1966).

5.3.2.1 Maximum likelihood skattning

Parametrarna i den generaliserade linjära modellen skattas med Maximum Likelihood metoden (Olsson, 2002). Skattningen av parametrarna är de värden som maximerar log likelihoodfunktionen, som för en enskild observation kan skrivas som;

$$l = \log[L(y;\theta,\phi)] = \frac{(y\theta - b(\theta))}{a(\phi)} + c(y,\phi) \tag{5.21}$$

Parametrarna i modellen är en vektor av regressionskoefficienter **β** som i sin tur är en funktion av θ. Kedjeregeln används då l deriveras med avseende på β. Detta ger;

$$\frac{\partial l}{\partial \beta_j} = \frac{\partial l}{\partial \theta}\frac{d\theta}{d\mu}\frac{d\mu}{d\eta}\frac{\partial \eta}{\partial \beta_j} \tag{5.22}$$

Vi har att $b'(\theta) = \mu$ och $b''(\theta) = V$, variansfunktionen. Då är $\frac{\partial \mu}{\partial \theta} = V$. Den linjära komponenten är $\eta = \sum_j x_j \beta_j$, vilket ger $\frac{\partial \eta}{\partial \beta_j} = x_j$. Detta ger;

$$\frac{\partial l}{\partial \beta_j} = \frac{(y-\mu)}{a(\phi)}\frac{1}{V}\frac{d\mu}{d\eta}x_j = \frac{W}{a(\phi)}(y-\mu)\frac{d\eta}{d\mu}x_j \tag{5.23}$$

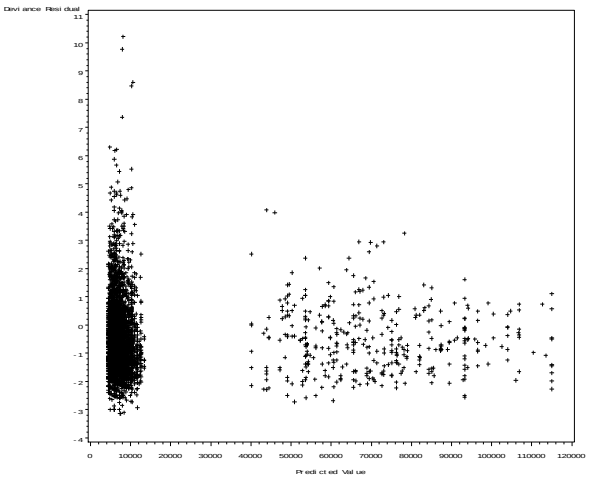
där $W = \left(\frac{d\mu}{d\eta}\right)^2 \frac{1}{V}$

Genom att summera över alla observationer erhålls maximum likelihood skattningen för en parameter β_j.

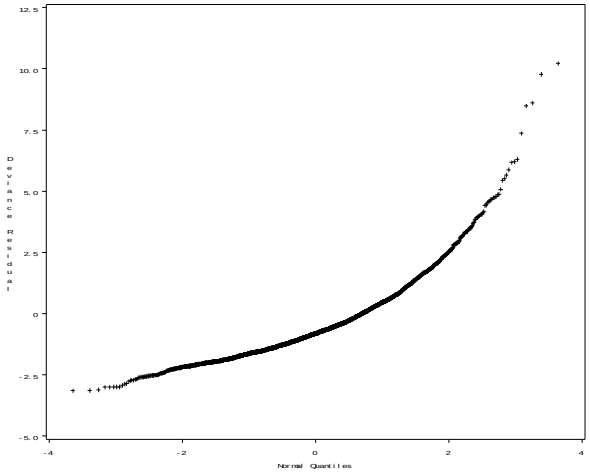
$$\sum_i \frac{W_i(y_i - \mu_i)}{a(\phi)}\frac{d\eta_i}{d\mu_i}x_{ij} = 0 \tag{5.24}$$

5.3.2.2 Validering

För att utvärdera modellens anpassning till materialet plottas deviance residualer mot de skattade värdena. Plotten, figur 6, visar att residualerna har konstant varians men att det finns ett antal uteliggare för individer med låga predikterade värden. Normal-kvartil diagrammet i figur 7 visar att residualerna inte är normalfördelade. Ett mått på hur väl anpassad modellen är till materialet är scaled deviance-residualer delat med sina frihetsgrader, vars kvot ej bör vara större än ett. Kvoten i modellen är 1,22 vilket indikerar någon form av misspecificering. Detta kan vara ett resultat av felaktig länkfunktion, uteliggare eller felaktigt val av förklarande variabler. Exempelvis skulle modellen kunna behöva fler variabler, interaktionstermer eller icke-linjära variabler.



Figur 6 *Residualplott, 2-part GLM*



Figur 7 *Normal-kvartil diagram*

6. Jämförelse av modellerna

Enligt beskrivning i avsnitt 2.2 skattas regressionsmodellerna i skattningsdelen. De skattade koefficienterna används därefter för att prediktera kostnader i valideringsdelen. Nedan följer en jämförelse mellan regressionsmodellerna och matrismodellen avseende effekter på resursfördelning och anpassning till materialet.

6.1 Jämförelse av regressionsmodeller och anpassning till materialet

I tabell 8 finns de skattade koefficienterna från skattningsdelen.

Variabel	Vanlig linjär regression	Two-part Del 1, logistisk regression	2-part del 2, OLS	2-part del 2, GLM
	Koefficient (p-värde)	*Koefficient (p-värde)*	*Koefficient (p-värde)*	*Koefficient (p-värde)*
Intercept	5549.39 (0.0136)	1.0474 (<.0001)	7.8841 (<.0001)	9.0891 (<.0001)
Åldersgrupp	1905.59 (0.0106)	0.3122 (<.0001)	0.1313 (0.0008)	0.2007 (<.0001)
Kön	-2033.74 (0.0067)	0.0669 (<.0001)	-0.1574 (<.0001)	-0.1014 (0.0061)
Civilstånd	103.64 (0.8988)	0.0696 (0.0009)	0.11623 (0.0075)	0.0737 (0.0662)
Sysselsättning	2619.72 (0.1528)	0.0412 (0.7951)	0.16115 (0.0980)	0.0605 (0.5108)
Inkomst1	2713.73 (0.1939)	1.0087 (<.0001)	0.27833 (0.0213)	0.0811 (0.4740)
Inkomst2	322.31 (0.8780)	0.1628 (<.0001)	0.04777 (0.6956)	-0.1700 (0.1368)
Boendetyp	-1460.19 (0.0834)	0.0738 (0.0076)	-0.12404 (0.0052)	-0.1355 (0.0009)
Utbildning1	-238.41 (0.8096)	-0.1397 (0.0982)	-0.08219 (0.1184)	-0.2901 (<.0001)
Utbildning2	-402.95 (0.6433)	-0.0204 (0.7879)	-0.07137 (0.1197)	-0.2020 (<.0001)
Vårdtung diagnos	61984 (<.0001)		2.35883 (<.0001)	2.2085 (<.0001)

Tabell 8 *Koefficienter från regressionsmodellerna*

De tre modellerna visar genomgående samma tecken. Skillnad finns för GLM-modellen och övriga modeller i variabeln Inkomst2. Denna variabel är dock inte signifikant i någon av modellerna vilket kan vara anledningen till att olika tecken erhålls. Det är dock svårt att göra en jämförelse mellan modellerna avseende variablers effekter på sjukvårdskostnader, dels på grund av att de modelleras med olika länkfunktion dels på grund av att two-part modellen består av två delar.

De tre modellerna visar stora skillnader vad gäller signifikanta variabler. Den vanliga linjära modellen har flest icke-signifikanta variabler medan two-part GLM modellen uppvisar flest signifikanta variabler. Det är endast variablerna åldersgrupp, kön, boendetyp och vårdtung diagnos som är signifikanta för alla modeller. Alla variabler i tabell 8 har dock av Region Skåne ansetts nödvändiga för att förklara behov vilket innebär att inga variabler utesluts ur analysen.

Sjukvårdskostnader innefattar ofta individer med extrema värden, vilka kan påverka de skattade koefficienterna drastiskt. Genom att logaritmera den beroende variabeln och använda GLM-modellen med log-länk minskas dock inflytandet av uteliggande observationer eftersom logaritmering ger en lägre toppighet. En undersökning av residualerna visar att den vanliga

regressionsmodellens residualer är kraftigt skeva och kurtotiska. Logaritmeringen i den OLS baserade two-part modellen ger residualer som inte är skeva eller kurtotiska medan residualerna i GLM modellen fortfarande är något skeva och kurtotiska. För GLM modellen innebär skevheten och toppigheten i residualerna att effektiviteten minskar och skattningar av koefficienter blir mindre precisa (Manning och Mullahy, 2001).

De skattade koefficienterna från estimeringsdelen anpassas till valideringsdelen och matrismodellen anpassas till valideringsdelen. För att utvärdera modellernas förmåga att prediktera kostnader används följande mått

$$\text{Mean squared prediction error } (MSPE) = \frac{1}{n}\sum_k \left(\hat{Y}_k - Y_k\right)^2$$

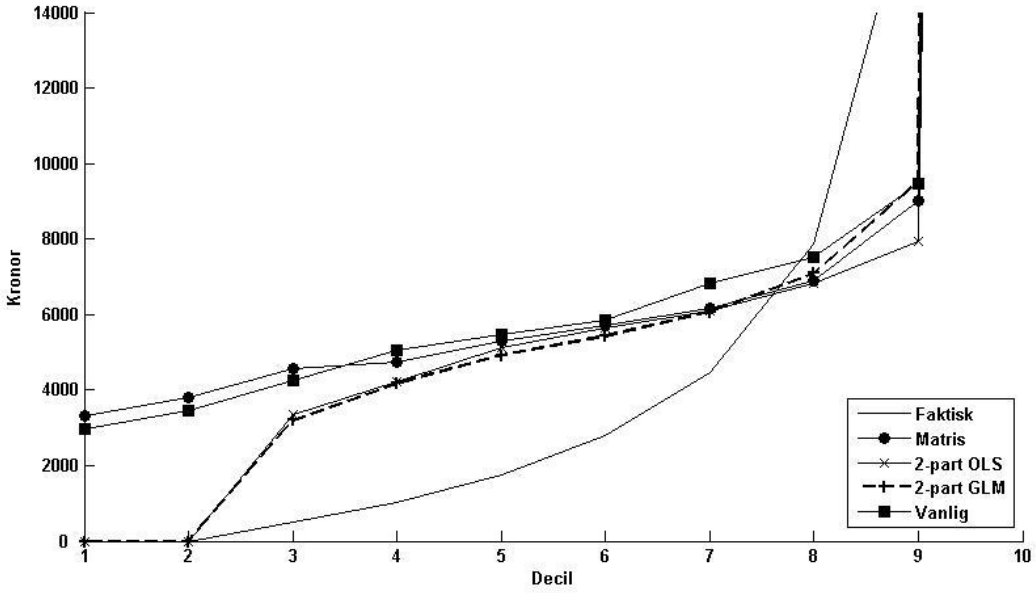
$$\text{Mean absolute prediction error } (MAPE) = \frac{1}{n}\sum_k \left|Y_k - \hat{Y}_k\right|$$

Höga värden på dessa mått indikerar en dålig förmåga att prediktera kostnader. Tabell 9 visar att jämförelsemåtten inte ger ett entydigt resultat. Matrismodellen predikterar kostnader bäst enligt MSPE medan de båda two-part modellerna är bättre enligt MAPE.

Modell	MSPE	MAPE
Matrismodell	1 077 021 124	9 748
Vanlig regression	1 101 642 481	9 904
2-part OLS	1 101 244 225	8 609
2-part GLM	1 111 355 569	8 519

Tabell 9 *Jämförelsemått*

Figur 8 visar predikterad kostnad per decil för alla modeller och för faktisk kostnad. Matrismodellen och den vanliga regressionsmodellen överpredikterar kostnader upp till den åttonde decilen och underpredikterar kostnader i den nionde och tionde decilen. Two-part modellerna överpredikterar kostnader från den andra till den åttonde decilen och underpredikterar därefter. Från den tredje decilen och uppåt ger dock alla modeller ungefär samma resultat.



Figur 8 *Predikterad kostnad per decil* (MATLAB version 7.3)

6.2 Jämförelse av effekter på resursfördelningen

För varje modell i valideringsdelen summeras prediktioner per sjukvårdsdistrikt. Resultatet av denna resursfördelning finns i tabell 10. Alla modeller fördelar procentuellt sett lika mycket resurser till varje distrikt. Den stora skillnaden mellan modellerna består av storleken på de totala predikterade kostnaderna, då de två two-part modellerna fördelar totalt sett mindre resurser än matrismodellen och den vanliga regressionsmodellen.

	Matris	**%**	**Vanlig**	**%**	**2-part OLS**	**%**	**2-part GLM**	**%**	**Faktisk**	**%**
Mellersta	649	24,9	661	25	547	24,7	515	24,4	629	24,1
Nordvästra	554	21,3	561	21,2	473	21,3	454	21,5	563	21,6
Nordöstra	396	15,2	404	15,3	338	15,3	325	15,4	413	15,9
Sydvästra	785	30,1	794	30	666	30	636	30	788	30,2
Ystad-Österlen	221	8,5	229	8,6	192	8,7	184	8,7	212	8,1
Summa, kr	2 605		2 649		2 216		2 114		2 605	

Tabell 10 *Resursfördelning, miljontals kronor*

För att undersöka hur modellerna predikterar kostnader för vårdtunga grupper görs en fördelning endast för vårdtunga grupper, vilken finns i tabell 11. Resultatet visar att two-part GLM modellen fördelar minst resurser av modellerna till sjukvårdsdistrikten. Även two-part OLS modellen fördelar mindre resurser än de faktiska kostnaderna. För regressionsmodeller kan det vara ett problem att skatta höga kostnader då det inte finns tillräckligt många individer med höga värden för att ge en bra skattning. Den vanliga linjära modellen skattar dock höga kostnader men har som tidigare beskrivits inte förutsättningarna för regressionsanalys uppfyllda.

	Matris	**%**	**Vanlig**	**%**	**2-part OLS**	**%**	**2-part GLM**	**%**	**Faktisk**	**%**
Mellersta	309	25,4	325	25,7	285	25,2	254	25	331	27,2
Nordvästra	246	20,2	255	20,2	230	20,3	207	20,4	241	19,8
Nordöstra	189	15,6	201	16	179	15,8	161	15,8	187	15,4
Sydvästra	362	29,8	363	28,7	333	29,4	299	29,4	357	29,4
Ystad-Österlen	110	9,0	119	9,4	105	9,3	95	9,4	99	8,1
Summa, kr	1 216		1 263		1 132		1 016		1 215	

Tabell 11 *Resursfördelning endast vårdtunga diagnoser, miljontals kronor*

7. Diskussion

Syftet med denna uppsats var att bilda regressionsmodeller för en behovsbaserad resursfördelning och jämföra dessa med matrismodellen avseende effekter på resursfördelning. Resultatet visar att det är främst storleken på resurserna som påverkas. Vidare är det svårt att avgöra vilken modell som ger den bästa anpassningen till materialet.

I denna uppsats har regressionsmodellerna baserats på ett urval av 6000 individer. Montez-Rath mfl (2006) visar genom sin studie, där olika stora urval från totalt 525 620 individer undersökts, att storleken på urvalet påverkar vilken modell som bäst predikterar kostnader då urvalet är relativt sett litet. En generalisering av resultatet från denna studie är därmed inte möjlig. Ett större urval ger mer precisa skattningar av populationens parametrar samtidigt som det också skattar höga kostnader bättre.

Ett problem med two-part modellen i kontexten resursfördelning är att alla individer som har noll kostnader också tilldelas noll kronor vid resursfördelningen. Ett grundläggande antagande i en behovsbaserad resursfördelning är dock att individer med samma karakteristiker har samma behov av vård. Utifrån perspektivet behov görs därmed ett implicit antagande i en two-part modell att individer med samma karakteristiker, men som har respektive inte har några sjukvårdskostnader, har olika behov av sjukvård. Detta innebär att individer med samma karaktäristiker kan komma att tilldelas olika resurser beroende på om de har kostnader eller inte. Detta implicita antagande motsäger syftet med en behovsbaserad resursfördelning och innebär att det blir problematiskt att använda two-part modellen vid resursfördelning.

För att utvärdera hur väl modellerna är anpassade till materialet, det vill säga hur väl de predikterar kostnader, användes korsvalidering. Jämförelsemåtten MSPE och MAPE gav dock inte ett entydigt resultat. Det är därför svårt att på grundval av dessa mått avgöra vilken av modellerna som predikterar kostnader bäst. Ett alternativ för att få konsekventa resultat hade varit att använda sig av bootstrapping, det vill säga göra flera urval och beräkna måtten för varje urval. Vidare visar figur 7 att alla modeller ger ungefär samma prediktioner för varje decil. Modellerna predikterar dock i snitt högre kostnader än de faktiska förutom för de allra högsta kostnaderna vilka grovt underpredikteras. Bristen på samt det tvetydiga resultatet gör det därmed svårt att avgöra hur bra modellerna är anpassade till materialet.

Fördelen med korsvalidering är att den gör det möjligt att utvärdera och jämföra modeller avseende prediktioner vilket annars är svårt då two-part modeller används, eftersom det inte går att använda sig av traditionella mått såsom R^2 måttet. Metoden skulle dock kunna användas av Region Skåne för att undvika att bilda en regressionsmodell för hela populationen. En potentiell nackdel är dock att det kan vara praktiskt komplicerat att utföra.

I denna uppsats har regressionsmodellerna gjorts lika matrismodellen vad gäller val av variabler. Det är dock möjligt att regressionsmodellerna hade behövts utvecklas för att ge en bättre anpassning till materialet, exempelvis genom att inkludera kvadratiska termer och/eller interaktionstermer. Fördelen med att använda regressionsanalys som metod är just denna flexibilitet i utformning av modeller i jämförelse med matrismodellen.

Syftet med att använda en two-part modell istället för en matrismodell eller vanlig multipel regression är att sjukvårdskostnader har en mixad fördelning. Det faktum att sjukvårdskostnader har en starkt skev fördelning korrigeras för genom att logaritmera samt att använda gammafördelningen. Resultatet visar att en mer symmetrisk fördelning uppnås då dessa korrigeringar används. Det är dock troligt att det finns uteliggare som eventuellt påverkar regressionskoefficienterna, något som inte har kontrollerats på grund av det stora antalet individer.

Resultatet visar att resursfördelningen påverkas beroende på vilken modell som används. Procentuellt sätt fördelar alla modeller i stort sätt lika men skillnad finns för storleken på resurserna. De båda two-part modellerna fördelar mindre resurser medan resursfördelningen med den vanliga multipla regressionsmodellen och matrismodellen i stort sätt är lika. Genom att fördela resurser till sjukvårdsdistrikten exakt efter föregående års förbrukning riskeras faktorer som invanda förbrukningsmönster och utbud bevaras, vilket leder till svaga incitament till kostnadskontroll. Att fördela resurser till sjukvårdsdistrikt exakt efter de totala faktiska kostnaderna är därmed inte att eftersträva. Genomsnittskostnaderna i matrismodellen samt koefficienterna i regressionsmodellerna är dock baserade på hela populationen vilket borde minska effekterna av sjukvårdsdistriktens lokala utbud. Matrismodellen fördelar dock totala resurser nära de totala faktiska kostnaderna. En närmare studie över huruvida utbud påverkar resursfördelning med matrismodellen borde därmed genomföras.

Denna uppsats visar att regressionsanalys vid en behovsbaserad resursfördelning blir komplicerad. En vanlig multipel regression ger en modell som avviker starkt från de grundläggande antagandena. Two-part modellen är bättre anpassad efter de förutsättningar som sjukvårdskostnader har men visar sig vara konceptuellt problematisk då individer med samma karakteristiker tilldelas olika resurser. Baserat på de modeller som har jämförts i denna uppsats tycks därmed matrismodellen vara den modell som är bäst lämpad att använda, på grund av dess enkelhet, teoretiskt och praktiskt.

8. Källförteckning

Blough K. David, Madden W. Carolyn, Hornbrook C. Mark (1999). 'Modeling risk using generalized linear models'. *Journal of Health Economics* 18:2 s. 153-171

Blough K. David, Ramsey D. Scott (2000). 'Using Generalized Linear Models to Assess Medical Care Costs'. *Health Services & Outcomes Research Methodology*. 1:2 s. 185-202

Buntin Beeuwkes Melinda, Zaslavsky M. Alan (2004). 'Too much ado about two-part models and transformation? Comparing methods of modelling Medicare expenditures'. *Journal of Health Economics* 23:3 s. 525-542

Diderichsen Finn, Varde Eva (1996). 'Konsten att fördela resurser efter behov. Stockholmsmodellens kriterier'. *Läkartidningen* nr 42 s. 3677-83

Dier P, Yanez D, Ash A, Hornbrook M, Lin D. Y (1999). 'Methods for analyzing health care utilization and costs'. *Annual Review of Public Health*. 20:1 s. 125-144

Duan Naihua, Manning G. Willard, Morris N. Carl, Newhouse P. Joseph. (1983). 'A comparison of alternative models for the demand for health care'. RAND Health Insurance Experiment Series http://www.rand.org/pubs/reports/2006/R2754.pdf

Duan Naihua (1983). 'Smearing Estimate: a nonparametric retransformation method'. *Journal of the American Statistical Association* 78:383 s. 605-610

Lithman Thor (2001). 'Underlag för resursfördelning för hälso- och sjukvård 2002'. Region Skåne; Regionkontoret. Kompetenscentrum Hälso- och sjukvård

Ljung Rickard, Wikström Max, Lundberg Michael, Ponce de Leon Antonio, Diderichsen Finn (2001). 'Förslag till behovsindex för sjukvårdsområden 2002-2004'. Enheten för Socialmedicin. (PM).

Manning G. Willard (1998). 'The logged dependent variable, heteroscedasticity, and the retransformation problem'. *Journal of Health Economics* 17:3 s. 283-295

Manning G. Willard, Mullahy John (2001). 'Estimating log models: to transform or not to transform?'. *Journal of Health Economics* 20:4 s. 461-494

Montez-Rath M., Christiansen C., Ettner S., Loveland S., Rosen A (2006). 'Performance of statistical models to predict mental health and substance abuse cost'. *BMC Medical Research Methodology* 6:53 http://www.biomedcentral.com/1471-2288/6/53

Olsson Ulf (2002). Generalized Linear Models. An applied Approach. Lund: Studentlitteratur

Rice Nigel, Smith Peter (1999). 'Approaches to capitation and risk adjustment in health care: an international survey'. The University of York; Centre of Health Economics. http://www.york.ac.uk/inst/che/pdf/op38.pdf

Smith C. Peter., Rice Nigel, Carr-Hill Roy. (2001) ‘Capitation funding in the public sector’. *Journal of the Royal Statistical Society* 164:2 s. 217-257.

SOU-1996:163. *Behov och Resurser i Vården – en analys*. Statens offentliga utredningar. Göteborg: Fritzes.

Walander Anders, Burström Bo (2005). ’Att fördela psykiatriresurser efter behov. Analyser av behovsindex, vårdutnyttjande och upplevd psykisk hälsa’. Stockholms läns landsting; Centrum för folkhälsa.

Sammanfattning

Att fördela resurser till de olika sjukvårdsdistrikten är en del av Region Skånes verksamhet. Som modell för resursfördelning använde Region Skåne mellan åren 1999-2002 en behovsbaserad resursfördelning. Den behovsbaserade resursfördelningen utgår från att individer med samma socioekonomiska och demografiska karakteristiker har samma behov av vård och ska därmed tilldelas samma mängd resurser. Under perioden med behovsbaserad resursfördelning användes som metod en så kallad matrismodell. I matrismodellen delas befolkningen upp i celler efter alla kombinationer av de socioekonomiska och demografiska variablerna. Därefter beräknas genomsnittliga kostnader i varje cell vilka summeras för varje sjukvårdsdistrikt.

Ett alternativ till matrismodellen är att använda sig av regressionsanalys. Den beroende variabeln sjukvårdskostnader karakteriseras dock av en stor andel individer med noll kostnader och ett fåtal med mycket höga kostnader vilket ger en starkt skev fördelning. Sjukvårdskostnader kan därför anses ha en mixad fördelning, det vill säga den är både diskret och kontinuerlig. För mixade fördelningar har så kallade two-part modeller utvecklas. Genom att använda sig av en two-part modell för resursfördelning kan därmed en mer precis resursfördelning uppnås.

I denna uppsats jämförs därmed matrismodellen med två olika specificeringar av two-part modellen samt med en vanlig multipel linjär regressionsmodell. Särskilt kontrolleras vilka effekter de olika modellerna får på resursfördelningen. Resultatet visar att two-part modellerna totalt sett fördelar mindre resurser än övriga modeller. Slutsatsen är att det finns få vinster med att använda sig av en two-part modell på grund av att analysen blir komplicerad, både teoretiskt och praktiskt.

Bilaga 1; Vanlig multipel linjär regression

The REG Procedure
Model: MODEL1
Dependent Variable: krtot

Number of Observations Read 6000
Number of Observations Used 6000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1.461681E12	1.461681E11	181.52	<.0001
Error	5989	4.822711E12	805261425		
Corrected Total	5999	6.284392E12			

Root MSE	28377	R-Square	0.2326
Dependent Mean	9392.86852	Adj R-Sq	0.2313
Coeff Var	302.11355		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	5549.38785	2248.97462	2.47	0.0136
vldersgrupp		1	1905.59052	745.07064	2.56	0.0106
kon	Kön	1	-2033.74193	749.76985	-2.71	0.0067
alone	Civilstånd	1	103.64431	814.57963	0.13	0.8988
ejarbet		1	2619.71805	1832.29156	1.43	0.1528
Boendetyp		1	-1460.18715	843.20222	-1.73	0.0834
Utbdummy1		1	-238.41356	989.72605	-0.24	0.8096
Utbdummy2		1	-402.94715	870.01014	-0.46	0.6433
Inkdummy1		1	2713.73108	2088.59261	1.30	0.1939
Inkdummy2		1	322.31475	2098.92249	0.15	0.8780
sjukdom	Vårdtunga grupper	1	61984	1503.01279	41.24	<.0001

Bilaga 2; Two-part modellen del 1

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	321.6901	9	<.0001
Score	316.6485	9	<.0001
Wald	297.3702	9	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.0474	0.1784	34.4707	<.0001
vldersgrupp	1	0.3122	0.0650	23.0386	<.0001
kon	1	-0.9611	0.0669	206.3090	<.0001
alone	1	-0.2316	0.0696	11.0769	0.0009
ejarbet	1	0.0412	0.1588	0.0675	0.7951
Boendetyp	1	-0.1970	0.0738	7.1322	0.0076
Utbdummy1	1	-0.1397	0.0845	2.7347	0.0982
Utbdummy2	1	-0.0204	0.0760	0.0724	0.7879
Inkdummy1	1	1.0087	0.1637	37.9636	<.0001
Inkdummy2	1	0.8500	0.1628	27.2644	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
vldersgrupp	1.366	1.203	1.552
kon	0.382	0.335	0.436
alone	0.793	0.692	0.909
ejarbet	1.042	0.763	1.423
Boendetyp	0.821	0.711	0.949
Utbdummy1	0.870	0.737	1.026
Utbdummy2	0.980	0.844	1.137
Inkdummy1	2.742	1.989	3.779
Inkdummy2	2.340	1.701	3.219

Bilaga 3; Two-part modellen OLS

The REG Procedure
Model: MODEL1
Dependent Variable: lnkrtot

Number of Observations Read	6000
Number of Observations Used	4641
Number of Observations with Missing Values	1359

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	2287.13673	228.71367	131.30	<.0001
Error	4630	8065.04432	1.74191		
Corrected Total	4640	10352			

Root MSE	1.31981	R-Square	0.2209
Dependent Mean	8.10822	Adj R-Sq	0.2193
Coeff Var	16.27749		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	7.88410	0.12855	61.33	<.0001
vldersgrupp		1	0.13126	0.03926	3.34	0.0008
kon	Kön	1	-0.15739	0.03983	-3.95	<.0001
alone	Civilstånd	1	0.11623	0.04345	2.67	0.0075
ejarbet		1	0.16115	0.09738	1.65	0.0980
Boendetyp		1	-0.12404	0.04441	-2.79	0.0052
Utbdummy1		1	-0.08219	0.05262	-1.56	0.1184
Utbdummy2		1	-0.07137	0.04585	-1.56	0.1197
Inkdummy1		1	0.27833	0.12079	2.30	0.0213
Inkdummy2		1	0.04777	0.12210	0.39	0.6956
sjukdom	Vårdtunga grupper	1	2.35883	0.07083	33.30	<.0001

Bilaga 4; Two-part modellen GLM

The GENMOD Procedure

Model Information

Data Set	XXX.TRAININGSET
	Predicted Values and Diagnostic Statistics
Distribution	Gamma
Link Function	Log
Dependent Variable	krtot

Number of Observations Read	6000
Number of Observations Used	4641
Number of Invalid Responses	1359

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	4630	8454.9288	1.8261
Scaled Deviance	4630	5646.7717	1.2196
Pearson Chi-Square	4630	24716.4631	5.3383
Scaled Pearson X2	4630	16507.3211	3.5653
Log Likelihood		-46203.9668	

Algorithm converged.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	9.0891	0.1210	8.8518	9.3263	5639.94	<.0001
vldersgrupp	1	0.2007	0.0368	0.1285	0.2728	29.73	<.0001
kon	1	-0.1014	0.0369	-0.1737	-0.0290	7.53	0.0061
alone	1	0.0737	0.0401	-0.0049	0.1523	3.38	0.0662
ejarbet	1	0.0605	0.0920	-0.1198	0.2409	0.43	0.5108
Boendetyp	1	-0.1355	0.0407	-0.2153	-0.0557	11.07	0.0009
Utbdummy1	1	-0.2901	0.0488	-0.3858	-0.1945	35.34	<.0001
Utbdummy2	1	-0.2020	0.0428	-0.2860	-0.1181	22.26	<.0001
Inkdummy1	1	0.0811	0.1133	-0.1409	0.3032	0.51	0.4740
Inkdummy2	1	-0.1700	0.1143	-0.3940	0.0539	2.21	0.1368
sjukdom	1	2.2085	0.0655	2.0801	2.3369	1136.46	<.0001
Scale	1	0.6679	0.0118	0.6452	0.6913		

NOTE: The scale parameter was estimated by maximum likelihood.

Bilaga 5; SAS-koder

Vanlig multipel linjär regression;

```
/* Vanlig mulitipel linjär regression */

proc reg data = xxx.trainingset outest = one  ;

model krtot = vldersgrupp kon alone ejarbet Boendetyp Utbdummy1 Utbdummy2
Inkdummy1 Inkdummy2 sjukdom;

output out = xxx.trainingset p=predone r=resone;
run ;

/* Validering */

 proc score data = yyy.validationset1 score=one out=validone type=parms
nostd predict ;
 var krtot vldersgrupp kon alone ejarbet Boendetyp Utbdummy1 Utbdummy2
Inkdummy1 Inkdummy2 sjukdom ;
run ;
```

Two-part modellen;

```
/* Del 1 two-part modellen */

proc logistic data=xxx.trainingset outmodel=onepart ;
model kostar (event='1')= vldersgrupp kon alone ejarbet Boendetyp Utbdummy1
Utbdummy2 Inkdummy1 Inkdummy2 ;
run ;

/* Validering del 1 */

proc logistic inmodel=onepart ;
score data=yyy.validationset1 out=yyy.validationset1 ;
run ;

/* Two-part OLS */

proc reg data= xxx.trainingset outest=two2;
model lnkrtot = vldersgrupp kon alone ejarbet Boendetyp Utbdummy1 Utbdummy2
Inkdummy1 Inkdummy2 sjukdom;

output out= xxx.trainingset p=predtwo r=restwo2;
run ;


/* Validering two-part OLS */

proc score data=yyy.validationset1 score=two2 out=validtwools type=parms
nostd predict ;
 var lnkrtot vldersgrupp kon alone ejarbet Boendetyp Utbdummy1 Utbdummy2
Inkdummy1 Inkdummy2 sjukdom ;
```

```
run ;

/*  Two-part GLM */

proc genmod data=xxx.trainingset ;
model krtot = vldersgrupp kon alone ejarbet Boendetyp Utbdummy1 Utbdummy2
Inkdummy1 Inkdummy2 sjukdom / dist=gamma link=log ;
 output out= xxx.trainingset  resdev=glmres pred=glmskatt;
run ;

/* Validering  two-part GLM */

data xxx.trainingset ;
set xxx.trainingset ;
weight = 1 ;
run ;

data yyy.validationset1 ;
set yyy.validationset1 ;
weight = 0 ;
run ;

data allaobs ;
set xxx.trainingset yyy.validationset1 ;
run ;

proc genmod data=allaobs ;
weight weight ;
model krtot = vldersgrupp kon alone ejarbet Boendetyp Utbdummy1 Utbdummy2
Inkdummy1 Inkdummy2 sjukdom / dist=gamma link=log ;
output out=allaobs p=glmskattad ;
run ;
```