Stockholms universitet

Regressionsanalys av antalet biståndstagare i svenska landsbygds- och storstadskommuner

Marika Lisinski

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2017:4
Matematisk statistik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2017:4**
http://www.math.su.se/matstat

Regressionsanalys av antalet biståndstagare i svenska landsbygds- och storstadskommuner

Marika Lisinski*

Juni 2017

Sammanfattning

Beslut om vem som får ekonomiskt bistånd i Sverige tas kommunalt. I detta arbete undersöker vi om det existerar något samband mellan antalet biståndstagare och typ av kommun, tillsammans med faktorer kön och ålder. Typer av kommuner som undersöks är landsbygds- och storstadskommuner och sker genom modellering med Poisson- och negativ binomialregression. Lämpligast modell väljs sedan ut med avseende på AIC, där låga värden innebär en bättre anpassning av data. Till skillnad från Poissonfördelningen behöver inte väntevärde och varians vara lika i den negativa binomialfördelningen och vi får att regressionsmodeller med negativ binomialfördelning är lämpligare för att beskriva data. Undersökning av förklaringsvariabler visar att årtal inte bör behandlas som en linjär variabel och vi behandlar den istället som en kategorisk. Denna modell tillåter ingen prediktion av framtida antal biståndstagare och alternativa metoder som splinregression eller tidsserieanalys för att hantera icke-linjäritet i årtal kan vara att föredra.

Vi ser att det finns signifikant skillnad på antalet biståndstagare i landsbygdskommuner och i storstadskommuner, denna skillnad är dock icke konsekvent beroende på fördelningsantagande. Vi observerar att det existerar fler underliggande faktorer inom kommunindelningarna och skillnaden i antalet biståndstagare beror på mer än endast om de är landsbygds- eller storstadskommuner. Vår slutgiltliga modell visar en minskning av antalet biståndstagare i högre åldrar samt att en lägre andel män tar ekonomiskt bistånd både på landsbygd och i storstad. Vi obseverar även en minskning av antalet biståndstagare över tiden vilket skiljer sig från tidigare studier inom ämnet vilket kan förklaras av vårt urval av kommuner.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: mali1218@student.su.se. Handledare: Martin Sköld.

Abstract

The decision on who is allowed financial aid in Sweden is taken separately in each county. In this essay we will examine the dependence between the number of people getting financial aid with type of county, together with age and gender. We will compare counties in the countryside with counties in the cities using Poisson- and negative binomial regression. The most suitable model will then be chosen with respect to its AIC value, where the lower value is preferable. Models with different types of explanatory variables are compared, where the negative binomial regression with age och year treated as categorical variables turns out to be the best suitable model to describe data. This model does not allow for prediction of future number of people getting financial aid and alternative methods such as spline regression or time serie analysis to handle the non-linearity in the year variable could be preferable.

Our final model shows that the number of people getting financial aid is decreasing when age is increasing and also that the proportion of men getting financial aid is lower than the proportion of women. We can also establish that there is a significant difference between the proportion of people getting financial aid in the city counties compared to the counties on the countryside. Although there are other underlying factors within the counties and the difference between them depends on more than just if they are in the city or on the countryside. We can observe a decreasing number of people getting financial aid with time which differ from earlier results on similar studies. This could be explained by our selection of counties.

Sammanfattning

Beslut om vem som får ekonomiskt bistånd i Sverige tas kommunalt. I detta arbete undersöker vi om det existerar något samband mellan antalet biståndstagare och typ av kommun, tillsammans med faktorer kön och ålder. Typer av kommuner som undersöks är landsbygds- och storstadskommuner och sker genom modellering med Poisson- och negativ binomialregression. Lämpligast modell väljs sedan ut med avseende på AIC, där låga värden innebär en bättre anpassning av data. Till skillnad från Poissonfördelningen behöver inte väntevärde och varians vara lika i den negativa binomialfördelningen och vi får att regressionsmodeller med negativ binomialfördelning är lämpligare för att beskriva data. Undersökning av förklaringsvariabler visar att årtal inte bör behandlas som en linjär variabel och vi behandlar den istället som en kategorisk. Denna modell tillåter ingen prediktion av framtida antal biståndstagare och alternativa metoder som splinregression eller tidsserieanalys för att hantera icke-linjäritet i årtal kan vara att föredra.

Vi ser att det finns signifikant skillnad på antalet biståndstagare i landsbygdskommuner och i storstadskommuner, denna skillnad är dock icke konsekvent beroende på fördelningsantagande. Vi observerar att det existerar fler underliggande faktorer inom kommunindelningarna och skillnaden i antalet biståndstagare beror på mer än endast om de är landsbygds- eller storstadskommuner. Vår slutgiltliga modell visar en minskning av antalet biståndstagare i högre åldrar samt att en lägre andel män tar ekonomiskt bistånd både på landsbygd och i storstad. Vi obseverar även en minskning av antalet biståndstagare över tiden vilket skiljer sig från tidigare studier inom ämnet vilket kan förklaras av vårt urval av kommuner.

Förord och tack

Detta arbete är gjort i samband med en kandidatexamen i matematisk statistik vid Stockholms Universitet och utgör 15 högskolepoäng. Ett stort tack till min handledare Martin Sköld för ett genomgående stöd och säker hjälp under arbetets gång.

Innehållsförteckning

1 Inledning **6**

2 Data **8**
2.1 Statistikdatabaser 8
2.2 Regionsindelning 8
2.3 Variabler 8

3 Teori **9**
3.1 Generaliserade Linjära Modeller 10
3.2 Poissonregression 10
3.3 Negativ binomialregression 11
3.4 EM-algoritmen 12
3.4.1 Tillämpning 12
3.4.2 Härledning 12
3.5 Akaikes Informationskriterium 13

4 Modellering **13**
4.1 Offsetvariabel 13
4.2 Poissonregression med EM-algoritmen 14
4.3 Negativ binomialregression med EM-algoritmen 14
4.4 Variabeltyp 15

5 Resultat **17**
5.1 Stad och land 17
5.2 Modelljämförelse 18
5.3 Koefficientskattningar 18

6 Diskussion **20**
6.1 Hantering av okända observationer 20
6.2 Alternativa modeller 20
6.3 Tolkning 21

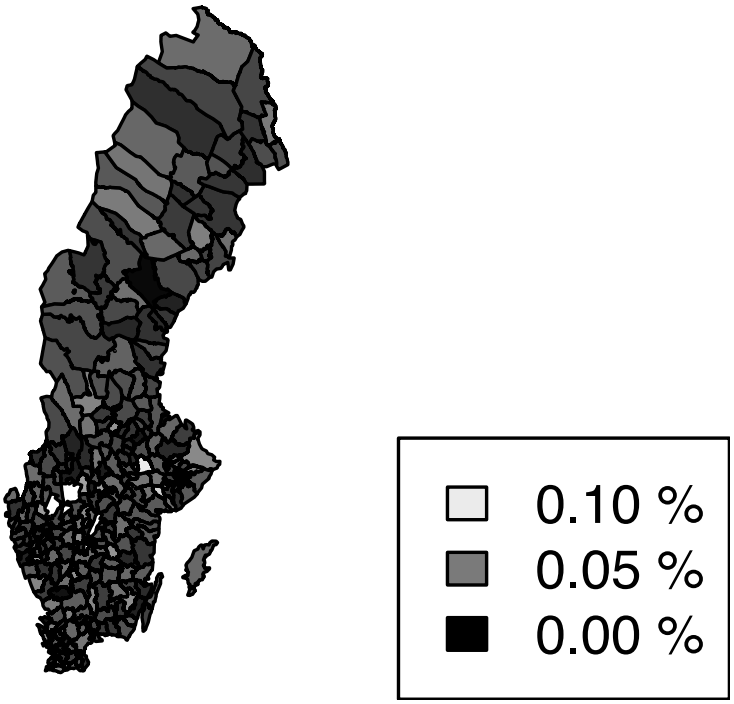
7 Appendix **22**
7.1 Data 22
7.2 Modellbyggnad 22
7.2.1 Koefficientskattningar 23
7.2.2 P-värden och residualer 24

8 Referenser **26**

9 Bilagor **27**

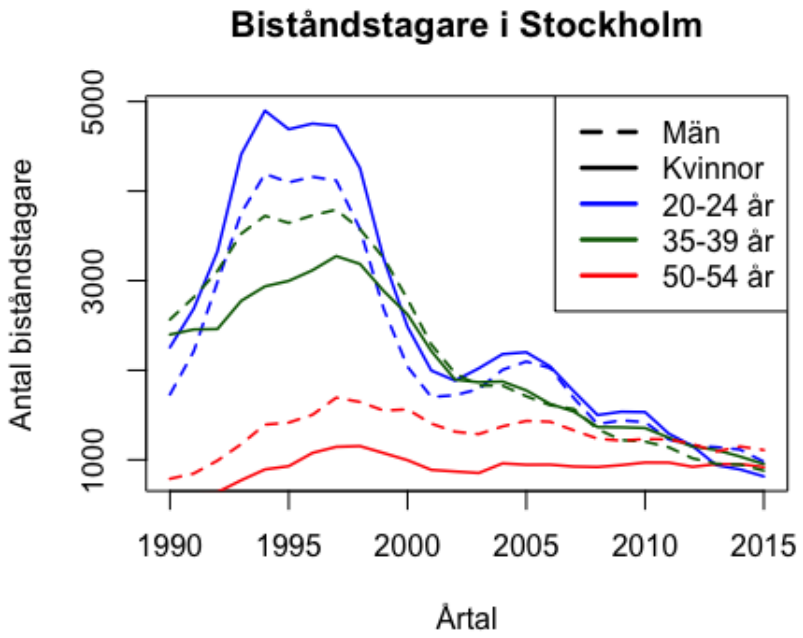
1 Inledning

Ekonomiskt bistånd i Sverige består av ett försörjningsstöd som i sig består dels av en riksnorm och dels av skäliga kostnader utanför riksnormen. Beslut om belopp och beviljande av ekonomiskt bistånd tas av socialtjänsten i vardera kommun. Riksnormen är i stort sett samma för samtliga biståndstagare medan bidraget för skäliga kostnader kan skilja sig åt beroende på handläggare och kommun (Socialstyrelsen, 2017). Vi vill undersöka om antalet biståndstagare skiljer sig åt mellan kommuner och kommer göra det genom att modellera detta antal med kommun, ålder och kön som förklarande variabler.



Figur 1: Andel biståndstagare kommunvis i Sverige.

I Figur 1 får vi en överblick av hur stor andel av befolkningen i vardera kommun som får ekonomiskt bistånd. Utan att se något klart samband mellan kommuner inriktar vi oss på landsbygds- och storstadskommuner för att se om och hur antalet biståndstagare skiljer sig åt mellan dessa. Kommunsindelningen får vi från Sveriges Kommuner och Landsting och väljer här att använda även storstadsnära kommuner tillsammans med storstadskommuner, se Tabell 9 och 10 i appendix 7.1 för specifikation av dessa.



Figur 2: Antalet manliga respektive kvinnliga biståndstagare i Stockholms kommun från 1990-2015 för tre olika åldersgrupper.

Förutom regionseffekt vill vi även undersöka hur kön och ålder påverkar antalet biståndstagare. I Figur 2 ser vi hur antalet biståndstagare varierar i Stockholms kommun från 1990 till 2015 för män och kvinnor i tre olika åldergrupper om fem år vardera. Vi observerar en stark ökning av antalet biståndstagare i de yngre åldergrupperna under 90-talet för att sedan stadigt minska med tiden. Observerar även ett lägre antal biståndstagare i högre åldrar samt att majoriteten biståndstagare växlar mellan att vara män och kvinnor för olika åldersgrupper. Vi kommer vidare i detta arbete undersöka om dessa faktorer har någon signifikant påverkan på antalet biståndstagare.

2 Data

Nedan följer en beskrivning samt en överblick av den insamlade data vi kommer använda för vidare analys. Samtlig datahantering och dataanalys i detta arbete sker i den statistiska programvaran R (R core team 2017).

2.1 Statistikdatabaser

Data som används är hämtad från dels statistiska centralbyrån och dels socialstyrelsens statistikdatabas över ekonomiskt bistånd. Observerar här att mellan åren 1993 och 2011 ingick även introduktionsersättning för flyktingar i statistiken över ekonomiskt bistånd. Vi ser i Figur 15 och 16 i bilagor att utbetalat biståndsbelopp inklusive respektive exklusive introduktionsersättningen skiljer sig relativt lite åt och vi antar att antalet personer med introduktionsersättning under denna tidsperiod är litet. Inga åtgärder tas i vidare analys för att urskilja dessa ersättningar från övriga.

Från socialstyrelsens statistikdatabas över ekonomiskt bistånd använder vi data över antalet myndiga biståndstagare i sverige under tidsperioden 1990 till 2015. Data är indelad på kommuns-, köns- och åldersnivå. Vi exkluderar bortfallskommuner där inga mätningar har gjorts ur vidare analys. Från statistiska centralbyrån använder vi data över folkmängden i Sverige under samma tidsperiod och motsvarande kommuns-, köns- och åldersindelning.

2.2 Regionsindelning

Vi vill undersöka hur antalet biståndstagare skiljer sig åt beroende på typ av kommun. Vi delar därför upp data i två grupper, en med landsbygdskommuner och en med storstadskommuner, data reduceras därmed till sammanlagt 86 kommuner. Vi tar sedan fram modeller separat för respektive grupp för att undersöka möjliga skillnader mellan dem. Utöver dessa gör vi även en modell där vi istället för specifika kommuner använder en binär variabel vilken beskriver om en kommun är landsbygds- eller storstadskommun. Detta för att undersöka om variation mellan kommuner kan beskrivas enbart av om de är just landsbygdskommuner eller storstadskommuner.

2.3 Variabler

Då ovan nämnda bortfallskommuner exkluderas ur data och vi endast ser till landsbygds- och storstadskommuner har vi kvar $n = 48\ 257$ observationer av variabler vilka redogörs för i Tabell 1 nedan. Dessa observationer antas vara oberoende i vidare analys.

Tabell 1: Variabler.

Variabel	**Utfallsrum**
Antal (personer med ekonomiskt bistånd)	0,X,4,5,...
Årtal	$1990, ..., 2015$
Kommun	40 landsbygds- respektive 46 storstadskommuner
Kön	män, kvinnor
Ålder	18-19 år, 20-24 år, 25-30, år, ... , 65+ år
Folkmängd	0,1,2,...
Stad	0 om landsbygdskommun, 1 om storstadskommun

Tabell 2: Utdrag ur datamaterial.

Årtal	**Kommun**	**Kön**	**Ålder**	**Antal**	**Folkmängd**	**Stad**
2015	Huddinge	Kvinnor	18-19 år	39	1218	1
2015	Botkyrka	Män	18-19 år	81	1152	1
2015	Botkyrka	Kvinnor	18-19 år	77	1069	1
2015	Salem	Män	18-19 år	X	183	1
2015	Salem	Kvinnor	18-19 år	X	212	1
2015	Haninge	Män	18-19 år	50	963	1

Observerar i Tabell 1 att utfallsrummet för antal personer med ekonomiskt bistånd inte innehåller 1, 2 eller 3. Detta är på grund av att data inte ska vara utlämnande på individnivå (Socialstyrelsen, 2017), observationer av tre eller färre biståndstagare ersätts därför med X.

I Tabell 2 får vi en överblick av hur datamaterialet är uppbyggt. Vi ser till exempel att år 2015 var 1152 män i åldern 18-19 år folkbokförda i Botkyrka kommun, vilken räknas som en storstadskommun, varav 81 stycken tog ekonomiskt bistånd. Observerar även att antal personer i åldern 18-19 år med ekonomiskt bistånd i Salems kommun år 2015 anges med X vilket betyder att det fanns tre eller färre biståndstagare bland kvinnor respektive män.

3 Teori

I följande avsnitt ges en grundläggande beskrivning av den teori vi kommer att använda oss av för att modellera data. Parameterskattningar görs genom maximum likelihood-metoden, se appendix 7.2. Som nämnts används R för att utföra denna modellering, vi går inte närmre in på i denna uppsats hur R implementerar den teori vi redogör för nedan.

3.1 Generaliserade Linjära Modeller

Följande information kommer från Agresti (2002;116-117). I allmänna linjära modeller antas observationer vara oberoende och normalfördelade. Generaliserade linjära modeller, GLM, är en generalisering av dessa där observationer istället antas komma från en fördelning tillhörande den naturliga exponentialfamiljen.

En GLM består av följande tre komponenter

- Stokastisk komponent
- Systematisk komponent
- Länkfunktion

Den stokastiska komponenten är responsvariabel Y med parameter θ_i och oberoende observationer $y = (y_1, ..., y_n)$ där n är antalet observationer. Parametern θ_i kan variera för olika $i = 1, ..., n$ beroende på värden av förklaringsvariabler. Då fördelningen för Y tillhör den naturliga exponentialfamiljen kan täthets- eller sannolikhetsfördelningen, beroende på om Y är kontinuerlig eller diskret, skrivas på formen

$$f_Y(y_i; \theta_i) = a(\theta_i)b(y_i)\exp(y_i Q(\theta_i)). \qquad (1)$$

Den systematiska komponenten beskriver en linjär modell av förklaringsvariabler. Anta att vi har p stycken förklaringsvariabler, den systematiska komponenten ges då av vektorn $\eta = (\eta_1, ..., \eta_n)$ där

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad \text{för } i = 1, ..., n,$$

där x_{ij} är värdet av förklaringsvariabel j i observation i och där β_j och intercept β_0 är okända parametrar.

Länkfunktionen länkar till sist samman den stokastiska och den systematiska komponenten genom en monoton och differentierbar funktion g sådan att $g(\mu_i) = \eta_i$ där $\mu_i = E[Y_i]$.

3.2 Poissonregression

Anta att vi har n stycken oberoende observationer av stokastisk variabel Y vilken endast kan anta icke-negativa heltalsvärden utan någon övre gräns. Ett enkelt antagande är då att Y kommer från en Poissonfördelning vilken endast beror på en parameter μ_i, där värdet på μ_i kan variera för olika $i = 1, ..., n$ beroende på värden på förklaringsvariabler. Väntevärde, varians och sannolikhetsfunktion för Y ges då av (Agresti 2002;7)

$$E[Y_i] = Var(Y_i) = \mu_i,$$

$$p(y_i;\mu_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \quad \text{för } y_i = 0,1,2,... \text{ och } i = 1,...,n. \quad (2)$$

Omskrivning av (2) ger

$$p(y_i;\mu_i) = \frac{1}{y_i!}\exp(-\mu_i)\exp(y_i \log \mu_i) \quad \text{för } y_i = 0,1,2,...\text{och } i = 1,...,n,$$

vilket är på samma form som (1) där $a(\mu_i) = \exp(-\mu_i)$, $b(y_i) = \frac{1}{y_i!}$ och $Q(\mu_i) = \log \mu_i$. Detta betyder att Poissonfördelningen är en del av den naturliga exponentialfamiljen och vi kan använda GLM för att modellera data. Vår stokastiska komponent är Poissonfördelad med parameter μ_i, den systematiska komponenten ges av $\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ och vår länkfunktion ges av den naturliga logaritmen. Vi får alltså den generaliserade linjära modellen

$$\log(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \quad \text{för } i = 1,...,n.$$

3.3 Negativ binomialregression

Följande information kommer från Agresti (2002;559-561). Vi såg i avsnittet ovan att för Poissonfördelningen gäller det att väntevärde och varians är ekvivalenta. I många fall kan detta vara ett naivt antagande och ett alternativ till Poissonfördelningen är den negativa binomialfördelningen. Denna fördelning kan tyckas mer komplicerad än Poissonfördelningen då den istället beror på två parametrar men kan i många fall ge en säkrare modellskattning. För en negativt binomialfördelad stokastisk variabel Y med parametrar μ och k, där k är en spridningsparameter, ser väntevärde, varians och sannolikhetsfunktion ut som följer

$$E[Y] = \mu,$$

$$Var(Y) = \mu + \frac{\mu^2}{k}, \qquad (3)$$

$$p(y;\mu,k) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y \quad \text{för } y = 0,1,2,... \quad (4)$$

Vi observerar i (3) att då $k \to \infty$ går fördelningen mot en Poissonfördelning. Vidare ger omskrivning av (4)

$$p(y;\mu,k) = \exp\left(\log\left(\frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\right) + k\log\left(\frac{k}{\mu+k}\right) + y\log\left(1-\frac{k}{\mu+k}\right)\right).$$

För ett fixt värde på k känner vi igen detta som motsvarande form som (1) och vi kan behandla detta som en GLM med den naturliga logaritmen som länkfunktion. Observerar

dock att för ett okänt värde på spridningsparametern k tillhör den negativa binomialfördelningen inte den naturliga exponentialfamiljen och uppfyller därför inte kraven för att GLM ska kunna användas.

3.4 EM-algoritmen

Följande information kommer från Held & Bové (2014; 34-37). Då man har okända observationer i ett dataset kan man hantera detta med EM-algoritmen. Detta är en iterativ likelihoodbaserad algoritm som använder sig av skattade väntevärden av de okända observationerna för att ta fram konvergerande ML-skattning av okänd parameter θ.

3.4.1 Tillämpning

Låt $\theta_{t=0}$ vara vår startgissning av parameter θ. Vi använder denna för att få ut nästa gissning genom att följa stegen nedan.

1. Beräkna $Q(\theta;\theta_t) = E_{\theta_t}[l(\theta;x,U)]$
2. $\theta_{t+1} = max_\theta Q(\theta;\theta_t)$
3. Upprepa steg 1. och 2. med $\theta_t = \theta_{t+1}$

Här är x observerad data och U okänd data. Stegen ovan upprepas tills $|\theta_t - \theta_{t+1}| < \epsilon$ för något litet $\epsilon > 0$, från vilket vi får att $\hat{\theta}_{ML} = \theta_t$. Vidare i detta arbete använder vi $\epsilon = 1.0e-5$ som gräns för att stoppa iterationen.

3.4.2 Härledning

Anta att vi har observerad data X och okänd data U. Den sammansatta täthets- eller sannolikhetsfunktionen för den fullständiga datan X och U följer då av

$$f(u|x) = \frac{f(x,u)}{f(x)} \iff f(x,u) = f(u|x)f(x).$$

Motsvarande loglikelihoodfunktion ges då av

$$l(\theta;x,u) = l(\theta;u|x) + l(\theta;x).$$

Då vi inte känner till u byter vi ut denna mot stokastisk variabel U. Vidare ges väntevärde med avseende på θ_t av höger- och vänsterled av

$$E_{\theta_t}[l(\theta;x,U)] = E_{\theta_t}[l(\theta;U|x)] + E_{\theta_t}[l(\theta;x)].$$

Låt $E_{\theta_t}[l(\theta;x,U)] = Q(\theta;\theta_t)$ och $E_{\theta_t}[l(\theta;U|x)] = C(\theta;\theta_t)$. Observerar även att $E_{\theta_t}[l(\theta;x)] =$

$l(\theta; x)$ då $l(\theta; x)$ inte beror på U. Man kan sedan visa att

$$Q(\theta; \theta_t) \geq Q(\theta_t; \theta_t) \Longrightarrow l(\theta) - l(\theta_t) \geq C(\theta_t; \theta_t) - C(\theta; \theta_t) \geq 0,$$

vilket innebär att för varje iteration av EM-algoritmen ökar loglikelihooden i värde. Att detta ger oss ML-skattningen av θ följer av att logaritmen är en strikt monoton function vilket betyder att

$$\hat{\theta}_{ML} = argmax_{\theta \in \Theta} l(\theta).$$

3.5 Akaikes Informationskriterium

Akaikes informationskriterium, AIC, är ett mått på hur väl en modell anpassar data relativt andra modeller. För en enskild modell säger dess AIC-värde oss ingenting om modellens anpassningsförmåga utan är endast relevant då det jämförs mellan flera möjliga modeller.

"Optimal model is the one that tends to have fit closest to reality" Agresti (2002). Alltså optimal modell ges av den vars anpassade värden är så nära de observerade värdena som möjligt. Akaike visade att den mest lämpliga modellen är den med lägst AIC-värde, där AIC för modell M ges av Held & Bové (2014;224)

$$AIC(M) = -2l(M) + 2p(M).$$

Här är $l(M)$ den maximerade loglikelihoodfunktionen i modell M och $p(M)$ är antal parametrar i samma modell. Ett högt antal parametrar i en modell straffar sig alltså med att AIC-värdet ökar.

4 Modellering

Nedan redogör vi för hur de generaliserade linjära modellerna beskrivna ovan byggs upp tillsammans med EM-algoritmen för att modellera data beskriven i avsnitt 2. Dessa används sedan för att modellera olika typer av förklaringsvariabler vilka redogörs för i avsnitt 4.4 och 5.

4.1 Offsetvariabel

För att modellera sambandet mellan kommun och antal biståndstagare måste vi ta hänsyn till folkmängden i vardera kommun. Vi använder oss av en offsetvariabel för att modellera andelen biståndstagare (Agresti, 2002). Då vår länkfunktion är den naturliga logaritmen betyder detta att vi får det linjära sambandet

$$\log\left(\frac{\mu_i}{b_i}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \Longleftrightarrow \log(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \log(b_i),$$

där $\log(b_i)$ är vår offsetvariabel och b_i är folkmängden i observation i.

4.2 Poissonregression med EM-algoritmen

Då antalet biståndstagare är en icke-negativ heltalsvärd variabel utan någon övre gräns kan man anta att denna är Poissonfördelad med parameter μ_i där

$$\mu_i = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \log(b_i)\right) \quad , i = 1, ..., n.$$

För att hantera den ospecifierade datan över tre eller färre biståndstagare använder vi EM-algoritmen för att skatta modellen. Låt $y = (y_1, ..., y_n)$ vara vår fullständiga observationsvektor. Vi delar sedan in denna i $u = \{y_i; y_i = X, i = 1, ..., n\}$ och $z = \{y_i; y_i \neq X, i = 1, ..., n\}$ där u är okända observationer från stokastisk variabel U och z är kända observationer. Låt sedan $\theta = (\beta_0, ..., \beta_p)$ vara vår parametervektor samt $x_i = (1, x_{i1}, ..., x_{ip})$ en vektor med våra förklaringsvariabler i observation i. Av detta får vi

$$Q(\theta; \theta_t) = \sum_{a \in A} z_a \log(\mu_a) - \mu_a + \sum_{b \in B} E_{\theta_t}[U_b] \log(\mu_b) - \mu_b,$$

där $A = \{i; y_i \neq X, i = 1, ..., n\}$ och $B = \{i; y_i = X, i = 1, ..., n\}$. Vidare fås $E_{\theta_t}[U_b]$ genom

$$p_{U_b}(u) = \frac{p_{Y_b}(u)}{p_{Y_b}(1) + p_{Y_b}(2) + p_{Y_b}(3)} \iff E_{\theta_t}[U_b] = \frac{p_{Y_b}(1) + 2p_{Y_b}(2) + 3p_{Y_b}(3)}{p_{Y_b}(1) + p_{Y_b}(2) + p_{Y_b}(3)},$$

där $Y_b \sim Po(\exp(\theta_t x_b + \log(b_b)))$. Vi låter startgissningen θ_0 vara ML-skattningen vi får då vi sätter de ospecifierade låga värdena till medelvärdet av möjliga utfall, alltså $X = 2$. Vi maximerar sedan $Q(\theta; \theta_t)$ genom funktion glm i R där vi använder förväntade värden av U i responsvektorn. Då $Q(\theta; \theta_t)$ är på samma struktur som loglikelihoodfunktionen för en Poissonfördelad stokastisk variabel anger vi denna som Poissonfördelad för att få ut motsvarande maximum likelihood-skattningar. Vi stoppar iterationen då $|\theta_t - \theta_{t+1}| < 1.0e-5$.

4.3 Negativ binomialregression med EM-algoritmen

Det kan tyckas rimligt att väntevärde och varians är olika för antalet biståndstagare varför negativ binomialfördelning vara ett rimligare fördelningsantagande. Då vi antar negativ binomialfördelning måste vi även skatta spridningsparametern k. Vi skattar denna med R-funktion $glm.nb$ för data där vi sätter $X = 2$ och använder sedan denna som fixt värde på k under iteration med EM-algoritmen. På motsvarande sätt som för Poissonregression får vi då

$$Q(\theta; \theta_t) = \sum_{a \in A} k \log\left(\frac{k}{\mu_a + k}\right) + z_a \log\left(1 - \frac{k}{\mu_a + k}\right) + \sum_{b \in B} k \log\left(\frac{k}{\mu_b + k}\right) + E_{\theta_t}[U_b] \log\left(1 - \frac{k}{\mu_b + k}\right),$$

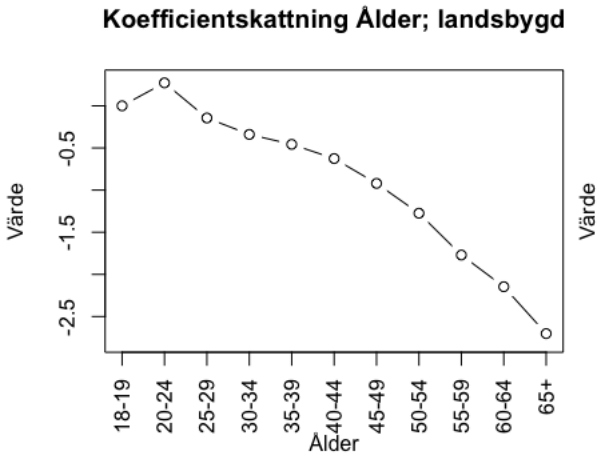
där väntevärdet av U_b fås av

$$E_{\theta_t}[U_b] = \frac{p_{Y_b}(1) + 2p_{Y_b}(2) + 3p_{Y_b}(3)}{p_{Y_b}(1) + p_{Y_b}(2) + p_{Y_b}(3)},$$

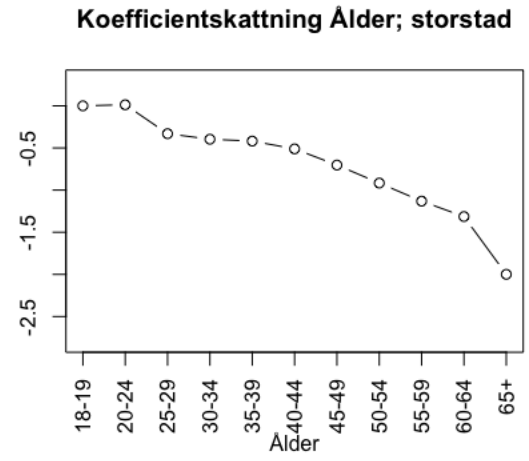
och $Y_b \sim NB(\exp(\theta_t x_b + \log(b_b)), k)$. Steg 2 i algoritmen fås sedan på motsvarande sätt som ovan med skillnaden att vi här anger $Q(\theta; \theta_t)$ som negativt binomialfördelad med spridningsparamer k.

4.4 Variabeltyp

Vi använder variablerna vi såg i Tabell 1 för att modellera data. Av dessa använder vi region och kön som kategoriska variabler. Ett första antagande är att även behandla åldersgrupp som en kategorisk variabel, samt årtal som en numerisk. Då vi undersöker dessa närmre ser vi dock att det finns anledning att tro att de bör behandlas annorlunda.



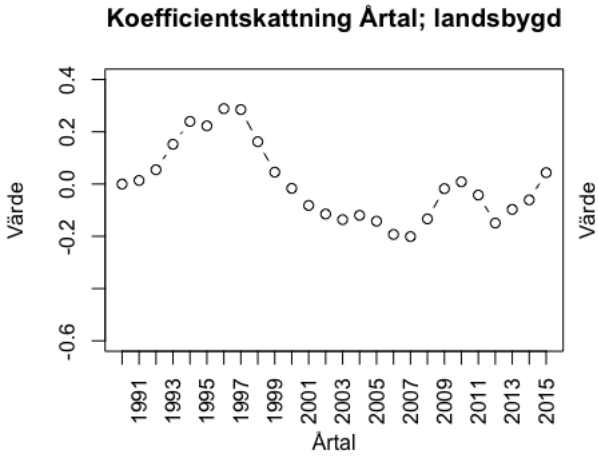
Figur 3:
Poissonregression med ålder som kategorisk variabel



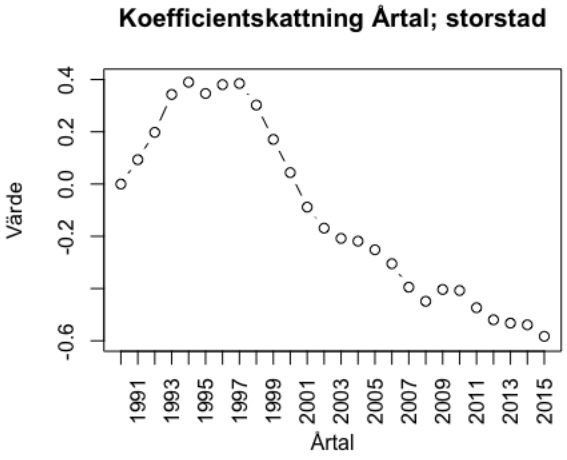
Figur 4:
Poissonregression med ålder som kategorisk variabel

I Figur 3 och 4 ovan ser vi koefficientskattningar av de olika åldersgrupperna då vi använder Poissonregression för landsbygdskommuner respektive storstadskommuner. Vi ser att skattningarna skiljer sig åt en aning mellan landsbygd och storstad men att i båda fallen kan vi ana ett linjärt negativt samband mellan ålder och det logaritmerade antalet biståndstagare. Samma tendenser ser vi då vi använder negativ binomialregression, se Figur 9 och 10 i appendix 7.2.1.

Vi testar vårt antagande om att årtal är en linjär variabel på motsvarande sätt. I Figur 5 och 6 nedan ser vi koefficientskattningar då årtal behandlas som en kategorisk variabel i Poissonregression över landsbygds- respektive storstadskommuner. Vi ser att dessa inte ser ut att följa ett linjärt samband. Vi observerar även skillnader mellan kommunstyperna framförallt från år 2002 och framåt. Se Figur 11 och 12 i appendix 7.2.1 för motsvarande koefficientskattningar för negativ binomialregression.



Figur 5:
Poissonregression med årtal som kategorisk variabel



Figur 6:
Poissonregression med årtal som kategorisk variabel

5 Resultat

Vi använder modellerna beskrivna ovan för att testa olika typer av förklaringsvariabler. Bäst lämpade modeller väljs sedan ut med avseende på AIC-värde.

5.1 Stad och land

Då vi använder binär variabel *Stad* från Tabell 1 istället för kategorisk variabel *Kommun* ser vi att det är signifikant skillnad mellan landsbygds- och storstadskommuner på alla signifikansnivåer. Effekten skiljer sig dock t beroende p frdelningsantagande.

Tabell 3: Skattning av binär regionsvariabel med Poisson- respektive negativ binomialregression.

Fördelning	Koefficientskattning stad	Standardavvikelse	P-värde
Poisson	0.1148	0.00157	0
Negativ Binomial	−0.1621	0.00576	0

Tabell 4: AIC för modell med region som binär respektive kategorisk förklaringsvariabel.

Region	AIC Poisson	AIC negativ binomial
Binär	1262296	388379
Kategorisk	700143	**362486**

I Tabell 3 ser vi att då vi antar Poissonfördelning får vi att storstadskommuner bidrar med en skattad effekt på 0.1148, med en standardavvikelse på ca. 0.002, på det logaritmerade antalet biståndstagare. Om vi däremot antar negativ binomialfördelning blir motsvarande effekt −0.1621, här med en standardavvikelse på ca. 0.006. I Tabell 4 ser vi att vi får betydligt lägre AIC-värde för den negativa binomialmodellen och väljer därför den framför Poissonregressionen.

I Tabell 4 ser vi också att vi får ett lägre AIC-värde då vi använder alla specifika kommuner som nivåer, vilket betyder att det finns variation inom landsbygds- och storstadskommunerna som beror på mer än endast om de är landsbygd- eller storstadskommuner. Vi väljer därför att betrakta region som en kategorisk variabel och gör separata modeller för landsbygds- respektive storstadskommuner.

5.2 Modelljämförelse

Vi såg i avsnitt 4.4 att det finns anledning att tro att årtal inte bör behandlas som en linjär variabel. Som alternativ till detta använder vi istället årtal som en kategorisk variabel med 26 nivåer. Detta är inte ett optimalt alternativ då det betyder att vi inte kan prediktera framtida antal biståndstagare, se avsnitt 6.1 för förslag på alternativa metoder. Vi såg också att ålder visade linjära tendenser, vi testar därför att använda ålder som en numerisk variabel där vi sätter åldersvariabeln till medelvärdet av möjliga åldrar i respektive åldersgrupp.

Tabell 5: AIC för Poissonregressionsmodeller med årtal och ålder som linjära respektive kategoriska förklaringsvariabler.

Årtal	Ålder	AIC landsbygd	AIC storstad
Linjär	Kategorisk	158925	488894
Kategorisk	Kategorisk	**152191**	**406880**
Kategorisk	Linjär	164239	450847
Linjär	Linjär	171108	532753

Tabell 6: AIC för negativa binomialmodeller med årtal och ålder som linjära respektive kategoriska förklaringsvariabler.

Årtal	Ålder	AIC landsbygd	AIC storstad
Linjär	Kategorisk	142207	213559
Kategorisk	Kategorisk	**140151**	**210119**
Kategorisk	Linjär	144410	211446
Linjär	Linjär	146093	214747

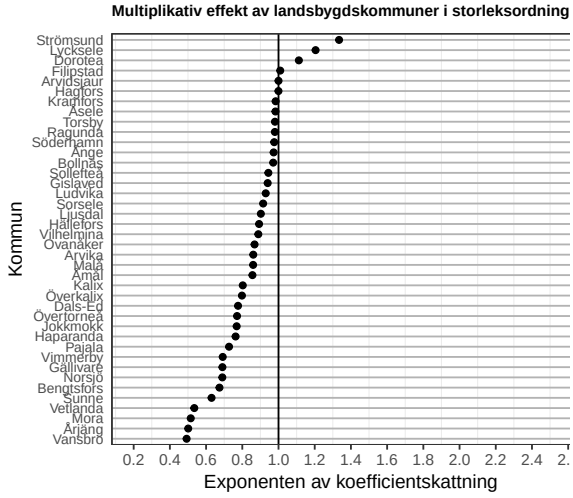
I Tabell 5 och 6 ovan ser vi att vi får lägst AIC i de modeller där vi använder både årtal och ålder som kategoriska variabler, oavsett fördelningsantagande. Detta gäller både för landsbygds- och storstadskommuner. Observerar även att modeller med negativ binomialregression har betydligt lägre AIC-värde än motsvarande modeller med Poissonregression.

5.3 Koefficientskattningar

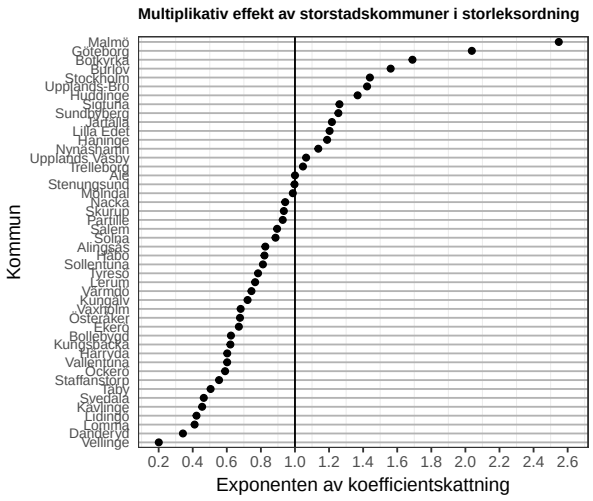
Vår bäst lämpade modell för både landsbygds- och storstadskommuner är alltså negativ binomialmodell med ålder och årtal som kategoriska variabler. I Figur 9-12, appendix 7.2.1, har vi observerat effekten ålder och årtal har på det logaritmerade antalet biståndstagare och r är även intresserade av vad kön och kommun bidrar med för effekt.

Tabell 7: Koefficientskattning kön med tillhörande standardavvikelser och p-värden för landsbygd respektive storstad.

	Kön Man	Standardavvikelse	P-värde
Landsbygd	-0.03338	0.00554	$1.65e-09$
Storstad	-0.01790	0.00501	$3.52e-4$



Figur 7:
Multiplikativ effekt av landsbygdskommuner



Figur 8:
Multiplikativ effekt av storstadskommuner

I Tabell 7 ser vi att det logaritmerade antalet biståndstagare minskar något för män både i landsbygds- och storstadskommuner, med en större negativ effekt i landsbygdskommuner. I Figur 7 och 8 ser vi de skattade multiplikativa effekterna på antalet biståndstagare av de olika kommunerna. Vi observerar en större spridning i storstadskommunerna där Malmö och Vellinge kommun är extrema åt vart sitt håll. Vi ser att endast ett fåtal av landsbygdkommunerna bidrar till ett ökat antal biståndstagare. Observerar här att i båda modeller förekommer ett fåtal osignifikanta parametrar, se Tabell 11 och 12 i appendix 7.2.2 för specifikation av dessa.

6 Diskussion

6.1 Hantering av okända observationer

För att hantera den ospecifierade datan över tre eller färre biståndstagare har vi använt oss av EM-algoritmen där vi ersätter denna okända data med motsvarande förväntade värden. Dessa förväntade värden är givna med avseende på observerade förklaringsvariabler och anpassas alltså enskilt för vardera observation. Ett alternativ till detta är att använda oss av medelvärdet av möjliga utfall av de okända observationerna. Vi vet att antalet biståndstagare inte specifieras då de är tre eller färre vilket betyder att vi med denna metod skulle sätta samtliga okända obsevationer till $X = 2$.

Tabell 8: AIC med EM-algoritmen respektive med $X = 2$.

Slutgiltlig modell	AIC: EM	AIC: $X = 2$
Landsbygd	**140151**	141373
Storstad	**210119**	210859

I Tabell 8 ser vi att vi får lägre AIC-värden då vi använder EM-algoritmen jämfört med då vi använder medelvärdet av möjliga utfall. EM-algoritmen ger alltså bättre lämpade modeller vilket är vad man kan förvänta sig då de okända värdena skattas med större noggrannhet. Då vi undersöker modellerna närmre ser vi dock att koefficientskattningarna skiljer sig åt som mest på tusendels- respektive hundradelsnivå i de anpassade landsbygds- respektive storstadsmodellerna. Beroende på hur stor noggrannhet man kräver kan man därför argumentera för att använda medelvärdet av möjliga utfall för de okända observationerna för att på det sättet spara arbetskraft.

6.2 Alternativa modeller

I Figur 13 och 14 i appendix 7.2.2 ser vi de summerade årsresidualerna för samtliga åldersgrupper i en landsbygdskommun, Mora, respektive en storstadskommun, Göteborg. Residualerna mellan de två kommunerna skiljer sig tydligt åt, dock ser vi en autokorrelation hos båda vilket tyder på att teori för tidsserie bör tillämpas för vardera kommun. Vi kan anta att detsamma gäller för resterande kommuner. Detta betyder att p-värden i ovanstående analys inte är pålitliga. En tidsserieanalys skulle möjliggöra prediktion av framtida antal biståndstagare.

Ett annat alternativ är att använda polynomregression då vi saknar linjäritet hos en eller flera förklaringsvariabler. Som alternativ till detta kan man då använda splineregression vilken går ut på att skapa flera polynom för den icke-linjära förklaringsvariabeln för olika intervall av denna (J.S.Racin, 2012). På det sättet kan trender i årtalen upptäckas och en prediktion kan utföras.

6.3 Tolkning

I Figur 11 och 12, appendix 7.2.1, ser vi hur det logaritmerade antalet biståndstagare förändras under åren 1990 till 2015 i landsbygds- respektive storstadskommuner. I storstadskommunerna ser vi att antalet stadigt minskar från slutet av 90-talet. I ett masterarbete vid Göteborgs universitet skriver Åsa Lingonblad om ett ökat utbetalat belopp av ekonomiskt bistånd och säger angående Figur 17 i bilagor *"Som vi kan se i ovanstående diagram har utvecklingen av utbetalt ekonomiskt bistånd per hushåll nästintill dubblerats under en period på drygt ett decennium."*. Utbetalat belopp har alltså ökat kraftigt medan vi observerar en minskning av antalet biståndstagare. I landsbygdskommunerna ser vi inte en lika stadig minskning, men inte heller en ökning motsvarande ökningen av utbetalat belopp observerad i Figur 17. Tidigare studier visar att ven antalet biståndstagare i Sverige har ökat markant sedan 1992 (Eardley, 1996) vilket skiljer sig från vad vi sett i detta arbete. Att resultaten skiljer sig t kan bero p att vi inte undersker samtliga kommuner i Sverige utan endast r intresserade av landsbygds- och storstadskommuner.

Problematik med att modellera ekonomiskt bistånd nationellt ligger till stor del i lokala faktorer (Saraceno, 2002). Att försöka hitta en samband inom landsbygds- respektive storstadskommuner kan vara svårt då antalet biståndstagare kan variera mycket mellan kommunerna inom dessa uppdelningar. Samtidigt som ett möjligt upptäckt samband kan bero på underliggande faktorer inom kommunerna. Detta gör det svårt att tolka våra modellskattningar.

7 Appendix

7.1 Data

Tabell 9: Landsbygdskommuner.

Gislaved	Vetlanda	Vimmerby	Dals Ed	Bengtfors
Åmål	Torsby	Årjäng	Sunne	Filipstad
Hagfors	Arvika	Hällefors	Vansbro	Mora
Ludvika	Ovanåker	Ljusdal	Söderhamn	Bollnäs
ånge	Kramfors	Sollefteå	Ragunda	Strömsund
Norsjö	Malå	Sorsele	Dorotea	Vilhelmina
Åsele	Lycksele	Arvidsjaur	Jokkmokk	Överkalix
Kalix	Övertorneå	Pajala	Gällivare	Haparanda

Tabell 10: Storstads- och storstadsnära kommuner.

Stockholm	Malmö	Göteborg	Upplands Väsby	Vallentuna
Österåker	Värmdö	Järfälla	Ekerö	Huddinge
Botkyrka	Salem	Haninge	Tyresö	Upplands Bro
Täby	Danderyd	Sollentuna	Nacka	Sundbyberg
Solna	Lidingö	Vaxholm	Sigtuna	Nynäshamn
Håbo	Staffanstorp	Burlöv	Vellinge	Kävlinge
Lomma	Svedala	Skurup	Trelleborg	Kungsbacka
Härryda	Partille	Öckerö	Stenungsund	Ale
Lerum	Bollebygd	Lilla Edet	Mölndal	Kungälv
Alingsås				

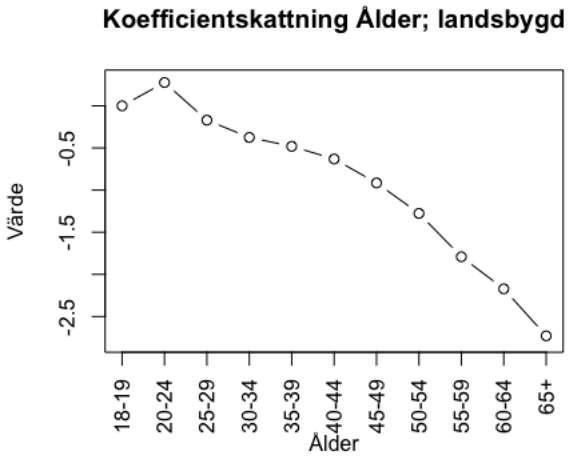
7.2 Modellbyggnad

Definition maximum likelihood-metoden [6]:
Maximum likelihood-skattningen $\hat{\theta}_{ML}$ av en parameter θ ges av att maximera likelihood-funktionen

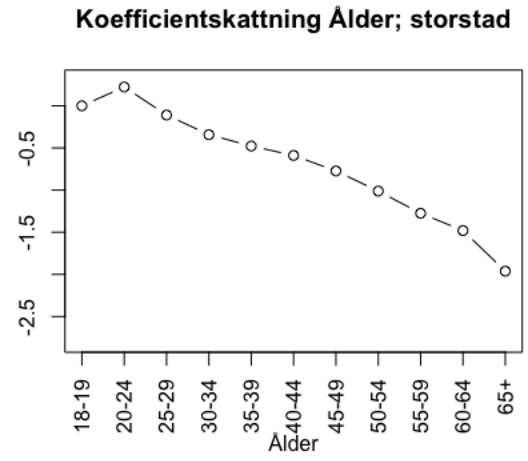
$$\hat{\theta}_{ML} = argmax_{\theta \in \Theta} L(\theta),$$

där $L(\theta)$ är likelihoodfunktion av parameter $\theta \in \Theta$.

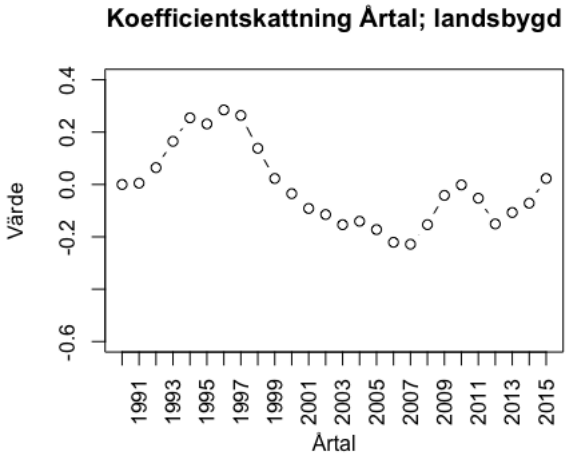
7.2.1 Koefficientskattningar



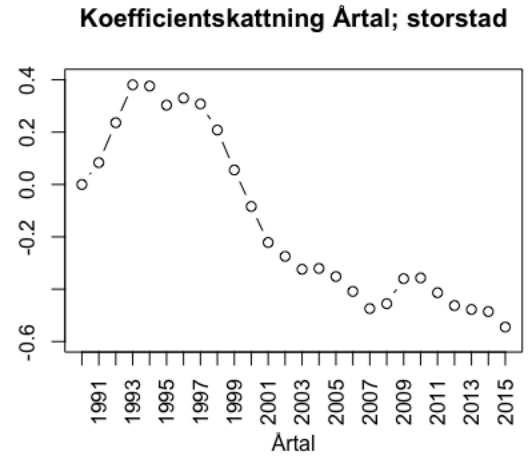
Figur 9:
Negativ binomialregression med ålder som kategorisk variabel



Figur 10:
Negativ binomialregression med ålder som kategorisk variabel



Figur 11:
Negativ binomialregression med årtal som kategorisk variabel



Figur 12:
Negativ binomialregression med årtal som kategorisk variabel

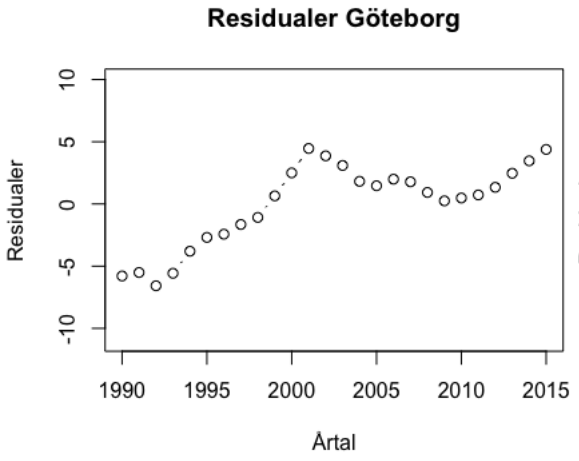
7.2.2 P-värden och residualer

Tabell 11: Osignifikanta kommuner i slutgiltlig landsbygds- respektive storstadsmodell.

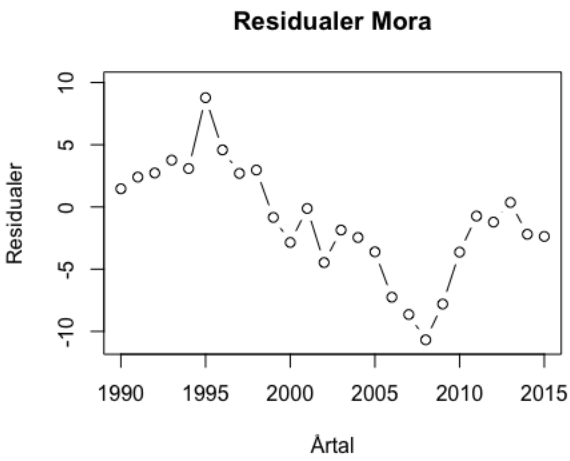
Landsbygd	P-värde	Storstad	P-värde
Bollnäs	0.2028	Mölndal	0.5753
Filipstad	0.7002	Stenungsund	0.9105
Hagfors	0.9922		
Kramfors	0.5015		
Ragunda	0.4459		
Söderhamn	0.3031		
Torsby	0.4197		
Ånge	0.2729		
Åsele	0.5519		

Tabell 12: Osignifikanta årtal i slutgiltlig landsbygdsmodell.

Årtal	P-värde
2015	0.2532
2010	0.9344
2000	0.0780
1999	0.2508
1991	0.8006



Figur 13:
Årsresidualer Göteborg



Figur 14:
Årsresidualer Mora

8 Referenser

Svenska Kommuner och Landsting.
https://skl.se/tjanster/kommunerlandsting/faktakommunerochlandsting/kommungruppsindelning.2051.html [2017-03-27]

Socialstryrelsen, Ekonomiskt Bistånd
http://www.socialstyrelsen.se/statistik/statistikdatabas/ekonomisktbistand [2017-03-24]

Statistiska Centralbyrån, Folkmängd
http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/Befolkning/Ny/rxid=f45f90b6-7345-4877-ba25-9b43e6c6e299 [2017-03-24]

Alan Agresti, 2002 *Categorical Data Analysis 2nd edition*, John Wiley & Sons

Leonard Held & Daniel Sabanés Bové, 2014 *Applied Statistical Inference*, Springer

The R Project for Statistical Computing, 2017
https://www.r-project.org/

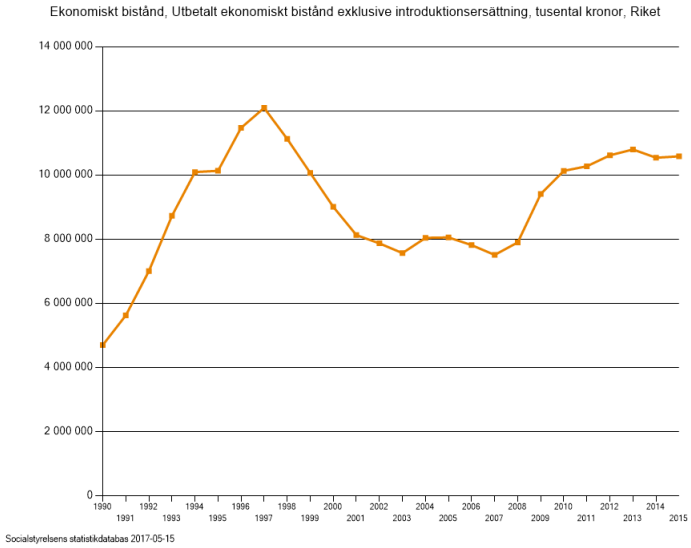
Åsa Lingonblad, 2014 *Ekonomiskt bistånd - en kvantitativ studie om bidragande bakgrundsfaktorer*, Göteborgs Universitet

Chiara Saraceno, 2002 *Social Assistance Dynamics in Europe*, The Policy Press

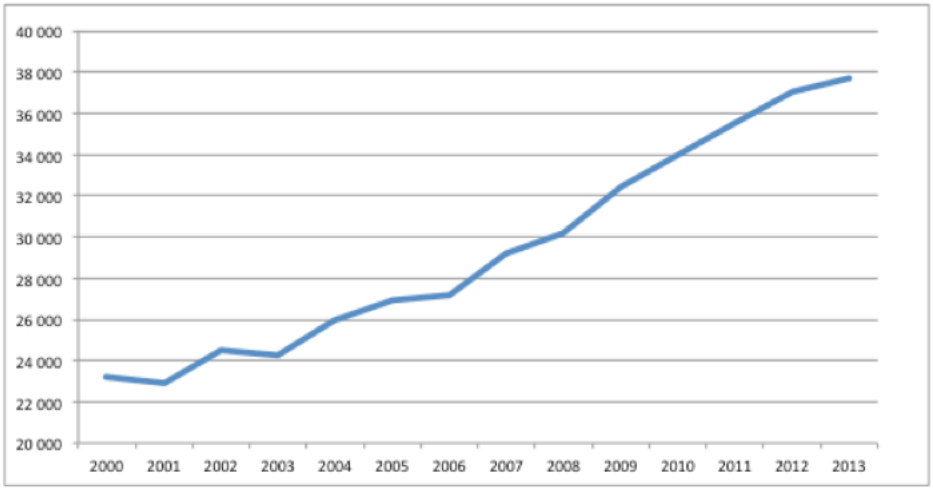
Tony Eardley, Jonathan Bradshaw, John Ditch, Ian Gough and Peter Whiteford, 1996 *Social Assistance in OECD Countries*, Journal of European Social Policy

Jeffrey S. Racine, 2012 *A primer on regression splines*, The R Project

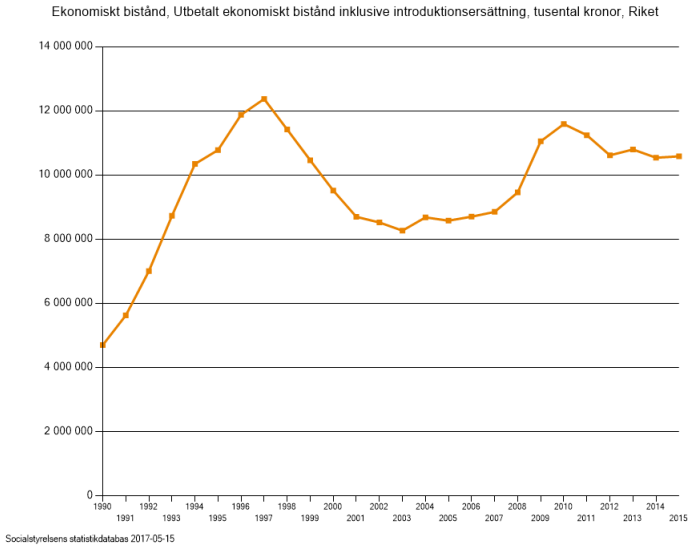
9 Bilagor



Figur 15: Utbetalat ekonomiskt bistånd exklusive introduktionsersättning (Socialstyrelsen, 2017)



Figur 17: Sveriges kommuners medelvärde för utbetalt ekonomiskt bistånd kronor per hushåll över åren 2000 - 2013 (Lingonblad, 2014)



Figur 16: Utbetalat ekonomiskt bistånd inklusive introduktionsersättning (Socialstyrelsen, 2017)