Stockholms universitet

Varför vissa kommuner bygger fler bostäder än andra

Molly Lindenau Stokke

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2016:12
Matematisk statistik
Juni 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2016:12**
http://www.math.su.se/matstat

Varför vissa kommuner bygger fler bostäder än andra

Molly Lindenau Stokke*

Juni 2016

Sammanfattning

Det råder bostadsbrist i stora delar av Sverige. I detta arbete undersöker vi därför vilka faktorer som avgör hur många lägenheter som byggs i en kommun. Datamaterialet omfattar alla Sveriges kommuner. Eftersom det i snitt tar åtta år från beslut till färdig bostad är data för förklaringsvariablerna från början av 2000-talet medan responsvariabeln beskriver antalet byggda bostäder åtta år senare. I analysen används till en början multipel linjär regression, Poissonregression och negativ binomialregression. På grund av att data är överspridd visar det sig dock att Poissonregression inte kan användas. De andra två metoderna får liknande resultat och visar båda att stor folkökning i en kommun bidrar mest till bostadsbyggandet, men att även stor andel improduktiv skogsmark och högerstyrd kommunstyrelse har positiv inverkan. Hög skatt och hög medelålder hos befolkningen har däremot negativ effekt. För att undersöka befolkningstäthetens påverkan behöver vi anpassa en B-spline och det visar sig att befolkningstäthetens effekt varierar mycket. Resultatet av undersökningen är dock osäkert, främst på grund av att tiden från beslut till färdigbyggd bostad varierar mycket.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: molly.stokke@gmail.com. Handledare: Martin Sköld och Mikael Petersson.

Sammanfattning

Det råder bostadsbrist i stora delar av Sverige. I detta arbete undersöker vi därför vilka faktorer som avgör hur många lägenheter som byggs i en kommun. Datamaterialet omfattar alla Sveriges kommuner. Eftersom det i snitt tar åtta år från beslut till färdig bostad är data för förklaringsvariablerna från början av 2000-talet medan responsvariabeln beskriver antalet byggda bostäder åtta år senare. I analysen används till en början multipel linjär regression, Poissonregression och negativ binomialregression. På grund av att data är överspridd visar det sig dock att Poissonregression inte kan användas. De andra två metoderna får liknande resultat och visar båda att stor folkökning i en kommun bidrar mest till bostadsbyggandet, men att även stor andel improduktiv skogsmark och högerstyrd kommunstyrelse har positiv inverkan. Hög skatt och hög medelålder hos befolkningen har däremot negativ effekt. För att undersöka befolkningstäthetens påverkan behöver vi anpassa en B-spline och det visar sig att befolkningstäthetens effekt varierar mycket. Resultatet av undersökningen är dock osäkert, främst på grund av att tiden från beslut till färdigbyggd bostad varierar mycket.

Tack

Jag vill tacka mina handledare Martin Sköld och Mikael Petersson för att de kontinuerligt har svarat på frågor och gett feedback. Jag vill även tacka Håkan Stokke och Linda Lindenau för att de har läst igenom arbetet och kommit med värdefulla synpunkter. Slutligen vill jag tacka David Schofield för hans stöd.

Innehåll

1 Inledning

Av Sveriges 290 kommuner bedömer 240 att det råder underskott på bostäder (Boverket, 2016). Bostadsbristen är något som drabbar och engagerar många och det har genomförts flera undersökningar i ämnet. I Lind (2003) konstateras det bland annat att dålig markåtkomst och långdragna planeringsprocesser är orsaker till att det byggs mindre än vad det borde. Länsstyrelsen i Stockholms län (2012) tar även upp överklaganden och höga produktionskostnader som hinder.

Gemensamt för dessa undersökningar är dels att man har undersökt kommuners gemensamma förutsättningar och dels att man har tagit reda på vad som är anledningarna till att man inte bygger tillräckligt. Dessutom har man låtit kommuner och andra själva uppge vad de ser för hinder.

I denna undersökning gör vi lite annorlunda. Istället för att se till vad som hindrar alla kommuner från att bygga i den utsträckning som behövs undersöker vi vilka faktorer som gör att vissa bygger mer än andra. Vi har även undersökt mer underliggande faktorer. Om dålig markåtkomst är ett problem borde kommuner med mycket outnyttjad mark bygga mer. Att höga produktionskostnader är ett hinder antyder att kommuner med hög medelinkomst och skatt borde ha ett större byggande. Vi ställer oss också frågan om man faktiskt kan se något samband mellan folkökning och ökat byggande. Denna undersökning är också en renodlat statistisk undersökning. Med hjälp av regression tar vi reda på vad som ger stort respektive litet byggande. Regressionsmetoderna som har använts är linjär, Poisson- samt negativ binomialregression.

Vi kommer att i Avsnitt 2 börja med den för studien nödvändiga teorin. Därefter kommer i Avsnitt 3 en beskrivning av det använda datamaterialet med förklaringar till varför respektive variabel används. I Avsnitt 4 sker bearbetning av datamaterialet samt själva modelleringen. I Avsnitt 5 sker slutligen en diskussion om undersökningens resultat samt förslag på hur en eventuell framtida studie skulle kunna förbättras.

2 Teori

Regression används om man har en responsvariabel och en eller flera förklarande variabler och vill ta reda på om responsvariabeln beror på de förklarande variablerna. Målet är att, med ett så litet fel som möjligt, kunna beskriva responsvariabeln med hjälp av de förklarande variablerna. Regression används om man vill förstå vad som ligger bakom ett visst fenomen och/eller kunna prediktera framtida resultat.

2.1 Multipel linjär regression

Teorin om multipel linjär regression bygger på Andersson & Tyrcha (2014).

Den kanske vanligaste formen av regression är multipel linjär regression. Antag att vi har k förklaringsvariabler och med dessa vill förklara variationen i en responsvariabel $\boldsymbol{Y}$. Då kan Y_i beskrivas som

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i, \qquad i = 1, \ldots, n$$

där Y_i är element i i $\boldsymbol{Y}$ och x_{ji} är element i i förklaringsvariabel j. Vidare är β_j den koefficientskattning som hör till förklaringsvariabel j och där β_0 kallas intercept och är det värde som antas då alla förklarande variabler har värdet 0. Slutligen är ε_i störningstermer som är oberoende och $N(0, \sigma_\varepsilon^2)$-fördelade. I matrisform kan detta skrivas som

$$\boldsymbol{Y} = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

där $\boldsymbol{Y}$ är en kolonnvektor med alla Y_i, $\boldsymbol{\beta}$ är en kolonnvektor med alla β-skattningar (inklusive β_0), $\boldsymbol{x}$ är en $(k+1) \times n$-matris där första kolonnen består av enbart ettor (representerar interceptet) och x_{ji} i övrigt är element i av förklaringsvariabel j samt $\boldsymbol{\varepsilon}$ är en radvektor med alla störningstermer.

2.1.1 Minstakvadratskattning

Teorin om minstakvadratskattning bygger på Miller (u.å.).

Väntevärdet för $\boldsymbol{Y}$ är $E[\boldsymbol{Y}] = \boldsymbol{\mu} = \boldsymbol{x}\boldsymbol{\beta}$ men eftersom β-parametrarna är okända måste dessa skattas. Detta görs vanligtvis med hjälp av minstakvadratmetoden.

Antag att vi har ett stickprov av storlek n från någon fördelning Z. Stickprovsvariansen definieras då som

$$\widehat{\sigma}_z^2 = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \bar{z})^2.$$

Skillnaderna mellan det observerade värdet och det skattade värdet kallas residualer. Residualen för observation i är därför $\widehat{\varepsilon}_i = y_i - \widehat{\mu}_i$ där $\widehat{\mu}_i = \sum_{j=1}^{k} \beta_j x_{ji}$ är det skattade värdet för observation i. Vi kan alltså se våra residualer $(y_1 - \widehat{\mu}_1, \ldots, y_n - \widehat{\mu}_n)$ som ett stickprov av storlek n. Enligt antagande i linjär regression är residualernas väntevärde 0 och i en bra anpassning bör deras medelvärde vara försumbart nära 0. Residualernas skattade varians är då

$$\widehat{\sigma}_\varepsilon^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2.$$

För att hitta den bästa uppsättningen β-parametrar behöver vi alltså minimera residualvariansen. Eftersom $1/(n-1)$ är en konstant är detta detsamma som att minimera $\sum_{i=1}^{n}(y_i - \widehat{\mu}_i)^2$. Det går att visa att $\boldsymbol{\beta}$-skattningen som minimerar summan av avstånden är $(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}$.

Om väntevärdet av en skattning är detsamma som det sanna väntevärdet kallas skattningen väntevärdesriktig (Alm & Britton, 2008). Det går att visa att av alla väntevärdesriktiga skattningar av $\boldsymbol{\beta}$ är minstakvadratskattningen den med minst varians vilket medför att det är den mest säkra skattningen.

2.2 Andra typer av förklaringsvariabler

En förutsättning för regression är att att $\boldsymbol{Y}$ kan beskrivas av linjärkombinationer av förklaringsvariablerna. Ibland beskrivs $\boldsymbol{Y}$ dock bättre om någon eller några av förklaringsvariablerna $\boldsymbol{x}_j$ är i form av ett polynom. Linjär regression går dock fortfarande att använda eftersom vi då inför nya variabler $\boldsymbol{x}_j^2$, $\boldsymbol{x}_j^3$ och så vidare och därefter skattar β-parametrar för dessa.

Förklaringsvariablerna behöver inte heller vara numeriska utan kan även vara kategoriska. Då beskriver vi variabeln med en eller flera så kallade dummy-variabler. En dummy-variabel kan bara anta värdet 0 eller 1, där 0 anger att en viss situation inte råder medan 1 anger att situationen råder. Om vi till exempel vill beskriva hur mycket en person tjänar utifrån en förklaringsvariabel som beskriver om hen bor i villa, radhus eller lägenhet kan vi ha en dummy-variabel som antar värdet 1 om personen bor i villa (och 0 annars) och en dummy-variabel som antar värdet 1 om personen bor i radhus. Vi behöver ingen dummy-variabel för att beskriva om personen bor i lägenhet ty om personen varken bor i villa eller radhus (vilket innebär att både villa- och radhus-dummyn är 0) måste det innebära att hen bor i lägenhet. Kategorin "personen bor i lägenhet" kallas då referenskategori och de β-skattningar vi får beskriver då hur mycket personen tjänar i förhållande till de som bor i lägenhet. Vill vi istället undersöka skillnaden mot exempelvis de som bor i villa måste vi ändra referenskategori.

2.3 Spline-regression

Teorin om spline-regression bygger på Fahrmeir et al. (2013), Maindonald & Braun (2010) och Racine (2014).

Vi har konstaterat att om en förklaringsvariabel inte har en linjärt samband med responsen kan detta ofta lösas genom att införa termer av högre grad för variabeln. I vissa fall måste man dock anpassa ett polynom av väldigt hög grad för att det ska följa data, vilket inte är realistiskt.

När data inte ser ut att följa ett polynom av grad 1, 2 eller 3 är det bättre att anpassa en spline-kurva. Istället för ett polynom består en spline-kurva av flera segmentvis sammanlänkade polynom. Punkterna där polynomen

möts kallas knutar. En funktion för så kallade naturliga splines kan skrivas som

$$\widehat{\mu} = b_0 P_0(x)\mathbb{1}_{[t_0,t_1]}(x) + \ldots + b_N P_N(x)\mathbb{1}_{(t_N,t_{N+1}]}(x)$$

där b_i är koefficienterna, $P_i(x)$ är polynomet som beskriver kurvans utseende i intervallet $(t_i, t_{i+1}]$ och $\mathbb{1}_{(t_i,t_{i+1}]}(x)$ är en indikatorfunktion som antar värdet 1 om x är i intervallet $(t_i, t_{i+1}]$ och 0 annars. Ett problem med naturliga splines är att om anpassningen för ett polynom ändras måste anpassningen för alla polynom ändras. För att undvika detta kan man använda så kallade B-splines. B står för "bas" och en B-spline är en linjärkombination av $N+n+1$ stycken basfunktioner. Vidare är den $n-1$ gånger kontinuerligt deriverbar, även i knutarna. En B-spline definieras av sin ordning $n+1$ och dess knutsekvens $\boldsymbol{t}$. I $\boldsymbol{t}$ är knutarna ordnade i stigande ordning. En B-spline definieras vidare som

$$\widehat{\mu} = b_0 B_{0,n}(x) + \ldots + b_{N+n} B_{N+n,n}(x).$$

Här är n spline-kurvans grad och B_j är basfunktionerna. Dessa konstrueras rekursivt enligt

$$B_{i,0}(x) = \mathbb{1}_{(t_i,t_{i+1}]}(x),$$

$$B_{i,p}(x) = \frac{t_{i+p} - t_i}{x - t_i} B_{i,p-1}(x) + \frac{t_{i+p+1} - x}{t_{i+p+1} - t_{i+1}} B_{i+1,p-1}(x),$$

där $p = 0, \ldots, n$.

2.4 Generaliserade linjära modeller

Teorin om generaliserade linjära modeller bygger på Agresti (2013) och Cameron & Trivedi (2013).

Det finns fall då linjär regression inte är tillämpbart. Linjär regression förutsätter till exempel att residualerna är normalfördelade, vilket inte alltid är fallet. Då använder man ofta generaliserade linjära modeller, GLM.

Man brukar tala om att en GLM består av tre komponenter: en slumpmässig, en systematisk och en länkfunktion.

Den slumpmässiga komponenten är responsvariabeln $\boldsymbol{Y}$. För att vi ska kunna anpassa en GLM måste $\boldsymbol{Y}$ komma från en fördelning som tillhör den naturliga exponentialfamiljen. En fördelning tillhör den naturliga exponentialfamiljen om dess täthetsfunktion kan skrivas på formen

$$f(y_i|\theta_i) = a(\theta_i)b(y_i)\exp(y_i Q(\theta_i)). \tag{1}$$

Här kallas $Q(\theta_i)$ för den naturliga parametern.

Den systematiska komponenten $\boldsymbol{\eta}$ beskriver det linjära sambandet genom

$$\boldsymbol{\eta} = \boldsymbol{x}\boldsymbol{\beta}$$

Detta ser ut som uttrycket för $\boldsymbol{\mu}$ i den linjära regressionen. I en GLM är $\boldsymbol{\eta}$ dock inte nödvändigtvis lika med väntevärdet. Sambandet mellan $\boldsymbol{\mu}$ och $\boldsymbol{\eta}$ beskrivs istället av de sista komponenten, länkfunktionen. Länkfunktionen g är den funktion av $\boldsymbol{\mu}$ som uppfyller $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$. Den kanske vanligaste länkfunktionen kallas för den kanoniska länken och är Q i (1). Länkfunktionen är monoton och kontinuerligt differentierbar. Länkfunktionen omvandlar alltså väntevärdet för det slumpmässiga komponenten till en linjär funktion av förklaringsvariablerna.

2.4.1 Poissonregression

När man behandlar räknedata antar man ofta att $\boldsymbol{Y}$ är Poissonfördelad. Poissonfördelningen tillhör den naturliga exponentialfamiljen eftersom dess täthetsfunktion kan skrivas på formen i (1) med $a(\theta_i) = \exp(-\mu_i)$, $b(y_i) = 1/y!$ och $Q(\theta_i) = \log(\mu_i)$. Den kanoniska länken är alltså log.

Poissonfördelningen har samma väntevärde som varians. I verkligheten är variansen dock ofta större än väntevärdet, vilket kallas överspridning. En anledning till detta kan vara att man behöver fler förklaringsvariabler. Om man utelämnar en variabel som kan förklara en del av variationen i $\boldsymbol{Y}$ blir variansen större eftersom de variabler som används inte kan förklara den variationen. I många fall kan man dock inte lägga till fler förklaringsvariabler.

Det finns flera sätt att ta reda på om man har överspridning. Ett är att plotta residualer mot anpassade värden (se Avsnitt 2.5.6). Det finns även flera test. Det vi ska använda oss av utnyttjar det faktum att variansen kan skrivas som $Var(Y) = \mu + \mu^2/k$ för något k. Om $1/k = 0$ är även $\mu^2/k = 0$ vilket motsvarar att variansen är lika med väntevärdet. Om $1/k$ är större än 0 har vi överspridning. Vi kan alltså testa om vi har överspridning genom att testa hypotesen $H_0 : 1/k = 0$ mot $H_1 : 1/k > 0$. Detta kan i sin tur göras med vanlig linjär regression där $1/k$ är parametern som ska skattas. Får vi ett signifikant p-värde för $1/k$ kan vi förkasta nollhypotesen.

2.4.2 Negativ binomialregression

Ett vanligt sätt att hantera överspridning är att övergå till en blandad modell. I Poissonregressionen antog vi att, för varje kombination av värden på förklaringsvariablerna, y_i kom från en Poissonfördelning med väntevärde $E[Y_i] = \lambda_i$. Om istället λ_i också är en slumpvariabel tillåts större varians. Vanligt är att man låter λ_i vara gammafördelad med parametrar μ_i och k. Den marginella fördelningen för Y_i blir då negativ binomial. Negativa binomialfördelningen har $E[Y_i] = \mu_i$ och $Var(Y_i) = \mu_i + \mu_i^2/k$. Termen $1/k$ kallas här spridningsparameter.

Spridningsparametern är okänd och måste skattas, exempelvis i överspridningstestet som vi introducerade i Avsnitt 2.4.1. Hade den varit fix hade vi kunnat skriva sannolikhetsfunktionen för negativa binomialfördelningen

på exponentialfamiljsform. Det kan vi inte nu. Däremot kan vi skriva den på en utökad form för exponentialfamiljen och den tillhör därmed den exponentiella spridningsfamiljen. I exponentiella spridningsfamiljen tillåts en spridningsparameter ϕ och en fördelning tillhör den exponentiella spridningsfamiljen om dess täthetsfunktion kan skrivas på formen

$$f(y_i|\theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right).$$

Här är θ_i den naturliga parametern. Negativa binomialfördelningens täthetsfunktion är

$$\begin{aligned} p(y_i|k, \mu_i) &= \frac{\Gamma(y_i + k)}{\Gamma(k)\Gamma(y_i + 1)}\left(\frac{k}{\mu_i + k}\right)^k\left(1 - \frac{k}{\mu_i + k}\right)^{y_i} \\ &= \exp\left(y_i \log\left(\frac{\mu_i}{\mu_i + k}\right) + k\log\left(1 + \frac{\mu_i}{\mu_i + k}\right) + \log\left(\frac{\Gamma(y_i + k)}{\Gamma(k)\Gamma(y_i + 1)}\right)\right) \end{aligned}$$

Här är $\theta_i = \log(\mu_i/(\mu_i + k))$.

2.4.3 Offset-variabel

När vi modellerar räknedata är det inte ovanligt att vi på förhand har information som gör att de olika observationerna av responsvariabeln har olika förutsättningar för att få ett högt eller lågt värde. I vårt fall är det till exempel väldigt stor skillnad på om Bjurholm, med ett invånarantal på 2618, eller Stockholm, med över 760 000 invånare, bygger 500 bostäder. Variabeln som avgör de olika förutsättningarna kallas offset-variabel och vi är mer intresserade av att modellera kvoten mellan respons- och offset-variabeln än enbart responsvariabeln. Detta leder dock till att responsen antar icke heltalsvärda värden vilket inte är möjligt då vi genomför exempelvis Poisson- eller negativ binomialregression. Om vi använder log-länk kan vi dock utnyttja sambandet

$$\begin{aligned} \log\left(\frac{\mathit{respons}_i}{\mathit{offset}_i}\right) &= \sum_{j=0}^{k}\beta_j x_{ji} \\ \Rightarrow \log(\mathit{respons}_i) - \log(\mathit{offset}_i) &= \sum_{j=0}^{k}\beta_j x_{ji} \\ \Rightarrow \log(\mathit{respons}_i) &= \log(\mathit{offset}_i) + \sum_{j=0}^{k}\beta_j x_{ji}. \end{aligned}$$

Logaritmen av offset-variabeln används alltså som en förklaringsvariabel med koefficienten satt till 1.

2.4.4 Maximum likelihood

När vi jobbar med GLM kan vi inte anpassa skattningar med minstakvadratmetoden. Istället använder vi maximum likelihoodmetoden.

Om vi har en stokastisk variabel $\boldsymbol{Y}$ med observerade värden $y_1, \ldots, y_n$ beskrivs dess likelihoodfunktion av

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \begin{cases} \prod_{i=1}^{n} p(y_i|\boldsymbol{\beta}), & \text{då } \boldsymbol{Y} \text{ är diskret,} \\ \prod_{i=1}^{n} f(y_i|\boldsymbol{\beta}), & \text{då } \boldsymbol{Y} \text{ är kontinuerlig.} \end{cases}$$

Denna funktion beskriver hur sannolikt det är att få de observerade värden vi har fått. Funktionen beror på β-parametrarna och bildar ett plan i $n+1$ dimensioner. Detta plan har en stationär punkt, som man kan visa är en maximipunkt, som alltså är den uppsättning β-variabler som är mest sannolikt med avseende på vårt utfall av $\boldsymbol{Y}$. För att hitta dessa β-skattningar deriverar vi alltså $L(\boldsymbol{\beta}|\boldsymbol{y})$ med avseende på respektive β_i och hittar dess nollställen.

2.5 Modellval

2.5.1 Multikolinjäritet

Teorin om multikolinjäritet bygger på Sundberg (2014) där inget annat anges.

När vi genomför en minstakvadratskattning ingår det ju att invertera $\boldsymbol{x}^T\boldsymbol{x}$. Detta kräver att $\boldsymbol{x}^T\boldsymbol{x}$ inte är singulär, vilket inträffar om en rad är en linjärkombination av en eller flera andra rader. Detta inträffar i sin tur om en förklaringsvariabel kan beskrivas av en eller flera andra förklaringsvariabler. Lösningen i detta fall blir att ta bort den variabeln som kan förklaras av andra variabler eftersom den är överflödig.

Detta händer dock aldrig i praktiken. Vad som däremot ofta händer är att en förklaringsvariabel till stor del kan beskrivas med hjälp av en eller flera av de andra förklaringsvariablerna. Detta kallas multikolinjäritet och gör tolkningen svår, eftersom om $\boldsymbol{x}_1$ och $\boldsymbol{x}_2$ är delvis beroende och vi upptäcker att $\boldsymbol{x}_1$ har en påverkan på $\boldsymbol{y}$, så vet vi inte det egentligen är $\boldsymbol{x}_2$ som påverkar. Ett exempel på en sådan situation är om man vill undersöka vad som påverkar försäljningspriset på en lägenhet. Möjliga förklaringsvariabler är lägenhetens storlek samt dess antal rum. Dessa är dock inte oberoende eftersom en större lägenhet tenderar att ha fler rum. För att lösa detta får man antingen eliminera en av variablerna eller göra någon typ av transformation. I detta fall skulle man till exempel kunna ändra variabeln som beskriver antal rum till att beskriva genomsnittlig storlek per rum. Det ska dock påpekas att om syftet med undersökningen enbart är att prediktera gör det inget om vi har multikolinjäritet. Om vi däremot vill kunna förstå vad som orsakar en förändring är multikolinjäritet ett problem.

Det finns flera sätt att upptäcka multikolinjäritet. Om en förklaringsvariabel delvis beskrivs av endast en annan förklaringsvariabel syns det på deras korrelationskoefficient. Korrelationskoefficienten definieras som

$$r_{12} = \frac{c_{12}}{\widehat{\sigma}_1 \widehat{\sigma}_2}$$

där $\widehat{\sigma}_j$ är den skattade standardavvikelsen för variabel x_j och c_{12} är kovariansen mellan x_1 och x_2, som definieras som

$$c_{12} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2).$$

Korrelationskoefficienten uppfyller $-1 \leq r \leq 1$ och mäter hur stort det linjära beroendet mellan $\boldsymbol{x}_1$ och $\boldsymbol{x}_2$ är. Om den ena variabeln helt och hållet kan beskrivas av den andra blir korrelationskoefficienten -1 eller 1 medan oberoende variabler ger en korrelationskoefficient på 0 (Alm & Britton, 2008).

Om det är tre eller fler förklaringsvariabler som tillsammans är beroende kan vi inte titta på korrelationskoefficienten. Då brukar man istället beräkna variationsinflationsfaktorn, VIF. Varje förklaringsvariabel har en egen VIF och den beror på vilka andra förklaringsvariabler vi använder oss av varför vi först måste välja vilka förklaringsvariabler som eventuellt ska ingå i modellen. VIF-faktorn för förklaringsvariabeln $\boldsymbol{x}_j$ definieras som

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

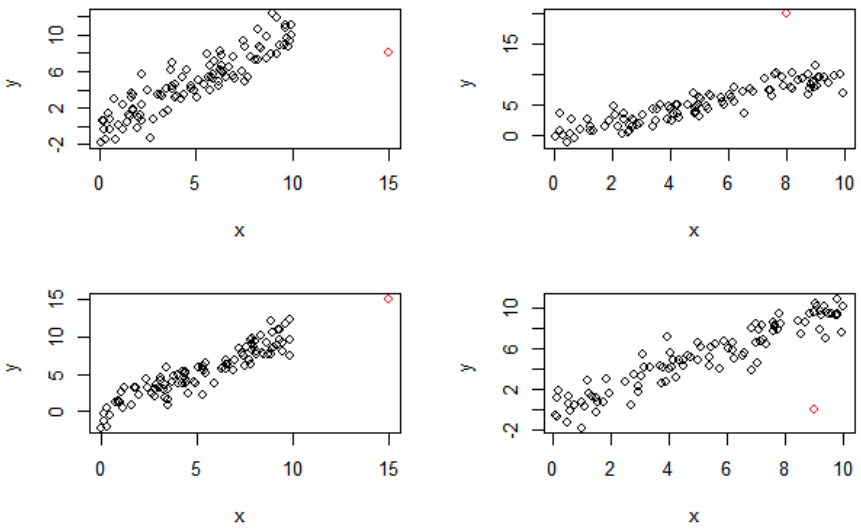
där R_j^2 uppfyller $0 \leq R_j^2 \leq 1$ och anger hur mycket av variationen i $\boldsymbol{x}_j$ som kan förklaras av de övriga variablerna. Hur R_j^2 beräknas förklaras i Avsnitt 2.5.3. Vad man anser är ett för högt VIF-värde varierar, men ofta sätts gränsen vid 5 eller 10 vilket motsvarar ett R_j^2 på 0.8 respektive 0.9.

2.5.2 Outliers

Teorin om outliers bygger på Sundberg (2014).

Förhoppningen när man genomför linjär regression är att responsvariabeln ska kunna beskrivas som en linjär funktion av förklaringsvariablerna. Ibland tyder de flesta observationer på ett linjärt samband mellan responsen och förklaringsvariablerna medan en eller några observationer ligger långt från de andra. Sådana observationer kallas outliers och kan påverka den skattade regressionslinjen ganska mycket. Det finns fyra sorters outliers, som visas i Figur 1.

I plottarna är outlierna markerade i rött. I den första plotten är outliern extrem i sitt x-värde men inte sitt y-värde, i andra plotten är outliern extrem



Figur 1: Olika typer av outliers − observationer som avviker från mönstret.

i sitt y-värde men inte sitt x-värde, i tredje plotten är outliern extrem i både sitt x- och sitt y-värde och i fjärde plotten är outliern varken extrem i sitt x- eller y-värde. I alla fall utom det i den tredje plotten påverkar outliern regressionsskattningen. Även då observationen är extrem i både x- och y-värde kan det dock påverka skattningen, om den exempelvis är mycket extrem i x-värde men bara lite extrem i y-värde.

Om man upptäcker en outlier bör man undersöka om den observationen till exempel har samlats in på ett annat sätt och därför skulle kunna ha ett avvikande värde. Ofta finns det dock ingen uppenbar förklaring till en outliers extrema värde och man får avgöra från fall till fall om man bör utesluta observationen eller inte.

2.5.3 Förklaringsgrad

Teorin om förklaringsgrad bygger på Andersson & Tyrcha (2014) samt Sundberg (2014).

När vi har en modell är det av intresse att få veta hur mycket av variationen i responsvariabeln som egentligen beskrivs av modellen. När vi använder linjär regression kan vi få reda på detta från förklaringsgraden, som betecknas R^2. I definitionen för R^2 används två kvadratsummor, RSS och TSS, som definieras som

$$RSS = \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2,$$

och

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

där RSS är summan av residualerna i kvadrat och TSS är kvadratsumman av variationen kring totalmedelvärdet. Med hjälp av dessa definierar vi nu R^2 som

$$R^2 = 1 - \frac{RSS}{TSS}.$$

Förklaringsgraden uppfyller $0 \leq R^2 \leq 1$ där 0 anger att inget av variationen i $\boldsymbol{y}$ förklaras av modellen och 1 anger att all variation i $\boldsymbol{y}$ kan förklaras av modellen.

Ett problem med R^2 är att det alltid ökar om vi tillför en extra förklarande variabel. Detta kan göra det svårt att veta om tillförandet av en extra förklaringsvariabel egentligen bidrar till modellen. Därför är det ofta bättre att istället undersöka den korrigerade förklaringsgraden, R^2_{adj}. Den korrigerade förklaringsgraden beskriver förändringen i residualernas (gemensamma) varians istället för deras sammanlagda avstånd till regressionslinjen. Den korrigerade förklaringsgraden definieras som

$$R^2_{adj} = 1 - \frac{\widehat{\sigma}^2}{\widehat{\sigma}^2_0}.$$

Här är $\widehat{\sigma}^2$ den skattade standardavvikelsen i modellen och definieras som $RSS/(n-k)$, där k är antalet förklarande variabler i modellen. Vidare är $\widehat{\sigma}^2_0$ den skattade standardavvikelsen i modellen utan några förklaringsvariabler alls och definieras som $TSS/(n-1)$. Med hjälp av dessa definitioner ser vi att vi kan skriva

$$R^2_{adj} = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - R^2 \cdot \frac{n-1}{n-k}$$

vilket gör det enklare att förstå varför R^2_{adj} inte nödvändigtvis ökar när antalet förklarande variabler ökar.

Varken den vanliga eller den korrigerade förklaringsgraden kan användas i generaliserade linjära modeller, utan endast i linjär regression.

2.5.4 Akaikes informationskriterium

Teorin om Akaike information criterion bygger på Agresti (2013).

Ett vanligt verktyg för modelljämförelse är Akaikes informationskriterium, AIC (från engelskans "Akaike Information Criterion"). AIC bedömer både hur bra variationen i $\boldsymbol{y}$ förklaras av förklaringsvariablerna men också hur många förklarande variabler som ingår i modellen. Bra beskrivning av $\boldsymbol{y}$ höjer AIC, men fler parametrar sänker det. Precis som R^2_{adj} hjälper alltså

AIC till att hitta den modellen som är bäst i förhållande till antalet förklarande variabler den innehåller. AIC är dock inte begränsat till linjär regression utan kan användas även i GLM. AIC definieras som

$$\text{AIC} = -2(\log(L_{max}(\boldsymbol{\beta}|\boldsymbol{y})) - k)$$

där $L_{max}(\boldsymbol{\beta}|\boldsymbol{y})$ är maximum av likelihood-funktionen och k är antalet parametrar. Ett lägre AIC antyder en bättre modell. AIC säger dock ingenting om hur bra en modell är överlag utan kan bara användas för att jämföra modeller med samma responsvariabel. Om alla modeller har dålig anpassning avslöjar alltså inte AIC något om det.

2.5.5 Stegvis variabelselektion

Teorin om stegvis variabelselektion bygger på Fahrmeir et al. (2013).

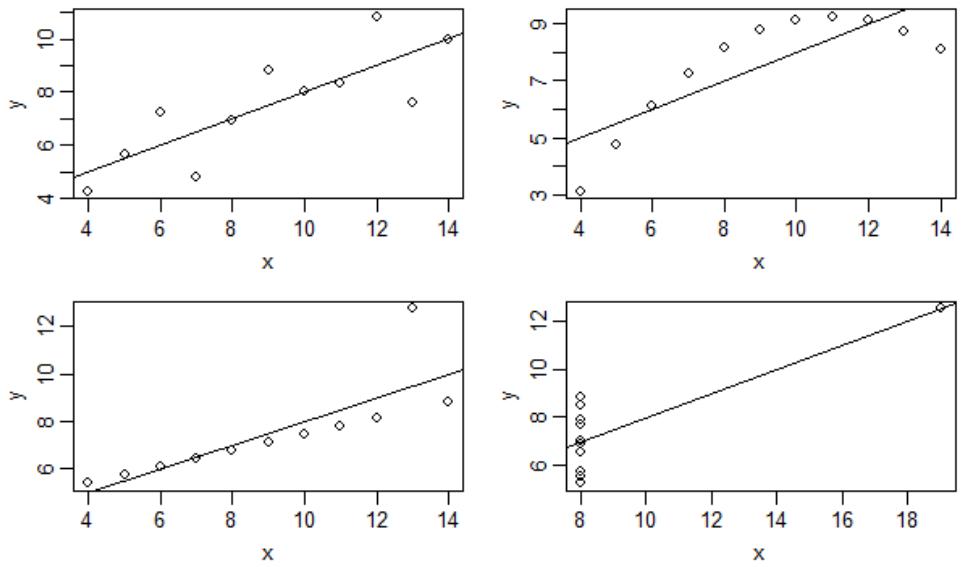
En av utmaningarna i regressionsanalys är att avgöra vilka förklarande variabler som ska ingå i modellen. En stor del av svårigheten ligger i att två eller fler variabler kan i sig förklara variationen i $\boldsymbol{y}$ dåligt, men tillsammans förklara den bra. Ett vanligt hjälpmedel till detta är stegvis variabelselektion. Stegvis variabelselektion går ut på att man testar att ta bort respektive lägga till en variabel i taget och undersöka vilken förändring som ger störst förbättring av modellen. När man inte längre kan få en bättre modell genom att lägga till eller ta bort en variabel slutar man. I denna undersökning kommer vi att använda AIC som mått på modellens förbättring/försämring.

När man genomför stegvis variabelselektion kan man starta med vilken modell som helst. Vi kommer i detta arbete att starta med dels den tomma modellen, utan några förklarande variabler, dels modellen som innehåller alla våra potentiella förklaringsvariabler.

2.5.6 Residualer och plottar

Teorin om Pearsonresidualer bygger på Cameron & Trivedi (2013).

Ett viktigt steg för att undersöka om man har kommit fram till en bra modell är att plotta anpassningen för att se om modellen verkligen beskriver data tillfredsställande. Ett klassiskt exempel på varför det är viktigt att plotta anpassningen och de observerade värdena är Anscombe's kvartett som syns i Figur 2. Dessa fyra dataset har alla samma medelvärde $\bar{x}$, samma medelvärde $\bar{y}$, samma stickprovsvarians $\widehat{\sigma}_x$, samma stickprovsvarians $\widehat{\sigma}_y$, samma korrelation mellan x och y samt samma skattade regressionslinje, men som synes skiljer sig anpassningen väldigt mycket. Det första fallet är ett exempel på en bra regressionsskattning. I den andra plotten ser vi att en andragradsterm skulle behöva läggas till. I tredje plotten ser vi en outlier och



Figur 2: Anscombe's kvartett, fyra datamaterial som har samma statistiska egenskaper men är väldigt olika när de plottas.

i sista plotten har för få unika x-värden observerats och en regressionsanalys är inte riktigt genomförbar.

En annan sak som vi kan upptäcka när vi plottar residualer mot anpassade värden är om vi har heteroskedasticitet. Heteroskedasticitet innebär att att residualerna inte har samma varians. En vanlig form av heteroskedasticitet är att variansen ökar med ökande väntevärde. Om vi arbetar med Poissonregression ska dock residualerna ha just ökande varians med ökande väntevärde, eftersom Poissonfördelningen har egenskapen att variansen är lika med väntevärdet. För att då undersöka om residualerna varierar mer än tillåtet används så kallade Pearsonresidualer. Pearsonresidualen definieras som

$$e_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\mu}_i}}$$

Anledningen till att vi dividerar med $\sqrt{\widehat{\mu}_i}$ är att vi vill dividera med standardavvikelsen för att på så sätt få bort Poissonregressionens naturliga heteroskedasticitet. Om Y_i är en Poissonfördelad slumpvariabel med $E[Y_i] = \mu_i$ är även $Var(Y_i) = \mu_i$. Eftersom μ_i är en konstant är då $Var(Y_i - \mu_i) = \mu_i$. Då μ_i är okänt och vi undersöker $y_i - \widehat{\mu}_i$ är det naturligt att skatta variansen till $\widehat{\mu}_i$.

Förutom att plotta residualer mot anpassade värden är det även bra att plotta dem mot varje förklaringsvariabel, både de som är med i modellen och de som valdes bort. Då kan man exempelvis upptäcka att en variabel bör vara inkluderad i en annan form, till exempel i form av ett polynom.

3 Beskrivning av data

Datamaterialet är inhämtat från två tidsperioder. Responsvariabeln är över åren 2007-2014 medan kovariaterna beskriver perioden 1999-2006. Detta är eftersom det tar i snitt åtta år från beslut till färdig bostad (Sundell, 2014; Cederfelt, 2013 och Bostads AB Poseidon, 2013). Det relevanta är nämligen vilka förhållanden som råder när det fattas beslut om bostad och inte när den faktiskt färdigställs.

Alla variabler, utom *höger*, *vänster* och *bostäder*, är i genomsnitt per år under perioden de är tagna från. Detta leder bland annat till att *folkmängd* inte bara kan anta heltal. Anledningen till att *bostäder* är det totala antalet byggda lägenheter under perioden är att den annars kan anta andra tal än heltal, och under Poisson- och negativa binomialregressionen kommer det att vara en förutsättning att den bara antar heltal.

I datamaterialet utgör en kommun en observation. Det finns 290 kommuner i Sverige. Knivsta kommun bildades dock först år 2003. Eftersom kommunen inte existerar under hela perioden vi undersöker är den inte med i datamaterialet. Det som blev Knivsta var tidigare en del av Uppsala kommun. Förutsättningarna för Uppsala kommun har alltså ändrats under perioden. Av den anledningen är inte heller Uppsala kommun med i datamaterialet. Detta gör att vi har 288 observationer.

Tillgängliga variabler är:

bostäder - antalet färdigställda lägenheter. I linjära regressionen används antal byggda bostäder per 1000 invånare som responsvariabel. Anledningen till att vi inte tittar på det totala antalet byggda bostäder är att en stor kommun med många invånare förmodligen bygger mer än en liten kommun med få invånare vilket skulle ge oss ett missvisande resultat. Denna variabel sträcker sig från 0 till 31856 med ett medelvärde på 692.85.

folkmängd - antalet invånare i kommunen. Denna används i responsen som divisor till *bostäder*. Denna variabel sträcker sig från 2618.1 till 760979.4 med ett medelvärde på 30373.36.

befolkningstäthet - antal invånare per kvadratkilometer. Här finns det två motsägande hypoteser. Den ena är att en kommun med hög befolkningstäthet inte har så mycket plats att bygga fler bostäder på. Den andra är att en kommun med hög befolkningstäthet antagligen är positivt inställd till många invånare och därför fortsätter bygga. Denna variabel sträcker sig från 0.275 till 4052.8 med ett medelvärde på 126.43.

folkökning - nettoinflyttning per 1000 invånare i kommunen. Tanken är att om befolkningen ökar så lär kommunen bygga fler bostäder. Anledningen till att vi använder enbart in- och utflyttningar och inte inkluderar födelser och dödsfall är att variabeln annars blir högt korrelerad med *ålder*, *inkomst* och *skatt*. Denna variabel sträcker sig från -12.2 till 16.0 med ett medelvärde på 0.905.

höger och *vänster* - antalet mandatperioder med högerstyrd respektive

vänsterstyrd kommunstyrelse. Detta gör att $2 - höger - vänster$ är antalet år med koalitionsstyre. Även om det är byggherrar som står för själva utformandet och byggandet av bostäder är det kommunpolitikerna som bestämmer om, var och hur mycket det ska byggas. Respektive av dessa variabler kan anta värdena 0, 1 och 2 och summan av dem kan vara högst 2.

inkomst - medianen av invånarnas förvärvsinkomst per år i tusen kronor. För denna variabel finns det också två hypoteser som motsäger varandra. Å ena sidan innebär en hög medianinkomst förmodligen att kommunen har bra ekonomi vilket i sin tur bör medföra att de bygger mycket. Å andra sidan bor höginkomsttagare i större utsträckning i villor. Eftersom nybyggnation tenderar att likna bebyggelsen som redan finns bör detta innebära att det byggs färre lägenheter i kommuner med hög medianinkomst. Som höginkomsttagare drabbas man inte heller lika mycket av bostadsbristen och är därför kanske inte lika positivt inställda till nybyggnation. Huruvida koefficienten framför *inkomst* är positiv eller negativ är alltså svårt att gissa. Denna variabel sträcker sig från 148.5 till 237.2 med ett medelvärde på 176.4.

kommungrupp - kommungruppsindelning enligt SKL. Sveriges kommuner och landsting har delat upp Sveriges kommuner i vid det aktuella tillfället 9 kategorier baserat på saker som folkmängd, pendling och försörjning. Kategorierna beskrivs i Tabell 1.

1	Storstäder
2	Förortskommuner
3	Större städer
4	Pendlingskommuner
5	Glesbygdskommuner
6	Varuproducerande kommuner
7	Övriga kommuner, mer än 25 000 inv.
8	Övriga kommuner, 12 500-25 000 inv.
9	Övriga kommuner, mindre än 12 500 inv.

Tabell 1: Kommungruppsindelning enligt Sveriges Kommuner och Landsting.

Som basvariabel väljer vi den grupp inom vilken det har byggts minst antal bostäder per invånare, vilket är grupp 5. Alla grupper har mellan 26 och 40 observationer utom grupp 1 som endast har 3 observationer.

skatt - kommunalskatt i procent. Även här är hypotesen tudelad. Å ena sidan kan man tänka sig att en kommun med hög skatt har höga skatteintäkter och därför har råd att bygga mycket. Å andra sidan brukar höginkomsttagare rösta för lägre skatt och vise versa. Det senare talar även för att *skatt* lär vara högt korrelerad med *inkomst*. Denna variabel sträcker sig från 28.56 till 33.57 med ett medelvärde på 31.58.

skog - procent av kommunens landyta som består av improduktiv skogsmark. Mark kan användas till mycket, inte bara bostäder. En kommun med låg befolkningstäthet skulle ändå kunna ha mycket mark som används, till exempelvis industri eller odling. Variabeln *skog* mäter alltså hur stor andel av kommunens yta som inte används till något. Hypotesen är att en kommun med hög andel improduktiv skogsmark bygger mer eftersom de har mer plats att bygga på. Enligt hypotesen skulle alltså koefficienten framför *skog* vara positiv. Denna variabel sträcker sig från 0.00 till 43.32 med ett medelvärde på 9.52.

valdeltagande - hur stor procent av befolkningen som röstat i kommunvalet. Högt valdeltagande tyder på stort politiskt engagemang. Vad det har för påverkan på bostadsbyggandet är dock svårt att säga. Det troligaste är nog ändå att *valdeltagande* har negativ påverkan på responsen. Tanken bakom detta är att de som redan bor i kommunen uppenbarligen redan har en bostad och kanske istället prioriterar stora grönområden och liknande medan de som är positivt inställda till nybyggnation i kommunen inte ännu bor där. Denna variabel sträcker sig från 59.10 till 87.55 med ett medelvärde på 78.63.

ålder - medelålder i år hos befolkningen i kommunen. Ett hinder mot bostadsbyggande är att det kommer in många överklaganden. En hypotes är att äldre människor har större benägenhet att överklaga byggprojekt då de har ett tryggt och stadigt liv, vill ha lugn och ro och märker av förändringar mer eftersom de ofta bor länge på samma ställe. Yngre personer bör enligt hypotesen däremot vara mer positiva till nybyggnation, främst för att de känner av bostadsbristen mer men också för att yngre människor tenderar att vara mer positivt inställda till stadsmiljö och förändringar. Enligt hypotesen skulle alltså variabeln *ålder* ha en negativ koefficient. Denna variabel sträcker sig från 36.2 till 46.8 med ett medelvärde på 41.8.

Allt datamaterial utom *höger*, *vänster* och *kommungrupp* kommer från Statistiska centralbyråns statistikdatabas. Data gällande politiskt styre och kommungruppsindelning har istället hämtats från Sveriges Kommuner och Landsting. Exakta URL-adresser till variablerna finns i Avsnitt 7.2.

Många variabler är korrelerade, men inte så mycket att det bör ställa till något problem. Den enda oroväckande korrelationen är den mellan *ålder* och *inkomst* som är -0.724. Vi vill dock inte ta bort någon av kovariaterna och det finns ingen given transformation. Vi behåller alltså båda dessa i den formen de är men är extra vaksamma på deras skattningar och signifikanser.

4 Analys

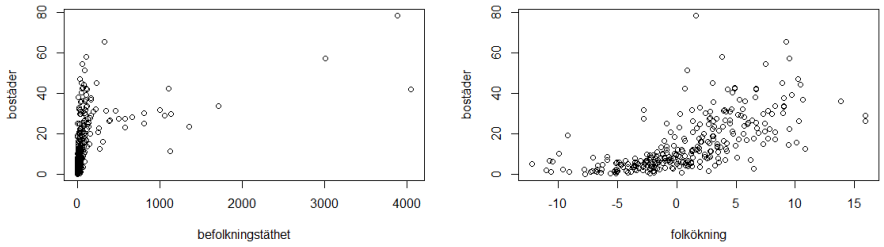
4.1 Bearbetning av datamaterialet

I den multipla linjära regressionen kommer vi att använda $\frac{bostäder}{folkmängd} \cdot 1000$, alltså antalet byggda bostäder per 1000 invånare, som responsvariabel. Den-

na kommer därför att refereras till som endast *bostäder* i detta avsnitt.

En förutsättning man gör när man arbetar med regression är att de olika observationerna av responsvariabeln är oberoende. Detta innebär att ett visst värde på en observation inte gör något värde på en annan observation varken mer eller mindre sannolikt. I vårt fall är det inte helt säkert att observationerna i responsen är oberoende men vi kommer ändå att anta det i analysen. Se Avsnitt 5.2 för vidare diskussion kring observationernas oberoende.

Det första vi bör fråga oss är om det räcker med att låta alla variabler vara i sin grundform eller om vi borde transformera någon. Vi undersöker detta genom att göra en scatterplot över *bostäder* och respektive kovariat. För varje plott undersöker vi därefter visuellt om sambandet verkar ickelinjärt. Det visar sig att det finns två variabler som eventuellt skulle behöva transformeras, nämligen *befolkningstäthet* och *folkökning*, se Figur 3.

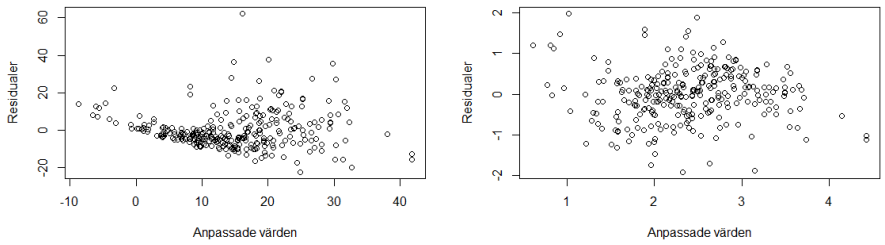


Figur 3: Responsvariabeln mot *befolkningstäthet* respektive *folkökning*.

Sambandet mellan *folkökning* och *bostäder* verkar vara kvadratiskt. Vidare verkar *befolkningstäthet* vara skevt fördelad. Ett vanligt sätt att hantera denna typ av problem är att logaritmera eller dra roten ur variabeln. För *befolkningstäthet* ger logaritmering en mycket jämnare spridning. Sambandet ser nu även här kvadratiskt ut varför vi bör lägga till en kvadratterm i analysen.

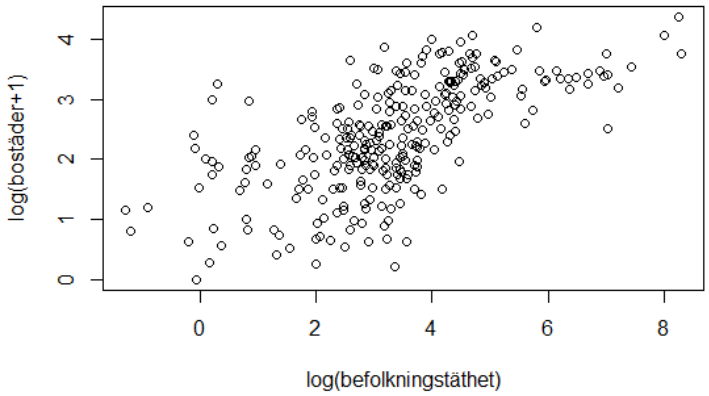
Ett av antagandena man gör i linjär regression är att residualerna är homoskedastiska. Om så inte är fallet kan man ändå i vissa fall lösa situationen genom någon transformation av responsvariablen. För att undersöka om det förekommer heteroskedasticitet i vårt fall genomför vi en enkel regression med den kovariat som ensam bäst beskriver responsvariabeln, vilket är *folkökning*. Om vi genomför den regressionen och därefter ritar upp en plott över residualer mot anpassade värden, som syns i Figur 4, ser vi att variansen verkar öka med ökande väntevärde. Sätt att lösa denna typ av problem är att exempelvis logaritmera responsen. Vi har en observation av *bostäder* som har värdet 0 och $\log(0)$ är odefinierat. Däremot kan vi lägga till 1 till alla observationer och alltså transformera variabeln som $\log(\textit{bostäder}+1)$. Om vi gör detta och på samma sätt genomför enkel regression med *folkökning* som

förklarande variabel och plottar residualer mot anpassade värden ser resultatet mycket bättre ut, se Figur 4. Vi väljer alltså att använda log(*bostäder*+1) som responsvariabel.



Figur 4: Residualer mot anpassade värden då en regression av *bostäder* respektive log(*bostäder* + 1) mot *folkökning* genomförts.

Eftersom vi har transformerat responsvariabeln är det inte längre säkert att transformationerna av *befolkningstäthet* och *folkökning* gäller. Dessutom kan någon annan variabel nu behöva transformeras. Därför gör vi om det vi gjorde tidigare och ritar upp scatterplots över log(*bostäder*+1) mot respektive förklaringsvariabel. Plottarna vi får fram nu är dock lika dem vi fick när vi använde *bostäder* som respons, men det visar sig att vi inte längre behöver transformera *folkökning*. Däremot ger fortfarande en log-transformation av *befolkningstäthet* en mycket jämnare spridning. Nu ser dock sambandet inte kvadratiskt ut och vi konstaterar att vi bör anpassa en spline-kurva för att beskriva förändringen (Figur 5).



Figur 5: En scatterplot över log(*bostäder*+1) mot log(*befolkningstäthet*).

Det är rimligt att anta responsen är korrelerad med kovariaten *folkökning* på så vis att ju mer befolkningen ökar desto fler bostäder bygger man. En

första undersökning tydde också på detta. Frågan är dock hur det ser ut när befolkningen minskar. Responsvariabeln är ju antalet byggda bostäder och inte nettotillskottet av bostäder; den kan alltså inte anta negativa värden. Det är rimligt att anta att man när befolkningen minskar inte bygger några nya bostäder, oavsett hur stor eller liten minskningen är. Vi gör därför om variabeln *folkökning* till två variabler: *folkökningTot*, som är samma som den gamla *folkökning*, och *folkökningPos*, som är samma som *folkökning* men med alla negativa värden satta till 0. För att testa vilken av dessa som beskriver data bäst genomför vi två enkla linjära regressioner med respektive av *folkökningPos* och *folkökningTot* och jämför därefter deras förklaringsgrad.

Det ser ut som *folkökningPos* beskrivs bäst av ett andragradspolynom. Det spelar dock ingen roll för när vi genomför regression med *folkökningTot* som förklarande variabel får vi $R^2 = 0.5181$ medan vi för *folkökningPos* får $R^2 = 0.4783$, trots att vi tekniskt sett har en extra förklarande variabel. Vi konstaterar därför att *folkökningTot* är den bästa varianten av *folkökning* och vi återgår till att referera till den som bara *folkökning*.

4.2 Multipel linjär regression

Vi kom ju tidigare fram till att vi behöver anpassa en spline-kurva till log(*befolkningstäthet*). Bästa sättet att komma fram till hur många knutar vi bör ha och var de bör sitta är att helt enkelt testa oss fram. Redan från början hamnar vi dock lite i ett dödläge. Vilken spline-kurva som passar bäst beror nämligen på vilka andra kovariater som vi har med men vilka kovariater som bör vara med kan bero på vilken spline-kurva vi använder. Eftersom spline-kurvan beror mer på vilka övriga kovariater som är med än vad valet av variabler beror på spline-kurvan väljer vi dock först, med hjälp av stegvisa procedurer, ut vilka variabler som beskriver log(*bostäder*+1) bra och därefter anpassar vi en spline-kurva till log(*befolkningstäthet*).

Oavsett om vi använder framåt- eller bakåtproceduren väljs *folkökning*, *höger*, *skatt*, *skog* och *ålder* ut förutom log(*befolkningstäthet*). Eftersom *höger* och *vänster* måste förekomma tillsammans eller inte alls, och det ger bättre AIC att lägga till *vänster* än att ta bort *höger*, låter vi även *vänster* vara med i modellen.

För att nu hitta den bästa knutsekvensen genomgår vi några steg. Till att börja med placerar vi ut en inre knut vid 0.5-kvantilen och undersöker vad vi får för korrigerad förklaringsgrad. Därefter testar vi att ha två knutar, vid 0.33- respektive 0.67-kvantilen, och undersöker R^2_{adj}. Vi fortsätter på detta vis tills vi inte längre får en höjning av R^2_{adj}. Därefter testar vi att ta bort en knut i taget, undersöker förändringen i den korrigerade förklaringsgraden och tar permanent bort den knut vars uteslutning ger störst höjning. När vi inte längre kan få en höjning av R^2_{dj} avslutar vi det steget. Slutligen justerar vi knutarna. För varje knut testar vi att flytta den ett steg bakåt respektive framåt och väljer den justering som ger störst höjning av R^2_{adj}.

När vi inte längre kan få en höjning gör vi samma sak fast med steglängd 0.5 och därefter steglängd 0.1. Den knutsekvens vi har kommit fram till då blir den vi väljer.

Den knutsekvens som vi får fram genom att använd denna metod är $(2.0, 3.1, 3.3, 3.5)$. Det finns förstås ingen garanti för att detta är den bästa möjliga knutsekvensen, men den bör passa data hyfsat bra.

För säkerhets skulle genomför vi nu nya stegvisa procedurer för att se om samma förklarande variabler väljs ut när log(*befolkningstäthet*) har fått en spline-kurva. Det visar sig dock att även nu väljs log(*befolkningstäthet*), *folkökning*, *höger*, *skatt*, *skog* och *ålder* ut. Precis som tidigare visar det sig även att det är bättre att lägga till *vänster* än att ta bort *höger*, så vi inkluderar även *vänster* i modellen.

Vi undersöker koefficientskattningarna i modellen. De flesta är signifikant nollskilda på nivån 5%. Undantagen är *vänster* och vissa av splinefunktionerna. Dessutom är *höger* endast signifikant på nivån 10%. Att *vänster* är insignifikant är inte förvånande − den valdes ju inte med i de stegvisa procedurerna. Faktum är att den får så lågt skattad koefficient ($\widehat{\beta} = -0.0035$) och extremt dålig signifikans ($p = 0.941$) att det antyder att det inte är någon skillnad mellan vänster- och koalitionsstyre när det kommer till bostadsbyggande. Vi bestämmer oss därför att ändå låta *höger* stå själv i modellen. Nu skiljer vi alltså bara på "högerstyre" och "inte högerstyre". Eftersom *vänster* inte bidrog mycket till modellen blir det nya resultatet liknande, men nu är även *höger* signifikant på nivån 5%.

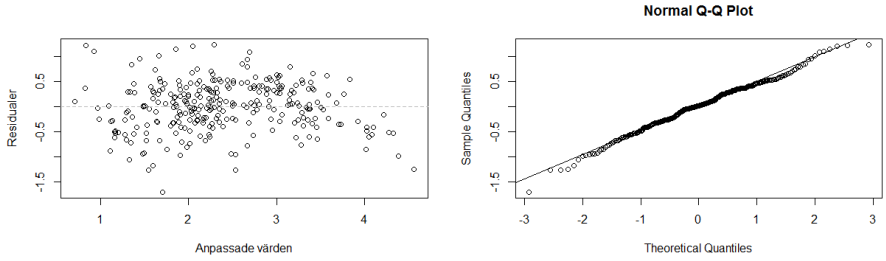
Att vi fick några insignifikanta resultat i vår splineregression är inte något problem. Det skulle kunna tyckas att detta tyder på att vi inte borde ha så många knutar men enligt Charpentier (2016) kan ett litet antal signifikanta spline-funktioner vara mycket sämre än ett större antal funktioner där vissa får ickesignifikanta skattningar. Eftersom vi kom fram till denna sekvens genom att eftersöka ett bra värde på R^2_{adj} får vi ändå anta att spline-kurvan i fråga beskriver sambandet mellan log(*befolkningstäthet*) och log(*bostäder*+1) bra.

I Figur 6 har vi plottat residualer mot anpassade värden respektive en normalfördelningsplott av residualerna. Vi ser att residualerna är jämnt spridda kring medelvärdet. Dessutom ser normalfördelningsplotten bra ut.

4.3 Generaliserade linjära modeller

Eftersom vår responsvariabel räknar någonting är det relevant att anta att den kommer från en räknefördelning. Den vanligaste räknefördelningen är Poisson och därför testar vi Poissonregression. Vår responsvariabel är nu det totala antalet byggda bostäder per kommun och vi använder *folkmängd* som offsetvariabel.

Det skulle kunna vara så att andra variabeltransformationer än de från den linjära regressionen är lämpliga nu. Vi väljer dock att ha förklarings-



Figur 6: Residualer mot anpassade värden respektive normalfördelningsplott över residualer i den slutgiltiga modellen i multipel linjär regression.

variablerna i samma form som i den linjära regressionen. Detta innebär att vi använder samma knutsekvens till log(*befolkningstäthet*).

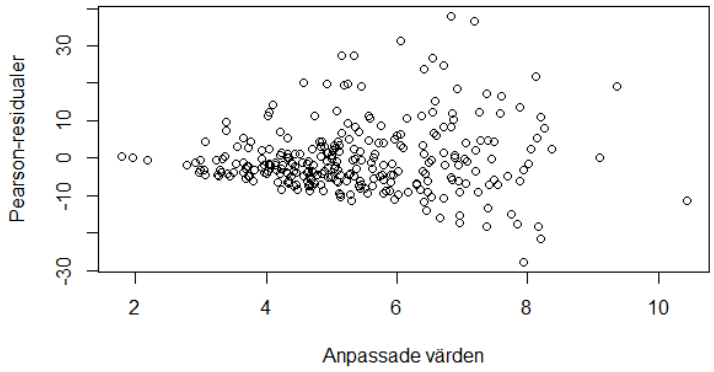
Nu undersöker vi med hjälp av stegvisa procedurer om vi kan utesluta några förklarande variabler. Både när vi använder framåt- och bakåtproceduren väljs dock alla förklaringsvariabler ut. Tittar vi på koefficientskattningar ser resultatet underligt ut. Nästan alla skattningar är signifikant skilda från 0 på nivån 0.001%. Många är dessutom signifikanta på den minsta nivån som datorn kan räkna med, ε_{mach}. Det är visserligen bra att våra skattningar är signifikanta, men det är inte realistiskt att alla förklaringsvariabler ger ett så signifikant resultat.

Vi har ett AIC på 24268.73 men det säger ingenting om hur bra modellen är överlag utan kan bara användas i jämförelse med andra modeller. När vi arbetar med GLM kan vi inte heller beräkna korrelationskoefficienten R^2. Det finns dock flera olika pseudo-varianter som ger ett mer eller mindre liknande resultat. Faraway (2006) föreslår $1-\frac{residual\ deviance}{null\ deviance}$. Om vi använder den varianten får vi att $R^2_{pseudo} = 0.7292$ vilket antyder att modellen ändå förklarar variationen bra.

En av följderna om man har data som är överspridd men ändå modellerar den med hjälp av Poissonregression är just att skattningarnas standardavvikelser underskattas, vilket innebär att koefficienter verkar vara signifikanta när de egentligen inte är det (Hilbe, 2014). Detta antyder starkt att vi har överspridning i data.

Om vi genomför ett överspridningstest får vi p-värdet $4.077 \cdot 10^{-5}$ vilket betyder att vi på alla rimliga signifikansnivåer kan förkasta nollhypotesen om att vi inte har överspridning. Detta bekräftas när vi plottar Pearsonresidualer mot anpassade värden i Figur 7. Notera gärna även skalan på y-axeln. Residualerna är väldigt stora. Vi bör därför anpassa en annan typ av modell.

Om vi testar att därför genomföra en negativ binomialregression med alla kovariater får vi ett AIC på 3485.81 så redan det är väldigt mycket bättre än modellen vi fick fram med hjälp av Poissonregression.



Figur 7: Pearsonresidualer mot anpassade värden i modellen i Poissonregression.

Använder vi sedan stegvisa procedurer väljs samma variabler som i linjära regressionen ut, alltså log(*befolkningstäthet*), *folkökning*, *höger*, *skatt*, *skog* samt *ålder*, och precis som i linjära regressionen får vi ett lägre värde på AIC om vi lägger till *vänster* än om vi tar bort *höger*. Även här är dock *vänster* mycket insignifikant. Vi drar samma slutsats som tidigare, alltså att det inte är någon skillnad på vänster- och koalitionsstyre, och låter därför *höger* vara själv i modellen.

Både koefficientskattningarna och deras signifikanser blir mycket lika de vi fick fram i linjära regressionen.

5 Diskussion

5.1 Resultat

Vi kan inte lita på Poissonmodellen, eftersom den inte kan ta hänsyn till överspridningen, vilket gör att vi har den multipla linjära modellen och negativa binomialmodellen att välja mellan för tolkningen. Det finns argument för och emot båda men det spelar faktiskt inte så stor eftersom samma variabler ingår i båda modellerna och dessutom fick båda modeller snarlika skattningar.

Nu använde vi visserligen olika responsvariabler i de två olika regressionerna. I negativa binomialregressionen modellerade vi log(*bostäder*) med log(*folkmängd*) som offset, vilket är detsamma som att modellera $\log(\frac{bostäder}{folkmängd})$. I linjära regressionen använde vi däremot $\log(\frac{bostäder}{folkmängd} \cdot 1000)$. Enligt loga-

ritmlagar har vi dock att

$$\log\left(\frac{bostäder}{folkmängd}\cdot 1000\right) = \log\left(\frac{bostäder}{folkmängd}\right) + \log\left(1000\right).$$

Vidare är $\log(1000) \approx 6.9$. Alltså borde interceptskattningen för linjära modellen vara ungefär 6.9 mer än den för negativa binomialmodellen. Intercepten skattas till 11.23 i linjära modellen och 4.36 i negativa binomialmodellen. Skillnaden mellan dessa är 6.87 så det stämmer bra.

Eftersom det vi har modellerat är $\log(\frac{bostäder}{folkmängd}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ och det vi är intresserade av är $\frac{bostäder}{folkmängd}$ får vi att

$$\begin{aligned}\frac{bostäder}{folkmängd} &= \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \\ &= \exp(\beta_0)\cdot\exp(\beta_1 x_1)\cdot\ldots\cdot\exp(\beta_k x_k).\end{aligned}$$

Vi har alltså en multiplikativ modell vilket innebär att när någon variabel x_i ökar med 1 ökar responsen multiplikativt med $\exp(\beta_i)$.

Interceptet vi kom fram till ska tolkas som att det är hur många bostäder som byggs då alla förklaringsvariabler är 0. Det kan tyckas väldigt högt eftersom det motsvarar att det i ett sådant fall byggs över 80 bostäder per invånare och åttaårsperiod. Värt att notera är dock att det inte är en realistisk situation eftersom det till exempel skulle innebära att medelåldern i kommunen i fråga skulle vara 0 år, det vill säga att kommunens invånare endast skulle bestå av spädbarn. Vidare skulle även skatten ligga på 0% vilket också är orimligt, även om man inte kan kräva särskilt mycket skatt av bebisar.

Nu skulle man ändå kunna argumentera för att interceptet är orimligt stort eftersom det används även när förklaringsvariablerna har värden som inte är 0. Borde inte det göra så att vi alltid får en skattning som säger att det byggs absurt många bostäder? Om vi undersöker vad de olika variablerna antar för värden (nämns i Avsnitt 3) och deras koefficientskattningar (gås igenom härnäst) så märker vi dock att variablerna med negativ skattning både antar större värden och har koefficientskattningar med större absolutbelopp. De drar alltså ner resultatet tillräckligt mycket för att det ska vara rimligt inom de intervall vi har undersökt.

5.1.1 Tolkning av koefficienterna

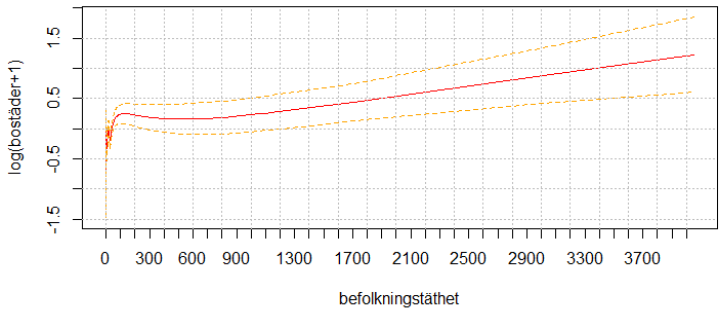
Det kanske intressantaste resultatet är det för *skog*. En scatterplot över variabeln och responsen antyder ett mycket vagt samband mellan dem och att om det finns något är det negativt. Plotten går att se som Figur 10 i Appendix. Intrycket förstärks av att de har korrelationskoefficienten $r = -0.155$. I alla modeller som har testats har dock *skog* fått hög signifikans och skattningen har dessutom alltid varit positiv. I linjära regressionen blev koefficientskattningen 0.0214 och i negativa binomial-regressionen blev skattningen

0.0250, vilket exponentierat blir 1.0216 respektive 1.0253. Detta innebär att en procentenhet mer improduktiv skogsmark leder till drygt två procent fler byggda lägenheter.

Även *höger* fick en positiv koefficientskattning. Den exponentierade skattningen blev 1.1009 för den linjära regressionen och 1.1298 för negativa binomialregressionen. Under en mandatperiod med högerstyre byggs det alltså drygt 10% fler bostäder än om det hade varit vänster- eller koalitionsstyre.

Vidare visar det sig att hög skatt och hög medelålder inverkar negativt på antalet byggda bostäder. För varje procentenhet skatten ökar minskar bostadsbyggandet med 12 procent enligt linjära modellen och 13 procent enligt negativa binomialmodellen. För åldern fick vi liknande resultat. Ett år högre medelålder ger 13 (linjära modellen) eller 14 (negativa binomialmodellen) procent färre lägenheter.

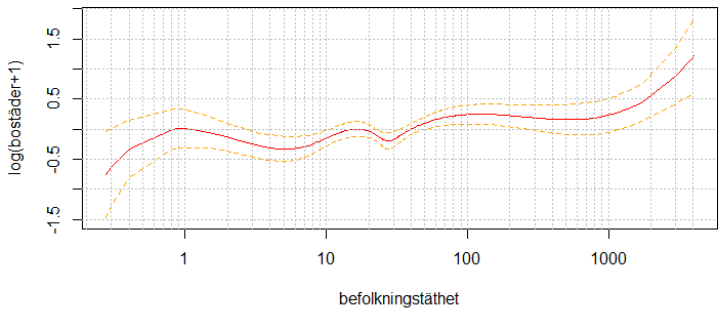
Effekten av variabeln *befolkningstäthet* är mer svårtolkad. Här kan vi inte bara titta på koefficientskattningar. Det bästa sättet att undersöka effekten är istället att helt enkelt rita upp den anpassade spline-kurvan. Även här fick vi snarlika resultat i de två olika regressionerna så vi visar bara kurvan från linjära regressionen, som syns i Figur 8, där vi även har lagt till ett 95%-igt konfidensintervall och ett rutnät som stöd för ögat. För den intresserade läsaren finns motsvarande kurva från negativa binomialregressionen som Figur 11 i Appendix.



Figur 8: Spline-kurvan för *befolkningstäthet* från linjära regressionen med 95%-igt konfidensintervall.

Vi ser att överlag verkar hög befolkningstäthet bidra till fler byggda bostäder. När befolkningstätheten är ungefär 125 invånare per kvadratkilometer minskar dock befolkningen lite innan den åter börjar öka vid ungefär 500 invånare per kvadratkilometer. För låga värden på *befolkningstäthet* beter sig kurvan dock underligt och gör väldigt kraftiga svängningar. Det är över huvud taget svårt att se hur kurvan ser ut bland de låga värdena. Vi undersöker därför hur kurvan ser ut då *befolkningstäthet* är logaritmerad,

vilket syns i Figur 9. Även här går versionen från negativa binomialregressionen att se i Appendix, i Figur 12.



Figur 9: Spline-kurvan för log(*befolkningstäthet*) från linjära regressionen med 95%-igt konfidensintervall.

Här syns förändringen mycket bättre. Svängningarna upp och ner är många. Responsen ökar ganska kraftigt till ungefär 1 invånare per kvadratkilometer. Därefter minskar den till ungefär 5 invånare per kvadratkilometer då det börjar öka igen. Mellan ungefär 15 och 30 invånare per kvadratkilometer är den en ganska kraftig dipp men därefter blir det en kraftig ökning till ungefär 125 invånare per kvadratkilometer. På konfidensintervallet kan vi dock se att vi på nivån 5% inte kan säkerställa några av nedgångarna. Kom ihåg att multiplikativiteten i modellen även gäller spline-kurvan. När *befolkningstäthet* ökar ökar/minskar alltså responsen multiplikativt enligt spline-kurvan. Detta innebär till exempel att om två kommuner har en befolkningstäthet på 1 respektive 40 invånare per kvadratkilometer och i övrigt samma värden på förklaringsvariablerna byggs det också approximativt lika många bostäder i dem.

Men hur var det nu med folkökningen? Bygger man mer ju mer befolkningen ökar? Ja. Koefficienten framför *folkökning* skattas till 0.0726 av linjära modellen och 0.0820 av negativa binomialmodellen. Exponentierat blir det 1.0753 respektive 1.0855. Variabeln representerade ju antalet inflyttade per 1000 invånare vilket innebär att när befolkningen ökar med en promille byggs ungefär 8 procent fler lägenheter. Vidare byggs även ungefär 8 procent färre lägenheter för varje promille befolkningen minskar. Detta kan verka som att för varje person som flyttar in i en kommun bygger man 80 bostäder, men riktigt så är det inte.

Enklaste sättet att beskriva denna förändring är nog genom ett exempel. Det genomsnittliga antalet invånare i en kommun är ungefär 30 000. Vidare bygger en genomsnittlig kommun ungefär 720 bostäder per åttaårsperiod, vilket motsvarar 3 bostäder per 1000 invånare och år. Om det sedan flyttar

in 30 personer i kommunen per år motsvarar det 1 person per 1000 invånare och år. Då ökar alltså bostadsbyggandet med 8 procent. Detta innebär att man nu bygger $3 \cdot 1.08 = 3.24$ bostäder per år. Till en inflyttad person i denna exempelkommun bygger man alltså knappt en fjärdedels bostad.

5.2 Förslag på förbättringar

Ett problem med studien är valet av perioder som vi har tittat på. Vi har antagit att det tar i snitt åtta år från beslut till färdig bostad. Källorna för detta är dock svaga och även om det stämmer är det bara ett snittvärde som kan variera mycket. Bostäder som byggts under perioden för responsen kan det ha fattats beslut om både innan och efter perioden för kovariaterna. Detta gör hela analysen mycket osäker. Ett bättre sätt hade varit att undersöka hur många lägenheter det fattas beslut om för att på så sätt kunna ha kovariater och responsvariabel från samma period. Främsta anledningen till att detta inte har gjorts är svårigheter med att hitta data rörande beslut om bostadsbyggande.

En av förutsättningarna när man genomför regressionsanalys är att de olika observationerna av responsvariabeln är oberoende av varandra. I vårt datamaterial är det inte säkert att det är uppfyllt eftersom det kan byggas bostadsområden som sträcker sig över kommungränser. Detta är emellertid mycket ovanligt så våra observationer bör vara så gott som oberoende. Det bästa vore förstås ändå om vi hade tillgång till data över byggnadsprojekt och huruvida de byggs i en eller flera kommuner för att därefter kunna ta ställning till åtgärder om det uppstår fall då något projekt sträcker sig över flera kommuner.

En annan sak som man kan ställa sig kritisk till är spline-kurvan för *befolkningstäthet*. Den har många upp- och nedgångar och frågan är om det verkligen ser ut så i verkligheten. Ett av argumenten för spline-regression är just att man vill undvika att få en kurva som går upp och ner orealistiskt många gånger, vilket är precis vad vi ändå fick nu.

En tanke fanns också om att inkludera tvåfaktorsamspel. Det mest relevanta hade då varit att låta *kommungrupp* samspela med olika variabler. Då *kommungrupp* inte inkluderades i någon modell blev detta dock inte aktuellt. Nu för tiden är kommungruppsindelningarna dock andra vilket kanske skulle ha förändrat resultatet. En annan möjlighet hade varit att slå ihop vissa kommungrupper för att kanske få variabeln att ge ett mer signifikant resultat.

I analysen kom vi fram till tre olika modeller med hjälp av tre olika regressionsmetoder; multipel linjär regression, Poissonregression samt negativ binomialregression. Alla har sin för- och nackdelar.

I den linjära regressionen förutsätter vi att responsvariabeln är normalfördelad runt väntevärdet. Detta uppfyllde inte vår responsvariabel som i sin grundform enbart kan anta naturliga tal. Responsens natur föreslår

dock att den skulle kunna vara Poissonfördelad och en slumpvariabel från en Poissonfördelning med en parameter μ som är större än ungefär 15 kan approximeras med just en normalfördelning (Alm & Britton, 2008). Över 90% av kommunerna byggde fler än 15 bostäder så det skulle kunna gå att motivera att använda linjär regression för de högre värdena på *bostäder* om den används i sin grundform. Den normalfördelning som kan användas som approximation är dock $N(\mu, \mu)$ vilket skulle leda till heteroskedasticitet.

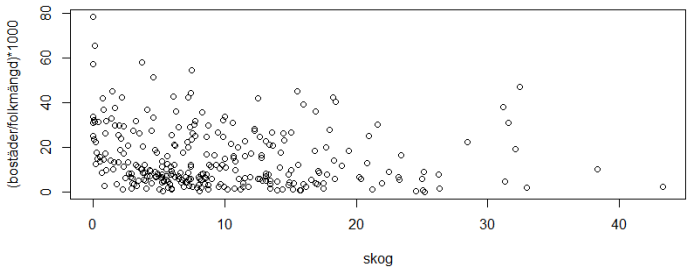
Vi undersökte emellertid antalet byggda bostäder dividerat med antalet invånare. Då kan alla ickenegativa tal antas. Normalfördelning är då mer intuitivt försvarbart, även om det är problematiskt att negativa värden inte kan antas.

Det vi slutligen använde som responsvariabel i den linjära regressionen var $\log(\frac{\textit{antal byggda bostäder}}{\textit{antal invånare}} + 1)$ som också bara kan anta positiva tal, eftersom $\log(1) = 0$. För att få responsvariabeln att kunna sträcka sig över hela reella tallinjen hade istället $\log(\frac{\textit{antal byggda bostäder}+1}{\textit{antal invånare}})$ kunnat användas. Skälet till att den inte användes är att insikten om den möjligheten kom för sent.

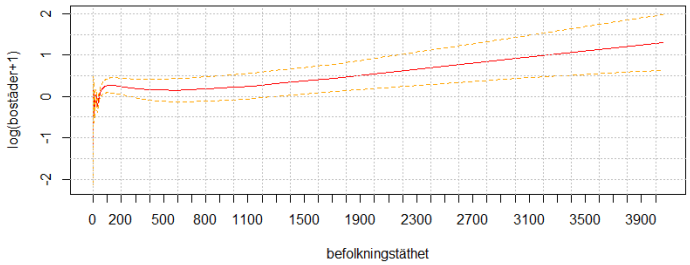
I linjära och negativa binomialregressionen valdes flera och samma förklaringsvariabler bort. I Poissonregressionen blev däremot resultatet att AIC skulle bli beaktansvärt sämre om någon förklarande variabel alls skulle tas bort. Dessutom fick vi extremt hög signifikans på koefficientskattningarna. Detta beror på att vår data är överspridd. En Poissonmodell kan inte räkna med spridningen då Poissonfördelningen endast beror på en parameter. Därför var vi tvungna att förkasta Poissonmodellen.

Resultatmässigt är det svårt att säga vilken modell som blev bäst av linjära modellen och negativa binomialmodellen. Däremot ger en linjär modell enklare beräkningar än en negativ binomialmodell varför en linjär modell bör lämpa sig bättre om man genomför en ny undersökning.

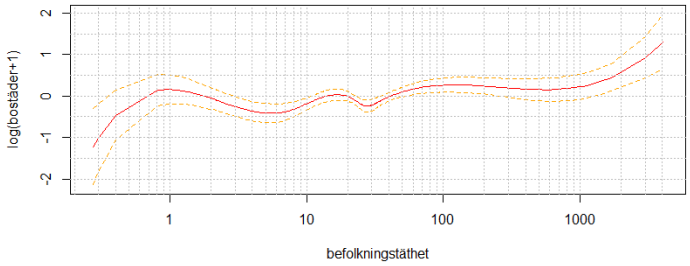
6 Appendix



Figur 10: Scatterplot över *skog* och $(bostäder/folkmängd) \cdot 1000$.



Figur 11: Spline-kurvan för *befolkningstäthet* från negativa binomialregressionen med 95%-igt konfidensintervall.



Figur 12: Spline-kurvan för log(*befolkningstäthet*) från negativa binomialregressionen med 95%-igt konfidensintervall.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.70573 -0.30944  0.01665  0.35113  1.23315

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     11.232478   1.457524   7.707 2.38e-13 ***
folkökning                                       0.072552   0.008935   8.120 1.60e-14 ***
ålder                                           -0.138047   0.019203  -7.189 6.22e-12 ***
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))1  1.606867   0.618305   2.599 0.009861 **
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))2 -0.346722   0.429994  -0.806 0.420746
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))3  0.973268   0.393494   2.473 0.013990 *
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))4  0.509283   0.367080   1.387 0.166450
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))5  1.775353   0.491686   3.611 0.000363 ***
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))6 -0.090999   0.562680  -0.162 0.871642
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))7  1.988161   0.493769   4.027 7.33e-05 ***
skog                                             0.021410   0.004930   4.343 1.98e-05 ***
skatt                                           -0.132054   0.038611  -3.420 0.000721 ***
höger                                            0.096113   0.038568   2.492 0.013292 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4946 on 274 degrees of freedom
Multiple R-squared:  0.7212, Adjusted R-squared:  0.709
F-statistic: 59.06 on 12 and 274 DF,  p-value: < 2.2e-16
```

Figur 13: Utskrift från linjära modellen.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.3094  -0.8368  -0.1621   0.4840   2.7608

Coefficients:
                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                      4.363925   1.622747   2.689 0.007162 **
ålder                                           -0.145732   0.021048  -6.924 4.39e-12 ***
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))1  2.781850   0.762935   3.646 0.000266 ***
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))2 -0.362490   0.520711  -0.696 0.486338
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))3  1.553443   0.497215   3.124 0.001782 **
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))4  0.912800   0.468170   1.950 0.051209 .
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))5  2.375114   0.590649   4.021 5.79e-05 ***
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))6  0.237700   0.657193   0.362 0.717584
bs(log(befolkningstäthet), knots = c(2, 3.1, 3.3, 3.5))7  2.530906   0.589931   4.290 1.79e-05 ***
höger                                            0.122046   0.042350   2.882 0.003954 **
skog                                             0.025027   0.005463   4.582 4.62e-06 ***
skatt                                           -0.139689   0.042448  -3.291 0.000999 ***
folkökning                                       0.082048   0.009834   8.344  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(3.52) family taken to be 1)

    Null deviance: 915.48  on 286  degrees of freedom
Residual deviance: 312.39  on 274  degrees of freedom
AIC: 3476.4

Number of Fisher Scoring iterations: 1


              Theta:  3.520
          Std. Err.:  0.306

 2 x log-likelihood:  -3448.393
```

Figur 14: Utskrift från negativa binomialmodellen.

7 Referenser

7.1 Teoretiska referenser

Agresti, Alan (2013), *Categorical Data Analysis*. 3 uppl. Hoboken: John Wiley & Sons.

Alm, Sven Erick & Britton, Tom (2008), *Stokastik - Sannolikhetsteori och statistikteori med tillämpningar*. Stockholm: Liber.

Andersson, Patrik & Tyrcha, Joanna (2014), *Notes in Econometrics*. Opublicerat manuskript. Stockholms universitet, Matematiska institutionen.

Bostads AB Poseidon (2013), *Från idé till boende*.
http://www.poseidonarsredovisning13.se/verksamhet/fastigheter
/franidtillboende/fran-id-till-boende.html (Hämtad februari 2016)

Boverket (2016) *240 kommuner har underskott på bostäder – men byggandet ökar*.
http://www.boverket.se/sv/om-boverket/publicerat-av-boverket/nyheter
/kraftig-okning-av-underskott-pa-bostader–men-byggandet-okar/ (Hämtad maj 2016).

Cameron, A. Colin & Trivedi, Pravin K. (2013), *Regression Analysis of Count Data*. 2 uppl. Cambridge: Cambridge University Press.

Charpentier, Arthur (2016), *Regression With Splines: Should We Care About Non-Significant Components?*
https://dzone.com/articles/regression-with-splines-should-we-care-about-non-s (Hämtad april 2016)

Cederfelt, Margareta (2013), Nu ska bostadsbyggen komma igång snabbare. *Svenska Dagbladet*, 24 januari.
http://www.svd.se/nu-ska-bostadsbyggen-komma-igang-snabbare/om/debatt
(Hämtad februari 2016)

Fahrmeir, Ludwig, Kneib, Thomas, Lang, Stefan & Marx, Brian (2013), *Regression - Models, Methods and Applications*. Berlin Heidelberg: Springer.

Faraway, Julian J. (2006), *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Boca Raton: Chapman & Hall/CRC.

Hilbe, Joseph M. (2014), *Modeling count data*. Cambridge: Cambridge Uni-

versity Press.

Lind, Hans (2003), *Bostadsbyggandets hinderbana – en ESO-rapport om utrvecklingen 1995-2001*. Stockholm: Finansdepartementet. (Ds 2003:6)

Länsstyrelsen i Stockholms län (2012), *Redovisning av kommunernas arbete med bostadsförsörjningen*. Stockholm.

Maindonald, John & Braun, W. John (2010), *Data Analysis and Graphics Using R - an Example-Based Approach*. 3 uppl. Cambridge: Cambridge University Press.

Miller, Steven J. (u.å.), *The Method of Least Squares*.
https://web.williams.edu/Mathematics/sjmiller/public_html/BrownClasses/54/handouts/MethodLeastSquares.pdf (Hämtad maj 2016)

Racine, Jeffrey S. (2014), *A Primer on Regression Splines*.
https://cran.r-project.org/web/packages/crs/vignettes/spline_primer.pdf (Hämtad mars 2016)

Sundberg, Rolf (2014),. *Lineära Statistiska Modeller*. Opublicerat manuskript. Stockholms universitet, Matematiska institutionen.

Sundell, Ulf (2014), Slopad detaljplan fel väg att öka byggandet. *Svenska Dagbladet*. 13 maj.
http://www.svd.se/slopad-detaljplan-fel-vag-att-oka-byggandet/om/debatt (Hämtad februari 2016)

7.2 Källor till data

bostäder: http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__BO__BO0101__BO0101A/LghReHtypLtAr/table/tableViewLayout1/?rxid=267992d4-92f5-40e5-b08a-01f657dc1ba9

folkmängd och *folkökning*: http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__BE__BE0101__BE0101A/BefolkningNy/table/tableViewLayout1/?rxid=6bcccba6-1418-4dc9-81bc-7d98d9d427d4

befolkningstäthet: http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__BE__BE0101__BE0101C/BefArealTathetKon/table/tableViewLayout1/?rxid=6bcccba6-1418-4dc9-81bc-7d98d9d427d4

inkomst: http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__HE__HE0110__HE0110A/SamForvInk1/table/tableViewLayout1

/?rxid=7edeb258-3c23-47df-83a1-349a1a98b847

skatt: http://www.statistikdatabasen.scb.se/pxweb/sv/ssd
/START__OE__OE0101/Kommunalskatt/table/tableViewLayout1
/?rxid=273d7535-bf6f-4349-93b5-e269523a2f0d
och
http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__OE__OE0101
/Kommunalskatter2000/table/tableViewLayout1/?rxid=273d7535-bf6f-4349-
93b5-e269523a2f0d

skog: http://www.statistikdatabasen.scb.se/pxweb/sv/ssd
/START__MI__MI0803__MI0803A/MarkanvJbSk/table/tableViewLayout1
/?rxid=81a9b46a-39ec-48a0-be61-8dbbf41c9d6f,
http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__MI__MI0802
/Landareal/table/tableViewLayout1/?rxid=81a9b46a-39ec-48a0-be61-
8dbbf41c9d6f
och
http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__MI__MI0802
/Areal/table/tableViewLayout1/?rxid=81a9b46a-39ec-48a0-be61-8dbbf41c9d6f

valdeltagande: http://www.statistikdatabasen.scb.se/pxweb/sv/ssd
/START__ME__ME0104__ME0104D/ME0104T4/table/tableViewLayout1
/?rxid=e1e2169b-3a52-402a-8bb2-9e1843d899f4

ålder: http://www.statistikdatabasen.scb.se/pxweb/sv/ssd
/START__BE__BE0101__BE0101B/BefolkningMedelAlder/table
/tableViewLayout1/?rxid=e1e2169b-3a52-402a-8bb2-9e1843d899f4

höger och *vänster*: http://skl.se/demokratiledningstyrning
/valmaktfordelning/valresultatmaktfordelning2014
/valresultatochmaktfordelningsammanstallning19942014.370.html

kommungrupp: Data gällande kommungruppsindelning har erhållits genom direkt korrespondens med Sveriges Kommuner och Landsting.