Stockholms universitet

Vädrets påverkan på löss

Emilia Olofsson

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2010:8
Matematisk statistik
Juni 2010

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2010:8**,
http://www.math.su.se/matstat

Vädrets påverkan på löss

Emilia Olofsson*

Juni 2010

Sammanfattning

Den här uppsatsen syftar till att undersöka huruvida det finns samband mellan vädret och spridningen av huvudlöss. Länge har kunskaperna om löss varit knapphändiga men på senare tid har påtryckningar kommit från framförallt skola efter mer forskning på området. Insamlade data över löss och vädret har lett till misstankar om korrelation mellan vädret och försäljningen av lusmedel. Via analys av detta kan medel för att prediktera försäljningen erhållas och den allmänna kunskapen om löss breddas. Data kommer från Danmark och består av totala försäljningen av lusmedel per månad under perioden december 2002 t.o.m. mars 2010, medeltemperatur och antal dagar med nederbörd per månad. I denna uppsats används regressionsanalys som främsta metod. Förmågan att prediktera har prövats med hjälp av korsvalidering. Både linjär och loglinjär regression utfördes. I den loglinjära antog vi data komma från en negativ binomialfördelning istället för en Poissonf ördelning p.g.a. överspridning. Resultatet av analysen är att temperaturen har ett positivt signifikant samband med försäljningen nästkommande månad och att månader med skollov också påverkar försäljningen positivt.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige. E-post: emilia.olofsson@hotmail.com. Handledare:Mikael Andersson.

Vädrets påverkan på löss

Emilia Olofsson

13 juni 2010

Sammanfattning

Den här uppsatsen syftar till att undersöka huruvida det finns samband mellan vädret och spridningen av huvudlöss.
Länge har kunskaperna om löss varit knapphändiga men på senare tid har påtryckningar kommit från framförallt skola efter mer forskning på området. Insamlade data över löss och vädret har lett till misstankar om korrelation mellan vädret och försäljningen av lusmedel. Via analys av detta kan medel för att prediktera försäljningen erhållas och den allmänna kunskapen om löss breddas.
Data kommer från Danmark och består av totala försäljningen av lusmedel per månad under perioden december 2002 t.o.m. mars 2010, medeltemperatur och antal dagar med nederbörd per månad.
I denna uppsats används regressionsanalys som främsta metod. Förmågan att prediktera har prövats med hjälp av korsvalidering.
Både linjär och loglinjär regression utfördes. I den loglinjära antog vi data komma från en negativ binomialfördelning istället för en Poissonfördelning p.g.a. överspridning.
Resultatet av analysen är att temperaturen har ett positivt signifikant samband med försäljningen nästkommande månad och att månader med skollov också påverkar försäljningen positivt.

Abstract

The aim of this thesis is to investigate whether there are significant correlations between the spread of lice and the weather.
For a long time the knowledge about lice has been limited, but the last few years mainly schools has started to question this. Since scientists started research on this area we have come to suspect that the sale of louseproducts and the weather are correlated. They are indirect since the lice do not depend on the weather, only of the hair of a human being. What makes them spread are instead human relations and how we interact physically. By analysing this we can achieve a model for predicting the sale in the future and even improve the common knowledge about louse.
The observations are from Denmark and contains the total number of sold louse products per month, mean temperature and the number of days with rain per month.
The methods used in this thesis is regression analyses. The regression models' power to predict are also evaluated with cross validation.
Both linear and loglinear models have been designed. In the loglinear models we assumed that the data came from a negative binomial distribution instead of a Poisson distribution due to overdispersion.
The conclusion is that there is a positive significant correlation between the temperature and the sale the next month. Even the months with schoolholidays affects positively.

Förord

Denna uppsats utgör ett självständigt arbete om 15 hp vilket leder till en kandidatexamen i matematisk statistik vid Stockholms Universitet.
Arbetet har utförts på uppdrag av Smittskyddsinstitutet i samarbete med danska KSL Consulting.

Jag vill rikta ett stort tack till min handledare Mikael Andersson, universitetslektor vid Stockholms Universitet, som väglett mig genom detta arbete. Jag vill även tacka Johan Nilsson, biolog vid Smittskyddsinstitutet och Kim Søholt Larsen, konsult och utvecklare av lusmedel, som båda har bistått med information rörande de biologiska frågorna.

Innehåll

1 Inledning

Sedan urminnes tider har befolkningen i Norden drabbats av huvudlöss. Idag är det främst ett problem hos barn i tidig skolålder och på förskola. Det medför egentligen inte någon medicinsk risk att drabbas av huvudlöss[1], troligtvis är det den största orsaken till att det ej har varit något stort ämne för forskning.

De senaste 10-15 åren har intresset ökat för mer kunskap om denna ohyra, i huvudsak via skola och barnomsorg. Det är som sagt främst barn som drabbas av löss, men även vuxna kan få det. De är vanligast hos flickor, huvudsakligen p.g.a. deras sätt att leka och vara med varandra; i allmänhet har flickor en närmare fysisk kontakt med varandra än pojkar. Då lössen varken kan hoppa eller flyga är det i första hand det sociala samspelet som styr spridningen av lössen. Att leka nära, fixa med varandras hår o.s.v. ökar risken för spridning. Det diskuteras huruvida det finns perioder då spridningen är extra omfattande; finns det faktorer som påverkar det sociala samspelet så att spridningen av löss gynnas eller missgynnas.

Lössen lever i hårbotten där de livnär sig på blod, ofta innebär det ganska svår klåda för bäraren av lössen. De överlever som högst 24 timmar utanför hårbotten[2] och då de trivs som bäst i minst 30°C och gärna fuktigt så är nära huvudkontakt den vanligaste orsaken till spridning. Risken att de sprids via mössor och annat är relativt liten eftersom det krävs en ganska hög värme för att de skall överleva. Blir det för torrt resulterar det i att de torkar ut.

Sedan några år tillbaka har Johan Nilsson och Kim Søholt Larsen forskat på detta område. Johan Nilsson är biolog på Smittskyddsinstituet i Stockholm och forskar på lössen i laboratorium. Kim Søholt Larsen är konsult och utvecklare av lusmedel i Danmark. Søholt Larsen är biolog i grunden, nu är han även verksam som föreläsare m.m.
Genom insamling av data i Danmark har de försökt utröna hur lössens spridning sker. Då man ej kan mäta hur många löss som finns eller hur många som är drabbade så förklaras spridningen av hur mycket avlusningsprodukter per månad som säljs. Avlusningsprodukterna består av luskammar och medel i form av oljor eller schampoon. Insamlingen av data har i sin tur lett till att de tror sig se samband mellan vad det är för väder och om förekomsten av löss ökar eller ej. Vädret definieras av månadens medeltemperatur och antal dagar med nederbörd.

Målet med det här arbetet är således att analysera insamlade data för att

[1] Sjukvårdsrådgivningen
[2] Lusguiden

avgöra om signifikanta samband föreligger.

2 Syfte och metod

Syftet är i första hand att undersöka huruvida det finns signifikanta samband mellan olika förklarande variabler och responsen; försäljningen av lusmedel. Vilka faktorer kan sägas påverka spridningen mest? De här frågorna ställs i första hand för att utvidga allmänhetens kunskaper om löss.

Det är även till viss del intressant att prediktera försäljningen av lusmedel. Att kunna prediktera skulle i första hand vara apotek och andra försäljningsplatser till nytta då man kan försöka förutspå åtgången. Däremot kan vi ej styra vädret eller hur barn leker med varandra, på så vis är det svårt att påverka förekomsten av löss. Samtidigt är det av vikt att kunna föra vidare kunskap om löss till skolor och föräldrar för att öka den allmänna kunskapen och ha bättre möjligheter att förhindra och stävja spridning.

Genom att utveckla en statistisk modell för data erhålls ett sätt att förklara det som redan observerats. Möjlighet fås att se samband och dra slutsatser om ämnet. Man skall dock komma ihåg att samband inte alltid är orsak och verkan.
I föreliggande arbete kommer regressionsanalys att tillämpas. Generaliserade linjära modeller grundat på antagande om två olika fördelningsfunktioner kommer att analyseras. För en första analys används antagande om additiva faktorer och att data är normalfördelat. Även regression med Poissonfördelningen kommer att tillämpas i form av loglinjära modeller. Det är rimligt att misstänka att responsen kan påverkas multiplikativt av de förklarande variablerna. De loglinjära modellerna tillåter sådana multiplikativa effekter. Loglinjära modeller är också intressant att undersöka av det skälet att antalet sålda lusartiklar per månad ej kan vara negativt vilket normalfördelningen antar. Loglinjära modeller medför endast positiva värden på responsen. Ett alternativ till Poissonfördelning för att göra loglinjära modeller är negativ binomialfördelning som bättre fångar upp stor variabilitet i data.

För att analysera modellernas riktighet används mått som Akaikes informationskriterium och residualkvadratsumma. Analyser om huruvida fördelningsantagandena stämmer görs med hjälp av grafer. Prediktionsanalys utförs i form av korsvalidering med *leave-one-out* metoden.
Programpaketet R används för att utföra beräkningarna.

3 Teori

3.1 Generaliserade linjära modeller

I denna uppsats används generaliserade linjära modeller (GLM). GLM tillåter en ytterligare utvidgning från endast normalfördelade responsvariabler med konstant varians, som i enkel linjär regression, till att kunna anta responsvariabeln komma från någon av fördelningarna tillhörande den exponentiella familjen.
En generaliserad linjär modell består av tre komponenter[3]:

1: *Slumpkomponenten* beskriver responsvariabeln Y och dess fördelningsfunktion som tillhör exponentialfamiljen, d.v.s. de oberoende observationerna $(y_1, ..., y_N)$ har en täthetsfunktion som kan skrivas på formen nedan vilken kallas exponentiella spridningsfamiljen[4]

$$f(y_i; \theta_i, \phi) = \exp\{[y_i\theta_i - b(\theta_i)]/a(\phi) + c(y_i, \phi)\} \tag{3.1}$$

Parametern θ_i kallas den naturliga parametern och kan variera för $i = 1, ..., N$, ϕ kallas för spridningsparametern. Vanligt förekommande fördelningar är exempelvis normalfördelningen, Poissonfördelningen, binomialfördelningen och gammafördelningen.

2: Den *systematiska komponenten* är den komponent som relaterar de förklarande variablerna till responsvariabeln. Det genom att specificera effekterna från just de oberoende förklarande variablerna på det förväntade y-värdet. Låt x_{ij} vara värdet på prediktorn j $(j = 1, ..., p)$ för i. Då gäller för varje observation $i = 1, ..., N$

$$\eta_i = \sum_j \beta_j x_{ij}$$

Linjärkombinationen av förklarande variabler ovan kallas för den linjära prediktorn.

3: *Länkfunktionen* är den sista komponenten som sammanlänkar slumpkomponenten och den systematiska komponenten. Låt $E(Y_i) = \mu_i$, då länkas väntevärdet μ_i till systematiska komponenten η_i av $\eta_i = g(\mu_i)$ där länkfunktionen g är en monoton, differentierbar funktion. Vi har alltså att g länkar väntevärdet till de förklarande variablerna via

$$g(\mu_i) = \sum_j \beta_j x_{ij}$$

där $i = 1, ..., N$ och $j = 1, ..., p$.

[3] Agresti, sid 116

[4] Agresti, sid 133. För mer information om exponentiella familjen se Lindgren, Avsnitt 6.11

Vid antagande om att observationerna är normalfördelade så är den så kallade *identitetslänken* vanligt förekommande och är även vad som tillämpas i detta arbete. Det innebär att länkfunktionen ger väntevärdet, $g(\mu) = \mu$. Identitetslänken har systematisk komponent $\eta = \mu$.
För fallet med endast positiva värden på responsvariabeln, d.v.s. man ställer kravet $\mu > 0$, kan man tillämpa Poissonfördelningen med en *log*-länk. Detta tvingar fram positiva värden på μ och vi får alltså att $\eta = \log \mu$. En modell av det slaget kan även kallas för en *loglinjär* modell. Det är dessa två varianter av GLM som tillämpas i detta arbete.

3.2 Maximum likelihoodmetoden

För att anpassa GLMs används Maximum likelihoodmetoden. Metoden innebär att ett dataset förklaras bäst av en parameter, d.v.s. de bästa skattningarna av data ges av $\hat{\theta}$. Värdet på $\hat{\theta}$ är värdet av θ som maximerar den så kallade likelihoodfunktionen $L(\theta)$[5]. Vanligast är att man använder sig av loglikelihoodfunktionen

$$\log(L(\theta|\mathbf{x})) = \sum \log(f(x_i|\theta))$$

Värdet $\hat{\theta}$ på parametern maximerar funktionen och med skattningen på parametern har man således erhållit den modell som beskriver data på bästa sätt.

3.3 Stegvis regression

När man genomför regression och har många förklarande variabler kan man vilja utesluta de variabler som visar sig ha mindre betydelse för anpassningen. Det kan dock vara svårt att avgöra vilka förklarande variabler som skall inkluderas i modellen och vilka som slutligen inte inkluderas. Den vanligast förekommande och också den metod som använts i denna uppsats kallas *stegvis regression*, som är en förfinad variant av *framlänges regression*.

Metoden[6] går till på så vis att man utgår från en tom modell där man succesivt inkluderar förklarande variabler en i taget. För att avgöra vilka variabler som skall inkluderas och i vilken ordning så tittar man på en full modell och sorterar sedan variablerna efter vilka som är mest signifikanta vid test av om parametern är noll. De som ej uppnår den på förhand bestämda signifikansnivån inkluderas ej i den tomma modellen. Den mest signifikanta variabeln inkluderas i den tomma modellen först och sedan den näst mest signifikanta o.s.v. Detta förfarande pågår tills alla variabler som var signifikanta i den fulla modellen inkluderats i den tomma. Vad som skiljer denna

[5]Lindgren, sid 225
[6]Sundberg, sid 70-71

metod från *framlänges regression* är att man efter varje steg kontrollerar att alla tidigare inkluderade variabler fortfarande är signifikanta. Om så ej är fallet tas variabeln återigen bort ur modellen.

Anledningen till att man kontrollerar att variablerna fortfarande är signifikanta efter varje steg är att t.ex. två variabler i kombination med varandra kan förklara data bra, men när en tredje inkluderas kan den beskriva data bättre än de två tidigare tillsammans.

3.4 Residualer

För att skapa sig en uppfattning om huruvida den anpassade modellen är bra eller ej så är residualer ett användbart redskap. Residualer är skillnaden mellan observerat värde och skattat värde och definieras enligt nedan[7]

$$e_i = y_i - (\hat{\alpha} + \hat{\beta} x_i) = y_i - \hat{y}_i$$

vilket är en skattning av det korrekta avståndet, ϵ_i mellan observation i och den korrekta regressionspunkten.

Vid antagande om att data är normalfördelat antas att residualerna är normalfördelade med väntevärde 0 och konstant varians, dvs $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. För att avgöra om detta antagande verkar vara någorlunda korrekt så undersöks grafer och normalfördelningsgrafer över residualerna.

På samma sätt vill man kunna kontrollera residualerna även vid antagande om annan fördelning än normalfördelningen. I sådana situationer ger residualerna definierade som ovan inte någon information om fördelningsantagandena utan för de fallen definieras *Pearson residualer*[8], betecknat med index P:

$$e_{Pi} = (y_i - \hat{y}_i)/\sqrt{Var(\hat{y}_i)}$$

Nu är residualen standardiserad och därmed approximativt normalfördelad vid stora väntevärden på $\hat{y}_i$. Nu kan man alltså göra normalfördelningsgrafer på dessa residualer och undersöka antagandets riktighet.

3.5 Akaikes informationskriterium

För att kunna jämföra modeller med varandra finns det flera olika sätt att gå tillväga. Ett sätt som använts i det här arbetet är att titta på *Akaikes informationskriterium*, AIC. Det är inte ett mått som testar någon speciell hypotes utan snarare ett sätt att betygsätta hur bra modellen passar till

[7] Blom & Holmquist, sid 216
[8] Agresti, sid 142

data och hur komplex den är. Om man då har flera modeller så kan denna ranking användas för att välja en modell. AIC definieras enligt[9]:

$$\text{AIC}=-2(\text{maximerad log likelihood}-\text{antal parametrar i modellen})$$

Som synes så beräknas värdet på den maximerade loglikelihood-funktionen, kriteriet tar även hänsyn till hur många parametrar man har med i modellen. Eftersom man strävar efter ett lågt värde så gynnas man ej av att ha många parametrar i och med att värdet då blir högre. Att ha många paramterar försvårar tolkandet och användandet av modellen.

3.6 Prediktion

Då man är intresserad av att förutsäga ett kommande utfall använder man sig av prediktion. Genom att ha utvecklat en modell baserat på observerade data kan man sedan med nya värden på de förklarande variablerna erhålla ett nytt värde på responsvariabeln. Det är alltså det utfall man förutspår för de specifika värdena på x-variablerna.

Att prediktera innebär en del svårigheter eftersom det kan falla sig så att framöver kanske inte samma betingelser råder kring det man vill observera och därmed är den modell man utgår ifrån ej tillförlitlig. Det är också så att om man vill prediktera för x-värden utanför de intervall som innefattas av modellen är det ej heller tillförlitligt eftersom det där kan råda omständigheter som man ej haft möjlighet att ta hänsyn till i utvecklandet av modellen.

För att undersöka hur bra en modell är för prediktion används *korsvalidering*, i form av *leave-one-out* metoden[10]. Den går ut på att man tillfälligt tar bort en observation i taget ur sitt datamaterial, sedan gör man regression utan den observationen och använder resultatet för att prediktera den borttagna observationen. Prediktionsfelet för denna observation i är då $y_i - \hat{\mu}_{i,-i}$ där notationen $-i$ betyder att observation i var borttaget ur materialet för att utföra regressionen. Detta görs sedan för alla observationer $i = 1, ..., N$ och man får måttet Mean Squared Error of Prediction

$$MSEP = \frac{1}{N}\sum_{1}^{N}(y_i - \hat{\mu}_{i,-i})^2 \tag{3.2}$$

som då helst skall vara så litet som möjligt för bästa prediktionsförmåga. Om man drar roten ur beräknat värde erhålls ytterlige ett mått kallat Root Mean Squared Error of Prediction; $RMSEP = \sqrt{MSEP}$. Detta kan nu ses som en grov uppskattning av prediktionsfelet.

[9] Agresti, sid 216
[10] Sundberg, sid 69

3.7 Överspridning

Överspridning är ett problem som kan uppstå då man försöker anpassa sina data till en modell där data har större variation än modellen tillåter. Det är ett vanligt fenomen då man anpassar sina data till en Poissonmodell där data antas bero på endast en parameter. Spridningsparametern för en sådan modell är 1, (jfr. ekvation 3.1). Variansen förutsätts vara densamma som väntevärdet för en Poissonmodell. Om variansen är större får man överspridning.

För situationen att anpassa en normalfördelning drabbas man sällan av samma problem eftersom man då har två parametrar som beskriver data; en parameter som beskriver väntevärdet och en som beskriver variansen. Ett av de vanligaste sätten att ta hänsyn till överspridning är att göra regression med den *negativa binomialfördelningen* som är en utvidgning av Poissonfördelningen. Den innehåller två parametrar och kan därmed ta tillvara på variationerna på ett bättre sätt än Poissonfördelningen, d.v.s. tillåta att variansen är högre än väntevärdet.

3.8 Negativ Binomialfördelning

Den negativa binomialfördelningen har sannolikhetsfunktion[11]:

$$f(y;k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y$$

$y = 0, 1, 2, ...$ och k och μ är parametrar. Väntevärde och varians för denna fördelning är

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

Parametern k^{-1} kallas spridningsparameter, för en känd sådan så kan sannolikhetsfunktionen uttryckas på formen av den naturliga exponentiella familjen och är då en generaliserad linjär modell[12]. Parametern är dock oftast inte känd utan den skattas, därmed kan man skaffa sig en uppfattning om överspridningens omfattning.

Kopplingen till Poissonfördelningen ligger i att då $k^{-1} \to 0$ så konvergerar den negativa binomialfördelningen till Poissonfördelning. Även här används en *log*-länk, precis som för Poissonfördelningen i GLM. Därmed erhålls loglinjär regression både vid antagande om negativ binomialfördelning och Poissonfördelning.

[11] Agresti, sid 131

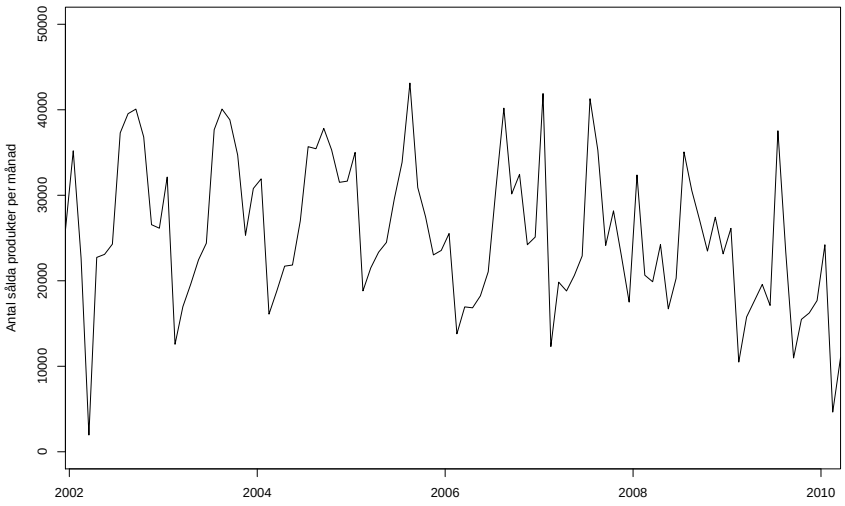
[12] För mer information om exponentiella familjen och generaliserade linjära modeller se Lindgren, avsnitt 6.11 respektive Agresti sid 116

4 Data

Datamaterialet är hämtat från Danmark som är det land i Norden som är mest drabbat av löss. Då vädret i Danmark är relativt homogent och så även utbredningen av löss så det är rimligt att göra en modell för hela landet.
Data består av totala antalet sålda avlusningsartiklar, d.v.s. lusmedel och luskam, per månad. Det är dessa som agerar responsvariabel. De förklarande variabler som används i denna uppsats är temperatur och nederbördsmängd. Dessa är uppmätta som medeltemperatur per månad och antal dagar med nederbörd per månad.

Kim Søholt Larsen har sammanställt Danmarks totala försäljningssiffror från december 2001 t.o.m. mars 2010, allt som allt 100 observationer. Data över vädret är hämtat från Danmarks Meteorologiske Institut som definierar dag med nederbörd som de dagar då nederbördsmängden är större än $0,1$mm.

Det är även av stor vikt att ta hänsyn till tiden, eftersom det redan på förhand är känt att spridningen oftast ökar då barn återgår till skola och förskola efter längre lov.



Figur 1: *Antalet sålda avlusningsprodukter per månad för hela tidsperioden.*

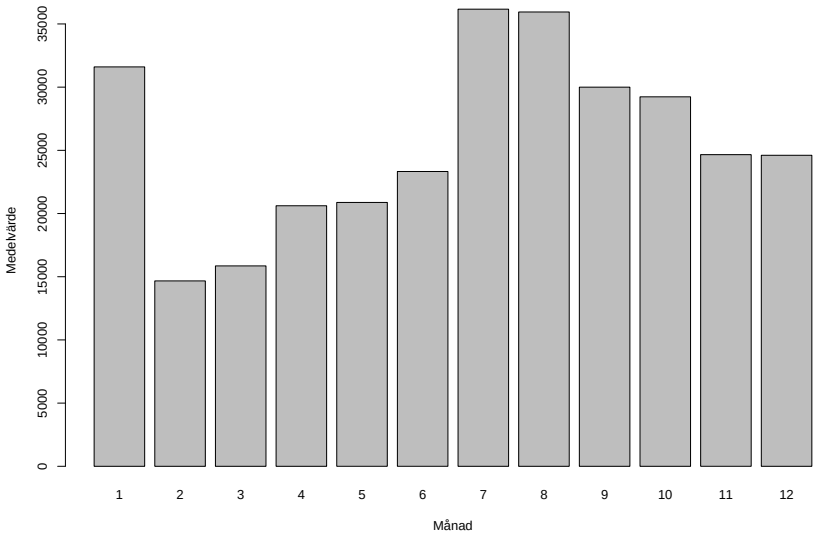
5 Analys

5.1 Undersökning av datamaterialet

Datamaterialet består av antal sålda avlusningsprodukter per månad, medeltemperatur per månad och antal dagar med regn per månad. I Figur 1 så ses antal sålda avlusningsprodukter per månad för hela perioden, där ses att det finns några i jämförelse lite mer extrema värden i början och i slutet av datasetet. De två mest dramatiska avstickarna identifieras till mars 2002 och februari 2010 som hade förhållandevis väldigt låga försäljningssiffror. Man ser också att de sista tre åren sjunker försäljningen något. I övrigt så verkar det inte under denna period ha skett någon väldigt dramatisk förändring i antalet sålda avlusningsprodukter per månad.

Försäljningen varierar mycket under året. I Figur 2 ses månadsvisa medelvärden för hela materialet.

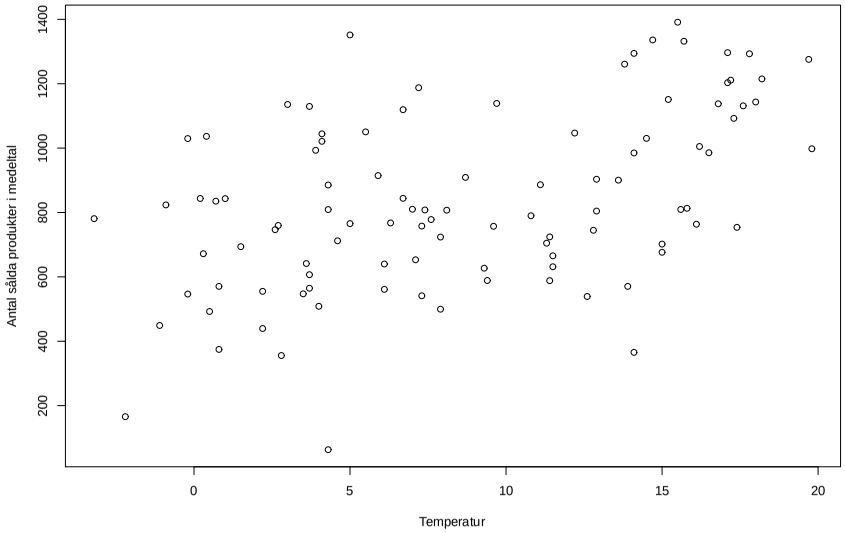
Då månaderna är olika långa har försäljningssiffrorna omvandlats till medeltal per dag för respektive månad och likadant med regn. D.v.s. värdena för antal sålda lusprodukter respektive antal dagar med nederbörd har dividerats med antalet dagar i varje månad. Hänsyn har tagits till februari 2004 och februari 2008 då det var skottår. Temperaturen är redan angiven som ett medelvärde så den lämnas orörd. Dessa nu beräknade medelvärden används i så stor utsträckning som möjligt.

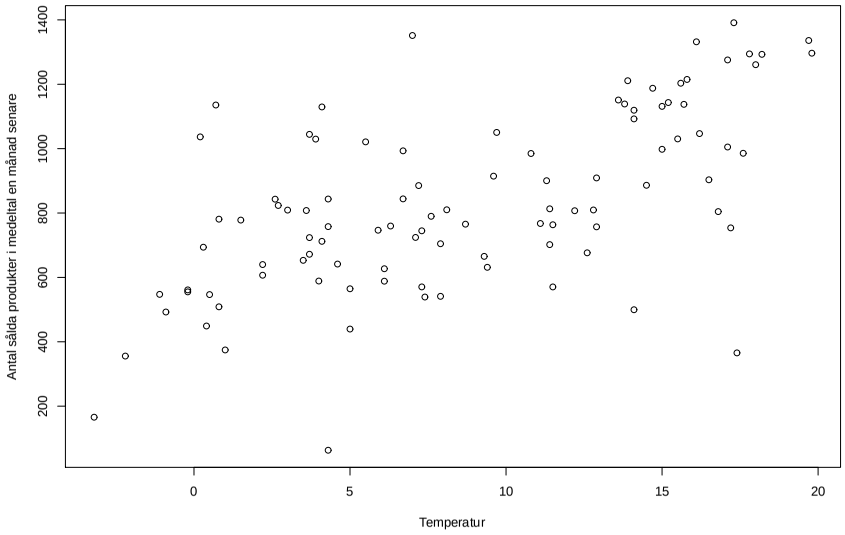


Figur 2: *Månadvisa medelvärden för hela datamaterialet.*

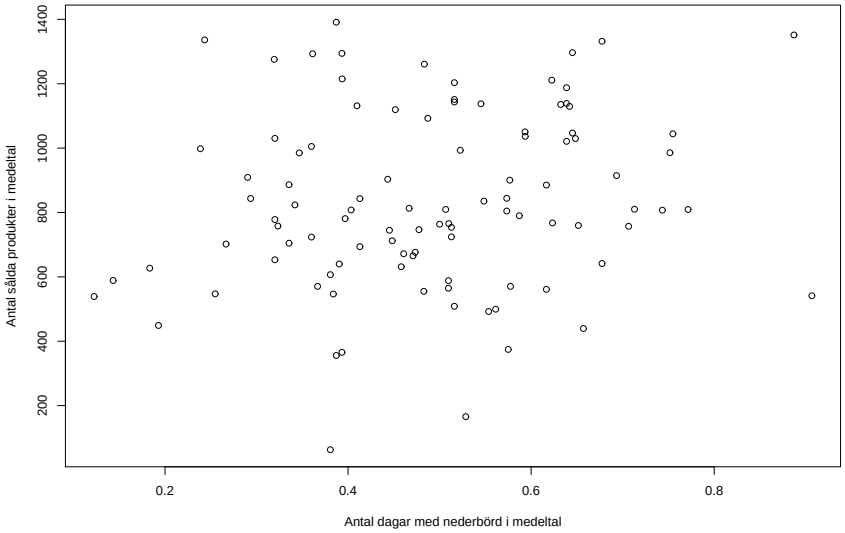
Med tanke på att mycket som styr spridningen av löss handlar om människors beteende har det varit intressant att undersöka en eventuell fördröjning. Ponera att en värmebölja påverkar spridningen av löss positivt, men det finns en fördröjning i upptäckten och därmed också när avlusningsprodukterna köps. Med de data som tillgås här finns ej möjlighet att urskilja samband förskjutna mindre än en månad. I Figur 3a ses sambandet mellan samma månads temperatur och försäljningssiffra, i Figur 3b sambandet mellan försäljningssiffra och föregående månads temperatur. Man kan ana en lite mer linjär trend i 3b. Det gäller även att den beräknade korrelationskoefficienten är högre för en månads förskjutning av temperaturen ($\rho = 0.32$) än för samma månads temperatur ($\rho = 0.44$).

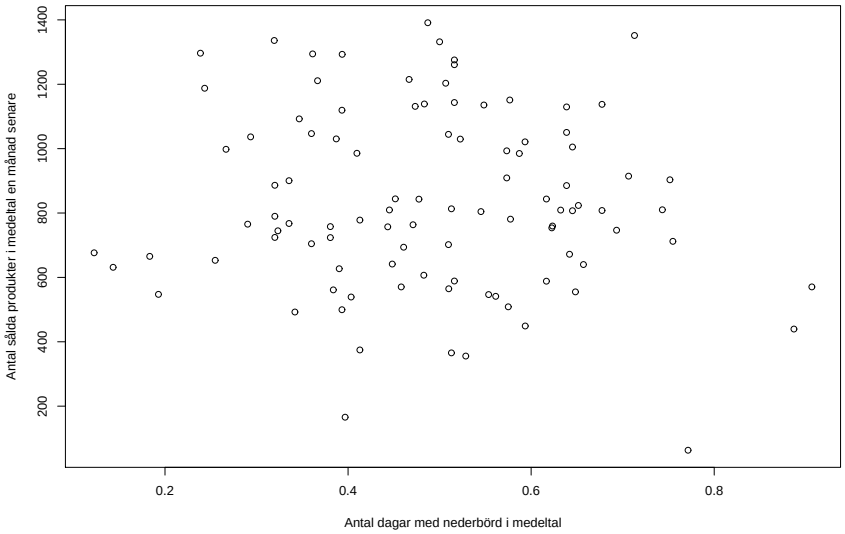
I Figur 4 så ses liknande grafer fast denna gång gällande sambandet mellan regn och försäljning, i a) kommer siffrorna från samma månad, i b) är det försäljning mot föregående månads regn. Här är det svårt att se någon form av samband över huvudtaget, varken samma månad eller med en månads förskjutning. De beräknade korrelationskoefficienterna för dessa samband ger att för samma månads nederbörd så gäller att $\rho = 0.14$ och för föregående månads nederbörd så gäller att $\rho = -0.02$.



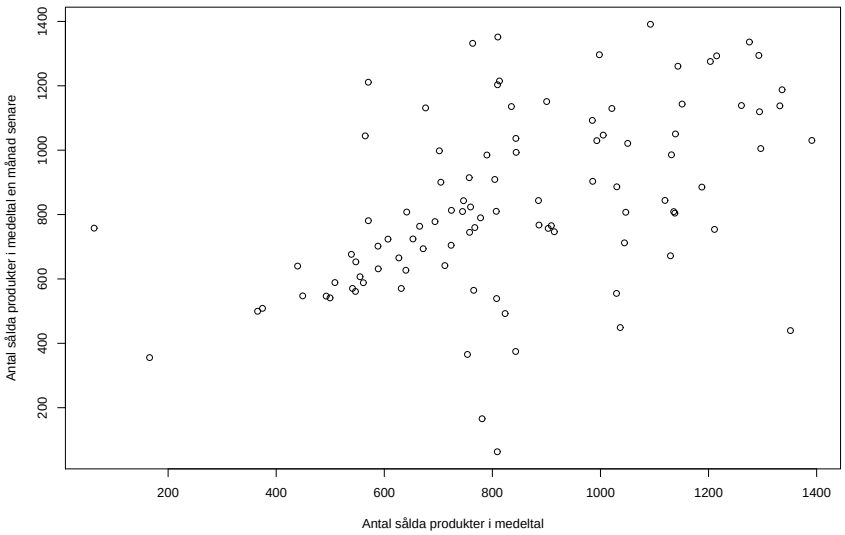


Figur 3: *a) försäljning mot samma månads temperatur, b) försäljning mot föregående månads temperatur.*





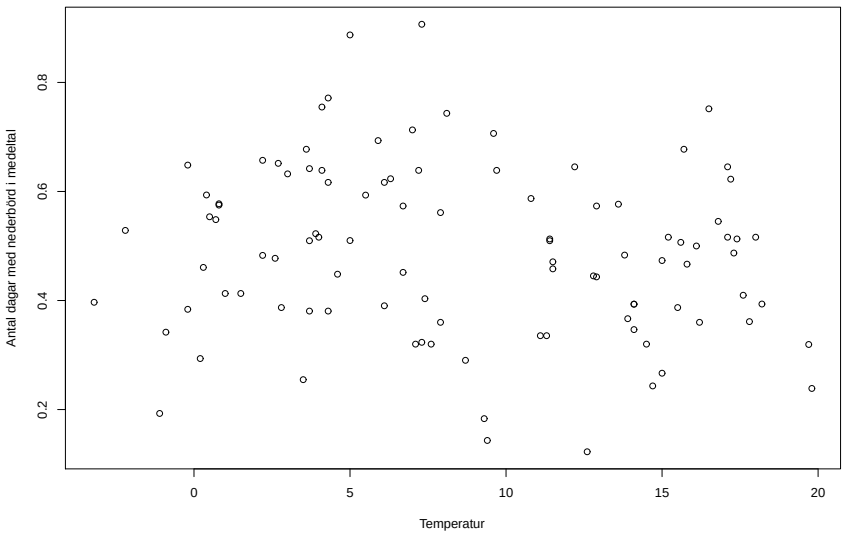
Figur 4: *a) försäljning mot samma månads nederbörd, b) försäljning mot föregående månads nederbörd.*



Figur 5: *Samband mellan försäljningsiffran en månad och nästa månads säljsiffra.*

Precis som nämnts tidigare så spelar människors sociala beteende en stor roll i detta. Ett vanligt scenario är att när en skola eller en förskola drabbas av ett lusutbrott så eskalerar det ganska fort. När lössen efterhand upptäcks så börjar informationen spridas och ännu fler upptäcker att deras barn drabbats. Med detta som anledning undersöks om det finns samband mellan antal sålda lusprodukter och föregående månads säljsiffra. I Figur 5 så ses ett linjärt samband vid lägre säljsiffror förutom den allra lägsta siffran (redan nämnda februari 2010), samtidigt som det vid höga värden på antalet sålda produkter är högst oklart om det finns något samband. Korrelationskoefficienten är här beräknad till 0.38.

Vad som också är värt att undersöka är om de två förklarande variablerna temperatur och regn är korrelerade. I Figur 6 så ses en graf av regn mot temperatur. På månadsbasis, som ju är vad data är givet i, så kan man ej se några samband de två variablerna emellan, korrelationskoefficienten $\rho = -0.08$.



Figur 6: *Undersökning av eventuell korrelation mellan temperatur och regn.*

5.2 Modellkonstruktion

Avsikten är att konstruera en modell som anses beskriva datamaterialet på ett tillförlitligt sätt. Genomgående har stegvis regression använts, se Avsnitt 3.3, med signifikansnivån 5%. Förutom att multipel linjär regression utförts, d.v.s. regression under antagande om att data kommer från en normalfördelning, så har även loglinjär regression utförts. Det grundat på tidigare resonemang om multiplikativa effekter och att responsen ej kan vara negativ enligt data. Den loglinjära regressionen innebar att data antogs komma från en Poissonfördelning för att sedan ersättas av en negativ binomialfördelning på grund av överspridning. För att finna den enligt våra mått mätt bästa modellen har några olika varianter av modeller testats.

Innan själva konstruerandet kunde sätta igång på riktigt ansågs det viktigt att ta hänsyn till säsongsvariationen, därför skapades ett antal förklarande variabler för det ändamålet.

5.2.1 Införande av säsongsvariabler

Förändringar under året är något som tros spela en stor roll, d.v.s. det finns anledning att ha någon eller några variabler som förklarar vilken tid på året det är. Ett alternativ var att använda en dummy-variabel för tidpunkter på

året då det är skollov. En 1/0-variabel har alltså definierats som

$$lov = \begin{cases} 1 & \text{om månad = jan, juni, juli, aug, dec} \\ 0 & \text{annars} \end{cases}$$

Ett andra alternativ har varit att ta hänsyn till årstiderna. En variabel *årstid* skapades som nedan

$$årstid = \begin{cases} 1 & \text{om vår = mars, april, maj} \\ 2 & \text{om sommar = juni, juli, augusti} \\ 3 & \text{om höst = september, oktober, november} \\ 4 & \text{om vinter = december, januari, februari} \end{cases}$$

På liknande vis skapades en variabel *månad* för att kunna ta hänsyn till varje månad

$$månad = 1, 2, ..., 12 \text{ för månad = jan, feb,...,dec}$$

Den mer långsiktiga årsvariationen, d.v.s. att försäljningen varierar från år till år vilket syns tydligast de tre sista åren, behöver också tas hänsyn till. Det skapades två variabler; *trend* och *medeltrend* som är medelvärdet av de senaste tolv månadernas totala försäljning för varje månad respektive medelvärdet av de senaste tolv månadernas medelförsäljning för varje månad.

5.2.2 Normalfördelningsantagande

Här har generaliserade linjära modeller under antagande om normalfördelning konstruerats. Väntevärdesmodellen har alltså sett ut som nedan

$$\mu = \alpha + \beta_1 x_1 + ... + \beta_n x_n$$

I Tabell 1 ses de olika modellerna som regressionen resulterade i, både samma och föregående månads väder har prövats som förklarande variabler och likaså de olika sätten att ta hänsyn till års- och säsongsvariationen. Modellerna har utformats med stegvis regression och därmed har exempelvis variabeln *månad* uteslutits ur samtliga modeller. Även nederbörden utesluts många gånger som förklarande variabel. Akaikes informationskriterium används för att kunna jämföra modellerna. För enkelhetens skull så betecknas en variabel där föregående månads värde avses med indexeringen -1.

Eftersom regn inte sett ut att ha något klart samband med responsen har fokus lagts på modeller där värdena på temperatur och regn kommer från samma månad för att underlätta tolkandet och nyttjandet av modellen. Variablerna *medeltrend* och *medelregn* innebär att det tagits medelvärde över varje månad som redovisats i Avsnitt 5.1 och ovan gällande trenden.

Modell	Förklarande variabler	AIC
1	*temp, medelregn, medeltrend, årstid*	1177
2	$temp_{-1}$, *medeltrend*	1176
3	*temp, medelregn, medeltrend*	1189
4	$temp_{-1}$, $medelregn_{-1}$, *medeltrend*	1176
5	*temp, medelregn, medeltrend, lov*	1171
6	$temp_{-1}$, *medeltrend, lov*	1157

Tabell 1: *Förteckning över de modeller som den stegvisa regressionen resulterade i under antagande om normalfördelning. AIC angivet för att kunna jämföra modellerna.*

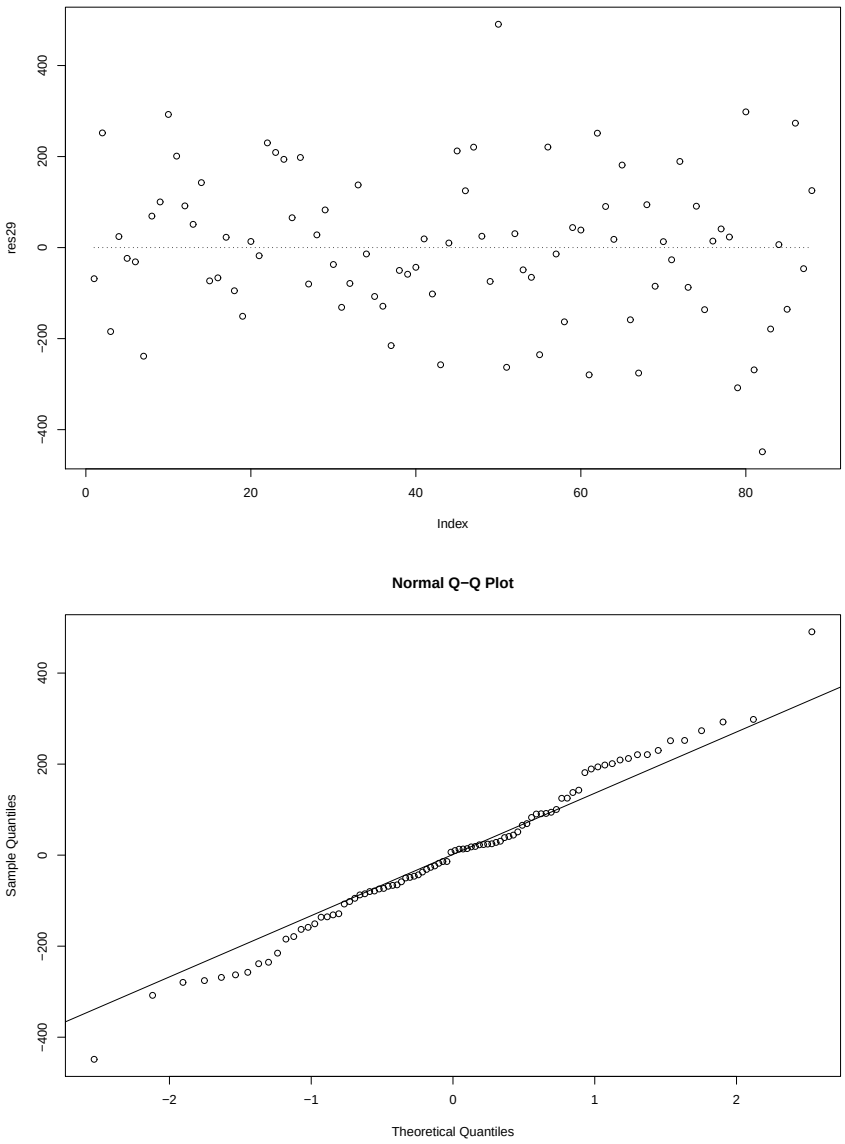
Grundat på Akaikes informationskriterium undersöks modell 6 närmare, notera att nederbörden ej haft signifikant effekt enligt denna modell. I första hand undersöks hur väl modellen verkar stämma överens med antagandena om data. I Figur 7a syns en graf över residualerna och i Figur 7b en normalfördelningsgraf för residualerna. Det är två värden som avviker från övriga. Dessa observationer identifieras till att vara januari 2007 och september 2009. Dessa punkter har högt respektive lågt värde på försäljningssiffran och har underskattats respektive överskattats i modellen. Föregående månads temperatur var förhållandevis varm för januari 2007 men temperaturen föregående månad för september 2009 var inte anmärkningsvärd på något vis.

Det undersöks om det finns några samband mellan residualerna och de olika förklarande variablerna. I Appendix, Figur 9 till 11, ses grafer med residualerna mot temperatur, regn och försäljningssiffra i medeltal. Det som utläses är att residualerna har ett linjärt positivt samband med försäljningssiffra samma månad. Det ser ut som att vid höga värden på antalet sålda lusprodukter så tenderar modellen att underskatta och vid låga värden så tenderar modellen att överskatta. Då man tittar på Figur 10a så finns antydningar till samband mellan nederbörden samma månad och residualerna. Det är ej helt självklart men skulle innebära att modellen tenderar att underskatta då det regnar mycket och överskatta då det regnar lite.

Hur bra är då denna modell på att prediktera? Det undersöks med hjälp av korsvalidering och *leave-one-out* metoden. Måttet RMSEP blir för denna modell, enl. Avsnitt 3.6, som nedan

$$RMSEP_6 = 172$$

vilket då är ett medelvärde på prediktionsfelet i kvadrat.



Figur 7: *Undersökning av residualerna för modell 6; a) residualerna i tidsföljd, b) en normalfördelningsgraf.*

5.2.3 Poissonfördelningsantagande

Eftersom det är rimligt att tänka sig att det finns multiplikativa effekter så har även generaliserade linjära modeller precis som ovan skapats fast denna gång under antagande om att data kommer från en Poissonfördelning och

med en *log*-länk. Väntevärdesfunktionen för dessa modeller ser ut som nedan

$$\mu = \exp(\alpha + \beta_1 x_1 + ... + \beta_n x_n)$$

Under detta antagande har dock inte medelvärden över månad för säljsiffrorna använts eftersom modellen kräver heltal som respons då Poissonfördelningen är en diskret fördelning. Det är mer korrekt att låta försäljningssiffrorna komma från månader som skiljer en dag i längd än att avrunda de beräknade medelvärdena. Dock finns inga hinder till att fortsätta använda de beräknade medelvärdena över nederbördsmängden för varje månad.

Dessa modeller skilde sig markant från de tidigare linjära modellerna. Det på så vis att de signifikanta variablerna var olika och signifikansnivåerna stämde inte överens. Misstanke om att det rörde sig om ett fall av överspridning väcktes. Det här bekräftades också av att de månadsvisa medelvärdena var avsevärt lägre än varianserna för respektive månad vilket ses i Tabell 2. D.v.s. en Poissonfördelning där variansen är densamma som väntevärdet är ej en bra fördelning för att förklara dessa data.

Månad	*Stickprovsmedelvärde*	*Stickprovsvarians*
Januari	31602	31625274
Februari	14669	30940685
Mars	15856	36469157
April	20615	7646035
Maj	20877	6773534
Juni	23326	15357970
Juli	36169	9390242
Augusti	35947	41080934
September	29998	92391849
Oktober	29235	51559805
November	24658	19346353
December	24609	24220670

Tabell 2: *Månadsvisa medelvärden och varianser för de observerade punkterna.*

5.2.4 Antagande om negativ binomialfördelning

För att ta hänsyn till den ovan konstaterade överspridningen användes den negativa binomialfördelningen som innehåller två parametrar och därmed tar mer hänsyn till variabiliteten i data. Väntevärdesfunktionen är densamma som för Poissonfördelningen då vi även här använt en *log*-länk. Spridningsparametern k^{-1} skattas i modellerna och värden på väntevärdenas varianser kan beräknas. I Tabell 3 ses modellerna som den stegvisa regressionen

Modell	Förklarande variabler	AIC
7	*temp, medelregn, trend, årstid*	1795
8	$temp_{-1}$, *trend*	1792
9	*temp, medelregn, trend*	1804
10	$temp_{-1}$, $medelregn_{-1}$, *trend*	1793
11	*temp, medelregn, trend, lov*	1785
12	$temp_{-1}$, *trend, lov*	1770

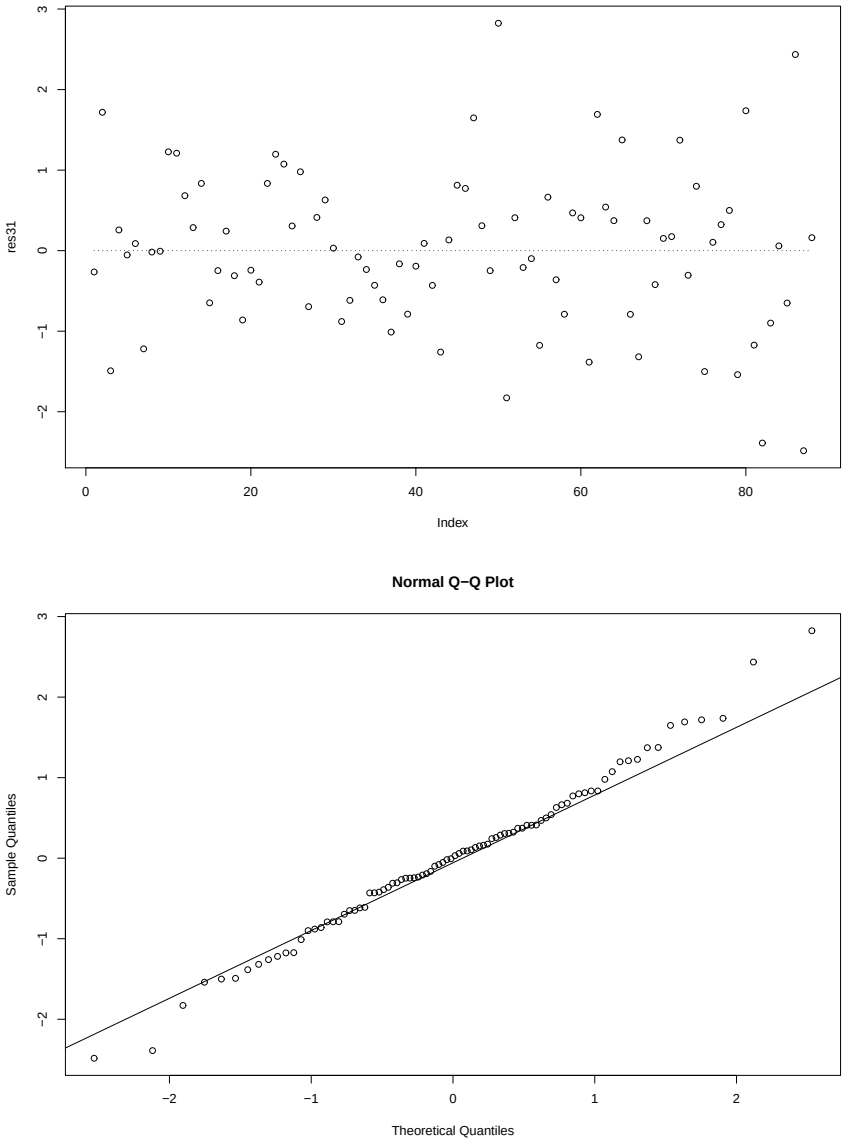
Tabell 3: *Förteckning över de modeller som den stegvisa regression resulterade i under antagande om negativ binomialfördelning. AIC angivet för att kunna jämföra modellerna.*

resulterade i och värde på AIC för respektive modell. Även här betecknas de variabler som avser värdet föregående månad med index -1. Precis som vid antagande om normalfördelning resulterade den stegvisa regressionen i att variabeln *månad* aldrig hade signfikant effekt.

Med Tabell 3 som grund valdes modell 12, som har klart lägre värde på AIC än övriga i tabellen, för närmare granskning. De beräknade Pearsonresidualerna ses i Figur 8, inte heller i denna modell har nederbörden haft signifikant effekt.

I residualerna för modell 12, Figur 8, så ses några avvikande värden; närmare bestämt är det fyra stycken som utmärker sig. De identifieras till januari 2007, september 2009, januari och februari 2010. De två första är även de som avviker i modell 6. För både januari och februari 2010 gäller att det observerade värdet är väldigt lågt jämförelsevis och i modellen har det överskattats. Temperaturerna föregående månader för dessa två värden var även de förhållandevis låga.

Det undersöks också om det finns något systematiskt samband mellan någon variabel och residualerna. Grafer över detta ses i Appendix, Figur 12 till 14. Man ser att för försäljningen samma månad så finns ett linjärt samband med residualerna. Precis som för modell 6 ser det ut som att vid höga värden på antalet sålda lusprodukter så tenderar modellen att underskatta och vid låga värden så tenderar modellen att överskatta. Även för denna modell ses en antydan till ett svagt positivt samband med nederbördsmängden samma månad.



Figur 8: *Undersökning av Pearsonresidualerna för modell 12; a) residualerna i tidsföljd, b) en normalfördelningsgraf.*

För att undersöka hur bra denna modell är på att prediktera så beräknades RMSEP. Då denna modell har responsen total försäljning per månad så divideras skillnaden mellan observerat och predikterat med antal dagar i månaden så att måttet blir jämförbart med det för modell 6 (jfr. ekvation 3.2, Avsnitt 3.6).

$$RMSEP_{12} = \sqrt{\frac{1}{N}\sum_{1}^{N}\left(\frac{y_i - \hat{\mu}_{i,-i}}{antal\ dagar\ i\ månaden_i}\right)^2} = 181$$

5.3 Resultat

Normalfördelning och negativ binomialfördelning anses beskriva data på ett tillförlitligt sätt. Enligt graferna med residualerna för modell 6 och modell 12, se Figur 7 och 8, så är båda residualerna någorlunda normalfördelade med undantag för några avvikande värden.

Poissonfördelningen som också tagits upp var inte en passande fördelning för datamaterialet som innehöll mycket variabilitet.

Resultatet av de två regressionerna för modell 6 och 12, som anses vara de bästa modellerna, ses i Tabell 4. I tabellen redovisas de signifikanta variablerna och även AIC och MSEP. Ytterligare ett jämförelsemått har beräknats, nämligen residualkvadratsumman, RSS, som är residualerna för modellen i kvadrat och sedan summerade. För modell 12 har varje residual dividerats med antalet dagar i respektive månad för att bli jämförbart med modell 6. Ett så lågt värde som möjligt är önskvärt.
Notera att man ej kan jämföra AIC dessa modeller emellan eftersom de grundar sig på antagande om olika fördelningar (jfr. Avsnitt 3.5). Inte heller skattningarna av koefficienterna är direkt jämförbara eftersom modell 6 är additiv och modell 12 multiplikativ.

Då modell 6 har lägre värde på både RSS och på prediktionsmåttet är det den modellen som väljs att representera data. Modellen ser ut som nedan där index -1 avser föregående månads värde

$$y = -422.98 + 22.48 \cdot temp_{-1} + 1.08 \cdot medeltrend + 179.94 \cdot lov$$

Modell 6

Variabel	Skattning	Std($\hat{\beta}$)
α	-422.98	195.94
$temp_{-1}$	22.48	3.08
medeltrend	1.08	0.21
lov	179.94	37.52

AIC	**RSS**	**RMSEP**
1157	2312402	172

Modell 12

Variabel	Skattning	Std($\hat{\beta}$)
α	8.31	0.26
$temp_{-1}$	0.03	$4.09 \cdot 10^{-3}$
trend	$5.10 \cdot 10^{-5}$	$9.01 \cdot 10^{-6}$
lov	0.26	0.05

AIC	**RSS**	**RMSEP**
1770	2587861	181

Tabell 4: *De två modellerna 6 och 12.*

För ändamålet prediktion är det möjligt att använda måttet RMSEP för att ge oss en uppfattning om storleksordningen på felen i prediktionerna. Värdet representerar en grov skattning av prediktionsfelen för antalet sålda lusprodukter per dag (eftersom det beräknats medelvärde på totala månadssiffran). Värdet är allt från 17% till 34% av antalet sålda lusartiklar i medeltal beroende på vilken månad man avser. För månader med högre säljsiffror är en felprediktion i den här storleken inte lika allvarlig som att erhålla detta prediktionsfel för en månad med låga försäljningsvärden.

6 Slutsatser

Slutsatsen är att temperaturen har ett positivt signifikant samband med försäljningen av lusmedel. Försäljningen påverkas som mest av temperaturen den föregående månaden. Det går inte att säga mer exakt när effekten är som störst men det har visat sig genom arbetet att föregående månad är mer signifikant än samma månad som försäljningen.

Förutom temperaturen har de månader då skollov infaller positiv signifikant effekt, d.v.s. försäljningen går upp under dessa månader. Detta sammanfaller delvis med temperaturen på så vis att de månader då temperaturen föregående månad är hög såsom juli, augusti är det också skollov. Den mer övergripande årsvariationen visade sig också vara av betydelse, trenden blev klart signifikant.

Mängden nederbörd har inte visat sig ha någon signifikant effekt på antalet sålda lusprodukter. Däremot ses antydan till ett svagt samband mellan residualerna och nederbörden samma månad. Det kan alltså finnas en korrelation däremellan men det är för svagt för att bli signifikant i modellen. Sambandet ser ut som att vid lite nederbörd tenderar modellen att överskatta värdena och vid mycket regn tenderar modellen att underskatta försäljningssiffrorna. Det här sambandet skulle alltså tyda på att försäljningen ökar de månader då det regnar mycket men det är ej en signifikant effekt eftersom temperaturen föregående månad beskriver data så pass mycket bättre.

Enligt grafen över samband mellan residualerna och försäljningssiffror (Figur 11) så innehåller modellen inte lika mycket variation som data. Man ser att vid höga värden på försäljningssiffrorna så tenderar residualerna att bli höga och vid låga värden på försäljningssiffrorna så tenderar residualerna att bli låga. Modellen klarar alltså inte riktigt av att skatta så pass höga eller låga värden som observerats.

Det är framförallt två punkter som tenderat att få höga residualer. De är januari 2007 som underskattas och september 2009 som överskattas. För dessa två punkter kan man se att den observerade säljsiffran är avsevärt högre respektive lägre än medelvärdet för vardera månad. Om det är något speciellt som orsakar felskattningen är dock svårt att säga eftersom januari som underskattas har ett värde på temperaturen föregående månad som är ovanligt högt och för september 2009 kan man ej notera något anmärkningsvärt kring vädret föregående månad.

7 Diskussion

Statistiska modeller är ett användbart redskap för att analysera samband mellan olika variabler. Man skall dock komma ihåg att en signifikant variabel inte nödvändigtvis betyder att det är den variabeln som är själva orsaken.

Den här uppsatsen har just handlat om att skapa en modell med förklarande variabler som vi vet egentligen inte är direkta orsaken till förändringar i responsvariabeln. Genom forskning vet man att löss sprids genom sociala kontakter och att spridningen inte direkt beror på vädret utan det är hur barnen agerar med varandra. Barnens sociala beteende styrs i sin tur till viss del av vädret och årstid. Syftet med det här arbetet var just att analysera eventuella samband med vädret.

Det finns alltså ändock ett värde i att skapa en statistisk modell som inom datasetets ramar beskriver samband variablerna emellan. Sådana modeller kan vara till nytta som i detta fall t.ex. för apotek som ska köpa in lusprodukter för vidare försäljning. Det är också ett steg i att utvidga kunskaperna om huvudlusen som länge varit ett ganska outforskat djur.

Den slutgiltiga modellen visar på ett positivt linjärt samband mellan temperaturen och antalet sålda lusprodukter. Rent intuitivt trodde vi från början att försäljningen skulle öka då temperaturen var ganska låg, säg $0-7^oC$, och sjunka vid kallare eller varmare temperaturer. Det eftersom vid kallt väder går man gärna ut mer och leker och vid varmt och fint väder går man också ut. Däremot då det är plusgrader men inte speciellt varmt är det ofta grått och blött och barnen stannar inomhus.
Eftersom det visade sig att temperaturen istället har en positiv linjär effekt föranleds vi att tro att temperaturen påverkar spridningen av löss mer än vad man hittills har trott, möjligtvis genom en ökad aktivitet hos lössen vid högre temperaturer.

Några av de månader med höga värden på temperaturen föregående månad sammanfaller som sagt med månader som är definierade som skollov. Förutom sommarmånaderna har även december och januari definierats som månader med skollov. Alltså beror januari månads i snitt höga försäljning rimligtvis mer på att den definierats som skollov än temperaturen föregående månad. September som har höga försäljningssiffror men ej är definierad som skollov förklaras då istället av att föregående månads temperatur är hög.

Vid utvidgning av denna analys hade man gärna tittat på fler möjliga förklarande variabler. Skulle t.ex. luftfuktigheten kunna ha betydelse, eller om det vore möjligt att definiera nederbörden på annat vis. I nuläget vet man ju ej hur mycket det regnat eller om det är nattetid eller dagtid.

Eftersom det finns antydningar till samband mellan antal dagar med nederbörd och försäljningssiffrorna samma månad skulle det vara intressant att undersöka det vidare, definierat på annat vis kanske regnet skulle ha betydelse.
Ytterligare en aspekt är antal barn på förskola/dagis. Det är inte helt omrimligt att tänka sig att en sådan variabel skulle få signifikant effekt.

En del av syftet med analysen var att kunna använda modellen för prediktion. Felmarginalerna för modellen är tyvärr ganska stora och prediktion är alltså inte helt tillförlitligt. Den som håller det i åtanke kan ändock använda modellen för att skapa sig en ungefärlig uppfattning om framtiden. Detta förutsatt att man håller variablerna inom ramarna för den här modellen.

Referenser

Agresti, Alan: *Categorical data analysis.* Second edition, 2002, John Wiley & Sons, Inc., Hoboken, New Jersey.

Blom, Gunnar & Holmquist, Björn: *Statistikteori med tillämpningar.* Tredje upplagan, 1998, Studentlitteratur

Lindgren, Bernard W: *Statistical theory.* Fourth edition, 1993, Chapman & Hall/CRC

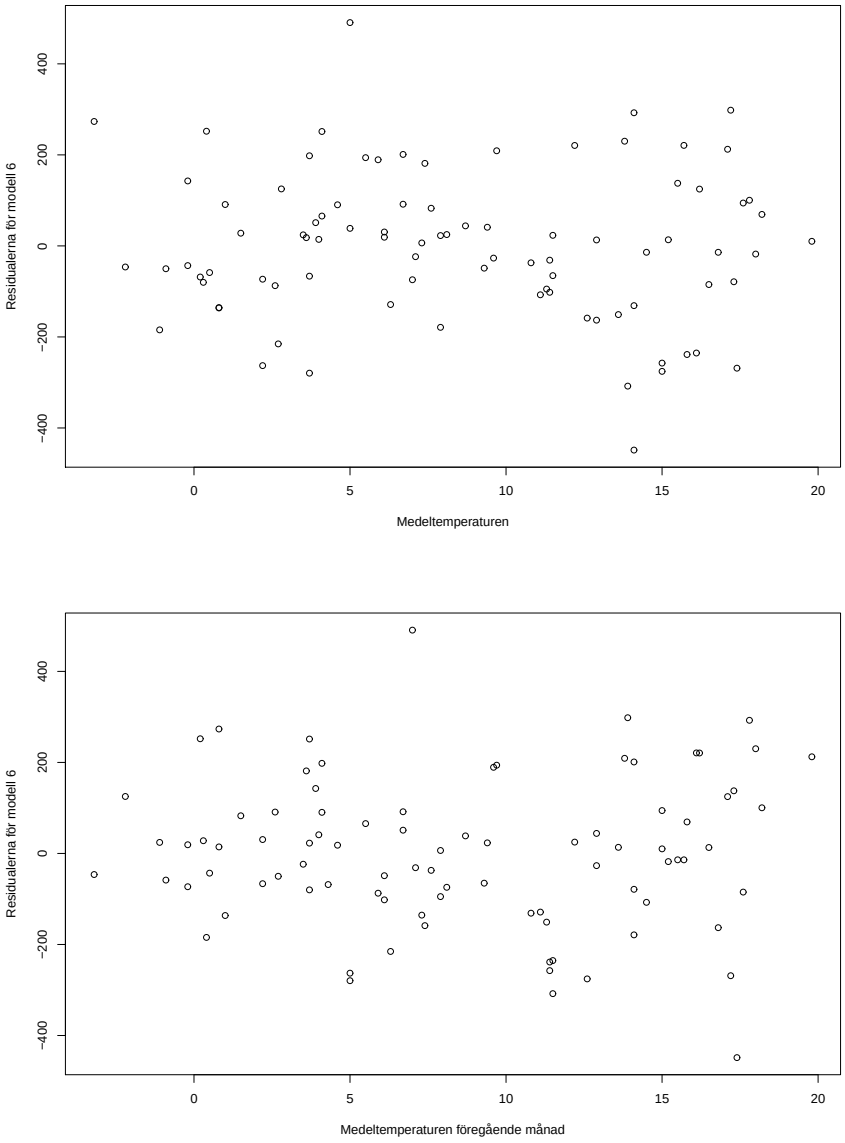
Sundberg, Rolf: *Kompendium i tillämpad matematisk statistik.* Reviderad version från hösten 2009.

Danmarks Meteorologiske Institut

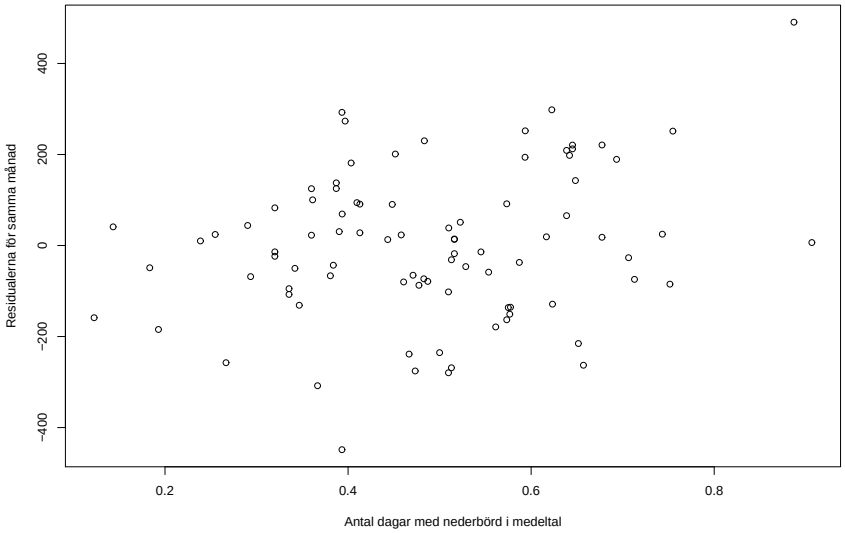
Sjukvårdsrådgivningen, 2010-05-12
http://www.sjukvardsradgivning.se/artikel.asp?CategoryID=27607

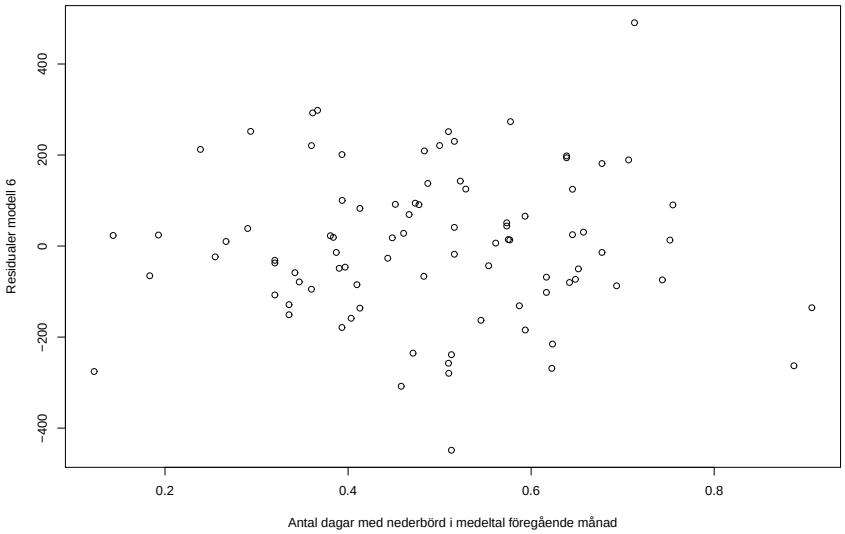
Lusguiden, 2010-05-12
http://www.lusguiden.se/om-loss/om-loss/

Appendix

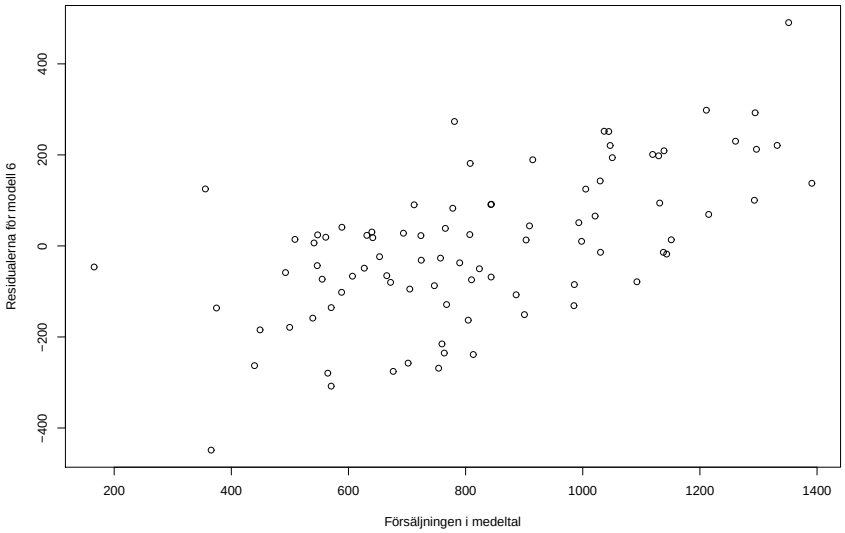


Figur 9: *Undersökning av huruvida residualerna för modell 6 har något samband med temperaturen (a) samma eller (b) föregående månad.*

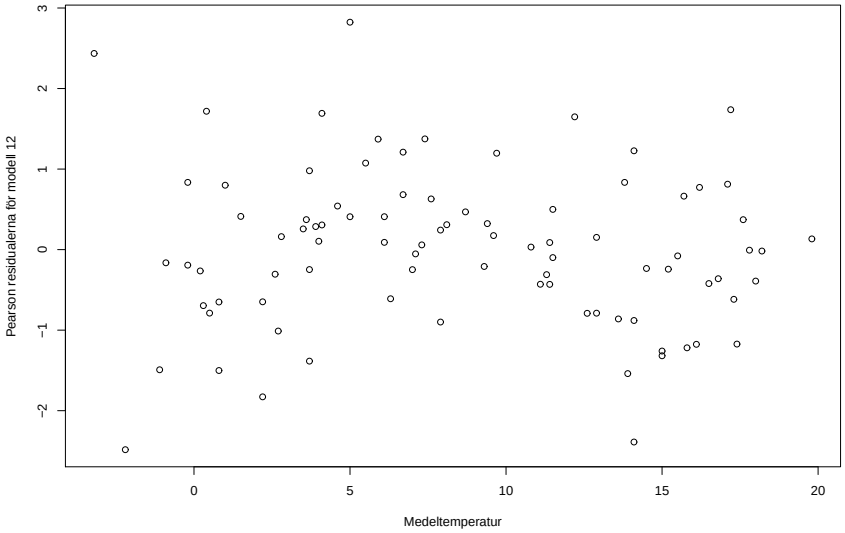


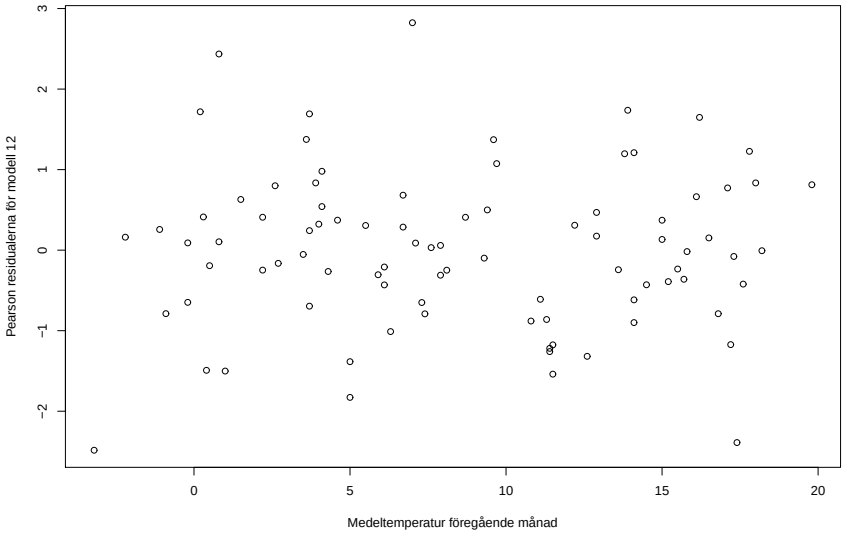


Figur 10: *Undersökning av huruvida residualerna för modell 6 har något samband med nederbörden (a) samma eller (b) föregående månad.*

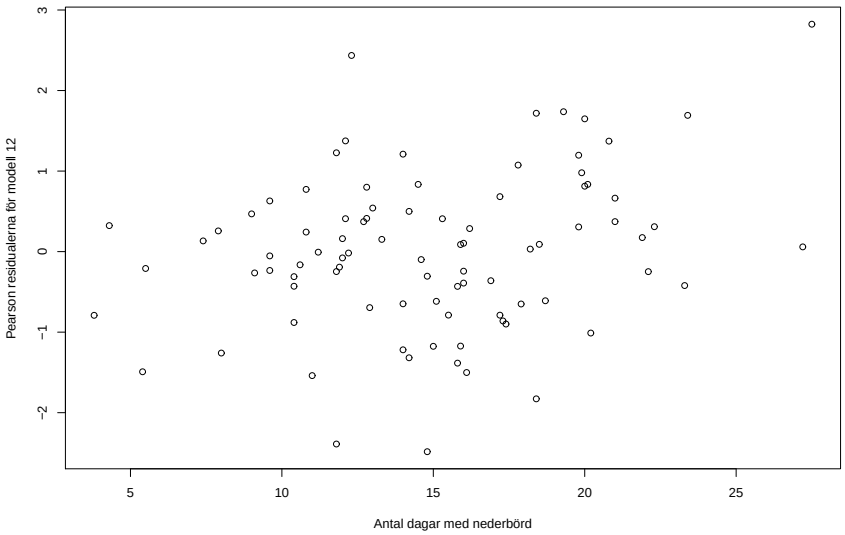


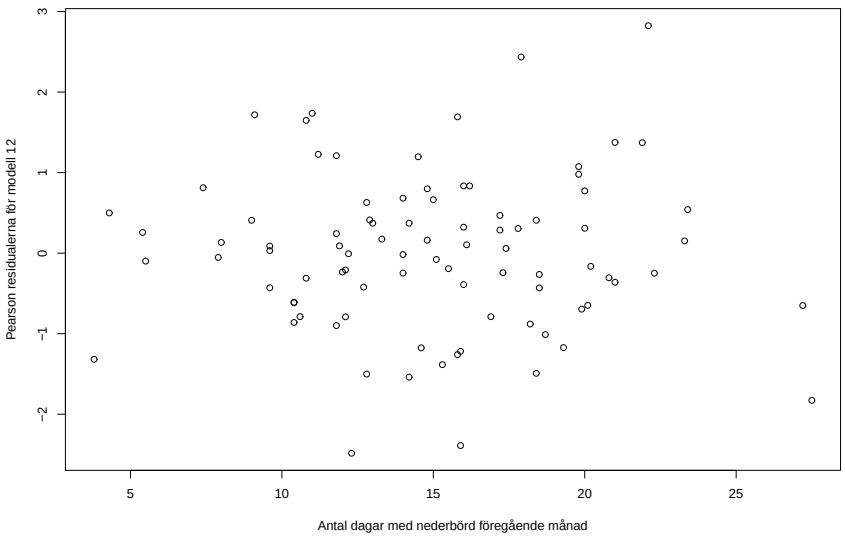
Figur 11: *Undersökning av huruvida residualerna för modell 6 har något samband med försäljningen samma månad.*



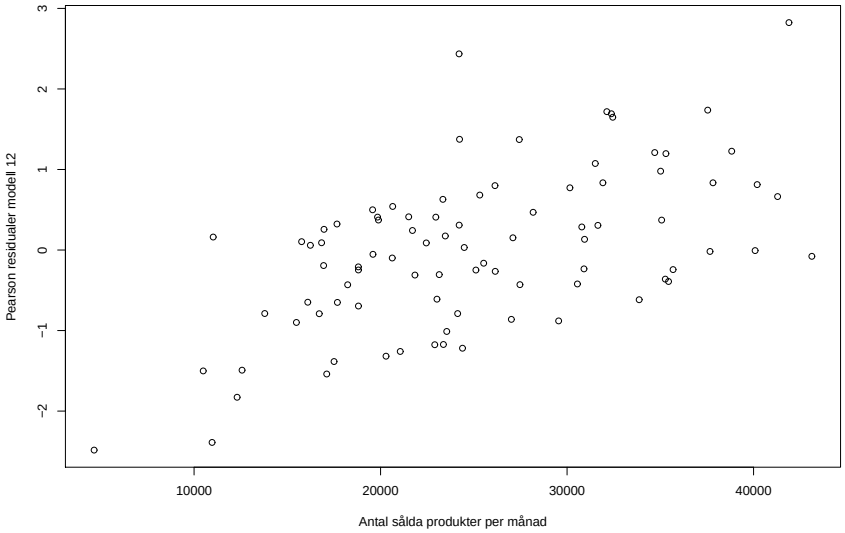


Figur 12: *Undersökning av huruvida residualerna för modell 12 har något samband med temperaturen (a) samma eller (b) föregående månad.*





Figur 13: *Undersökning av huruvida residualerna för modell 12 har något samband med nederbörden samma (a) eller (b) föregående månad.*



Figur 14: *Undersökning av huruvida residualerna för modell 12 har något samband med säljsiffrorna samma månad.*