LUNDS UNIVERSITET
Statistiska Institutionen

Sambanden mellan inandningsbara och fina partiklar i luften och insjuknande i stroke – *poissonregressionsmodeller*

Jenny Hillström & Joselyne Nsabimana

Uppsats i Statistik
15 högskolepoäng
Nivå 91-120 högskolepoäng
April 2009

Handledare: Mats Hagnell

ETT STORT TACK TILL

Kristina Jakobsson, Anna Oudin och Emelie Stroh på Arbets- och miljömedicin vid Universitetssjukhuset i Lund och Susanna Gustafsson på Miljöförvaltningen i Malmö.

ABSTRACT

The purpose of this Master's thesis is to model the short-term association between the number of strokes recorded per day and an increased daily mean level of particulate matter, PM. A short-term association means that the occurrence of stroke is related to the content of PM in the air the same day and up to seven days prior to the stroke (lag 0 to lag 7). Previous studies show that there is a significant association between PM and the incidence of stroke. Poisson regression models are used to explain the relationship in this study.

Three different types of data from Scania are used: daily records of stroke occurrences, daily mean levels of $PM_{2.5}$ and PM_{10} and the daily mean temperature. First Scania is divided into eight zones and two zones are chosen for further analyses. That is zone 3, which includes Helsingborg, Höganäs and Landskrona and zone 6, which includes Malmö. According to previous studies, only patients over 65 years old are included and the data is divided into cold and warm seasons. The cold season ranges from October to April and the warm season from May to September. This study is delimited to study the association at lag 0, 1, 5, 7 according to previous studies results and furthermore the correlations between the lags that follow each other are very high for the temperature.

Two main procedures, Proc GENMOD and Proc GAM in SAS software are used to estimate the models. For zone 3, one model for each season which includes PM_{10} and temperature are estimated. While for zone 6, three models for each season are estimated. The first two include each of the particles one by one and the temperature and the third includes all of the three independent variables. Proc GENMOD is first used to estimate the models, but the residuals show that the models are inadequate. Therefore further analyses are done in Proc GAM to discover nonlinear terms that need to be included in Proc GENMOD in order to improve the models. A scatter plot smoother in Proc GAM is also used to help visualise the nonlinear relationship.

To sum, a significant positive relationship is only shown during the cold season 2004-2005 in zone 6. The final estimated models in Proc GENMOD show significant relationship at lag 0, lag 1 and a quadratic term at lag 1 for both $PM_{2.5}$ and PM_{10}. Lag 0 for $PM_{2.5}$ and PM_{10} show a negative relationship to stroke. Lag 1 gives a much greater impact than lag 0, therefore the summed impact on stroke is positive. An increase with 1 $\mu g/m^3$ for $PM_{2.5}$ at lag 0, lag 1 and $(\text{lag } 1)^2$ implies an estimated increase in stroke occurrence with 33.9 %. Respective value for PM_{10} is an increase with 12.3 %. However the values estimated in this study cannot literally be compared to the ones estimated in previous studies. Since fewer independent variables and all cases of ischemic stroke are investigated, instead of fatal stroke in previous studies. A model with $PM_{2.5}$, PM_{10} and temperature shows that only $PM_{2.5}$ is significant and it gives the same model as the model with $PM_{2.5}$ and temperature.

Results from residual analyses show that the residuals are not normally distributed and that the variance is not constant and therefore are the estimated models inadequate. This indicates that the models estimated in this study cannot be used to estimate future stroke occurrences.

INNEHÅLLSFÖRTECKNING

1. INLEDNING

Den här magisteruppsatsen behandlar sambandet mellan antalet personer som insjuknar i stroke och mängden partiklar i luften och temperaturen i Skåne. Flera tidigare studier har påvisat samband mellan höga partikelhalter och insjuknande av stroke.

Partiklar delas in i olika grupper beroende på deras storlek. I denna studie kommer data över de inandningsbara och de fina partiklarna att användas. De inandningsbara partiklarna, PM_{10}, har en diameter mindre än 10 μm och de fina, $PM_{2.5}$, har en diameter mindre än 2,5 μm. En stor del av de inandningsbara partiklarna kommer från trafiken i form av damm som virvlar upp i samband med slitage av vägbanan, däck, bromsningar etc. Mindre partiklar härstammar ofta från avgaser från bl.a. trafik, energiproduktion och vedeldning. (Länsstyrelsen i Stockholms län 2003) Desto mindre partiklarna är, desto skadligare anses de vara. Större partiklar fastnar i slemhinnorna, medan partiklar under 10 μm i diameter tränger ner i luftvägarna. De fina partiklarna kan ta sig ända in i alveolerna. (SAD 2009)

Stroke innebär att skador på hjärnan uppkommer i samband med blodpropp eller en blödning i hjärnan. Det vanligaste fallet är att en blodpropp täpper till blodcirkulationen i en del av hjärnan, denna typ av stroke kallas ischemisk stroke. Varianten med hjärnblödning kallas också hemorrhagisk stroke och beror antingen på en blödning från blodkärl inne i hjärnan eller på hjärnans yta. (Sjukvårdsrådgivning 2008)

Sambanden mellan partiklar och stroke har också tidigare studerats i vår kandidatuppsats (Hillström et al. 2008) och anledningen till att vi fortsätter med samma ämne är för att se om resultaten skiljer sig åt vid användandet av olika metoder. I kandidatuppsatsen användes transferfunktionsmodeller för att illustrera sambanden och denna gång ämnar vi använda oss av poissonregressionsmodeller. I föregående uppsats konstaterades att strokedata tycks följa en poissonfördelning, därav grunden till byte av metod.

Precis som kandidatuppsatsen skrivs även magisteruppsatsen på uppdrag av Arbets- och miljömedicin vid Universitetssjukhuset i Lund. I denna uppsats används endast partiklar och temperatur som oberoende variabler. Andra variabler som vindhastighet, luftfuktighet, influensaperioder och så vidare anses vara mer relevanta när endast dödlig stroke studeras. I denna uppsats kommer endast ischemisk stroke att studeras. Genom att inte ta med så många variabler i regressionsmodellerna blir det enklare att urskilja varje variabels enskilda effekt på den beroende variabeln.

1.1 Syfte

Syftet med denna uppsats är att med hjälp av metoder för poissonfördelad data ta fram regressionsmodeller som beskriver det kortsiktiga sambandet mellan partiklar i luften och temperaturen och antalet insjuknande i stroke per dygn. Med det kortsiktiga sambandet menas hur antalet insjuknande i stroke en given dag påverkas av partikelhalten i luften och temperaturen på den dagen och upp till sju dagar innan. I fortsättningen kommer den aktuella dagen betecknas lag 0, dagen innan lag 1 osv. upp till sju dagar innan som benämns lag 7.

1.2 Tidigare studier

Det finns flera tidigare studier där sambandet mellan stroke och partiklar analyserats. På senare tid har det blivit allt vanligare att använda sig av generaliserade additiva modeller vid beskrivning av sambandet mellan olika typer av luftföroreningar och hälsoeffekter (Ramsay et al. 2003).

I en studie från Seoul i Sydkorea har sambandet mellan dödlig stroke och olika luftföroreningar, däribland PM_{10}, studerats. Studien löper över fyra år, från 1995 till 1998. Seoul är en stad med höga partikelhalter och har också en hög dödlighet av stroke. Under studieperioden hade staden ett medelvärde av PM_{10} på 71,1 $\mu g/m^3$ och i genomsnitt drabbades 15,3 personer av dödlig stroke varje dag. Sambandet mellan stroke och luftföroreningar beskrivs av en generaliserad additiv modell för respektive luftförorening. Även veckodag, meterologiska variabler och variabler för trender och säsongsskillnader inkluderades i modellen. För att eliminera serialkorrelationen i residualerna inkluderas även autoregressiva termer i modellen. Sambandet studerades upp till fem dagar innan inträffad stroke och de olika lageffekterna bestämdes genom att jämföra de relativa riskerna för varje lagmodell med varandra. Resultatet visar att sambanden mellan PM_{10} och ozon och dödlig stroke är högst på lag 0. Då är den uppskattade ökningen 1,5 % i dödlig stroke för varje interkvartil räckviddsökning i PM_{10} och ozon. (Hong et al. 2002)

En liknande studie har också gjorts i Shanghai i Kina. Där har stroke blivit en av de vanligaste dödsorsakerna och under studieperioden 2001-2002 dog i genomsnitt 3,32 personer av stroke varje dag. Den relativa risken mellan exponering för luftföroreningar, bl.a. PM_{10}, och dödligheten av stroke beräknades med loglinjära modeller. Huvudanalysen genomfördes med en semiparametrisk generaliserad additiv modell och först anpassades icke-parametriska utjämningstermer för trender, temperatur, relativ luftfuktighet, daggpunkten och dummy-variabler för veckodag. Sedan inkluderades luftföroreningarna en åt gången i modellen. Flera modeller gjordes med olika kombinationer av luftföroreningar för att kunna urskilja de individuella effekterna av respektive luftförorening. Hänsyn togs även till de laggade effekterna av de meteorologiska variablerna och luftföroreningarna i modellerna. Resultatet av denna studie visade på signifikanta samband på lag 1. Varje ökning av 10 $\mu g/m^3$ i PM_{10} motsvarar en ökning med 0,8 % av den relativa risken att drabbas av dödlig stroke. Men i modellerna med flera luftföroreningar försvagades istället varje enskilds effekt på strokerisken när flera variabler togs med i modellen. (Kan et al. 2003)

Även i Hong Kong i Kina har olika luftföroreningars effekt på hälsan studerats. I en studie baserad på data från perioden 1995-1997 undersöktes sambandet mellan den dagliga dödligheten och fyra olika luftföroreningar, däribland PM_{10}. Den dagliga dödligheten inkluderade inte de dödsfall som skett på grund av olyckor. Vidare delades den beroende variabeln upp i tre grupper; dödsfall av personer med kardiovaskulära respektive respiratoriska sjukdomar och övriga icke-olycksfallsrelaterade dödsfall. Poissonregression användes för att åskådliggöra sambanden mellan luftföroreningarna och varje sorts dödsfall i tre modeller. I varje modell inkluderades även icke-parametriska utjämningstermer för trender på dagar, säsongsvariationer, temperatur och luftfuktighet liksom dummyvariabler för veckodagar, helgdagar och influensaperioder. Dessutom inkluderades även de

laggade effekterna av de meteorologiska variablerna i modellerna. Residualanalys genomfördes på modellerna och både överspridning och autokorrelation justerades för i S-plus. De laggade effekterna togs med upp till fem och tre dagar på ozon respektive övriga luftföroreningar. För att kunna studera varje luftförorenings enskilda effekt inkluderades utjämningsfunktionen Loess för att justera för icke-linjära effekter från de övriga luftföreningarna. För att utföra de fortsatta analyserna beräknades det förväntade antalet dödsfall för modellerna på samtliga säsonger. Sedan anpassades en poissonregression på antalet dödsfall med koncentrationen av luftföroreningarna för att få den relativa risken. Slutligen studerades också sambandet mellan exponering och påverkan grafiskt med hjälp av generaliserade additiva modeller.

När hela året studerades fick NO_2, SO_2 och PM_{10} liknande effekt på samtliga dödsorsaker. Den relativa risken ökade från lag 0 till att anta ett max på antingen lag 1 eller lag 2 för att sedan minska igen på lag 3. Den relativa risken på den bäst laggade dagen var endast signifikant för de respiratoriska sjukdomarna för PM_{10}, men när justeringar gjordes med andra föroreningar i modellen förblev inte den relativa risken signifikant. Vid säsongsvis analys hittades inga signifikanta samband på varm säsong, maj till september. Under kall säsong, oktober till april, däremot var samtliga relativa risker signifikanta på de bäst laggade dagarna för alla typer av dödsorsaker. När justeringar gjordes för samvariation med andra föroreningar blir inte relativa risken för PM_{10} längre signifikant för någon dödsorsak. Vidare kunde inget tydligt samband mellan exponering och påverkan påträffas för någon av de tre dödsorsakerna under varm säsong för PM_{10}. Däremot fanns ett positivt samband under den kalla säsongen för respiratoriska sjukdomar på koncentrationer av PM_{10} upp till 80 $\mu g/m^3$. (Wong et al. 2001)

2. DATA

Den här studien baseras på tre olika typer av datamaterial; partikeldata, strokedata och temperaturdata. Partikeldata har hämtats från IVL Svenska Miljöinstitutets hemsida (IVL 2008). Data över personer som insjuknat i stroke kommer från Riksstroke och är insamlat från sjukhusen i Skåne. Temperaturdata över Helsingborg kommer från SMHI och samma data över Malmö har tillhandahållits från Miljöförvaltningen i Malmö.

Först delas Skånes 33 kommuner in i åtta zoner, där Malmö stad utgör en egen zon. Anledningen till att Malmö stad är en egen zon är för att samma data som i kandidatuppsatsen ska kunna användas. Det gör också att resultat från de olika metoderna blir jämförbara. Zonindelningen görs geografiskt och med hänsyn till vilket sjukhus upptagningsområde respektive kommun tillhör och blir som följer, se även figur 2.1 nedan:

Zon 1
Bjuv, Båstad, Klippan, Perstorp, Åstorp, Ängelholm och Örkelljunga

Zon 2
Bromölla, Hässleholm, Kristianstad, Osby och Östra Göinge

Zon 3
Helsingborg, Höganäs och Landskrona

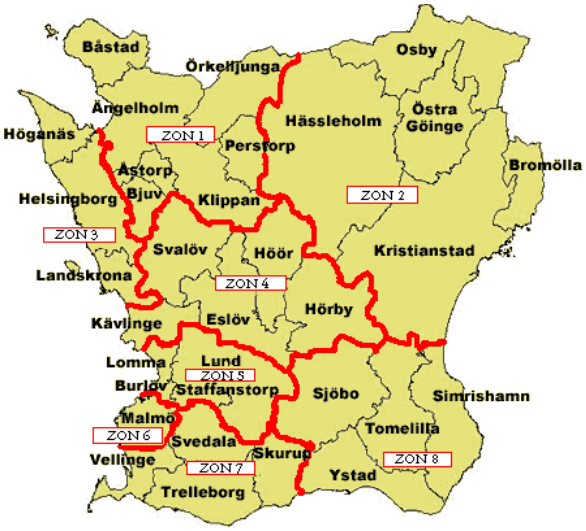
Zon 4
Eslöv, Hörby, Höör, Kävlinge och Svalöv

Zon 5
Burlöv, Lomma, Lund och Staffanstorp

Zon 6
Malmö stad

Zon 7
Skurup, Svedala, Trelleborg och Vellinge

Zon 8
Simrishamn, Sjöbo, Tomelilla och Ystad



Figur 2.1 Karta över Skåne uppdelat på åtta zoner

2.1 Partikeldata

Partikeldata kommer från en eller flera mätstationer i respektive kommun. Det finns tre typer av omgivning som mätstationerna i kommunerna i denna studie befinner sig i; bakgrund, urban bakgrund och gaturum. Bakgrund är utanför tätort medan urban bakgrund och gaturum finns inne i tätort. Mätstationer i urban bakgrund är placerade i takhöjd och där har partiklarna ett bättre flöde i luften. Stationer i gaturummet är placerade på platser där de omges av byggnader. Därav följer att av två stationer på närliggande plats uppmäter oftast den i urban bakgrund en lägre partikelhalt än den i gaturummet. I denna studie har urban bakgrund valts framför gaturum eftersom mätningarna i Malmö stad är gjorda i urban bakgrund.

Data består av uppmätta partikelhalter per timme och enheten är mikrogram per kubikmeter ($\mu g/m^3$). I zon 6 finns det sedan tidigare partikeldata över både PM_{10} och $PM_{2.5}$. I de flesta andra kommuner saknas det mätningar på $PM_{2.5}$ och där används enbart PM_{10}.

Först görs data om till dygnsmedelvärden genom att medelvärdet på timsobservationerna beräknas dag för dag. Om en dag saknar uppmätta värden på mer än 25 % av timmarna sätts hela den dagen som saknad. Därefter sammanställs data för respektive zon och det visar sig att inte alla kommuner har uppmätta värden. Zon 8 saknar partikeldata i samtliga kommuner och kan inte användas. I zon 4 finns det endast hela säsonger på Vavihill, i Svalövs kommun. Tyvärr visar det sig att data inte är tillförlitliga eftersom det har varit problem med mätinstrumenten på den stationen. Likaså har mätstationen i Kristianstad visat brist på tillförlitlighet då högre värden uppmätts på $PM_{2.5}$ än PM_{10} vilket är omöjligt eftersom PM_{10} innehåller $PM_{2.5}$. På grund av detta kommer inte heller zon 2 att användas.

När data sammanställts för respektive zon är nästa steg att skapa kompletta säsonger som kan användas i analysen. Med hjälp av partikeldata från de olika mätstationerna i respektive zon kan ett medelvärde som är representativt för hela zonen beräknas. Nedan följer en utförligare beskrivning av hur beräkningarna görs.

På samtliga säsonger och i alla zoner är det i denna studie aktuellt med data från högst två mätstationer. I de fall där partikeldata hämtas från enbart en station, ersätts de saknade dagarna med medelvärdet av värdena dagen innan och dagen efter. När det finns två mätstationer i samma zon beräknas först ett kvotvärde mellan de båda stationerna dag för dag. Därefter tas ett medelvärde på kvotvärdena. Kvotmedelvärdet används sedan för att skatta värdena på de saknade dagarna. Om det finns dagar som saknar uppmätta värden för båda mätstationerna, beräknas ett medelvärde för dessa dagar med hjälp av observationerna dagen innan och dagen efter. Slutligen tas ett medelvärde av de två mätstationerna och dessa används i analysen.

De säsonger som uppvisar komplett partikeldata redovisas i tabell 2.1 nedan och där kan också utläsas i vilka zoner det är och hur många personer som insjuknar i stroke i varje kommun under respektive säsong.

2.2 Strokedata

Våra strokedata sträcker sig från 2001 till 2006 och det innehåller strokepatienter som vårdats för ischemisk stroke på något av sjukhusen i Skåne. Data kommer från Riksstroke som samordnar ett register över strokepatienter som rapporteras från de olika sjukhusen. Registret uppskattas ha en täckningsgrad på ca 90 % i Skåne (Riksstroke 2009).

Data innehåller variabler som insjukningsdatum, födelseår, om patienten haft stroke tidigare, om patienten bor i Malmö stad och slutligen vilken kommun patienten bor i. Däremot finns inga uppgifter på om patienterna drabbats av dödlig stroke eller inte. Patienter som dör i stroke innan de hunnit fram till sjukhuset inkluderas inte i Riksstrokes register.

Data sorteras först på ålder, i den här studien inkluderas patienter från 65 år och uppåt, se avsnitt 2.6. Därefter görs data om till dygnsdata istället för att som ursprungligen baseras på insjukningsdatum. Slutligen delas data in zonvis och efter de kompletta säsonger som framkom av partikeldata.

KALL SÄSONG 2002-2003			KALL SÄSONG 2004-2005			KALL SÄSONG 2005-2006			VARM SÄSONG 2005			VARM SÄSONG 2006		
Zon	**Kommun**	**Antal**	**Zon**	**Kommun**	**Antal**	**Zon**	**Kommun**	**Antal**	**Zon**	**Kommun**	**Antal**	**Zon**	**Kommun**	**Antal**
1	Ljuv	10	**3**	Helsingborg	140	**3**	Helsingborg	144	**3**	Helsingborg	101	**3**	Helsingborg	90
	Båstad	17		Höganäs	26		Höganäs	31		Höganäs	27		Höganäs	20
	Klippan	19		Landskrona	49		Landskrona	23		Landskrona	39		Landskrona	23
	Perstorp	8		***Summa***	***215***		***Summa***	***198***		***Summa***	***167***		***Summa***	***133***
	Åstorp	15												
	Ängelholm	38	**6**	Malmö	221	**5**	Burlöv	7	**6**	Malmö	172	**5**	Burlöv	8
	Örkelljunga	4		***Summa***	***221***		Lomma	7		***Summa***	***172***		Lomma	11
	Summa	***111***					Lund	30					Lund	41
							Staffanstorp	10					Staffanstorp	6
3	Helsingborg	138					***Summa***	***54***					***Summa***	***66***
	Höganäs	23												
	Landskrona	52												
	Summa	***213***												

Tabell 2.1 Kompletta säsonger för respektive zon med antal patienter över 65 år som insjuknat i stroke

Från tabell 2.1 fås att zon 3 är representerad på fem säsonger och zon 3 är också en av de folktätaste zonerna och har därmed en högre frekvens av antal insjuknanden i stroke. På zon 7 fås inga kompletta säsonger överhuvudtaget. De zoner som fortsättningsvis används blir zon 3 och 6, se avsnitt 2.6 om avgränsningar.

2.3 Temperaturdata

Denna studie baseras på temperaturdata från Malmö under perioden oktober 2004 till september 2005 och från Helsingborg under perioderna oktober 2002 till april 2003 och oktober 2004 till september 2006. I zon 6 kommer följaktligen både stroke- och temperaturdata från Malmö medan temperaturdata från Helsingborg används för samtliga insjuknande i stroke i zon 3. Där förutsätts att temperaturen inte skiljer sig nämnvärt inom zonen och att data är representativt för hela zonen.

2.4 Beskrivande statistik

Tidsseriediagram över antal insjuknanden i stroke, partiklar och temperatur för zon 3 respektive zon 6 presenteras i bilaga A. I diagram A.1 skådas att kall säsong 2002-2003 har flera dagar med extremvärden på PM_{10} jämfört med de övriga två kalla säsongerna i zon 3. I övrigt finns inga uppenbara skillnader mellan de kalla respektive varma säsongerna över olika år. Vid presentation av beskrivande statistik och extremvärden slås därför de kalla säsongerna 2004-2005 och 2005-2006 ihop och detsamma görs med de varma säsongerna 2005 och 2006 i zon 3. I

zon 6 är extremvärden inte lika tydliga men även här anas de. Tabellen för beskrivande statistik finns i bilaga B och den beroende variabeln stroke har genomgående för alla säsonger ett medelvärde på nära ett insjuknande i stroke per dygn. Variansen för densamma är också ungefär ett, detta tyder på att stroke är en poissonfördelad variabel.

Slutligen ser vi på Pearsons korrelationskoefficient. I zon 3 är laggar som följer på varandra högt korrelerade för PM_{10} och temperatur var för sig och på samma laggar när de jämförs mot varandra. De högsta värdena är 0,73 för PM_{10}, 0,92 för temperatur och 0,32 mellan PM_{10} och temperatur. Motsvarande värden i zon 6 är 0,62, 0,88 och 0,50 men i zon 6 finns även $PM_{2.5}$. De högsta korrelationerna är 0,63 för $PM_{2.5}$, 0,88 mellan $PM_{2.5}$ och PM_{10} och slutligen 0,43 mellan $PM_{2.5}$ och temperatur.

2.5 Extremvärden

Nästa steg blir att identifiera och eliminera extremvärdena, detta görs för att de inte ska förvränga resultatet i den fortsatta analysen. Gränserna för extremvärdena beräknas enligt följande formler

$$\textit{Undre gränsen} = Q_1 - 1.5 * (Q_3 - Q_1) \qquad \textbf{(2.1)}$$

$$\textit{Övre gränsen} = Q_3 + 1.5 * (Q_3 - Q_1) \qquad \textbf{(2.2)}$$

där Q_1 står för den första kvartilen och Q_3 för den tredje. För beräkningarna hämtas Q_1 och Q_3 från bilaga B. Extremvärdena som identifieras ersätts med de övre respektive undre gränsvärdena. I tabell 2.2 nedan presenteras de övre gränsvärdena och antalet ersatta observationer.

		STROKE	TEMPERATUR	PM_{10}	$PM_{2.5}$
ZON 3					
KALL SÄSONG 2002-2003	Övre gräns	5	14,3	66,70	-
	Antal ersatta	0	0	12	-
KALLA SÄSONGER 2004-2005 & 2005-2006	Övre gräns	3	18,6	45,58	-
	Antal ersatta	10	0	17	-
VARMA SÄSONGER 2005 & 2006	Övre gräns	5	24,3	33,94	-
	Antal ersatta	0	0	13	-
ZON 6					
KALL SÄSONG 2004-2005	Övre gräns	5	16,6	32,39	20,56
	Antal ersatta	0	0	9	14
VARM SÄSONG 2005	Övre gräns	5	24,3	32,85	17,04
	Antal ersatta	1	0	4	7

Tabell 2.2 Extremvärden, övre gräns

Undre extremvärdena finns endast i zon 3 och samtliga sju hittas på temperatur, tre stycken på kall säsong 2002-2003 och fyra på varm säsong 2005 och 2006.

2.6 Avgränsningar

Den här studien omfattar Skåne som delas in i åtta zoner. Efter att de olika datamaterialen undersökts begränsas analysen till att omfatta zon 3 och zon 6. Anledningen till att zon 3 väljs är att det finns data på hela fem säsonger och den har relativt stor population. Zon 6 tas med för att resultat ska kunna jämföras med det från kandidatuppsatsen. Övriga zoner exkluderas då det inte ryms inom tidsramen att studera samtliga modeller. Ytterligare en avgränsning är att endast populationen som är äldre än 65 år studeras, det motiveras av att tidigare studier gjort en sådan avgränsning. Precis som i tidigare uppsats görs en uppdelning på kall och varm säsong och även det följer tidigare studiers upplägg. Kall säsong går från oktober till april och varm säsong går således från maj till september. I denna studie väljer vi att studera sambandet på lag 0, 1, 5 och 7, detta beslut baseras bl.a. på tidigare studiers upplägg och resultat. Dessutom är korrelationerna mellan laggarna höga för temperatur vilket ökar risken för multikolinjäritet om samtliga sju laggar inkluderas.

3. TEORI

Teorin som denna uppsats följer hämtas ifrån Dobsons bok *An Introduction to Generalized Linear Models* (2002) och Hastie och Tibshiranis *Generalized Additive Models* (1990). Även Agrestis *An introduction to Categorical Data Analysis* (2007) och information och programmeringskod ifrån SAS Institutets hemsida (SAS 2008) används.

3.1 Utforskande dataanalys

För att bekanta sig med data börjar analysen med en noggrann undersökning av respektive variabel. Detta görs för att kontrollera kvaliteten på data och för att få en uppfattning om vilka metoder och modeller som bör användas vid sambandsanalys. Vid undersökningen granskas variablernas skala, om de är kontinuerliga eller diskreta och vilken typ av fördelning de har.

Kontinuerliga variabler kan anta alla värden inom ett intervall och antas därför ofta vara normalfördelade. Diskreta variabler antar endast positiva heltalsvärden och antas ofta följa en poissonfördelning om sannolikheten för att en händelse ska inträffa är väldigt liten i en stor population. Karakteristiskt för poissonfördelning är att väntevärdet är lika med variansen.

Därefter väljs de metoder som ska användas i analysen baserat på den beroende variabelns och de oberoende variablernas skala. Om den beroende variabeln antar positiva heltalsvärden används poissonregressionsmodeller.

Ett vanligt problem med poissonregressionsmodeller är överspridning och det uppkommer när variansen är större än väntevärdet. En alternativ metod som tillåter överspridning för poissonfördelningen är negativ binomialfördelning. För negativ binomialfördelning är

$$\mathrm{var}(Y_i) = \phi E(Y_i) \qquad \textbf{(3.1)}$$

där $\phi > 1$ är en spridningsparameter som kan skattas. Skattningen kan göras i SAS med hjälp av proc GENMOD som beskrivs i avsnitt 3.3.

Innan modellformuleringen påbörjas kontrolleras Pearsons korrelationskoefficient mellan de oberoende variablerna och deras laggar. Först undersöks korrelationen inom varje variabel och dess laggar därefter mellan de olika oberoende variablerna och deras laggar. Höga korrelationer tyder på att variabeln och dess laggar eller de oberoende variablerna och deras laggar har liknande effekt på den beroende variabeln. Då korrelationen mellan två variabler är ungefär 0,9 eller högre inträffar multikolinjäritet om båda tas med samtidigt i modellen. Detta innebär att dessa variabler kan bli icke-signifikanta även om de skulle bli signifikanta i modeller där de studeras var för sig. Det räcker därför att endast ta med en av dem i modellen för att studera sambandet och den variabel som är mest signifikant väljs.

3.2 Modellformulering

Det som framkommit i utforskande dataanalys används som stöd för att formulera en modell. Poissonregressionsmodellen är en typ av generaliserad linjär modell och de består av tre komponenter. Den första är den beroende variabeln *Y* som här antas vara poissonfördelad, den andra är en linjär kombination av de oberoende variablerna och den tredje är en log-länkfunktion. Länkfunktionen ska visa sambandet mellan det förväntade värdet av *Y* och de oberoende variablerna. Formeln för poissonregressionsmodellen är

$$\log\left(E\left(Y\right)\right) = \alpha + \sum_{i=1}^{m} \beta_i x_i \qquad \textbf{(3.2)}$$

där $\alpha + \sum_{i=1}^{m} \beta_i x_i$ är den linjära kombinationen av x_i.

Väntevärdet uppfyller följande exponentiella samband

$$E(Y) = e^{(\alpha + \sum_{i=1}^{m} \beta_i x_i)} = e^{\alpha} \prod_{i=1}^{m} e^{\beta_i x_i} \qquad \textbf{(3.3)}$$

β_i tolkas här som partiella regressionskoefficienter och det betyder att en ökning i exempelvis x_1 ger en multiplikativ effekt av e^{β_1} på väntevärdet. Det innebär att väntevärdet av *Y* vid x_1+1 är lika med väntevärdet av *Y* vid x_1 gånger e^{β_1}, givet att övriga x_i hålls konstanta. Om β_1=0 blir $e^{\beta_1} = 1$ vilket betyder att väntevärdet av *Y* inte påverkas av förändringar i x_1. Då β_1>0 fås ett positivt samband och då β_1<0 fås ett negativt samband.

Om residualerna från de generaliserade linjära modellerna inte uppfyller kraven, se avsnitt 3.4, kan det bero på att det finns trender i de oberoende variablerna som inte upptäckts. Istället för att pröva sig fram till vilka transformationer av de oberoende variablerna som modellen behöver kompletteras med, används generaliserade additiva modeller. Dessa modeller använder sig av utjämningsfunktioner som är av icke-parametrisk natur. Antingen fås en helt icke-parametrisk modell där samtliga utjämningsfunktioner är signifikanta eller en semiparametrisk modell. I den semiparametriska modellen har en del variabler ett linjärt samband och en del en utjämningsfunktion. Utjämningsfunktionerna har två syften, dels att beskriva trenden för varje oberoende variabel och dels att beskriva sambandet mellan den beroende variabeln och samtliga oberoende variabler. Den generaliserade additiva modellen kan sägas vara en förlängning av den generaliserade linjära modellen.

För poissonregressionsmodellen fås den generaliserade additiva modellen genom att ersätta den linjära kombinationen av x_i i formel 3.2 med en additiv kombination av icke-parametriska funktioner av x_i. Den slutliga formeln blir

$$\log\left(E\left(Y\right)\right) = s_0 + \sum_{i=1}^{m} s_i\left(x_i\right) \qquad \textbf{(3.4)}$$

där s_0 motsvarar interceptet i formel 3.2 och där s_i är en utjämningsfunktion för respektive x_i. Den typ av utjämningsfunktion som används i denna studie kallas spline. Den är uppbyggd av flera delfunktioner som tvingats till en kontinuerlig funktion genom att sammanbinda knutpunkterna.

3.3 Modellskattning

När modellen ska skattas används de två procedurerna Proc GENMOD och Proc GAM i statistikdataprogrammet SAS. Proc GENMOD skattar generaliserade linjära modeller, i dessa modeller ingår såväl vanliga regressions- och ANOVA-modeller för kontinuerliga beroende variabler som loglinjära modeller för diskreta beroende variabler.

I syntaxen för proc GENMOD måste modellen, den beroende variabelns fördelning och länkfunktionen definieras för att modellen ska kunna skattas. Först skattas modellerna med poissonfördelning och därefter med negativ binomialfördelning för att kontrollera om det finns signifikant överspridning.

Det finns också flera olika tillägg som kan göras beroende på vilken information som efterfrågas. I detta fall är residualer och Maximum Likelihood-statistikan för typ1 och typ3 av intresse. Med typ1 får man fram effekten av den aktuella variabeln då endast den och föregående variabler tas med i modellen medan typ3 visar effekten av respektive variabel då samtliga inkluderas. Tillsammans ger typ1 och typ3 komplett information om respektive variabels effekt. Genom att enbart studera typ3 kan samtliga variabler som är signifikanta fås fram. Däremot finns risk för att variabler som egentligen skulle ha blivit signifikanta elimineras i förtid om enbart typ1 studeras. Om variablerna är oberoende av varandra borde typ1 och typ3 ge samma resultat.

Inledningsvis tas alla oberoende variabler med i modellen och därefter elimineras stegvis de parametrar som inte är signifikanta och först den variabeln med högst p-värde. Vid eliminering tas även hänsyn till typ1- och typ3-utfallen. Slutligen återstår endast interceptet och de signifikanta oberoende variablerna.

Om residualerna för modellerna som fås i proc GENMOD inte uppfyller kraven görs den fortsatta analysen med generaliserade additiva modeller i proc GAM för att skatta utjämningsfunktionerna.

I syntaxen för proc GAM måste modellen och den beroende variabelns fördelning anges medan ytterligare tillägg är valfria. De viktigaste tilläggen i denna studie är diagram över utjämningsfunktionerna, generaliserad korsvalidering och konvergenskriteriet. Diagrammen ger en överblick över trenderna hos respektive oberoende variabel. Från diagrammen kan man direkt se vilken typ av icke-parametrisk funktion variablerna har mot den beroende variabeln istället för att pröva sig fram med olika transformationer. De icke-parametriska funktionerna som föreslås ur diagrammen kan sedan läggas till i proc GENMOD för att få en komplett modell. Generaliserad korsvalidering är den metod som används för att välja parametrarna till utjämningsfunktionerna. Genom att använda generaliserad korsvalidering och att ange konvergenskriteriets standardvärde 10^{-8} fås bra resultat och risken för att förbise icke-linjära samband minskar.

3.4 Modellutvärdering

Residualanalys utgör en viktig del av modellutvärderingen, det finns flera kriterier residualerna måste uppfylla för en bra modell. För poissonregressionsmodeller är de standardiserade residualerna som är betydelsefulla i modellutvärderingen. De standardiserade residualerna ska vara oberoende och ungefär normalfördelade med väntevärdet noll och ha en konstant varians. De ska även vara oberoende av de förklarande variablerna.

För att kontrollera att kriterierna uppfylls beräknas först de standardiserade residualerna

$$r_i = \frac{y_i - \hat{\theta}_i}{\sqrt{\hat{\theta}_i}} \qquad \textbf{(3.5)}$$

där $E(Y_i) = \theta_i$ och $Y_i \sim \text{Poisson}(\theta_i)$.

De standardiserade residualerna jämförs med normalfördelningen för att bedöma om de är normalfördelade och för att identifiera extremvärden. För att kraven ska vara uppfyllda får inte mer än 5 % av observationerna anta värden som understiger -1,96 eller överstiger 1,96 och mer än 1 % av observationerna får inte överstiga ± 2,58.

De standardiserade residualerna plottas sedan mot samtliga förklarande variabler i modellen, en i taget. Om modellen beskriver effekten av variablerna på ett bra sätt, skall det inte finnas något tydligt mönster i spridningsdiagrammen. Om modellen däremot inte är lämplig, kan systematiska mönster i modellen skådas och det indikerar att flera eller andra variabler behöver inkluderas i modellen. De standardiserade residualerna plottas även mot de skattade värdena av θ_i för att undersöka om variansen är konstant.

Till slut görs även spridningsdiagram av residualerna mot tiden för att kontrollera att de är oberoende. Om residualerna är oberoende fördelas de slumpmässigt utan att uppvisa ett systematiskt mönster.

4. RESULTAT

Redovisningen av resultaten kommer att följa ordningen i teoriavsnittet och resultaten finns i bilaga C till F.

4.1 Resultat av utforskande dataanalys

Som framkommit i avsnitt 2.4 tycks den beroende variabeln vara poissonfördelad. De varma säsongerna i båda zonerna har dock en något högre varians än medelvärde. Därför skattas spridningsparametern i proc GENMOD, men varje gång utan signifikant resultat. Följaktligen finns ingen signifikant överspridning och därmed är stroke inte negativt binomialfördelad. Även kvartilerna tyder på poissonfördelning då det är många dagar med noll insjuknanden i stroke och medianen är ett.

4.2 Grundmodell

Resultatet från utforskande dataanalys bekräftar att den beroende variabeln är poissonfördelad och därför väljs poissonregressionsmodeller för att analysera data. I zon 3 utgås från en grundmodell där PM_{10} och temperatur inkluderas medan zon 6 har tre grundmodeller. I de två första inkluderas partiklarna var för sig tillsammans med temperatur och i den tredje tas samtliga tre oberoende variabler med. Varje modell innehåller lag 0, 1, 5, 7 på samtliga oberoende variabler.

4.3 Resultat av modellskattning

Modellerna skattas först i proc GENMOD men residualanalys antyder att modellerna är bristfälliga. Därför görs den fortsatta analysen i proc GAM för att undersöka om det finns trender i de oberoende variablerna. Med proc GAM skattas först en icke-parametrisk modell med utjämningsfunktioner, så kallade spline, för samtliga oberoende variabler. Om samtliga utjämningsfunktioner inte är signifikanta blir nästa steg att skatta en semiparametrisk modell. Därefter görs stegvis eliminering tills enbart signifikanta parametrar återstår. De utjämningsfunktioner som finns i den slutliga modellen som fås fram i proc GAM undersöks sedan grafiskt. Graferna ger en tydlig bild av vilken typ av trender som behöver kompletteras med i proc GENMOD.

I zon 3 fås inga signifikanta samband på någon av säsongerna och detsamma gäller för varm säsong 2005 för zon 6. Enbart kall säsong 2004-2005 för zon 6 uppvisar signifikanta samband.

I proc GAM är modellen med $PM_{2.5}$ signifikant för $PM_{2.5}$ på lag 0 och lag 1, se figur C.1 i bilaga C. Även $PM_{2.5}$ lag 7 är med i modellen fast den inte är signifikant då signifikansnivån är 5 %. Anledningen till det är att inga övriga variabler blir signifikanta om den elimineras och ett p-värde på 0,067 är inte heller särskilt högt. På lag 1 är även utjämningsfunktionen signifikant och från grafen i bilaga C.2 ses att funktionen är kvadratisk. Den kvadratiska termen och de övriga signifikanta variabler från proc GAM skattas sedan i proc GENMOD. Resultatet finns i bilaga C.3 och där ses att samtliga variabler är signifikanta. Den slutliga modellen blir

$$\log(E(Stroke)) = \quad \textbf{(4.1)}$$

$$-1{,}5424 - 0{,}0509PM_{2.5}(lag0) + 0{,}3558PM_{2.5}(lag1) - 0{,}0126(PM_{2.5}(lag1))^2$$

$$E(Stroke) = \quad \textbf{(4.2)}$$

$$\exp(-1{,}5424 - 0{,}0509PM_{2.5}(lag0) + 0{,}3558PM_{2.5}(lag1) - 0{,}0126(PM_{2.5}(lag1))^2)$$

Om $PM_{2.5}$ är lika med 1 μg/m^3 på samtliga laggar blir det förväntade antalet insjuknande i stroke

$$E(Stroke) = 0{,}214 * 0{,}950 * 1{,}427 * 0{,}987 = 0{,}214 * 1{,}339 = 0{,}287 \quad \textbf{(4.3)}$$

Den valda modellen för kall säsong 2005-2006 i zon 6 visar att 0,214 personer insjuknar i stroke om $PM_{2.5}$ är lika med noll. Den enskilda påverkan från respektive lag vid en ökning med 1 μg/m^3 i partikelhalten är en minskning med 5 % på lag 0, en ökning med 42,7 % på lag 1 och en minskning med 1,3 % på den kvadratiska funktionen av lag 1. Den sammanlagda effekten på insjuknanden i stroke är en ökning med 33,9 % då partikelhalten ökar med 1 μg/m^3 på samtliga laggar.

En minskning av antalet insjuknanden i stroke vid lag 0 motsvarar inte det förväntade resultatet och även typ1-analysen, se bilaga F.1, visar på att sambandet inte är signifikant. $PM_{2.5}$ lag 0 är inte signifikant när endast den och interceptet tas med i modellen. Först när ytterligare variabler inkluderas blir den signifikant, se typ3-analysen.

Resultaten för modellen med PM_{10} har liknande mönster som modellen för $PM_{2.5}$. I proc GAM blir PM_{10} lag 0 och lag 1 signifikanta, se bilaga D.1. Utjämningsfunktionen blir inte signifikant på 5 % -nivån, men p-värdet är lika med 0,0607 och inkluderas trots allt i modellen som skattas i proc GENMOD. I bilaga D.2 finns grafen över utjämningsfunktionen och den är av kvadratisk karaktär. Den slutliga modellen som fås från proc GENMOD, se bilaga D.3, blir således

$$\log\big(E\big(Stroke\big)\big) = \quad \textbf{(4.4)}$$

$$-0{,}9272 - 0{,}0254PM_{10}\big(lag0\big) + 0{,}1444PM_{10}\big(lag1\big) - 0{,}0031\big(PM_{10}\big(lag1\big)\big)^2$$

$$E\big(Stroke\big) = 0{,}396 * 0{,}975 * 1{,}155 * 0{,}997 = 0{,}396 * 1{,}123 = 0{,}446 \quad \textbf{(4.5)}$$

Det förväntade värdet av antal insjuknande i stroke blir 0,446 då partikelhalten är lika med 1 μg/m^3 på samtliga laggar. Totalt ger detta en ökning med 12,3 % då PM_{10} ökar med 1 μg/m^3 på respektive lag. De individuella effekterna blir en minskning med 2,5 % vid lag 0, en ökning med 15,5 % vid lag 1 och en minskning med 0,3 % på den kvadratiska funktionen för lag 1 då partikelhalten ökar med 1 μg/m^3. Då partikelhalten är lika med noll blir det förväntade antalet insjuknanden i stroke 0,396.

Slutligen skattas en modell som inkluderar $PM_{2.5}$, PM_{10} och temperatur. Resultatet visar att endast $PM_{2.5}$ blir signifikant och det ger samma modell som i formel 4.1.

4.4 Resultat av modellutvärdering

För att undersöka de valda modellernas lämplighet analyseras de skattade residualerna. Först undersöks om kravet för normalfördelning är uppfyllt dels grafiskt genom att studera normalitetsdiagram och histogram och dels görs normalitetstest. Dessa hittas i bilaga E för modellen med $PM_{2.5}$ på kall säsong 2004-2005 i zon 6. Ur figur E.1 kan normalfördelningskravet tyckas vara uppfyllt, sånär som på avvikelser från linjen vid sidorna. Även histogrammet i figur E.2 tycks uppfylla normalitetskravet. Normalitetstesten i tabell E.3 visar dock att residualerna inte är normalfördelade ty p-värdena är mindre än 5 %.

Vidare följer i bilaga E spridningsdiagram över de standardiserade residualerna mot de förklarande variablerna och mot de skattade värdena av den beroende variabeln. Slutligen finns även ett spridningsdiagram över residualerna mot tiden. De två första spridningsdiagrammen, E.4 och E.5, med de förklarande variablerna uppvisar inga tydliga mönster och verkar slumpvisa. I diagram E.6 ses att kravet för konstant varians inte är uppfyllt för de skattade värdena på stroke. Det sista diagrammet i E.7 uppvisar oberoende residualer ty inget systematiskt mönster kan urskiljas.

Residualanalysen för modellen med PM_{10} på kall säsong 2004-2005 i zon 6 följer samma upplägg som för modellen med $PM_{2.5}$. Resultaten presenteras i bilaga F. I normalitetsdiagrammet i figur F.1 blir det tydligt att residualerna inte är normalfördelade. Även histogrammet i figur F.2 visar på avvikanden från normalfördelningen. Slutligen bekräftar också normalitetstesten i tabell F.3 att kravet för normalfördelade residualer inte är uppfyllt, då p-värdena är små.

Därefter studeras spridningsdiagrammen, i diagram F.4 och F.5 plottas de standardiserade residualerna mot de förklarande variablerna var för sig. I båda diagrammen verkar residualerna slumpmässiga och uppvisar inget tydligt mönster. Precis som i föregående modell skådas i diagram F.6 att variansen inte är konstant då residualerna plottas mot de skattade värdena för stroke. I diagram F.7 plottas residualerna mot tiden och det finns inget i diagrammet som tyder på att residualerna är beroende då de är slumpmässigt fördelade.

Det sammanfattande resultatet blir att ingen av modellerna är tillförlitlig, då residualerna inte är normalfördelade och variansen inte är konstant.

5. SLUTSATS

Syftet med denna uppsats är att ta fram regressionsmodeller som beskriver sambandet mellan partikelhalten i luften och antal insjuknanden i stroke. I zon 3 har inga signifikanta samband kunnat påvisas. Det finns flera möjliga felkällor i zonen som kan ha förvrängt resultatet. För det första är det tre olika kommuner som slagits samman till en zon. Till skillnad från zon 6 där samtliga data kom från Malmö stad, används data från olika städer i zon 3. Partikeldata över Helsingborgs kommun saknas, istället används data från Höganäs och Landskrona. Dessa städer är mindre och antas ha lägre partikelhalter än Helsingborg.

I zon 6 fås signifikant samband på kall säsong 2004-2005 dels för modellen med $PM_{2.5}$ och dels för PM_{10}. Båda modellerna ger ett signifikant samband på lag 0, lag 1 och den kvadratiska termen av lag 1. Där sambandet på lag 0 är negativt för båda modellerna. Att antalet insjuknande i stroke minskar samma dag som partikelhalten ökar strider mot det förväntade resultatet. Sammantaget blir det emellertid en ökning på 33,9 % respektive 12,3 % av antalet insjuknanden i stroke då $PM_{2.5}$ respektive PM_{10} ökar med 1 $\mu g/m^3$ på samtliga laggar. Modellen med $PM_{2.5}$ ger en starkare påverkan på antal insjuknande i stroke. Det är ett resultat som stämmer väl med teorin om att mindre partiklar är skadligare än större.

I övrigt är resultaten inte helt jämförbara med tidigare studiers. Till exempel visar studien från Shanghai i Kina att risken för att dö i stroke ökar med 0,8 % då PM_{10} ökar med 10 $\mu g/m^3$ på lag 1. Motsvarande risk då PM_{10} ökar med 1 $\mu g/m^3$ skulle bli 0,08 %, vilket är betydligt lägre än 12,3 % som fås i denna studie. En möjlig förklaring till skillnaden är att antalet oberoende variabler är färre i denna uppsats. Eventuellt har viktiga variabler utelämnats, vilket innebär att modellen kanske är underspecificerad. Det medför att risken för att insjukna i stroke blir högre om endast PM_{10} inkluderas i modellen än om hänsyn tas till alla riskfaktorer. Samtidigt innebär det en risk för multikolinjäritet att ta med för många variabler, då de enskilda effekterna försvagas. En viktig skillnad är också att samtliga patienter som insjuknar i ischemisk stroke studeras i denna uppsats och inte enbart de som dör. Även det faktum att det är två länder med stora skillnader i partikelhalter, temperaturer och antal insjuknanden i stroke gör att de inte blir helt jämförbara.

Residualanalyserna för de båda modellerna ger liknande resultat och det konstateras att normalitetskraven för residualerna inte är uppfyllda. Det innebär att modellerna inte är tillförlitliga för att användas till att göra prognoser. I bilaga G ses de faktiska och de skattade värdena av antal insjuknanden i stroke i tidsseriediagrammen G.1 och G.2 för $PM_{2.5}$ respektive PM_{10}. Där blir tydligt att de framkomna modellerna inte ens kan skatta två insjuknanden i stroke. Vilket inte är konstigt eftersom partiklar bara är en faktor av många som leder till stroke.

Som nämnts i inledningen är denna magisteruppsats en utveckling av kandidatuppsaten. Därför är det intressant att undersöka om de olika metoderna ger ett likartat resultat. I kandidatuppsatsen studerades sambandet mellan partiklar och antal insjuknanden i stroke i Malmö stad under kall säsong 2004-2005 och varm säsong 2005. Metoden som användes för att illustrera sambanden var transferfunktionsmodeller. Vi fick signifikanta samband under den kalla säsongen på lag

0, 1 och 7 för modellen med ln $PM_{2.5}$ och på lag 1 för modellen med ln PM_{10}. För lag 0 i modellen med ln $PM_{2.5}$ var sambandet negativt, vilket stämmer överens med resultatet från magisteruppsatsen. Likaså överensstämmer resultat med lag 1 som blev positivt signifikant i samtliga modeller. Slutligen har vi lag 7 i modellen med ln $PM_{2.5}$, i magisteruppsatsen var lag 7 i $PM_{2.5}$–modellen inte signifikant i proc GENMOD. Men i proc GAM inses att den ändå är viktig för modellen då inga andra variabler blir signifikanta om lag 7 elimineras. I kandidatuppsatsen prövades olika transformationer, för att få konstant varians för de oberoende variablerna. De transformationer som sedan valdes var ln för både $PM_{2.5}$ och PM_{10}. I denna studie däremot lät vi proc GAM grafiskt åskådliggöra utjämnings-funktionen och därur kunde vi se att en kvadratisk funktion var mest lämplig. Det kan tänkas att lag 7 i modellen med ln $PM_{2.5}$ är en veckoeffekt, som elimineras i magisteruppsatsen med den kvadratiska funktionen.

Sammanfattningsvis tycks båda metoderna ge ungefär samma resultat. Men metoderna som använts i magisteruppsatsen är bättre anpassade för datamaterialet och resultatet är enklare att tolka.

REFERENSER

Tryckta källor

Agresti, A. (2007). *An Introduction to Categorical Data Analysis.* 2 uppl. Hoboken: Wiley.

Dobson, A. (2002). *An Introduction to Generalized Linear Models.* 2 uppl. Chapman & Hall.

Hastie, T.J. & Tibshirani, R.J. (1990). *Generalized Additive Models.* Chapman & Hall.

Hong, Y., Lee, J., Kim, H., Ha, E., Schwartz, J. & Christiani, D. (2002). ”Effects of Air Pollutants on Acute Stroke Mortality”. *Environmental Health Perspectives, Vol 110, s. 187-191.*

Kan, H., Jia, J. & Chen, B. (2003). ”Acute Stroke Mortality and Air Pollution: New Evidence from Shanghai, China”. *Journal of Occupational Health, Vol 45, s. 321-323.*

Ramsay, T., Burnett, R. & Krewski, D. (2003). ”Exploring bias in a generalized additive model for spatial air pollution data”. *Environmental Health Perspectives, Vol 111, s. 1283-1288.*

Wong, C-M., Ma, S., Johnson Hedley, A. & Lam, T-H. (2001). ”Effect of Air Pollution on Daily Mortality in Hong Kong”. *Environmental Health Perspectives, Vol 109, s. 335-340.*

Otryckta källor

Hillström, J. & Nsabimana, J. (2008). *Sambanden mellan inandningsbara, grova och fina partiklar i luften och strokeanfall i Malmö.* http://www.uppsatser.se/uppsats/986c787694/, hämtad 2009-04-29

IVL Svenska miljöinstitutet (2008), http://www3.ivl.se/db/plsql/dvst_pm10_st$.startup, hämtad 2008-12-11

Länsstyrelsen i Stockholms län (2003). *Mängden skadliga partiklar i luften måste minska. Halterna i länet överstiger lagstadgade normer,* http://www.ab.lst.se/upload/dokument/miljo_och_halsa/miljolagstiftning/MKN/faktabl_2003_03_partiklar_webb.pdf, hämtad 2009-01-04

Riksstroke (2009), www.riks-stroke.org/index.php?content=sjukhus#skane, hämtad 2009-04-11

SAD (2009), http://www.ytv.fi/SWE/luftkvalitet/effekter/halsoeffekter/hemsida.htm, hämtad 2009-04-15

SAS Institute (2008). *Fitting Generalized Additive Models with the GAM Procedure in SAS 9.2,* www2.sas.com/proceedings/forum2008/378-2008.pdf, hämtad 2009-04-12

Sjukvårdsrådgivning (2008), http://www.sjukvardsradgivning.se/artikel.asp?CategoryID=27011, hämtad 2009-04-29

BILAGA A

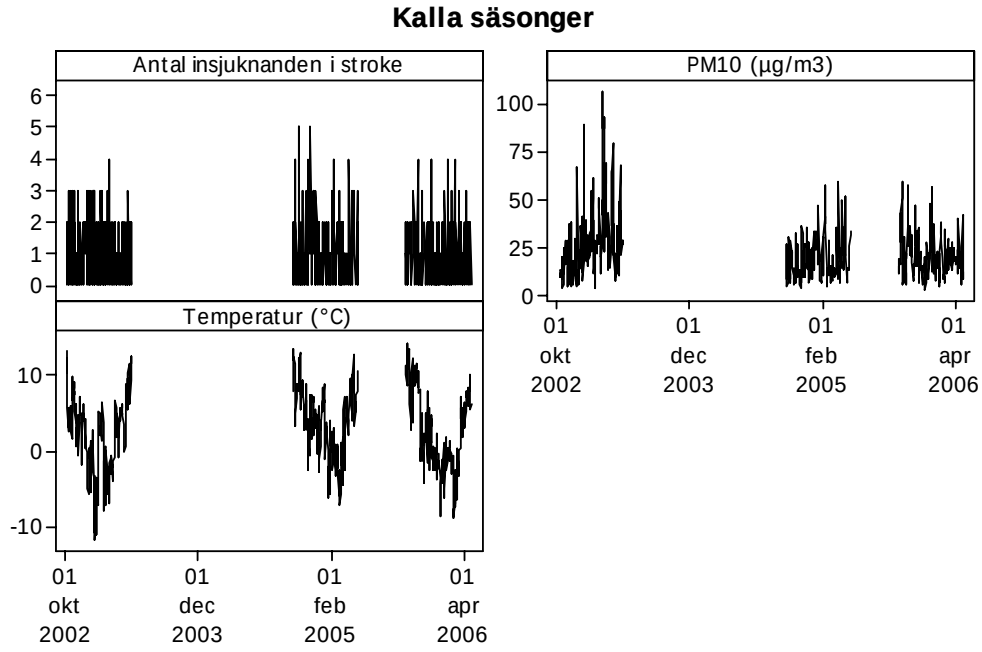


Diagram A.1 Tidsseriediagram. Kalla säsonger, zon 3

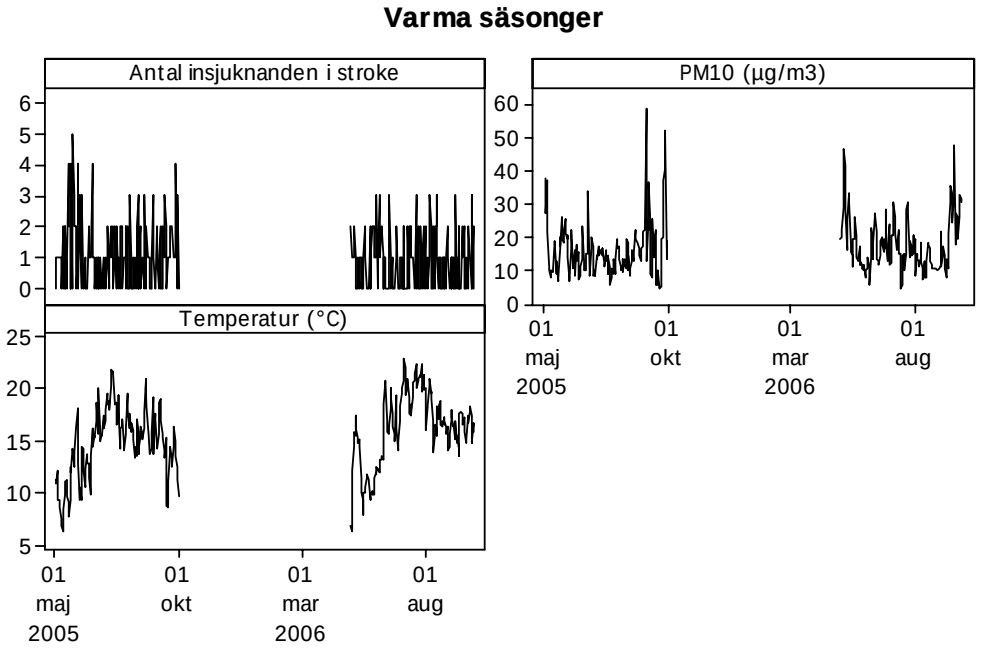


Diagram A.2 Tidsseriediagram. Varma säsonger, zon 3

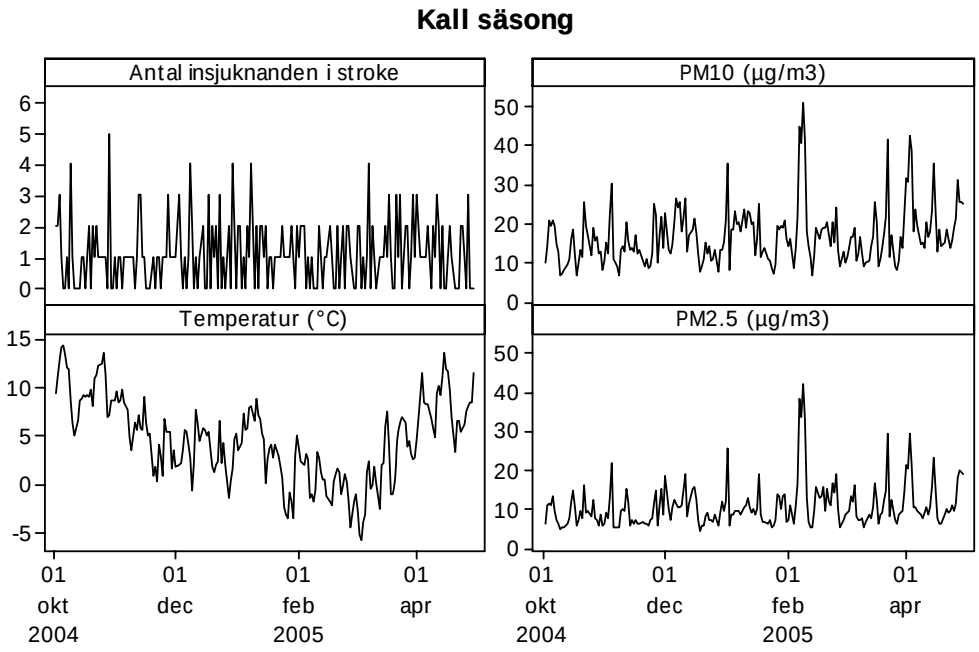


Diagram A.3 Tidsseriediagram. Kall säsong, zon 6

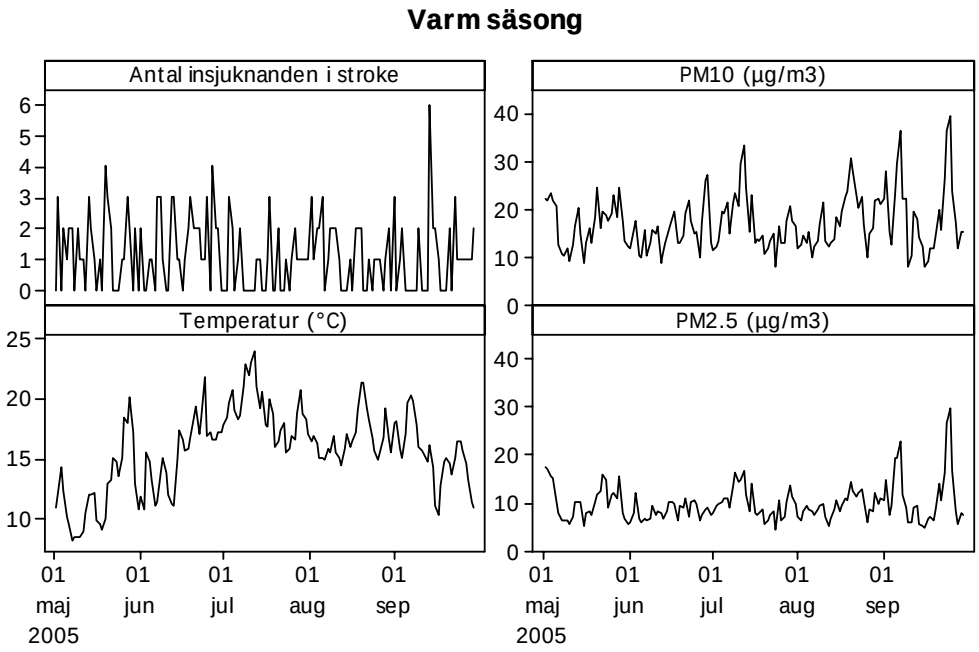


Diagram A.4 Tidsseriediagram. Varm säsong, zon 6

BILAGA B

Variabel	Antal dagar	Medelvärde	Medelfel	Varians	Minimum	Q1	Median	Q3	Maximum
ZON 3									
KALL SÄSONG 2002-2003									
STROKE	212	1,00	0,07	0,92	0	0	1	2	4
PM_{10}	212	28,06	1,33	360,28	4,15	15,95	23,84	36,25	106,20
TEMPERATUR	212	2,18	0,33	22,55	-11,70	-1,03	2,55	5,10	13,40
KALL SÄSONG 2004-2005 & 2005-2006									
STROKE	424	0,97	0,05	0,97	0	0	1	1	5
PM_{10}	424	20,05	0,53	117,22	3,30	12,20	17,90	25,55	59,60
TEMPERATUR	424	3,30	0,24	24,35	-8,90	-0,40	3,20	7,20	14,40
VARM SÄSONG 2005 & 2006									
STROKE	306	0,98	0,06	1,00	0	0	1	2	5
PM_{10}	306	16,87	0,48	68,23	4,30	10,88	14,75	20,10	58,40
TEMPERATUR	306	15,60	0,20	11,99	6,40	13,68	15,95	17,93	23,00
ZON 6									
KALL SÄSONG 2004-2005									
STROKE	212	1,10	0,07	1,08	0	0	1	2	5
PM_{10}	212	16,83	0,51	54,84	6,41	12,00	15,69	20,13	51,05
$PM_{2.5}$	212	11,12	0,39	32,24	4,89	7,41	9,74	12,66	42,38
TEMPERATUR	212	4,65	0,29	17,92	-5,67	1,83	4,92	7,70	14,37
VARM SÄSONG 2005									
STROKE	153	1,13	0,09	1,29	0	0	1	2	6
PM_{10}	153	17,20	0,48	34,72	7,94	12,94	15,83	20,91	39,46
$PM_{2.5}$	153	10,06	0,32	15,37	4,53	7,47	9,30	11,30	29,84
TEMPERATUR	153	15,78	0,27	11,20	8,25	13,81	15,94	18,01	24,02

Tabell B.1 Beskrivande statistik

BILAGA C

Iteration Summary and Fit Statistics

Number of local scoring iterations	6
Local scoring convergence criterion	4.474518E-12
Final Number of Backfitting Iterations	1
Final Backfitting Criterion	3.451299E-11
The Deviance of the Final Estimate	201.45028983

Regression Model Analysis
Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	-0.08153	0.29885	-0.27	0.7853
$PM_{2.5}$ lag 7	0.02907	0.01578	1.84	0.0670
Linear($PM_{2.5}$)	-0.05587	0.01830	-3.05	0.0026
Linear($PM_{2.5}$ lag 1)	0.04448	0.01968	2.26	0.0249

Smoothing Model Analysis
Fit Summary for Smoothing Components

Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline($PM_{2.5}$)	0.999992	3.000000	0.909769	176
Spline($PM_{2.5}$ lag 1)	0.999992	3.000000	0.910200	176

Smoothing Model Analysis
Analysis of Deviance

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline($PM_{2.5}$)	3.00000	3.750550	3.7505	0.2897
Spline($PM_{2.5}$ lag 1)	3.00000	18.346441	18.3464	0.0004

Figur C.1 $PM_{2.5}$; Proc GAM. Kall säsong 2004-2005, zon 6

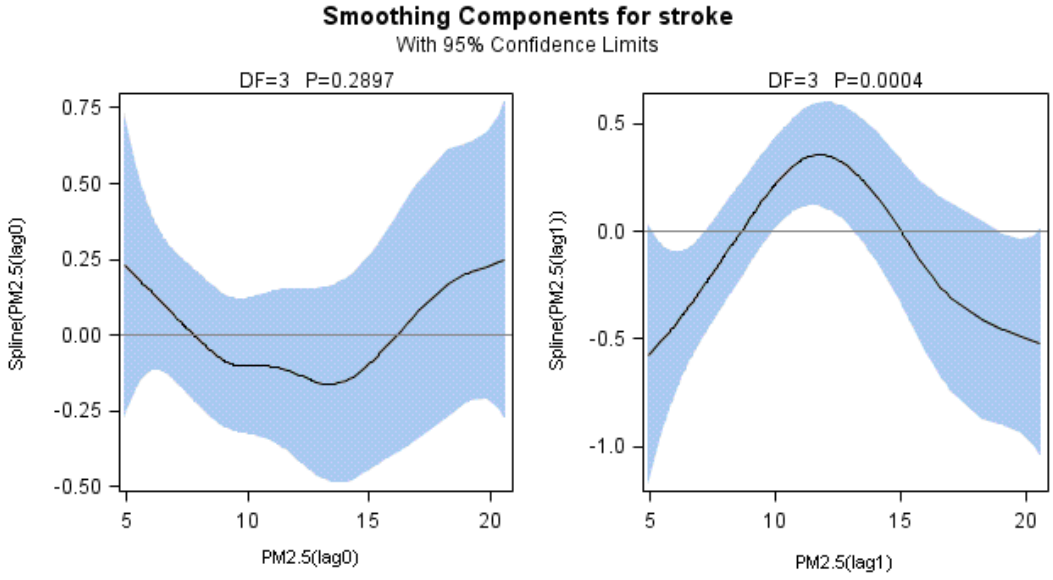


Diagram C.2 Utjämningsfunktionerna över $PM_{2.5}$ lag 0 och lag 1. Kall säsong 2004-2005, zon 6

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	207	221.9724	1.0723
Scaled Deviance	207	239.5433	1.1572
Pearson Chi-Square	207	191.8162	0.9266
Scaled Pearson X2	207	207.0000	1.0000
Log Likelihood		-217.8248	
Full Log Likelihood		-298.6263	
AIC (smaller is better)		605.2526	
AICC (smaller is better)		605.4468	
BIC (smaller is better)		618.6601	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.5424	0.5424	-2.6054	-0.4794	8.09	0.0045
$PM_{2.5}$	1	-0.0509	0.0188	-0.0876	-0.0141	7.34	0.0067
$PM_{2.5}$ lag 1	1	0.3558	0.0960	0.1675	0.5440	13.72	0.0002
$(PM_{2.5}$ lag 1$)^2$	1	-0.0126	0.0037	-0.0199	-0.0053	11.53	0.0007
Scale	0	0.9626	0.0000	0.9626	0.9626		

LR Statistics For Type 1 Analysis

Source	Deviance	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
Intercept	238.4398						
$PM_{2.5}$	236.5920	1	207	1.99	0.1594	1.99	0.1579
$PM_{2.5}$ lag 1	233.5287	1	207	3.31	0.0705	3.31	0.0690
$(PM_{2.5}$ lag 1$)^2$	221.9724	1	207	12.47	0.0005	12.47	0.0004

LR Statistics For Type 3 Analysis

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
$PM_{2.5}$	1	207	7.57	0.0065	7.57	0.0059
$PM_{2.5}$ lag 1	1	207	14.55	0.0002	14.55	0.0001
$(PM_{2.5}$ lag 1$)^2$	1	207	12.47	0.0005	12.47	0.0004

Figur C.3 $PM_{2.5}$; Proc GENMOD. Kall säsong 2004-2005, zon 6

BILAGA D

Iteration Summary and Fit Statistics

Number of local scoring iterations	6
Local scoring convergence criterion	6.663851E-13
Final Number of Backfitting Iterations	1
Final Backfitting Criterion	6.269188E-12
The Deviance of the Final Estimate	226.64591099

Regression Model Analysis
Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.05729	0.22419	0.26	0.7986
PM_{10}	-0.02549	0.01272	-2.00	0.0464
Linear(PM_{10} lag 1)	0.02753	0.01278	2.15	0.0324

Smoothing Model Analysis
Fit Summary for Smoothing Components

Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline(PM_{10} lag 1)	0.999994	3.000000	0.907535	195

Smoothing Model Analysis
Analysis of Deviance

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(PM_{10} lag 1)	3.00000	7.380249	7.3802	0.0607

Figur D.1 PM_{10}; Proc GAM. Kall säsong 2004-2005 zon 6

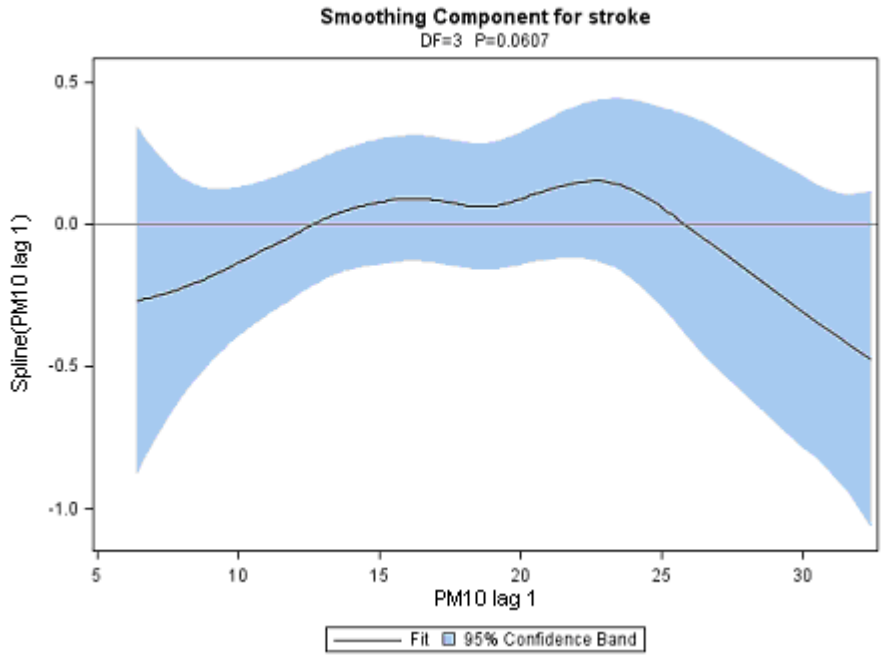


Diagram D.2 Utjämningsfunktionerna över PM_{10} lag 1. Kall säsong 2004-2005, zon 6

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	207	229.0578	1.1066
Scaled Deviance	207	243.9260	1.1784
Pearson Chi-Square	207	194.3826	0.9390
Scaled Pearson X2	207	207.0000	1.0000
Log Likelihood		-218.7216	
Full Log Likelihood		-298.4563	
AIC (smaller is better)		604.9126	
AICC (smaller is better)		605.1068	
BIC (smaller is better)		618.3201	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.9272	0.4848	-1.8774	0.0229	3.66	0.0558
PM_{10}	1	-0.0254	0.0125	-0.0499	-0.0010	4.15	0.0417
PM_{10} lag 1	1	0.1444	0.0552	0.0362	0.2526	6.84	0.0089
$(PM_{10}$ lag 1$)^2$	1	-0.0031	0.0014	-0.0058	-0.0004	4.93	0.0265
Scale	0	0.9690	0.0000	0.9690	0.9690		

LR Statistics For Type 1 Analysis

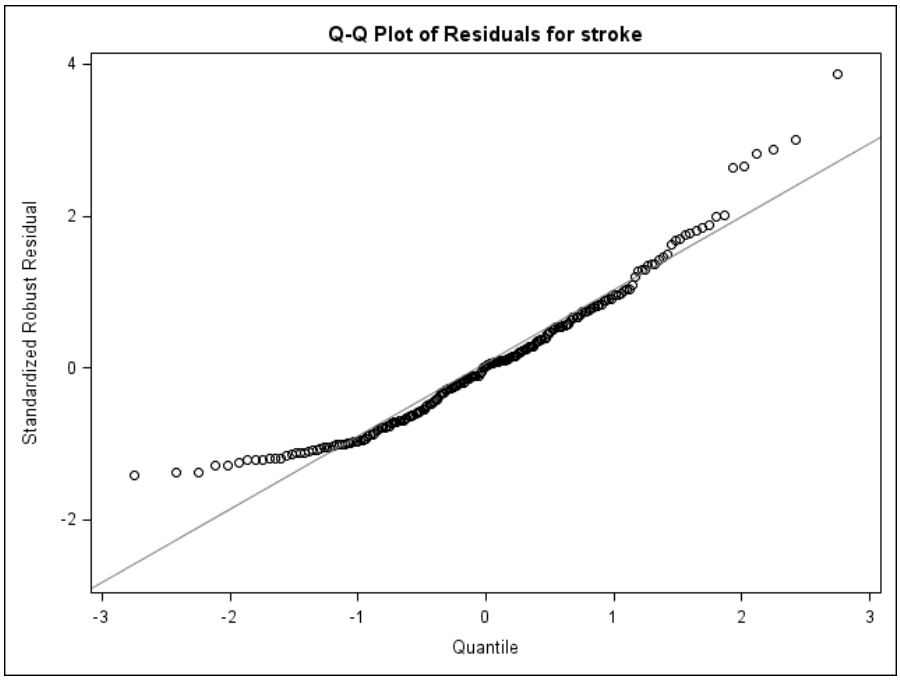
Source	Deviance	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
Intercept	238.4398						
PM_{10}	237.8440	1	207	0.63	0.4266	0.63	0.4257
PM_{10} lag 1	234.0262	1	207	4.07	0.0451	4.07	0.0438
$(PM_{10}$ lag 1$)^2$	229.0578	1	207	5.29	0.0224	5.29	0.0214

LR Statistics For Type 3 Analysis

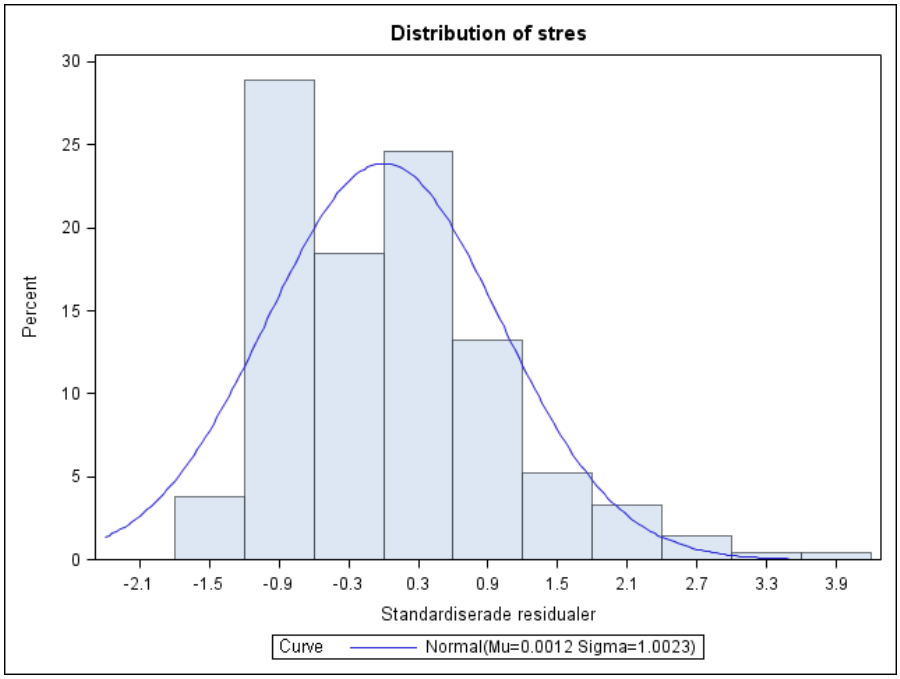
Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
PM_{10}	1	207	4.21	0.0415	4.21	0.0402
PM_{10} lag 1	1	207	7.28	0.0076	7.28	0.0070
$(PM_{10}$ lag 1$)^2$	1	207	5.29	0.0224	5.29	0.0214

Figur D.3 PM_{10}; Proc GENMOD. Kall säsong 2004-2005, zon 6

BILAGA E



Figur E.1 Normalitetsdiagram över residualerna för $PM_{2.5}$. Kall säsong 2004-2005, zon 6



Figur E.2 Histogram över residualerna för $PM_{2.5}$. Kall säsong 2004-2005, zon 6

Goodness-of-Fit Tests for Normal Distribution

Test	----Statistic-----		DF	------p Value------	
Kolmogorov-Smirnov	D	0.1184223		Pr > D	<0.010
Cramer-von Mises	W-Sq	0.4855380		Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	3.8062614		Pr > A-Sq	<0.005
Chi-Square	Chi-Sq	65.2403502	7	Pr > Chi-Sq	<0.001

Tabell E.3 Normalitetstest över residualerna för $PM_{2.5}$. Kall säsong 2004-2005, zon 6

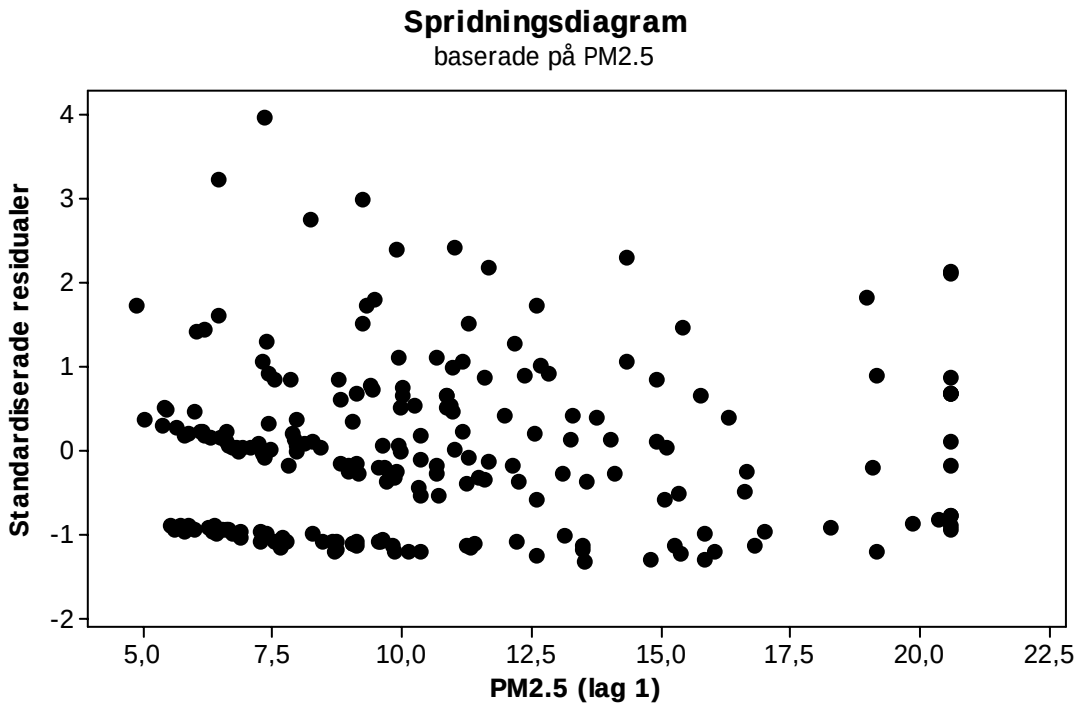


Diagram E.4 Residualdiagram för $PM_{2.5}$ lag 1. Kall säsong 2004-2005, zon 6

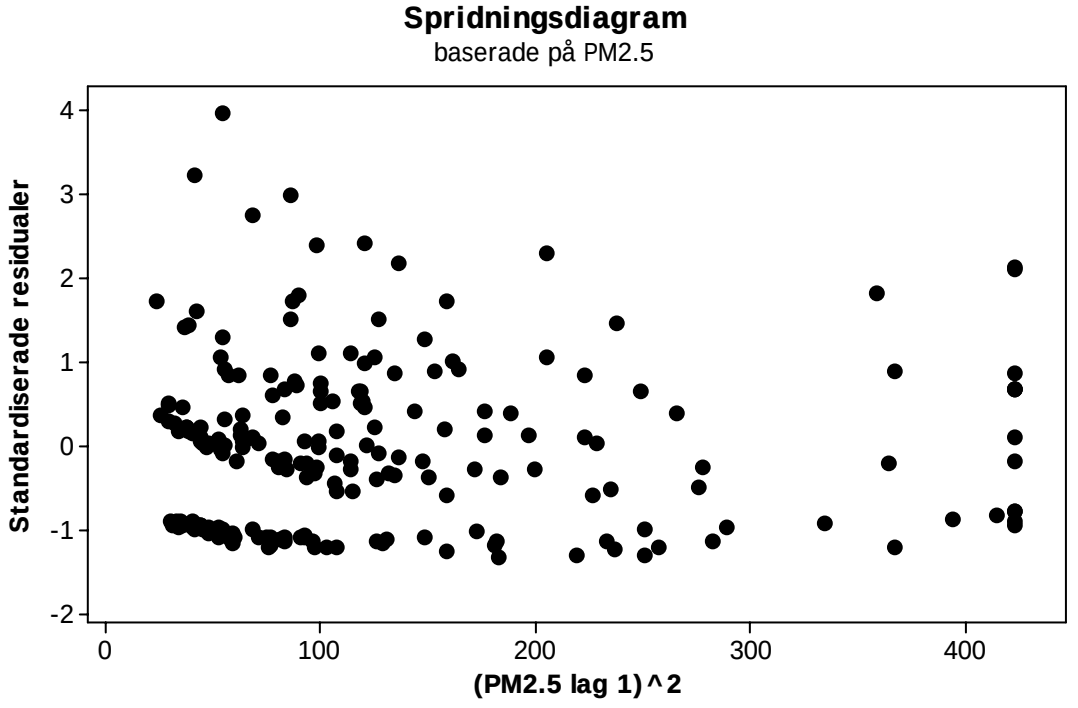


Diagram E.5 Residualdiagram för $(PM_{2.5}$ lag $1)^2$. Kall säsong 2004-2005, zon 6

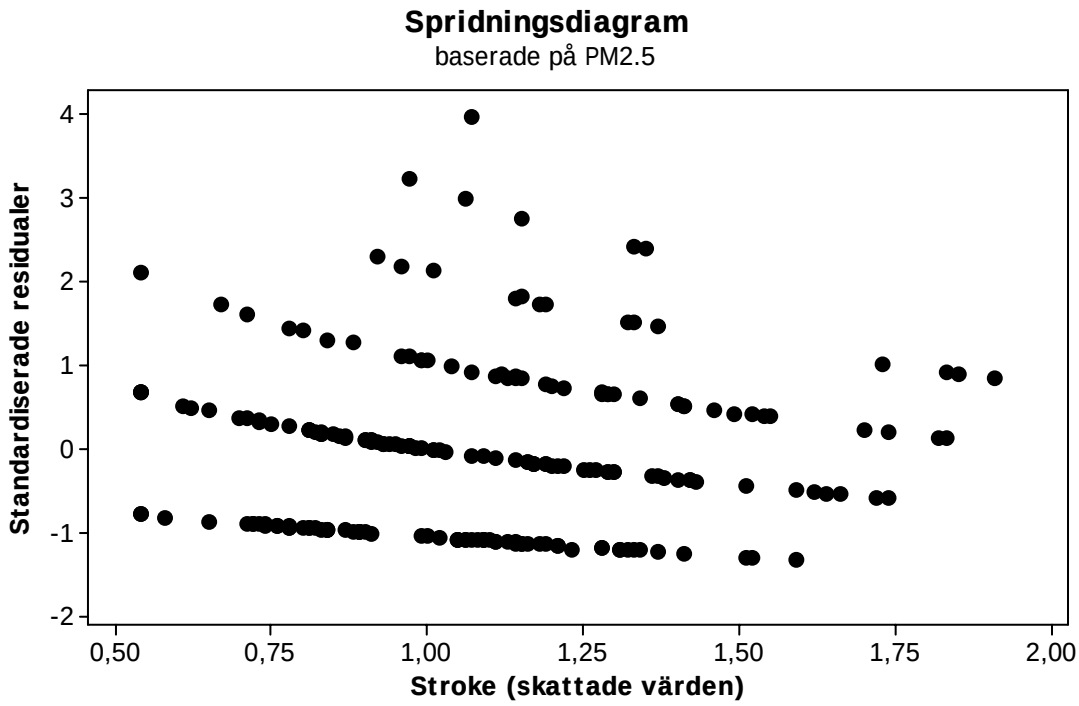


Diagram E.6 Residualdiagram för Stroke (skattade värden). Kall säsong 2004-2005, zon 6

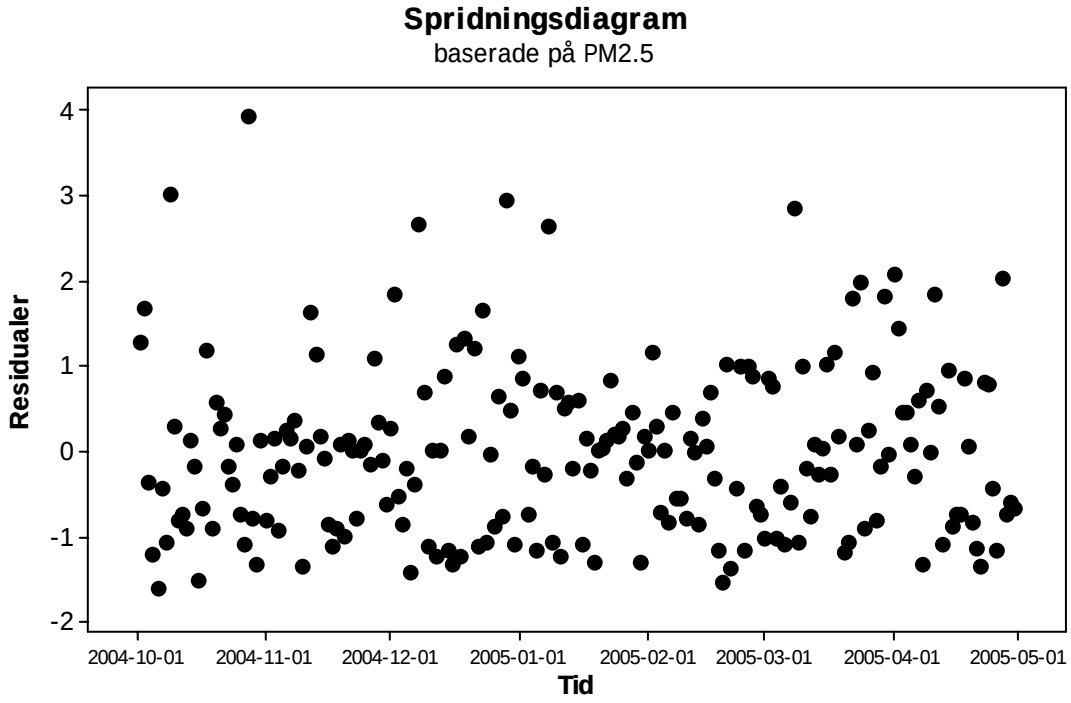
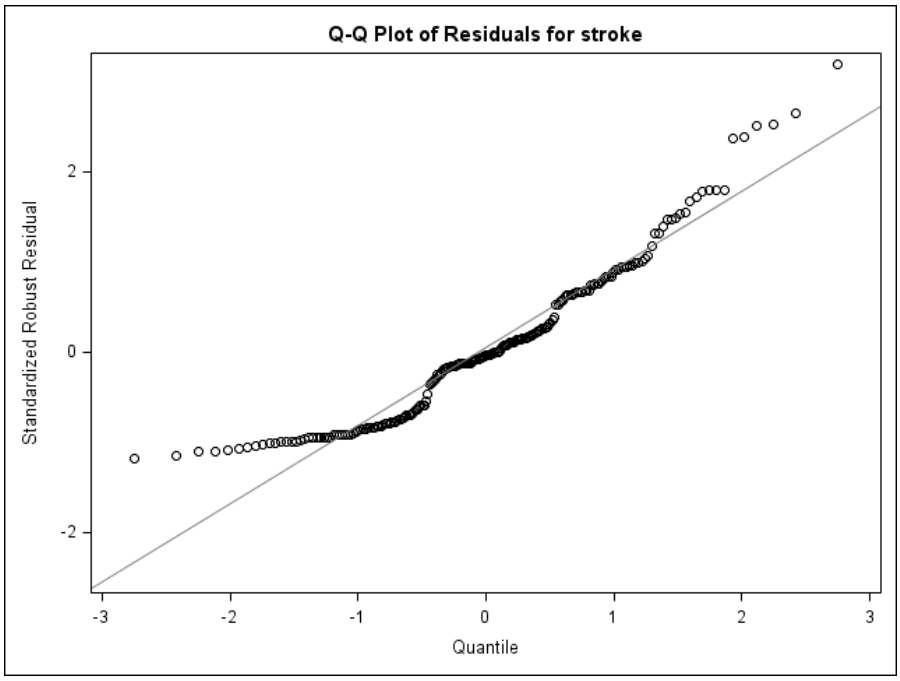
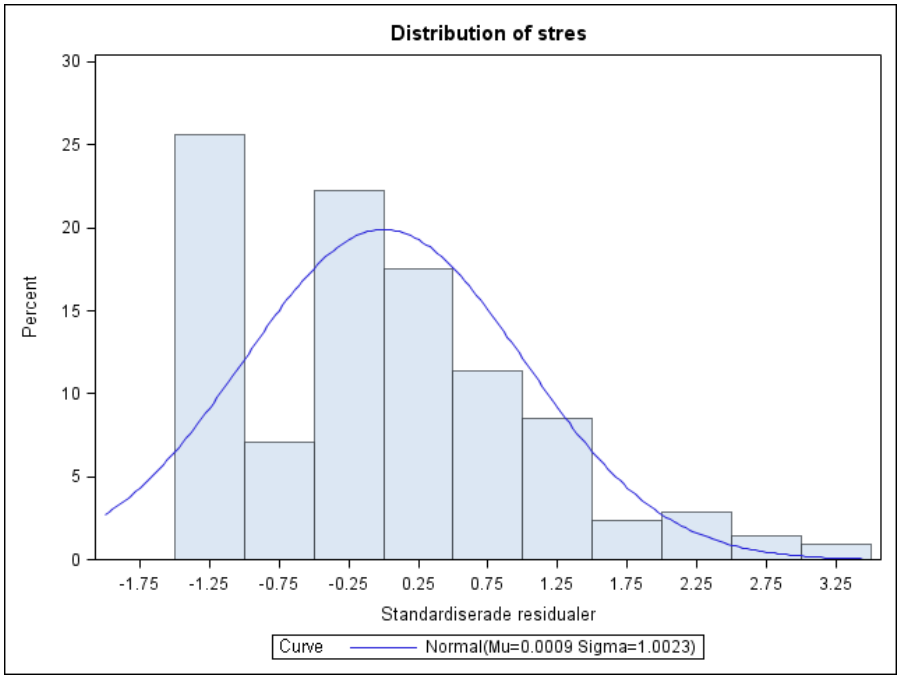


Diagram E.7 Residualdiagram för tiden. Kall säsong 2004-2005, zon 6

BILAGA F



Figur F.1 Normalitetsdiagram över residualerna för PM_{10}. Kall säsong 2004-2005, zon 6



Figur F.2 Histogram över residualerna för PM_{10}. Kall säsong 2004-2005, zon 6

Goodness-of-Fit Tests for Normal Distribution

Test	----Statistic-----		DF	------p Value------	
Kolmogorov-Smirnov	D	0.1350058		Pr > D	<0.010
Cramer-von Mises	W-Sq	0.5549343		Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	4.3019671		Pr > A-Sq	<0.005
Chi-Square	Chi-Sq	94.1136329	7	Pr > Chi-Sq	<0.001

Tabell F.3 Normalitetstest över residualerna för PM_{10}. Kall säsong 2004-2005, zon 6

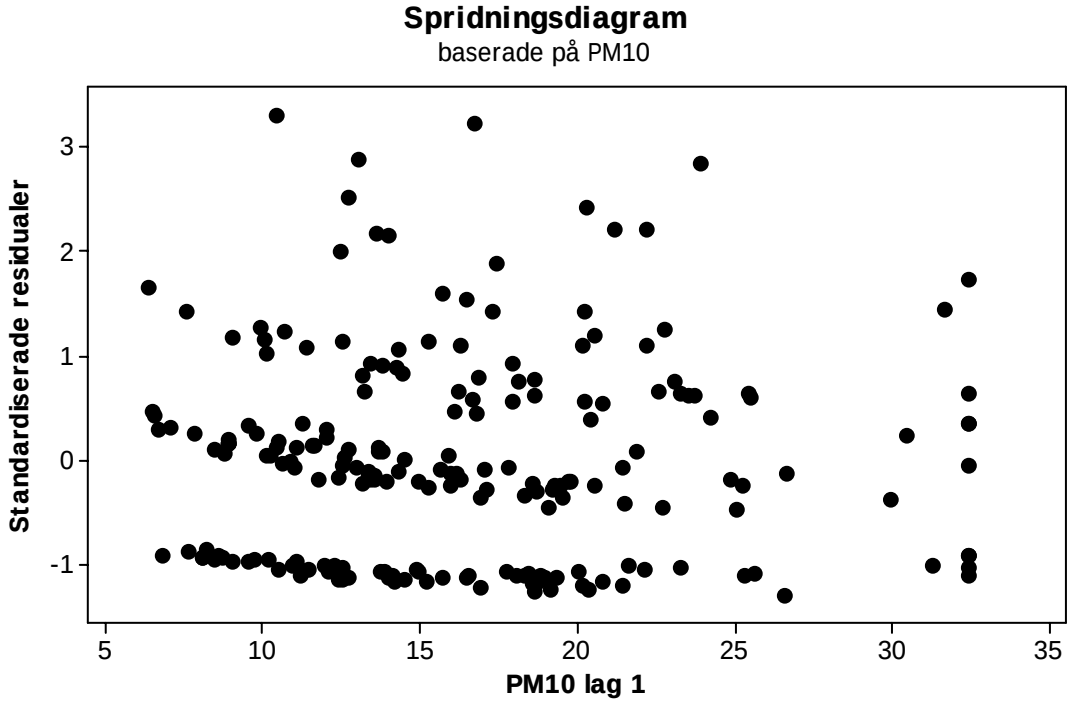


Diagram F.4 Residualdiagram för PM_{10} lag 1. Kall säsong 2004-2005, zon 6

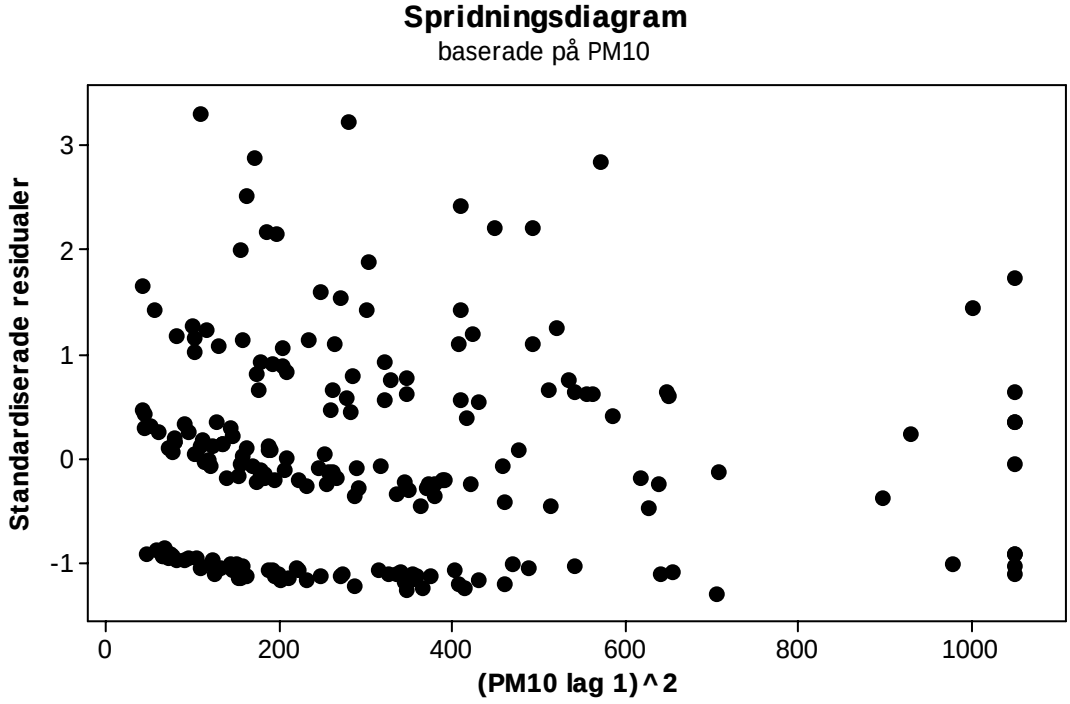


Diagram F.5 Residualdiagram för $(PM_{10}$ lag $1)^2$. Kall säsong 2004-2005, zon 6

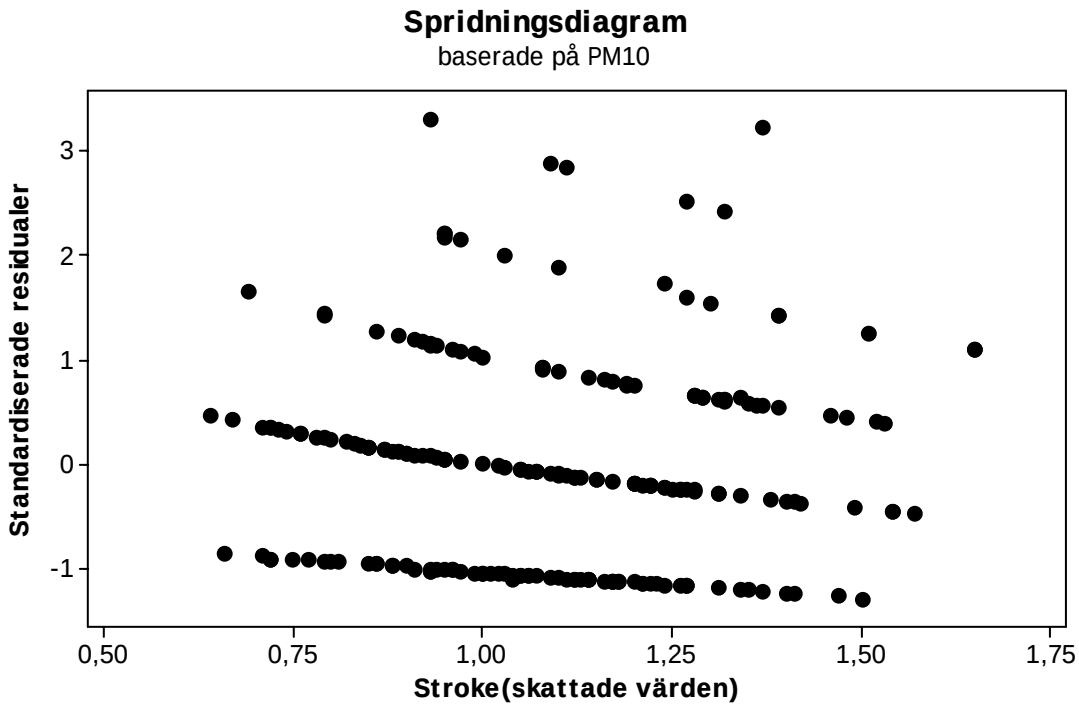


Diagram F.6 Residualdiagram för Stroke (skattade värden). Kall säsong 2004-2005, zon 6

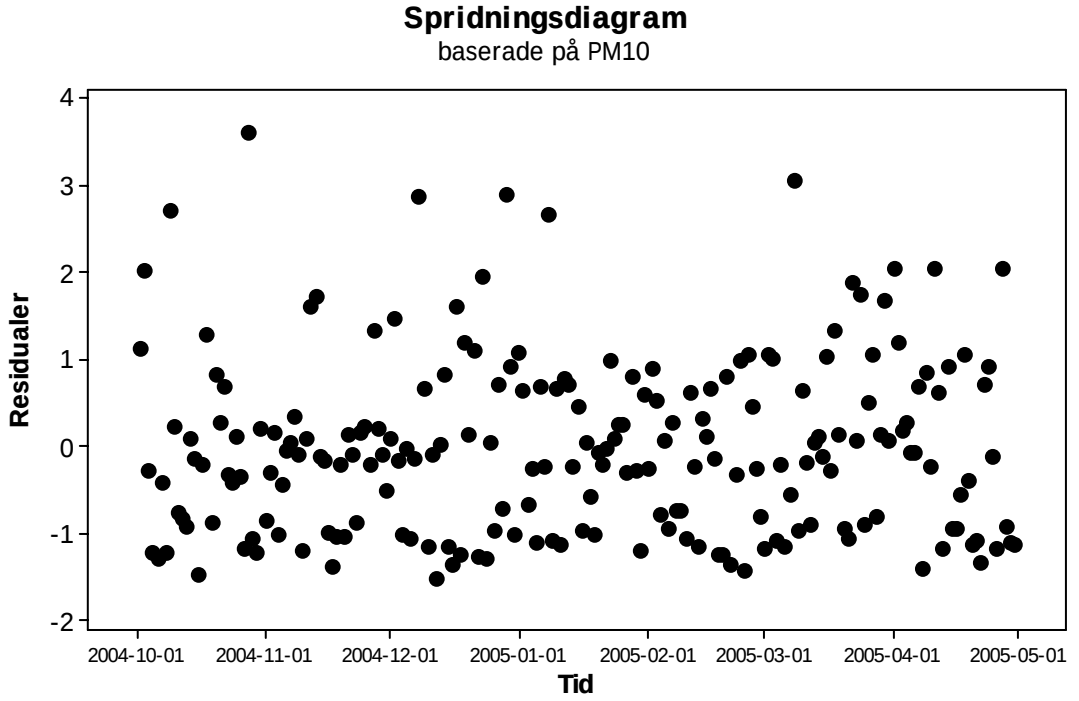


Diagram F.7 Residualdiagram för tiden. Kall säsong 2004-2005, zon 6

BILAGA G

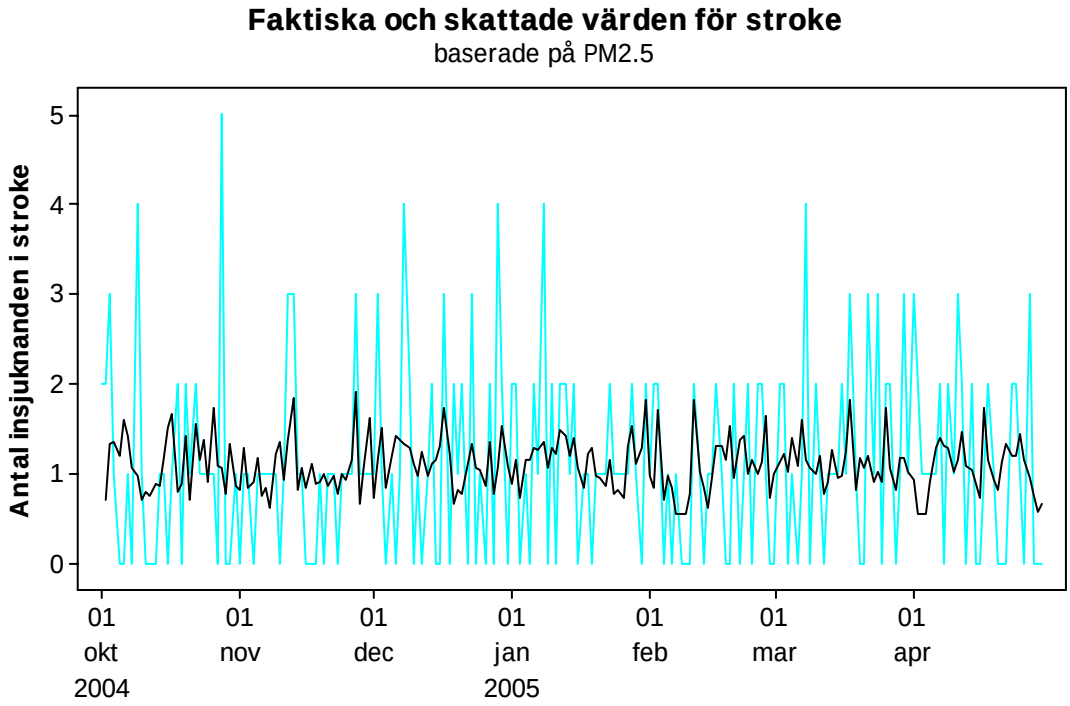


Diagram G.1 $PM_{2.5}$. Kall säsong 2004-2005, zon 6

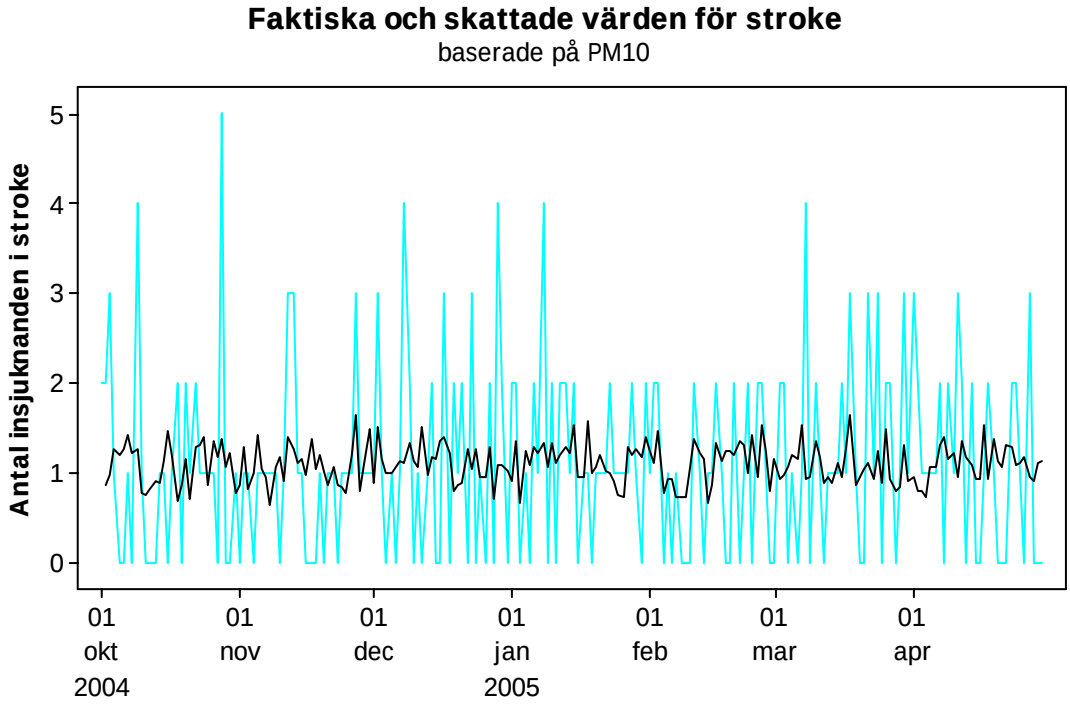


Diagram G.2 PM_{10}. Kall säsong 2004-2005, zon 6