STOCKHOLMS UNIVERSITET
Statistiska institutionen

# Kvällstidningens framtid

The future of tabloid newspapers

Anna Broman
Linus Nyberg Larsson

Uppsats inom Statistik III, ht 2011, 15 högskolepoäng
Handledare: Bertil Wegmann

## Förord

Stort tack till vår handledare Bertil Wegmann, som inspirerade oss att ta oss an ett Bayesianskt förhållningssätt i regressionsanalysen, och som stöttat oss genom arbetet. Vi vill också tacka Expressen som bidragit med datamaterial till uppsatsen.

## Sammanfattning

Syftet med detta arbete är att analysera utvecklingen av kvällstidningsförsäljning sedan internets uppkomst. Tidningen Expressen har fått tjäna som exempel för denna kategori tidningar. För att undersöka utvecklingen över tid utförs en tidsserieanalys för perioden 1995-2010 där exponentiella utjämningsmetoder samt ARIMA-modeller används för att hitta den modell som har bäst anpassning för tidsserien samt gör bäst prognoser. Holt-Winters multiplikativa metod visade sig vara mest användbar för detta syfte, och en prognos utifrån denna metod presenteras. Prognosen visade säsongsvariationer där fler tidningar såldes under sommaren och en nedåtgående trend. Vidare undersöks vilka variabler som kan förklara utvecklingen med hjälp av Bayesiansk regressionsanalys för perioden 2002-2009. Analysen görs med en icke-informativ priorfördelning över parametrarna i regressionsanalysen. I regressionsanalysen visar det sig att ett ökat antal besök på expressen.se och prisökningar leder till minskat antal sålda tidningar. Även regressionsanalysen visar att säsongen är viktig för att förklara antalet sålda tidningar då dummyvariabeln som representerar kvartal tre har en tydligt positiv effekt på tidningsförsäljningen.

## Abstract:

The purpose of this paper is to analyse the development of tabloid sales since the emergence of the Internet. The newspaper Expressen will serve as an example for this category of newspapers. To examine trends over time, a time series analysis for the period 1995-2010 is made, in which exponential smoothing techniques and ARIMA models are used to find the model that fits the time series and produces forecasts in the best way. Holt-Winters multiplicative method proved to be most useful for this purpose, and a forecast based on this method is presented. The prognosis showed a seasonal pattern where more papers are sold during the summer, and a down going trend. Furthermore, a Bayesian regression analysis is made in order to explain the causes of the tabloid sales of Expressen. The regression is made for the period 2002-2009. The regression analysis is made with a non-informative prior distribution for the parameter in the regression model. It shows that a greater number of visits to the website of the tabloid in addition to increases in the price level of the tabloid leads to a reduced number of copies sold. The regression analysis confirms the findings that the season is important for explaining the number of copies sold since the Dummy variable for the third quarter show a positive effect on the tabloid sales of Expressen.

# Innehållsförteckning

# 1. Inledning

## *1.1 Bakgrund*

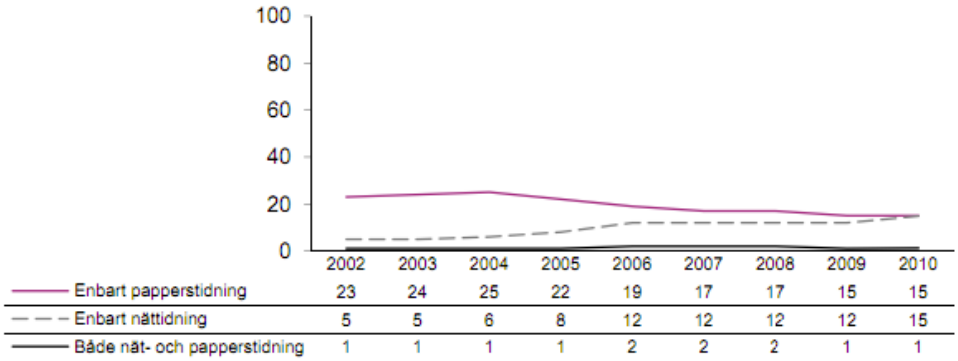
Under hela den moderna historien har den tryckta tidningen varit en självklar del i distributionen av nyheter. Människors intresse av att få veta vad som händer i världen och i den egna kommunen har drivit på efterfrågan, och tidningen har funnits på var mans köksbord. Emellertid har nyhetstidningen idag i många fall utvecklats till något annat än en pappersprodukt man håller i handen. Många väljer att hämta sina nyheter från någon tidnings hemsida på internet, för att snabbt tillgodogöra sig vad som har hänt de senaste timmarna. Enligt Färdigh och Westlund (2011) har nättidningen passerat papperstidningen i andel regelbundna läsare om man tittar på segmentet kvällstidningar, och därmed ökat kvällspressens sammanlagda räckvidd. Frågan är om den ökade användningen av nättidningar påverkat lösnummerförsäljningen av kvällstidningar negativt, eller om det exempelvis bara är ett sätt att få uppdateringar av det man läst i papperstidningen? Är tillgången till nyheter på internet ett komplement eller ett substitut för papperstidningar?

I detta arbete kommer försäljningssiffror för tidningen Expressen att analyseras för att försöka besvara dessa frågor. Expressen har valts som studieobjekt av flera anledningar. Studier (Internetbarometern 2010, 2011, Bergström och Wadbring, 2010, Färdigh och Westlund, 2011, m fl) indikerar att olika generationer gör olika val av tidningsformat. Unga människor är mer benägna att använda webbtidningar. Detta kan bero på att de inte etablerat lika starka medievanor, vilket gör dem mer benägna att ta till sig nya kanaler, och att de är uppvuxna i ett annat medielandskap. Unga människor (15-29 år) läser också kvällstidningar (om man slår ihop pappers och webb-versionerna) i högre utsträckning än äldre, som i högre utsträckning läser morgontidningar (Bergström och Wadbring, 2010, s 384). Detta indikerar att det är för kvällstidningar som en eventuell minskning av papperstidningar till förmån för webb-versioner kommer att synas först. Ett annat skäl att titta på just kvällstidningar är att kvällstidningen är lösnummersåld och inte prenumererad, vilket gör att formatet är mer känsligt för förändringar i användningsmönster än prenumererade morgontidningar (Färdigh och Westlund, 2011). I Sverige finns det två stora aktörer på kvällstidningsmarknaden; Expressen[1] och Aftonbladet (Färdigh och Westlund, 2011). Istället för att slå samman dessa till en variabel kommer endast försäljningen av Expressen att undersökas närmre för att möjliggöra en analys av hur ökad försäljning av den ena påverkar den andra. På så vis kan man se om det finns en korrelation mellan dessa, dvs. om inbördes konkurrens mellan papperstidningarna är det som påverkar försäljningen snarare än nättidningarnas framväxt.

Det finns indikationer på att ökat intresse på webben för ett medium också leder till ökat intresse för mediet utanför nätet; Studier av nedladdningens effekter på biobesök och DVD-försäljning/uthyrning har visat att när nedladdningarna ökat har antalet biobesök varit relativt konstant medan DVD-försäljningen ökat. Om man granskar hur stor andel av befolkningen som sökt sig till en biograf under året kan man till och med se en ökning sedan början av 90-talet. Orsaken kan tänkas vara att den ökade nedladdningen har lett till ett ökat filmintresse i allmänhet, (Ghersetti, 2008).

1 I den officiella statistiken ingår Expressen, KvP och GT i denna grupp. Dessa tidningar motsvarar de lokala utgåvorna av Aftonbladet och räknas alltså som samma tidning.

En studie av medievanor har gjorts av Nordicom. I deras Internetbarometer från 2010 (Nordicom-Sveriges Internetbarometer 2010, 2011) visar de att andelen av befolkningen som enbart läser kvällstidning på nätet är 16%, en genomsnittlig dag, vilket motsvarar andelen som endast läser kvällstidning i pappersform. I samma publikation konstaterar de en negativ trend för kvällstidningsläsandet i pappersform, samtidigt som de ser den omvända trenden för kvällstidningar på nätet. Detta visas i Figur 1.1, som är hämtad från ”Sveriges internetbarometer 2010”. Barometern görs utifrån telefonintervjuer med ett slumpmässigt individurval.




| | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
|---|---|---|---|---|---|---|---|---|---|
| Enbart papperstidning | 23 | 24 | 25 | 22 | 19 | 17 | 17 | 15 | 15 |
| Enbart nättidning | 5 | 5 | 6 | 8 | 12 | 12 | 12 | 12 | 15 |
| Både nät- och papperstidning | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 |

**Figur 1.1 Andelen av befolkningen som läser pappers- eller nättidningar regelbundet**

Figur 1.1 antyder att de effekter som finns mellan nedladdning och biobesök enligt Ghersetti (2008) inte har sin motsvarighet när det gäller tidningar. De som både läser tidningen i pappersform och på internet ser inte ut att öka. Papperstidningen är den distributionskanal som minskar. Andra studier (Dutta-Bergman, 2004) visar dock på att papperstidningen inte är på väg att försvinna och ersättas av nättidningar; istället går utvecklingen mot ett skifte där traditionell media blir ett komplement till internet för att leverera nyheter. Denna slutsats stöds av också av Färdigh och Westlund (2011).

I det här arbetet ställs frågan om hur utvecklingen har sett ut för papperstidningen och om det finns anledning att tro att den utvecklingen kommer att fortsätta. Det finns skäl att tro att det bakom en trend finns flera olika faktorer som driver på utvecklingen. Internets framväxt, hur konkurrensen ser ut, hur mycket tidningen kostar och kanske också vilken säsong det är, eftersom semestrar skulle kunna innebära att man köper mer tidningar. Vilken av dessa faktorer är viktigast?

I detta arbete vill vi bland annat med hjälp av tidsserieanalytiska verktyg modellera och prognosticera kvällstidningarnas i Sveriges månatliga utveckling vad gäller lösnummerförsäljning mellan 1995 och 2009. De flesta undersökningar kring medievanor och tidningsläsande bygger på värden baserade på intervjuer och självskattning för att bedöma utvecklingen (Bergström och Wadbring, 2010, Internetbarometern 2010, 2011, Färdigh och Westlund, 2011). I dessa fall används också företrädesvis årsdata för tidningskonsumtionen. I detta arbete används istället datamaterial över antalet sålda tidningar per månad. På detta sätt undviker vi felkällor som stammar från exempelvis under- eller överskattningar från respondenterna, ram/täcknings-problem eller bortfallsfel etc. Nackdelen är att man inte kan dra några

slutsatser om *vem* som köper tidningen, vilket hade funnits tillgängligt om surveydata hade använts. Användningen av månadsdata möjliggör också att undersöka om denna tidsserie består av trender och om det finns säsongsfaktorer som påverkar tidningsförsäljningen. Om dessa faktorer är funna kommer de att användas för att prognostisera tidningsförsäljningen. Detta utifrån antagandet att det är möjligt att förklara framtida utveckling på samma sätt som man förklarar det som redan hänt.

Utöver detta väljer vi att undersöka vilka faktorer som verkar förklara försäljningssiffror av Expressen genom att därefter att göra en Bayesiansk regressionsanalys. I analysen används ett antal oberoende variabler som kan vara viktiga för att förklara utvecklingen man sett i tidsserieanalysen. Regressionsanalysen kommer att begränsas till tidsperioden 2002-2009.

## *1.2 Syfte*

Syftet med uppsatsen är
att modellera hur försäljningen av Expressen har förändrats över tid sedan 1995 med lämpliga modeller för tidsserier.
att prognostisera försäljningens framtida utveckling med hjälp av den bäst lämpade tidsserieanalysmodellen.
att undersöka vilka faktorer som påverkat utvecklingen med hjälp av Bayesiansk regressionsanalys.

## *1.3 Avgränsning*

I detta arbete begränsas undersökningen av tidningsförsäljningen till kvällspress. I Sverige har vi två stora aktörer - Expressen och Aftonbladet. Tidsserieanalysen och regressionsanalysen kommer att utföras på Expressen. Försäljningssiffror av Aftonbladet kommer endast att användas i regressionsanalysen som förklaringsvariabel för att undersöka sambandet mellan de båda tidningarnas försäljningssiffror. Regressionsanalysen kommer att begränsas till tidsperioden 2002-2009. I regressionsanalysen har de oberoende variablerna konkurrentens försäljning, besök på tidningens hemsida expressen.se, pris på tidningen, besök på expressen.se och slutligen dummyvariabler för säsong, närmre bestämt för kvartal, valts ut. I detta arbete görs inga försök att undersöka det totala tidningsläsandet om man räknar in alla tidningsformer (webb, gratistidningar, kvällstidningar, morgontidningar etc), eller förändringar av detta över tid. För detta hänvisas till Bergström och Wadbring (2010). Det görs inte heller en bedömning av kvalitativa förändringar av tidningen över tid eller kvalitativa skillnader mellan tryckt press och webb-tidningar. För detta hänvisas till Ghersetti (2011).

## *1.4 Disposition*

I den här uppsatsen kommer vi i Kapitel 2 att redogöra för de metoder som används i arbetet, vilket är den teoretiska referensram vi kommer att använda oss av. Här går vi först igenom och definierar metoder inom tidsserieanalys som vi kommer att använda oss av för att utföra prognoser för försäljningssiffror. Här beskrivs också de kriterier och test vi kommer att använda oss av för att välja lämplig modell. Metodavsnittet fortsätter sedan med metoder för Bayesiansk regressionsanalys för att bestämma vilka faktorer som påverkar kvällstidningarnas omsättning.

I Kapitel 3 beskriver vi vårt datamaterial för tidsserien visuellt och beskriver också de transformeringar som är nödvändiga för att analysera tidsserien. Övriga variabler som används i den Bayesianska regressionsanalysen presenteras och motiveras. För dessa presenteras också olika transformeringar, exempelvis av veckodata till månadsdata etc.

Kapitel 4 är en redogörelse för de resultat som vår analys resulterat i. Här presenteras lämpliga modeller och prediktioner utifrån vår tidsserieanalys. I kapitlet presenteras också medelvärdet för parametrarna i den Bayesianska regressionsanalysen fram, och en modell föreslås med de signifikanta variablerna.

I Kapitel 5 ges en sammanfattning av och en redogörelse för våra slutsatser utifrån resultaten vi fått fram, samt förslag på vidare forskning.

# 2. Metod

Tidsserieanalys är ett verktyg för att undersöka mönster i utvecklingen av en variabel. Genom hela uppsatsen görs antagandet att tidsserien har genererats genom en stokastisk process i diskret tid, vilket innebär att man undersöker händelser med jämna tidsintervall som har slumpmässiga inslag. Alla observationer för $Y_t$ antas vara dragna slumpmässigt ur en sannolikhetsfördelning. Genom att beskriva hur tidsserien ser ut utifrån detta antagande kan man dra vissa slutsatser om sannolikheter för olika framtida skeenden, och på så vis göra prognoser för framtida värden. Man kan däremot inte ta fram den exakta sannolikhetsfunktionen, och således kan man bara skapa förenklade modeller som kan förklara det slumpmässiga beteendet (Pindyck & Rubinfeld, 1991, s 440).

Vanligtvis görs detta genom att dekomponera tidsserien till komponenterna trend, T, säsong, S, cykler, C, samt slumpfaktorer eller irregulariteter, I. Det finns additiva och multiplikativa modeller.

Om modellen lyckas förklara de mönster som syns i tidsserien genom de tre första komponenterna tillräckligt bra kommer slumpfaktorn eller feltermen, ε, endast att bestå av vitt brus. Feltermen antas vara normalfördelad med medelvärde 0 och varians $\sigma_\varepsilon^2$ och kovariansen mellan $\varepsilon_t$ och $\varepsilon_{t+k}$ antas vara 0 för k>0, där k = antalet tidsenheter (Montgomery m fl., 2008, s 171).

Vitt brus-processen, $Y_t = \varepsilon_t$, är stationär. Det innebär att den har samma förväntade medelvärde, varians och kovarians genom hela tidsserien oberoende av tidpunkt (Pindyck & Rubinfeld, 1991, s. 445).

För många typer av modeller som är vanliga inom tidsserieanalys är stationäritet en förutsättning. Då de flesta tidsserier inte genereras utifrån en stationär process, utan uppvisar trender, säsong etc måste dessa tidsserier ofta transformeras för att kunna prognostiseras genom en enkel linjär modell (Pindyck & Rubinfeld, 1991, s 444). Det finns skäl att misstänka att försäljningen av Expressen påvisar såväl trend som säsong, vilket i så fall kommer att behöva hanteras.

För att undersöka om en tidsserie är stationär kan man undersöka den så kallade autokorrelationsfunktionen, ACF, som beskriver den underliggande stokastiska processen. Autokorrelationen för lag k är korrelationen mellan två stokastiska

variabler, $Y_t$ och $Y_{t+k}$, för tidsserien på tidsavstånd k, dvs kovariansen delat på variansen, och skattas med

$$\hat{\rho}_k = \frac{\sum_{t=1}^{T-k}(Y_t-\bar{Y})(Y_{t+k}-\bar{Y})}{\sum_{t=1}^{T}(Y_t-\bar{Y})^2} \quad (2.1)$$

För en stationär tidsserie minskar $\hat{\rho}_k$ snabbt i takt med att tidsavstånden ökar. För den stationära tidsserien är ACF alltså endast en funktion av k och det spelar ingen roll var i tidsserien värdena ligger. Beroende på om man har korta eller långa tidsberoenden i serien kan dock ACF avta med olika takt - snabbare för korta beroenden. Detta är dock inte fallet för en icke-stationär tidsserie. Här ser man istället ett långsamt avtagande mönster där autokorrelationen minskar endast lite för varje tidsavstånd. Anledningen till detta är att autokorrelationen för flertalet laggar k blir hög för värden som ligger på en gemensam trend. Om serien visar sig vara icke stationär behöver den således transformeras. Ofta undersöks också partiella autokorrelationsfunktioner. Där mäter man exempelvis den partiella autokorrelationen mellan $Y_t$ och $Y_{t-2}$ när effekten av $Y_{t-1}$ är borttagen, eftersom de båda kringliggande värdena båda antas ha ett nära samband med mittenvärdet.

För en stationär tidsserie skall ACF endast vara en funktion av *k,* dvs. det skall inte spela någon roll *var* i tidsserien de två värdena ligger utan endast vilket tidsavstånd det är mellan dem (Montgomery m fl., 2008, s 28ff).

Andra delen av arbetet kommer att behandla frågan om vilka faktorer det är som har störst betydelse för utvecklingen som analyserats i arbetets första del. För detta kommer Bayesiansk regressionsanalys att användas. Bayesiansk inferens skiljer sig på flera sätt från klassisk inferens, och valet att använda sig av denna metodik än den klassiska bygger på just dessa skillnader.

Bayesiansk inferens betraktar parametrarna som stokastiska variabler, vilket ger möjligheten att ta fram fördelningar för parametrarna. Detta gör i sin tur att man kan uttala sig om sannolikheter för olika utfall, vilket man inte har möjlighet att göra i den klassiska inferensen.

Eftersom det finns teorier som indikerar såväl positiva som negativa effekter av exempelvis ökat antal besök på expressen.se har vi inte på förhand en uppfattning om vilka värden på parametrarna som är mest sannolika, utan använder oss av en likformig fördelning som utgångspunkt, så kallad icke-informativ (likformig) *priorfördelning*.

## *2.1 Exponentiell utjämning*

För att reducera volatiliteten i tidsserien används ofta olika utjämningstekniker. Exponentiell utjämning möjliggör enklare visuell analys av ett datamaterial. (Pindyck and Rubinfeld, 1991). Tre olika utjämningsmetoder kommer att testas. Dessa är relevanta vid olika typer av tidsserier. Eftersom de olika teknikerna som behandlas i detta avsnitt; EEU, Holts metod och Holt-Winters metod, är utformade för att hantera olika typer av serier (stationär tidsserie, tidsserie med trend respektive tidsserie med trend och säsong) kommer förmodligen endast en av metoderna att vara lämplig för just den tidsserie som analyseras här.

### 2.1.1 Enkel exponentiell utjämning (EEU)

EEU används framför allt för tidsserier som saknar trend eller säsongskomponenter. Metoden är användbar för att göra enkla enstegsprognoser, men då prognosen inte tar hänsyn till eventuell trend eller säsong blir prognosskattningen densamma för alla framtida tidpunkter. Utjämningsekvationen är följande:

$$\tilde{Y}_t = \alpha Y_t + (1-\alpha)\tilde{Y}_{t-1} \quad (2.1.1)$$

Ett mindre α ger större utjämning. Utjämningskonstanten α väljs genom att minimera följande uttryck:

$$min \sum (Y_t - \tilde{Y}_{t-1})^2 \quad (2.1.2)$$

Detta görs automatiskt i programvaran SAS Time Series Forecasting System som kommer att välja det optimala α, om man inte själv väljer ett α för att man vill ha starkare eller svagare utjämning.

Slumptermen, ε, är här exkluderad då den antas vara 0.

För att ta fram det första värdet, $\widehat{Y}_0$, som behövs för att göra utjämningen kan man till exempel använda sig av ett medelvärde av de första observationerna (Montgomery m fl., 2008). Beräkningen görs i SAS Time Series Forecasting System, och programmet använder sig av medelvärdet av några av de första observationerna.

### 2.1.2 Holts metod

Holts metod är en utveckling av EEU, och används när det finns trend i tidsserien (Pindyck & Rubinfeld, 1991). Metoden kallas ibland också för dubbel exponentiell utjämning. Här används två ekvationer som beror på två utjämningsparametrar, α och $\gamma$ som ligger mellan 0 och 1. I ekvationerna är r = trenden och $Y_t$= nivån:

$$\tilde{Y}_t = \alpha Y_t + (1-\alpha)(\tilde{Y}_{t-1} + r_{t-1}) \quad (2.1.3)$$

$$r = \gamma(\tilde{Y}_t - \tilde{Y}_{t-1}) + (1-\gamma)r_{t-1} \quad (2.1.4)$$

Ekvationen (2.1.3) ger således ett uttryck för nivån och (2.1.4) ger den utjämnade trenden. Startvärden skattas exempelvis genom att man gör en enkel linjär regression med Y som beroende variabel och tiden som oberoende variabel. För att bestämma $\tilde{Y}_0$ används värdet för det skattade interceptet och för $\hat{r}_0$ den skattade lutningen.

Prognosen med hjälp av Holts metod blir alltså

$$\hat{Y}_{T+l} = \tilde{Y}_T + lr_T \quad (2.1.5)$$

där l är prognospunkten. Exempelvis ger *l* = 1 prognosen för en tidsenhet framåt i tiden. Prognosen riskerar dock att bli något godtycklig, då det saknas en fullödig metod för att avgöra storleken på utjämningsparametrarna.

### 2.1.3 Holt-Winters metod

Holt-Winters metod inkluderar ytterligare en parameter för att hantera säsongs- eller cykliska inslag i datamaterialet. Det finns såväl en additiv som en multiplikativ metod som används beroende på om säsongsvariationerna är proportionerliga till nivån eller inte. Här används tre ekvationer: nivå (intercept), trend (lutning) och säsong. För den additiva metoden gäller följande ekvationer:

$$\tilde{Y}_t = \lambda_1(Y_t - S_{t-s}) + (1-\lambda_1)(\tilde{Y}_{t-1} + r_{t-1}) \quad (2.1.6)$$

$$r_t = \lambda_2(\tilde{Y}_t - \tilde{Y}_{t-1}) + (1-\lambda_2)r_{t-1} \quad (2.1.7)$$

$$S_t = \lambda_3(Y_t - \tilde{Y}_t) + (1-\lambda_3)S_{t-s} \quad (2.1.8)$$

där alla λ ligger mellan 0 och 1 och s = antalet säsonger. Precis som i Holts metod är ett vanligt sätt att skatta startvärden för $\tilde{Y}_0$ och $\hat{r}_0$ med hjälp av en enkel linjär regressionsmodell, men där man nu också lägger till s-1 dummyvariabler. I detta fall blir det alltså 11 dummyvariabler eftersom det är månadsdata som används. Startvärden för säsongen skattas genom dummyvariablernas parametrar eller genom att ta $Y_t/(T_t+C_t) = S_t+I_t$ vilket ger $\hat{S}_t$.

Prognosen blir

$$\hat{Y}_{T+l} = (\tilde{Y}_T + lr_T) + S_{t+l-s} \quad (2.1.9)$$

För den multiplikativa modellen ser ekvationerna istället ut som nedan, nivån och säsongen förändras något, medan trend-ekvationen förblir densamma.

$$\tilde{Y}_t = \lambda_1(\frac{Y_t}{S_{t-s}}) + (1-\lambda_1)(\tilde{Y}_{t-1} + r_{t-1}) \quad (2.1.10)$$

$$S_t = \lambda_3(\frac{Y_t}{\tilde{Y}_t}) + (1-\lambda_3)S_{t-s} \quad (2.1.11)$$

Prognosen blir

$$\hat{Y}_{T+l} = (\tilde{Y}_T + lr_T) \times S_{t+l-s} \quad (2.1.12)$$

## *2.2 ARIMA*

ARIMA står för Autoregressive Integrated Moving Average. För att ta fram en ARIMA-modell krävs att tidsserien är stationär, dvs. att den ska ha konstant väntevärde och varians. För att uppnå detta måste originalserien ofta differentieras ett antal gånger (Pindyck & Rubinfeld, s 472f, 1991). Tidsserien beskrivs alltså i denna modell som ett medelvärde och avvikelserna kring detta.

Om tidsserien som analyseras i detta arbete innehåller trend och säsong krävs olika metoder för att göra den icke-stationära tidsserien stationär. Utifrån den stationära tidsserien kommer sedan prognoser att utföras.

### 2.2.1 AR

Autoregressiva modeller, eller AR(p)-modeller, består av ett viktat medelvärde av tidigare observationer i tidsserien som går tillbaka p perioder i tiden, samt en felterm

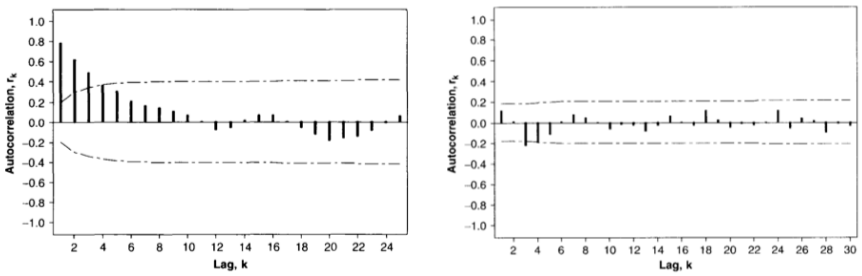
för den innevarande perioden. Den obestämda AR-modellen (Pindyck & Rubinfeld, s 478, 1991) skrivs

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \delta + \varepsilon_t \quad (2.2.1)$$

där $\phi_i$ är parametrar och $\varepsilon_t$ liksom tidigare är en slumpterm. I (2.2.1) förhåller sig δ till medelvärdet, μ, som

$$\mu = \frac{\delta}{1-\phi_1-\phi_2-\cdots-\phi_p} \quad (2.2.2)$$

μ måste vara ändlig om tidsserien är stationär. Alltså måste $\Phi_1+\Phi_2+\ldots+\Phi_p < 1$. För att vidare kontrollera för stationäritet används autokorrelationsfunktionen där man kan kontrollera visuellt att autokorrelationen avtar exponentiellt eller liknar en dämpad sinuskurva (Montgomery m fl. s256, 2008). Exempel på två olika stationära AR-modeller visas i Figur 2.1.




**Figur 2.1 Stationäritet i en AR-modell (Bilder tagna från Montgomery m fl, s 32f, 2008)**

Till och med för en AR(1)-modell, där man tittar på en period tillbaka i tiden ($Y_t = \Phi_1 Y_{t-1} + \delta + \varepsilon_t$) visar det sig att autokorrelationsfunktionen har ett oändligt minne, och alltså beror på samtliga tidigare observationer, även om betydelsen av dessa avtar över tid, vilket syns tydligast i den första bilden i figur 2.1. Däremot är värdet för den partiella autokorrelationen signifikant endast på tidsavstånd 1 i en AR(1)-modell, och 1 till och med p i en AR(p)-modell, därefter är den 0.

Autokorrelationen är autokovariansen, $\gamma_k$, genom variansen $\gamma_0$. Autokorrelationen beräknas för exempelvis en AR(2) modell (Pindyck m. fl. s 479, 1991) enligt nedan.

Först beräknas variansen på den stationära tidsserien som

$$\gamma_0 = E[(\phi_1 y_{t-1} + \varepsilon_t)^2] = E(\phi_1^2 y_{t-1}^2 + \varepsilon_t^2 + 2\phi_1 y_{t-1}\varepsilon_t) = \phi_1^2\gamma_0 + \sigma_\varepsilon^2 \quad (2.2.3)$$

$$\gamma_0 = \frac{\sigma_\varepsilon^2}{1-\phi_1^2} \quad (2.2.4)$$

Därefter kan autokovariansen beräknas för AR(2)-modellen som

$$\gamma_2 = E[y_{t-2}(\phi_1^2 y_{t-2} + \phi_1\varepsilon_{t-1} + \varepsilon_t)] = \phi_1^2\gamma_0 = \frac{\phi_1^2\sigma_\varepsilon^2}{1-\phi_1^2} \quad (2.2.5)$$

Den generella autokovariansen för en AR(p)-modell är alltså

$$\gamma_k = \phi_1^k \gamma_0 = \frac{\phi_1^k \sigma_\varepsilon^2}{1-\phi_1^2} \quad (2.2.6)$$

Autokorrelationsfunktionen är således

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \phi_1^k \quad (2.2.7)$$

### 2.2.2 MA

I glidande medelvärdesmodeller, eller MA(q)-modeller beskrivs processen som den viktade summan av nuvarande och laggade slumptermer som går tillbaka q perioder i tiden samt medelvärdet $\mu$. Då slumptermerna utgörs av vitt brus är dessa oberoende av varandra. MA-modeller är således alltid stationära. Den obestämda MA(q)-modellen (Pindyck & Rubinfeld, s 473, 1991) skrivs

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (2.2.8)$$

där varje felterm är en oberoende variabel där väntevärdet är 0, variansen är $\sigma_\varepsilon^2$ och kovariansen mellan $\varepsilon_t$ och $\varepsilon_{t+k}$ är 0 för $k \neq 0$. Processen beskrivs alltså med q + 2 parametrar; μ, $\sigma_\varepsilon^2$ och $\theta_1,\ldots,\theta_q$. För att tidsserien ska vara stationär krävs att summan av de kvadrerade $\theta_i$ är ändlig, dvs.

$$\sum_{i=1}^{q} \theta_i^2 < \infty$$

Om q går mot oändligheten krävs att summan ovan konvergerar. Detta kommer oftast att ske om $\theta_i$ blir mindre för tidpunkter längre tillbaka, det vill säga för större värden på laggar *i*. För en MA(1)-modell måste

$$|\theta_1| < 1$$

För en MA(2)-modell måste

$$\theta_1 + \theta_2 < 1$$
$$\theta_2 + \theta_1 < 1$$
$$|\theta_2| < 1$$

För en MA(q)-modell gäller det omvända mot en AR(p)-modell; autokorrelationsfunktionen antar signifikanta värden på tidsavstånd till och med q, och är därefter 0. (Montgomery m fl, s 236ff, 2008). Den partiella autokorrelationen å andra sidan avtar exponentiellt eller ser ut som en dämpad sinuskurva om modellen är invertibel[2] (Montgomery m fl. s256, 2008).

Autokorrelationen beräknas för en MA(2)-modell

2 En invertibel MA-process definieras som ett den kan skrivas om som en oändlig AR-process. (Montgomery m fl. s 251f, 2008)

$$\rho_y(1) = \frac{-\theta_1 + \theta_1\theta_2}{1 + \theta_1^2 + \theta_2^2}$$

$$\rho_y(2) = \frac{-\theta_2}{1 + \theta_1^2 + \theta_2^2}$$

$$\rho_y(k) = 0, k > 2 \qquad (2.2.9)$$

### 2.2.3 Att göra tidsserien stationär

För att bli av med trenden i tidsserien och göra den stationär är den vanligaste metoden att man differentierar serien (Montgomery m fl, s 36ff, 2008).

$$x_t = Y_t - Y_{t-1} = \Delta Y_t \qquad (2.2.10)$$

Där $\Delta$, som ofta skrivs (1-B), är bakåtoperatorn.

För tidsserier med säsong utvecklas differentieringsmetoden till

$$\Delta_d Y_t = (1 - B^d)Y_t = Y_t - Y_{t-d} \qquad (2.2.11)$$

I detta papper hanteras månadsdata vilket gör att d = 12. För att göra en sådan tidsserie stationär gör man först säsongsdifferentieringen för att ta bort säsongseffekterna. Därefter gör man den vanliga differentieringen från ekvation [2.2.10] för att ”avtrenda” tidsserien.

Vanligen räcker det med en eller två differentieringar för att göra tidsserien stationär (Montgomery m fl, s 256, 2008). Den differentierade tidsserien kommer genomgående att benämnas $w_t$.

Om variansen för tidsserien ökar över tid kan det i vissa fall också vara lämpligt att utöver differentiering logaritmera originalserien för att den ska bli stationär. Har man en exempelvis multiplikativ modell med säsong och trend skapar en logaritmering av tidsserien konstant varians för säsongsdelen, så att man genom en vanlig differentiering av den logaritmerade serien ofta uppnår stationäritet.

### 2.2.4 ARIMA

ARMA(p,q) är den enklaste formen av en kombination av AR- och MA-modellerna.

$$Y_t = \delta + \sum_{i=1}^{p} \phi_i Y_{t-i} + \varepsilon_t - \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \qquad (2.2.12)$$

För ARMA(p,q)-modeller avtar både autokorrelationsfunktionen, $\rho_k$, och den partiella autokorrelationsfunktionen, $\rho_{kk}$, exponentiellt eller uppvisar ett dämpat sinusmönster. Detta beror på att man i MA-delen bara får signifikanta utslag på tidsavstånd till och med q i autokorrelationsfunktionen, och samma sak gäller för tidsavstånd till och med p i AR-modeller för den partiella autokorrelationsfunktionen (Mongomery m fl, s 254f, 2008). Ett sinusmönster skulle dock göra modellen relativt svår att identifiera eftersom det inte blir tydligt att tidsserien har signifikanta värden upp till p eller q – dvs p och q blir svåra att identifiera. I detta fall ska flera tänkbara modeller att

undersökas för att utvärdera vilken som är mest lämplig utifrån diverse test (Montgomery m fl. s254f, 2008).

För att en ARMA-modell ska vara lämplig så krävs det att tidsserien är stationär. Om tidsserien inte är det krävs differentiering och modellen utvecklas till en ARIMA(p,d,q)-modell, där d:et står för antalet differentieringar som krävts för att göra modellen stationär. Det är således nödvändigt att det är samma antal differentieringar som gör båda delarna av modellen stationära. Det är såväl den allmänna trenden genom tidsserien som säsongstrenden som måste avtrendas.

$$\Phi(B)(1-B)^d Y_t \qquad (2.2.13)$$

När en lämplig ARIMA-modell väljs identifieras en modell, parametrarna skattas och därefter uppskattas modellens lämplighet genom diverse test då man kontrollerar om p, d och q har specificerats korrekt. (Montgomery m fl, s 265, 2008)

Modellen identifieras genom att titta på datamaterialet för att se om det finns en trend. Om trend är närvarande är tidsserien inte stationär och differentieringar bör göras. När man funnit det d där $\Delta^d Y_t$ är stationär ska p och q skattas. Detta görs genom att titta på autokorrelationsfunktionen och den partiella autokorrelationsfunktionen för stickprovet (data). I tidigare avsnitt har det konstaterats att för MA-modeller är autokorrelationen signifikant på tidsavstånd upp till och med q. För AR-modeller är den partiella autokorrelationsfunktionen signifikant på tidsavstånd till och med p. För en ARMA-modell avtar båda exponentiellt. För att i det fallet bestämma p och q undersöks om det finns spikar (signifikanta värden) i autokorrelationsfunktionen eller den partiella autokorrelationsfunktionen (Pindyck & Rubinfeld, s 492, 1991) Detta kallas ofta för Box-Jenkins-metoden för modellidentifiering. Tillsammans med denna visuella bedömning kommer vi med hjälp av de estimat som programvaran SAS Time Series Forecasting System producerar att anta värden på p och q. Genom de diagnostiska test som utförs i steg tre undersöks om de antaganden som gjorts är korrekta, eller om andra värden på p och q ska undersökas. (Montgomery m fl. s 267ff, 2008)

Se avsnitt 2.3 för diagnostiska test.

### 2.2.5 ARIMA med säsong

Om tidsserien uppvisar säsongsmönster, i detta fall handlar det om månadsdata och således månadssäsong, används en ARIMA-modell med en säsongskomponent. Eftersom det rör sig om månadsdata kommer s att vara 12 i detta fall.

Då en tidsserie med säsong inte kan vara stationär måste man transformera tidsserien och göra en differentiering för säsongsdelen (Montgomery m fl, s 282f, 2008). Säsongsberoendet minskar genom en logaritmering om tidsserien är multiplikativ och eventuell trend hanteras också den genom differentiering.

Vanlig differentiering görs som tidigare omnämnts

$$w_t = Y_t - Y_{t-1} = (1-B)Y_t = \Delta Y_t \qquad (2.2.14)$$

För säsongsdelen blir differentieringen istället

$$w_t = Y_t - Y_{t-s} = (1 - B^s)Y_t = \Delta_s Y_t \quad (2.2.15)$$

För en tidsserie som kräver differentiering för tidsserien och även för säsongsdelen får man alltså

$$w_t = (1 - B)(1 - B^s)Y_t = \Delta\Delta_s Y_t = Y_t - Y_{t-1} - Y_{t-s} + Y_{t-s-1} \quad (2.2.16)$$

Modellen skrivs nu ARIMA (p,d,q)(P,D,Q)$_s$ där den sista parentesen handlar om tidsförskjutningarna i AR och MA-delen för säsongsdelen, samt säsongsdifferentieringen. Dessa skattas på samma sätt som för den vanliga ARIMA-modellen genom att undersöka autokorrelationsfunktionen och den partiella autokorrelationsfunktionen för stickprovet (data), men där man nu undersöker effekter av tidsavstånd 12, 24 etc, för att se om det finns ett mönster där dessa sticker ut (Montgomery m fl, 2008, s 283).

## *2.3 Test*

### 2.3.1 Ljung-Box

Ljung-Box teststatistika används för att testa för autokorrelation i feltermerna, och är ett av de viktigaste diagnosiska testen för att testa lämpligheten av modellen. Om ARIMA-modellen skattats korrekt kommer autokorrelationen vid tidsavstånd k att ha väntevärde 0 (givet att modellen är stationär kring 0) och varians 1/T, där T är antalet observationer i tidsserien. Anledningen till detta är att residualerna består av vitt brus och alltså är oberoende normalfördelade variabler om modellen är korrekt. Testet utförs lämpligen på tidsavstånd större än 4 (Pindyck& Rubinfeld, s 505, 1991).

Ursprunget till testet hette från början Box-Pierce, men efter vissa modifieringar av Ljung och Box för att det även ska passa för små urval fick det istället det nuvarande namnet (Montgomery m fl, s 57, 2008).

Nollhypotesen i testet är att $\rho_1=\rho_2=...=\rho_K = 0$ i residualerna och alternativhypotesen blir således att åtminstone någon av termerna är skild från noll. Teststatistikan skrivs

$$Q_{LB} = T(T+2)\sum_{k=1}^{K}\left(\frac{1}{T+k}\right)r_k^2, \quad (2.3.1)$$

där $r_k^2$är Chi2-fördelad med K-p-q frihetsgrader;

$$r_k^2 \sim \chi^2(K-p-q)$$

där T = antalet observationer-differenser och k = tidsavstånd mellan residualerna och K = maximala tidsavståndet mellan residualerna; här kommer 20 att användas.

### 2.3.2 Test för prognosfel

Anpassningen kommer inte att göras utifrån all tillgänglig data, utan två år lämnas. 24 observationer används alltså för att kunna kontrollera hur väl modellen är anpassad för data, dvs en så kallad "out-of-sample"-analys för månaderna under de två sista åren. Genom att göra enstegsprognoser, det vill säga för en tidsperiod framåt i tiden

utifrån tidigare observationer, kan man få fram enstegsprognosfelen (Montgomery m fl, s 50ff, 2008), så som

Mean Absolute Error (MAD),

$$MAD = \frac{1}{n}\sum_{t=1}^{n} e_t\,(1), \tag{2.3.2}$$

där $\mathrm{e}_t(1) = Y_t - \hat{Y}_t$

Mean Squared Error (MSE),

$$MSE = \frac{1}{n}\sum_{t=1}^{n}[e_t\,(1)]^2 \tag{2.3.3}$$

och Mean absolute percent forecast error (MAPE)

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}|re_t\,(1)| \tag{2.3.4}$$

där $\mathrm{re}_t(1)$, relativt prognosfel i procent, är

$$\mathrm{re}_t(1) = \left(\frac{Y_t - \hat{Y}_t(t-1)}{Y_t}\right)100 = \left(\frac{e_t(1)}{Y_t}\right)100 \tag{2.3.5}$$

### 2.3.3 Goodness of fit-test

Test av modellerna utförs för att avgöra vilken modell som är bäst lämpad att prognostisera utifrån. Som nämnts tidigare kommer en så kallad "out of sample"-analys göras. För att undersöka hur väl en statistisk modell passar data är $R^2$-statistikan mycket vanlig. Målet med prognostisering är dock inte att modellen ska passa historisk data så bra som möjligt, utan istället att göra en bra prognos för framtida observationer. Därför kan överidentifiering, det vill säga att använda sig av fler parametrar än vad som är lämpligt, vara ett problem (Montgomery m fl, s 57, 2008). Att använda sig av alltför många parametrar ökar variansen för prognosfelet. Av den anledningen är det bättre att använda sig av det justerade $R^2$-måttet, där man justerar för antalet parametrar i modellen. Ett högt $R^2_{Adj}$-värde antyder att modellen är bra.

$$R^2_{Adj} = 1 - \frac{\sum_{t=1}^{T} e_t^2/(T-p)}{\sum_{t-1}^{T}(y_t - \bar{y})^2/(T-1)} \tag{2.3.6}$$

Akaikes informationskriterium, AIK, anses vara ett ännu bättre mått (Montgomery m fl, s 59, 2008) , då det innefattar ett straffmått för antalet parametrar i modellen och är asymptotiskt effektiv.

$$AIK = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \frac{2p}{T} \tag{2.3.7}$$

Det finns dock en risk att dessa kriterier inte straffar användandet av för många parametrar tillräckligt hårt vilket gör att man väljer en överidentifierad modell istället för en simplare modell som lämpar sig bättre (Montgomery m fl, s 59, 2008). Därför

kan det vara lämpligt att söka sig till andra test så som Schwarz' informationskriterium, SIK, som har ett större straff för användandet av onödigt många parametrar.

$$SIK = \ln\left(\frac{\sum_{t=1}^{T} e_t^2}{T}\right) + \frac{p\ln(T)}{T} \quad (2.3.8)$$

När det gäller bedömningar av informationskriteriemåtten ska man välja modeller som minimerar dessa (Montgomery m fl, s 58f, 2008). I detta arbete kommer alla prognosfelstest och goodness of fit-test att beaktas. I de fall dessa ger motstridigt resultat är framförallt informationskriterierna som kommer att prioriteras, och av dem bedöms SIK vara den viktigaste då denna är konsistent, dvs straffar tillräckligt hårt för tillagda parametrar.

En prognos för kommande månader, dvs. tiden efter november 2010 kommer att göras med de modeller som sammantaget klarat sig bäst i de genomförda testen. För prognoserna kommer allt tillgängligt data att användas.

## *2.4 Bayesiansk statistik*

I den här andra delen av arbetet kommer en regressionsanalys med Bayesiansk metodik genomföras för att undersöka vilka faktorer som påverkar storleken på upplagan.

### 2.4.1 Bayesianskt perspektiv

De stora fördelarna med att använda sig av Bayesiansk statistik är att man får mer information i inferensen än när man använder klassisk statistik. Man kan dra slutsatser om hypoteser på ett annat sätt. Exempelvis kan man uttala sig om att parametern $\theta \geq \theta_0$ med en viss sannolikhet. Anledningen till att man har möjlighet att dra denna slutsats är att man i Bayesiansk statistik betraktar parametrar som slumpmässiga variabler. Detta får i sin tur som konsekvens att man får en fördelning för varje parameter, vilket möjliggör mer långtgående slutsatser i termer av sannolikhet.

Grunden för Bayesiansk statistik lades av Thomas Bayes (1701-1761). Han definierade Bayes' sats, som

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)} \quad (2.4.1)$$

vilket visar att man genom inverterade sannolikheter kan hitta sannolikheter för värdet av en parameter. Dvs. att man med hjälp av kunskap från ett stickprov, x, givet en parameter $(x|\theta)$ kan dra slutsatser om parametervärdet; $(\theta|x)$. På många sätt utgör denna insikt grunden för statistisk analys. Målet är ofta att finna orsakerna genom att titta på effekterna, dvs. parametrarna utifrån observationerna. Med andra ord att dra inferens om en parameter, och att i sin tur kunna säga något om framtida observationer givet värdet på den parametern (Roberts, s. 8f, 2001).

Man söker *posteriorfördelningen*, vilket enkelt uttryckt innebär fördelningen för parametern, θ, givet observationerna, x. Den sammansatta täthetsfunktionen för θ och x skrivs $\varphi(\theta, x) = f(x|\theta)\pi(\theta)$, dvs. samplingfördelningen multiplicerat med *priorfördelningen*. Samplingfördelningen är detsamma som likelihood. Utifrån Bayes'

sats kan man utifrån denna produkt av förklaringar bestämma posteriorfördelningen (Gelman et al. s 7f, 2004), $\pi(\theta|x)$;

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{g(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta} \quad , \tag{2.4.2}$$

där g(x) är den marginella täthetsfunktionen för x.

$f(x|\theta)$, alltså täthetsfunktionen för de stokastiska variablerna givet parametrarna, kan också skrivas som likelihood-funktionen eftersom

$$L(\theta|x) = f(x|\theta), \tag{2.4.3}$$

där likelihood-funktionen är uttryckt som en funktion av parametrarna $\theta$.

### 2.4.2 Priorfördelning

Bayes menade att osäkerheten om en parameter i en modell kan modelleras med hälp av en priorfördelning, $\pi(\theta)$ (Robert, s9, 2001). Detta innebär att man utöver insamlade observationer, x, kan använda sig av andra informationskällor för att dra slutsatser, som exempelvis expertutlåtanden eller tidigare kunskap om parametern. Att man använder sig av den kunskap man har gör att två personer inte nödvändigtvis når samma slutsats, eftersom de tar med sig olika information in i processen. Således har man att göra med subjektiva sannolikheter.

En priorfördelning kan vara informativ eller icke-informativ. Vid en icke-informativ prior använder man sig ofta av en enklare fördelning, exempelvis en uniform fördelning, där man inte tar med sig någon bakgrundsinformation om parametern in i processen. Här grundar man alltså sina slutsatser enbart på data, så att ens inferens förblir opåverkade av extern information (Gelman et al. s 61, 2001). I detta arbete kommer en icke-informativ prior att användas.

## *2.5 Bayesiansk regression*

Här undersöks hur försäljningen av kvällstidningen Expressen, y, påverkas av en vektor av förklaringsvariabler, $x = (x_1, \dots x_k)$. En multipel linjär modell för den i:te observationen skrivs

$$E(y_i|\boldsymbol{\beta}, X) = \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \tag{2.5.1}$$

där i = (1,..., n), n är antalet observationer, och X är en n * k matris över n observationer för k förklaringsvariabler.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

Eftersom den första kolumnen antas vara 1 för alla i, där $x_{i1}$ är det i:te elementet i vektorn $x_{,1}$. Målet är att dra inferens om parametrarna givet X och y (Gelman, 2001, s 353f).

Låt $\boldsymbol{\beta}$ vara en vektor för alla koefficienter i regressionsmodellen, dvs. $\beta_i$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}$$

Eftersom en icke-informativ priorfördelning kommer att användas i regressionsanalysen kommer resultaten för medelvärdet i posteriorfördelningen att sammanfalla med de resultat man hade fått om man gjort en klassisk regressionsanalys. Genom posteriorn får man dock mer information i form av en fördelning för parametrarna $\boldsymbol{\beta}$, vilket möjliggör användandet av kredibilitetsintervall istället för konfidensintervall. Detta gör att man kan dra starkare inferens om hypotesen och uttala sig om sannolikheter, och inte bara i form av "säkerhet" som i klassiska konfidensintervall.

Den icke-informativa priorfördelningen är en uniform fördelning. Om man inte lägger in någon förhandsinformation antas alltså alla möjliga parametervärden vara lika sannolika. Eftersom vår data innehåller många observationer kommer en icke-informativ priorfördelning att ge liknande resultat som en någorlunda mer informativ priorfördelning (Gelman m fl. 2001, s 355f). Den uniforma priorfördelningen kan skrivas

$$p(\boldsymbol{\beta}, \sigma^2) = c\sigma^{-2} \tag{2.5.2}$$

där c är den normaliserande konstanten som gör att arean under priorfördelningen är 1. Alla kommande ekvationer detta avsnitt kommer att grunda sig på att det är priorn (2.5.2) som används.

### 2.5.1 Posteriorfördelning

Här kommer posteriorfördelningen för $\boldsymbol{\beta}$ givet $\sigma^2$ och den marginella posteriorfördelningen, dvs $(\sigma^2|y)$ att bestämmas. Den sammansatta posteriorfördelningen faktoriseras för $\boldsymbol{\beta}$ och $\sigma^2$som

$$\varphi(\boldsymbol{\beta}, \sigma^2|y) = f(\boldsymbol{\beta}|\sigma^2, y)\pi(\sigma^2|y) \tag{2.5.3}$$

och posteriorfördelningen för $\boldsymbol{\beta}$ givet $\sigma^2$blir

$$(\boldsymbol{\beta}|y, \sigma^2) \sim N\left(\widehat{\boldsymbol{\beta}}, V_\beta \sigma^2\right), \tag{2.5.4}$$

där

$$X^T = \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1k} & \cdots & x_{nk} \end{pmatrix}$$

och

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y \quad (2.5.5)$$

$$V_\beta = (X^T X)^{-1} \quad (2.5.6)$$

### 2.5.2 Marginella posteriorfördelningen

För att kunna dra värden från den sammansatta posteriorfördelningen behöver man först dra värden från den marginella posteriorfördelningen $(\sigma^2|y)$. Detta kan man göra genom att skriva om ekvationen (2.5.3) som

$$p(\sigma^2|y) = \frac{p(\boldsymbol{\beta},\sigma^2|y)}{p(\boldsymbol{\beta}|\sigma^2,y)}$$

Fördelningen har en inverterad Chi2-form

$$\sigma^2|y \sim Inv - \chi^2(n-k, s^2) \quad (2.5.7)$$
$$där, s^2 = \frac{1}{n-k}\left(y - X\widehat{\boldsymbol{\beta}}\right)^T (y - X\widehat{\boldsymbol{\beta}}) \quad (2.5.8)$$

### 2.5.3 Simulera värden från posteriorfördelningen

Dessa beräkningar och simuleringar kommer att göras med hjälp av programvaran R (Gelman m fl., 2001, s 357).

Tillvägagångssättet kommer att vara att man först gör vissa beräkningar. Man beräknar $\widehat{\boldsymbol{\beta}}$ med hjälp av ekvation (2.5.5), därefter beräknas $V_\beta$ från (2.5.6). Slutligen beräknas också $s^2$ från (2.5.8).

Urval från posteriorfördelningen dras sedan genom att man

1. Drar $\sigma^2$ från fördelningen i (2.5.7)
2. Drar $\boldsymbol{\beta}$ från fördelningen i (2.5.4) betingat på $\sigma^2$ i steg 1.

Varje gång man gör steg 1 och 2 får man en dragning. Eftersom steg 2 beror på steg 1 kommer dragningarna att loopas igenom i R. Vi drar 30 000 värden för att erhålla tillräckligt många dragningar över hela parameterrummet för $\boldsymbol{\beta}$. Utifrån dragningarna av $\boldsymbol{\beta}$ kan man sedan skapa ett histogram över de dragna värdena på respektive $\beta_i$, så att man kan dra slutsatser om fördelningen för parametern med exempelvis kredibilitetsintervall.

### 2.5.4 Kredibilitetsintervall

I klassisk statistik antas parametern θ vara en okänd konstant. Denna sanna parameter kommer att finnas med i 95 procent (om α = 5) av de klassiska konfidensintervallen som beräknas på olika urval. Utifrån detta kan man dra slutsatsen att med 95 procents *säkerhet* befinner sig den sanna parametern i det konfidensintervall som beräknats utifrån ett visst slumpmässigt urval. Man kan däremot inte säga att *sannolikheten* att

den sanna parametern ligger i konfidensintervallet är 95 procent, eftersom sannolikheten för det antingen är antingen 0 eller 1 (Jaynes, 1976, s207).

Den Bayesianska motsvarigheten till konfidensintervall kallas kredibilitetsintervall. I Bayesiansk statistik betraktas inte θ som en okänd konstant utan som en slumpvariabel med en viss fördelning, som beskrivits tidigare i arbetet. Detta leder till att man kan dra andra slutsatser utifrån dessa intervall än man kan från de klassiska konfidensintervallen. Kredibilitetsintervall skapas genom att man finner de värden i posteriorfördelningen mellan vilka 1- α av massan återfinns (KTH, 2011). Konsekvensen av detta är att man här kan säga att parametern θ ligger i kredibilitetsintervallet med 95 procents sannolikhet.

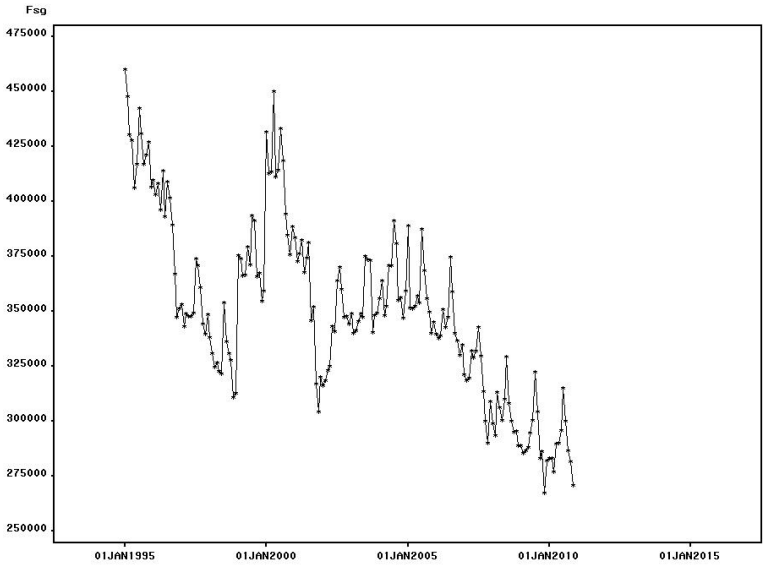
I de fall man använder sig av en uniform priorfördelning för att ta fram posteriorfördelningen, och posteriorfördelningen är normalfördelad kommer kredibilitetsintervallet att sammanfalla med det klassiska konfidensintervallet.

### 2.5.5 Residualanalys

Modellens lämplighet utvärderas genom en mindre residualanalys för att undersöka om antagandena om feltermerna är uppfyllda eller ej. Specifikt kommer vi att utvärdera om det föreligger oberoende mellan feltermen, $\varepsilon_t$, och förklaringsvariabeln som verkar vara mest betydelsefull, om feltermerna är normalfördelade och utifrån en plot över $\varepsilon_t$ och $\varepsilon_{t-1}$ undersöka om feltermerna är okorrelerade eller ej.

# 3. Datamaterial

Datamaterial för tidsserien över försäljningen av kvällstidningen Expressen är de officiella siffrorna och har erhållits från Expressen.

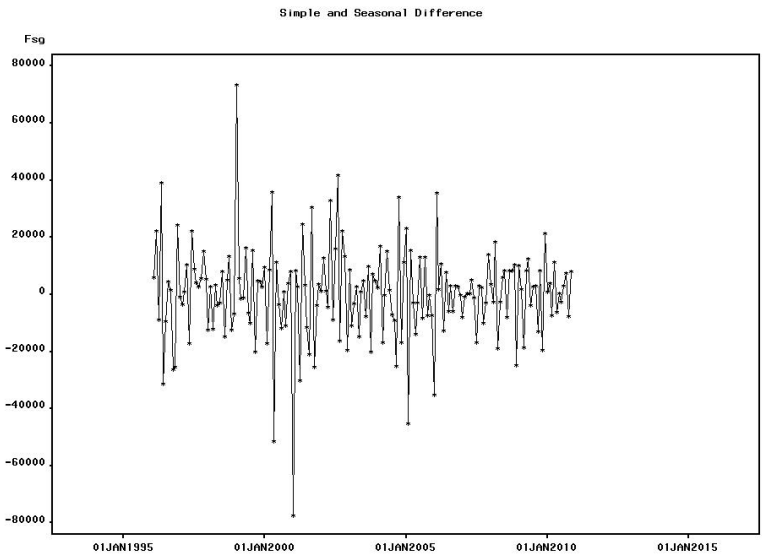



**Figur 3.1 Tidsserie över månadsdata för försäljningen av Expressen 1995-2010**

I figur 3.1 ovan visas observationerna i tidsserien innan några transformationer är gjorda. Man kan utifrån diagrammet konstatera att tidsserien uppvisar såväl trend som säsongskomponenter. Den övergripande trenden är negativ. Säsongsvariationen visar att tidningsförsäljningen ökar något under sommarmånaderna. Man kan också

konstatera en högre volatilitet under första halvan av mätperioden.

För att kunna anpassa lämpliga ARIMA-modeller kommer transformationer av tidsserien att behöva utföras. En enkel differentiering och en differentiering för säsongsdelen räcker för att tidsserien ska bli stationär, se figur 3.2 nedan. I figuren finns ett par outliers, men vi bedömer att dessa ensamma inte gör att man entydigt kan säga att kravet på konstant varians är otillfredsställt i den differentierade tidsserien. Tidsseriens autokorrelation avtar mycket långsamt utan någon differentiering, och när en enkel differentiering är gjord avtar säsongskomponenterna (tidsavstånd 1, 12, 24, 36 etc) långsamt vilket motiverar en säsongsdifferentiering. I figur 3.1 finns det ingenting som antyder ett multiplikativt förhållande i säsongsdelen då säsongsvariationerna inte ser ut att öka över tid, vilket antyder att en logaritmering av tidsserien inte verkar vara nödvändig.



**Figur 3.2 Differentierad tidsserie över försäljningen av Expressen 1995-2010**

Utöver den ovan presenterade tidsserien för Expressen, som används som beroende variabel, används flera andra datamaterial för att kunna utföra den Bayesianska regressionsanalysen. Dessa variabler har valts ut för att de kan tänkas ha en inverkan på tidningsförsäljningen på olika sätt.

Den första oberoende variabeln som kommer att användas i regressionsanalysen är besökare på expressen.se. Detta är en specifik variabel för att mäta internetanvändning då den fokuserar på användandet av tidningens egen hemsida. Användningen skulle kunna ha leda till en substitutionseffekt där papperstidningen överges, men också tjäna som en marknadsföringskanal och därmed öka tidningsförsäljningen. Från 2002 har vi fått tillgång till siffror över besökare på expressen.se. Data kommer från Expressen. Denna statistik visar dock besökare/vecka, vilket gör att vi måste anpassa detta data till månadsdata för att erhållas samma nivå på data för samtliga variabler i regressionsmodellen. Detta har utförts genom att för de veckor som spänner över två månader fördelas antalet

besökare på hemsidan den veckan mellan de två månaderna proportionellt. Vi har alltså gjort ett antagande om att besöken på hemsidan är linjära över veckan. Även dessa data är begränsat till en viss tidsperiod, nämligen 2002 till kvartal 3 2011. Detta leder till en begränsning i tiden för regressionsanalysen så att den börjar med data från 2002. Då det saknas värden för vissa veckor har metoden linjär interpolation i SAS använts för att fylla dessa. Det innebär att man antar en linjär trend inte bara över veckans dagar utan också mellan veckorna för att fylla tomrummen där det saknas data. Interpoleringsmetoden är den metod som föreslås vid imputering av saknade värden av SAS (SAS Institute Inc., 2011). Detta har utförts innan veckodata slogs samman till månadsdata. I de flesta fall användes flera kända veckovärden och bara någon eller några veckor med interpolerade värden. Detta gör att man erhåller man så små problem som möjligt som följd av den linjära interpolationen när man transformerar till månadsdata. Det är relativt få värden som fattas i detta datamaterial. De första fyra värdena saknades i veckodatan, och detta kan inte interpoleras eftersom det inte ligger mellan två andra värden. För första månadsobservationen har samma värde som för den andra månaden samma år använts.

Den andra oberoende variabeln är försäljning av Aftonbladet. Då båda tidningar finns hos återförsäljarna av tidningarna och kan antas vara substitut finns det anledning att tro att det finns en korrelation mellan denna variabel och den beroende variabeln. Datamaterialet för huvudkonkurrenten Aftonbladet har erhållits från Tidningsstatistik AB, som är en länk mellan säljare och köpare av annonser som förmedlar statistik över tidningsförsäljning (Tidningsstatistik AB, 2010). För Aftonbladet finns dock bara material från år 2000 till 2009, vilket gör att regressionsanalysen kommer att begränsas.

Den tredje oberoende variabeln som används är priset på Expressen. Det är rimligt att göra antagandet att tidningar är en ”vanlig vara” i det avseendet att om priset på tidningen går upp förväntas försäljningen gå ned[3]. Uppgifterna om pris på Expressen har erhållits från Expressen. Dessa siffror har sedan inflationsjusterats eftersom det är reala priser som ska undersökas. Detta har gjorts genom att Konsumentprisindex (1980 = 100), skuggindextal från SCB, använts. Utifrån dessa har priserna justerats. Det nominella priset för januari 2002 användes som bas. Indexvärdet var 268,8. Det reala priset för denna tidsenhet var 8 kronor. Därefter justerades alla nominella priser så att de följde samma index; (’index för tid i’/268,8) * ’nominellt pris för tid i’. På så vis blir alla följande priser i januari 2002 års priser.

Den fjärde oberoende variabeln är tid. Variabeln går från 1-96, och finns med för att upplagan kan ha ökat med tiden.

Utöver dessa variabler finns säsongsdummys. När människor har semester finns mer tid för tidningsläsande. För att undersöka om detta gör att säsongen spelar en viktig roll för förändringen i försäljningen inkluderas dummyvariabler för kvartal.

3 Med ”vanlig vara” avses det som inom mikroekonomi kallas ”ordinary good” för vilken efterfrågan ökar om priset sjunker och vice versa.

# 4. Resultat

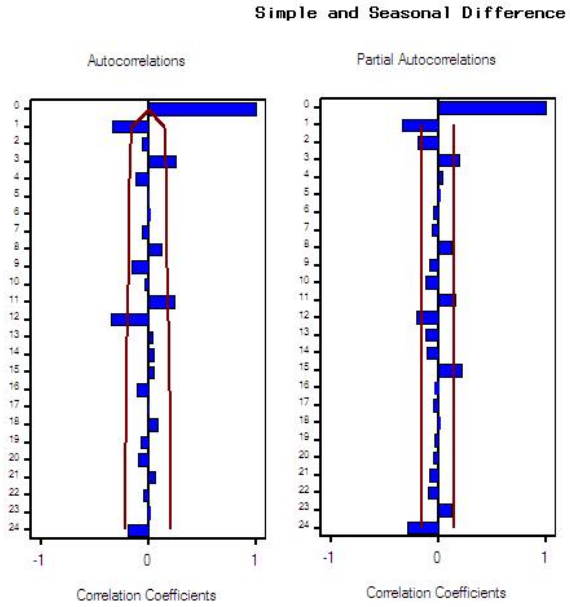
## *4.1 Modellbestämning tidsserie*

För alla modeller har det utförts prognoser där två år (24 observationer) lämnats utanför för att kunna göra en out of sample-analys. Modellen har alltså anpassats för tidsperioden januari 1995 - november 2008, och utvärderas för tidsperioden därefter, det vill säga december 2008 - november 2010. Målsättningen har varit att välja ut den bästa modellen att göra prognos utifrån. Metoden med ett ”hold-out-sample” ger oss tydliga indikationer på vilken modell som är bäst anpassad till datamaterialet eftersom prognosen kan jämföras med verkliga observerade värden. Prognosen kommer sedan att göras utifrån hela tidsserien med de modeller som tagits fram genom denna metod.

Varje gång utförs en enstegsprognos som testas för den del av tidsperioden som hålls utanför anpassningsperioden. För att välja modell används prognosfel för den utvärderade perioden samt AIK och SIK för att bestämma vilken modell som bör användas för prognoser.

För samtliga beräkningar har programvaran SAS 9.2 (Time Series Forecasting System) använts. Relevanta bilder som inte återfinns i arbetet finns för de modeller som bedömts vara mest riktiga i Appendix.

Tidsserien har undersökts visuellt genom granskning av autokorrelationsfunktionen och den partiella autokorrelationsfunktionen enligt Box-Jenkins-metoden för att bestämma vilka ARIMA-modeller som kan vara lämpliga att utvärdera utöver de exponentiella utjämningsmetoderna.




**Figur 4.1 Autokorrelation och partiell autokorrelation för den differentierade modellen**

I figur 4.1 kan man se att både autokorrelationen och den partiella autokorrelationen uppvisar ett dämpat sinusmönster snarare än att de avtar exponentiellt. Detta gör det svårt att entydigt bestämma vilken modell som är bäst lämpad och man bör därför testa olika modeller för att se vilka som ger bäst resultat för de test som utvärderas. Man kan dock i figuren skönja att det i den partiella autokorrelationsdelen finns en

liten spik, dvs. signifikant värde, på såväl 12 som 24. Både autokorrelationsdelen och den partiella autokorrelationsdelen visar en liten spik på tidsavstånd 1, och för den senare eventuellt även på tidsavstånd 2. Alla spikar är dock relativt små, och man kan därför inte dra alltför långtgående slutsatser endast från den visuella analysen då den inte ger en entydig indikation. Däremot kan denna användas som grund för att välja ut vilka ARIMA-modeller som ska undersökas närmre. Följande ARIMA-modeller utvärderas för att testa vilken av dem som är mest lämpad att använda för prognos:

ARIMA(1,1,1)(0,1,0)s
ARIMA(2,1,1)(0,1,0)s
ARIMA(2,1,1)(1,1,0)s
ARIMA(3,1,1,)(2,1,0)s
ARIMA(2,1,1)(2,1,0)s
ARIMA(1,1,1)(2,1,0)s
ARIMA(1,1,0)(2,1,0)s
ARIMA(2,1,0)(2,1,0)s

För att sedan utvärdera metoderna för exponentiell utjämning och de utvalda ARIMA-modellerna har de olika prognosfelsmåtten och informationskriterierna använts. Resultaten för dessa presenteras i tabell 4.1.

| METOD | MAD | MSE | MAPE | $R^2$adj | AIK | SIK |
|---|---|---|---|---|---|---|
| EEU | 9108,1 | 132456037 | 3,12 | 0,10 | 450,84 | 452,02 |
| Holt Additiv | 9132,7 | 131647566 | 3,13 | 0,07 | 452,70 | 455,05 |
| Holt-Winter Additiv | 6010,5 | 52622255 | 2,09 | 0,61 | 432,69 | 436,22 |
| Holt-Winter Multip. | 6027,3 | 49323659 | 2,09 | 0,63 | 431,13 | 434,67 |
| ARIMA(1,1,1)(0,1,0) | 7095,7 | 83321799 | 2,49 | 0,41 | 441,72 | 444,07 |
| ARIMA(2,1,1)(0,1,0) | 7419,9 | 89215934 | 2,60 | 0,34 | 445,36 | 448,89 |
| ARIMA(2,1,1)(1,1,0) | 6596,3 | 65496513 | 2,30 | 0,49 | 439,94 | 444,65 |
| ARIMA(3,1,1)(2,1,0) | 5917,5 | 57092221 | 2,07 | 0,51 | 440,64 | 447,71 |
| ARIMA(2,1,1)(2,1,0) | 5972,8 | 54978002 | 2,09 | 0,55 | 437,74 | 443,63 |
| ARIMA(1,1,1)(2,1,0) | 5508,4 | 50380775 | 1,93 | 0,61 | 433,64 | 438,36 |
| ARIMA(1,1,0)(2,1,0) | 5519,8 | 52158111 | 1,94 | 0,61 | 432,47 | 436,01 |
| ARIMA(2,1,0)(2,1,0) | 5519,6 | 49354281 | 1,93 | 0,62 | 433,15 | 437,86 |

**Tabell 4.1 Tabell över MAD, MSE, MAPE, justerat $R^2$, AIK och SIK för testade modeller**

Utifrån tabell 4.1 kan man dra slutsatsen att både den additiva och den multiplikativa Holt-Winter-modellen står sig mycket bra i jämförelsen med ARIMA-modellerna, även om ARIMA-modellerna längst ned ser ut att vara något mer lämpliga än den additiva att döma av tabellens första fyra felmått. Samtidigt står de sig mycket bra för både AIK- och SIK-värdena. Det finns viss motstridighet i resultaten i tabell 4.1 måtten emellan. Det är dock de fyra ARIMA-modellerna längst ned som klarar sig bäst vid en bedömning av SIK, vilket är det mått som prioriteras när det finns motstridigheter i modellvalen i och med att det är konsistent – dvs straffar tillräckligt hårt för överidentifiering.

För att undersöka om de fyra utvalda ARIMA-modellerna faktiskt är lämpliga för prognostisering används Ljung-Box-test. I en bra modell förklaras förändringarna i

tidsserien och autokorrelationen för residualerna blir i det fallet reducerat till vitt brus. Om man inte bara har vitt brus bör man kunna hitta en bättre modell som kan förklara mer av utvecklingen i tidsserien än den modell som föreligger. Nollhypotesen är att man har vitt brus upp till lag 20, dvs. $\rho_1=\rho_2=...=\rho_{20}=0$.

Vi beräknar p-värden för lag=K=20, men också för lag=1, lag=2, lag=3 och lag=10, för att få en bättre förståelse för utveklingen. Resultaten visas i tabell 4.2.

| | P-värde | | | | |
|---|---|---|---|---|---|
| **Modell** | **lag-20** | **lag-1** | **lag-2** | **lag-3** | **lag-10** |
| ARIMA(2,1,1)(2,1,0) | 0,305600 | 1,000000 | 1,000000 | 1,000000 | 0,324100 |
| ARIMA(1,1,1)(2,1,0) | 0,010900 | 1,000000 | 1,000000 | 1,000000 | 0,007800 |
| ARIMA(1,1,0)(2,1,0) | 0,003300 | 1,000000 | 1,000000 | 1,000000 | 0,003800 |
| ARIMA(2,1,0)(2,1,0) | 0,074400 | 1,000000 | 1,000000 | 1,000000 | 0,041700 |

**Tabell 4.2 P-värden från Ljung-Box-test för utvalda ARIMA-modeller**

Värdena i tabellen visar att man inte kan förkasta nollhypotesen för någon av modellerna vid K = 1-3. För modellen ARIMA(1,1,1)(2,1,0)s kan man däremot förkasta nollhypotesen vid både lag 10 och lag 20. Nollhypotesen för modell (2,1,1,)(2,1,0)s kan man inte förkasta på 5 % signifikansnivå på någon lag. ARIMA(1,1,1)(2,1,0)s är mera svårbedömd då man kan förkasta nollhypotesen för K = 20 på 5 % signifikansnivå men inte på 1 % signifikansnivå. För ARIMA(2,1,0)(2,1,0)s kan man inte förkasta nollhypotesen på lag 20, men däremot på lag 10 om man tittar på 5 % signifikansnivå.

ARIMA(1,1,0)(2,1,0)s som hade ett av de bästa resultaten i tabell 4.1, visade sig ge ett sämre resultat när man tar hänsyn till Ljung-Box testet. Där visade sig istället modell ARIMA(2,1,1)(2,1,0)s och modell (2,1,0)(2,1,0)s vara bättre på att förklara tidsserien eftersom det inte kan påvisas att autokorrelationen för residualerna på tidsavstånd 20 inte skulle vara skild från 0. Dessa två ARIMA-modeller anses därför vara de bästa modellerna att använda för prognostisering. Bilder över Ljung-Box testet återfinns i Appendix 1.

## *4.2 Prediktioner med bästa modellerna*

Efter att ha utvärderat de olika exponentiella modellerna och ARIMA-modellerna för ”out of sample”-analysen har följande modeller visat sig vara de bästa modellerna:
Holt-Winter (additiv)
Holt-Winter (multiplikativ)
ARIMA(2,1,1)(2,1,0)s
ARIMA(2,1,0)(2,1,0)s

Då dessa modeller i utvärderingen har visat sig hantera tidsserien bäst är det dessa som ska användas för att ta fram en prognos för den kommande tidsperioden. Prognosen görs för perioden december 2010 – december 2012. När prognosen görs används allt tillgängligt data, och inget ”hold-out sample” används. Det finns dock en viss möjlighet att en annan modell skulle vara bättre anpassad för hela datamaterialet än för den anpassningsperiod vi använt för att välja vilken modell som ska användas för prognos. ARIMA-modellerna är dock anpassade för en stationär tidsserie med konstant varians och väntevärde, varför resultatet inte bör ändras alltför mycket.

Tidsserien ser dessutom ut att följa samma mönster de närmsta åren före och efter november 2008, sett till figur 3.1.

| | Holt-Winter multip. | | Holt-Winter add | | ARIMA(2,1,1)(2,1,0)s | | ARIMA(2,1,0)(2,1,0)s | |
|---|---|---|---|---|---|---|---|---|
| | Predikt. | k.i./2 | Predikt. | k.i./2 | Predikt. | k.i./2 | Predikt. | k.i./2 |
| **dec-10** | 277742 | ±23203 | 276846 | ±22273 | 284317 | ±27614 | 283921 | ±26546 |
| **jan-11** | 283799 | ±28670 | 289325 | ±27378 | 280590 | ±33507 | 280677 | ±31689 |
| **feb-11** | 281028 | ±32933 | 281989 | ±31679 | 279280 | ±37301 | 278807 | ±35467 |
| **mar-11** | 280953 | ±36870 | 281213 | ±35470 | 282498 | ±43218 | 282445 | ±39776 |
| **apr-11** | 283463 | ±40696 | 285623 | ±38898 | 287134 | ±46825 | 286951 | ±43464 |
| **maj-11** | 281555 | ±43777 | 284670 | ±42054 | 286170 | ±50668 | 285893 | ±46792 |

**Tabell 4.3 Prognoser för 6 månader med de fyra bästa modellerna**

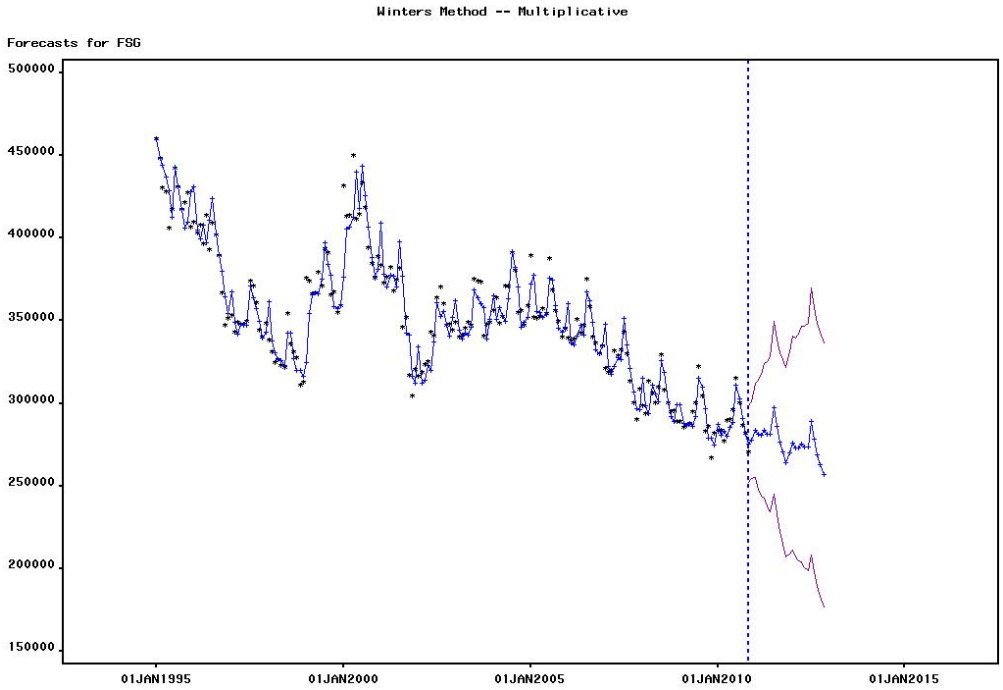
I tabell 4.3 syns prediktioner med de utvalda modellerna för de närmaste 6 månaderna. Prediktioner för en längre tidsperiod återfinns i Appendix 4. Prediktionerna uttrycks också visuellt i figur 4.4.

### 4.2.1 Prognoser med Holt-Winters metod

Holt-Winters multiplikativa och additiva metod visade sig vara mycket effektiv utifrån de värden som syns i tabell 4.1. Det är därför rimligt att misstänka att dessa modeller kommer att leverera de bästa prognoserna, utifrån gjorda antaganden om att det är möjligt att förklara framtida utveckling på samma sätt som man förklarar det som redan hänt.

Den multiplikativa Holt-Winter-modellen visade sig vara den modell som visade på de bästa värdena i tabell 4.1. Sannolikt är det alltså för denna modell som man får den bästa prognosen baserad på Holt-Winters metod också om man gör en anpassning för hela tidsperioden. Prognosen visas i figur 4.2. Den streckade linjen markerar den sista faktiska observationen, november 2010. Därefter syns en prognos och den övre och den undre konfidensgränsen. Prognostiserade värden, samt bredd på konfidensintervallet syns även i tabell 4.3 för de 6 första prognostiserade månaderna.

Den additiva Holt-Winter-modellen var nästan lika bra som den multiplikativa modellen, och hade bättre anpassning till tidsserien än nästan samtliga utvärderade ARIMA-modeller. En prognos utförd med denna modell visar liknande, men något högre, värden än den multiplikativa modellen. Prognostiserade värden återfinns i tabell 4.3.



Figur 4.2 Prognos med den multiplikativa Holt-Winter-modellen

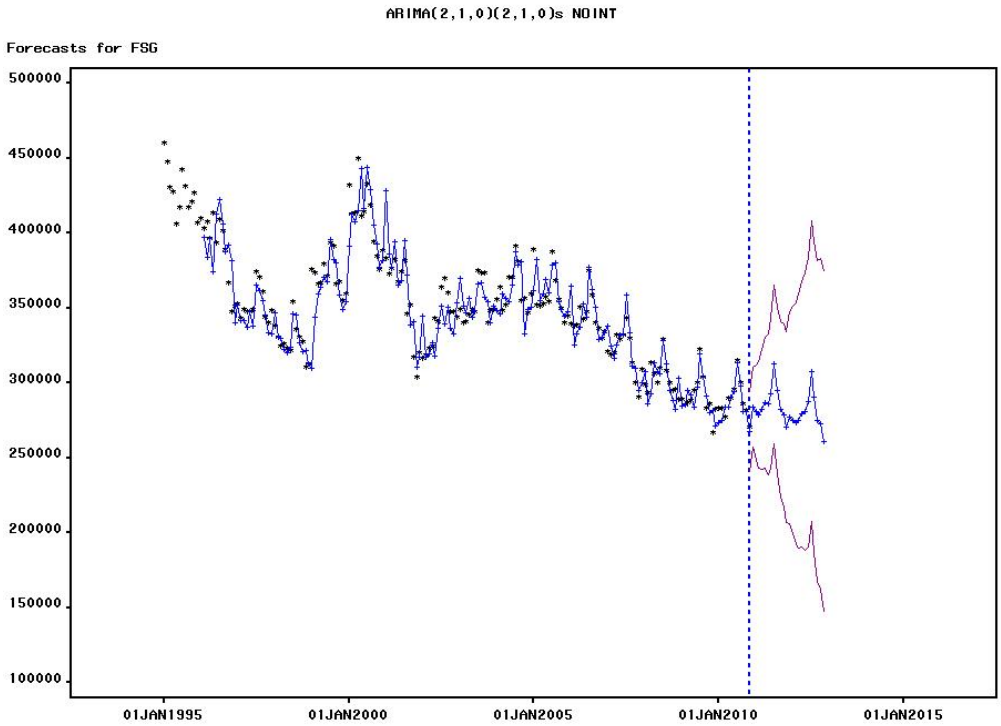
### 4.2.2 Prognoser med ARIMA

Vissa ARIMA-modeller uppvisade bra resultat vid en bedömning av deras förmåga att prognostisera, vilket syns i tabell 4.1. Emellertid visade det sig att vissa av modellerna som uppvisade bra värden där inte var helt lämpliga att använda utifrån värdena från Ljung-Box-testet i tabell 4.2. Prognoser med de två modeller som visade sig vara lämpliga i Ljung-Box-testet, ARIMA(2,1,0)(2,1,0)s och ARIMA(2,1,1)(2,1,0)s, presenteras i figur 4.3.

ARIMA(2,1,0)(2,1,0)s ser ut att vara den bästa av de två utvalda ARIMA-modellen om man enbart tittar på de felmått som återfinns i tabell 4.1. Med andra ord är detta den ARIMA-modell som producerar de bästa prognoserna. Modellen är dock något mindre tydlig i Ljung-Box-testet, eftersom detta test visar att det finns autokorrelation i residualerna för lag 10. Modellen klarar dock att förklara tillräckligt mycket av variationerna i tidsserien, och lyckas reducera $Y_t$ till vitt brus på lag 20. Prognostiserade värden finns i tabell 4.3. En bild över prognosen återfinns i figur 4.3 där den streckade linjen markerar den sista faktiska observationen. Därefter är de markerade värdena prognosen tillsammans med den övre och den undre konfidensgränsen.

ARIMA(2,1,1)(2,1,0)s är den modell som ser ut att ha högst förklaringsgrad om man utgår ifrån Ljung-Box-testet eftersom man efter detta test inte kan påvisa att det skulle finnas någon autokorrelation i residualerna vare sig på lag 10 eller lag 20. Modellen tycks således förklara mycket stora delar av variationerna i tidsserien. Denna modell visar dock på något sämre värden för samtliga test i tabell 4.1, vilket indikerar att den inte gör lika bra prognoser. Då detta ändå är den modell som visat sig mest lämpad av

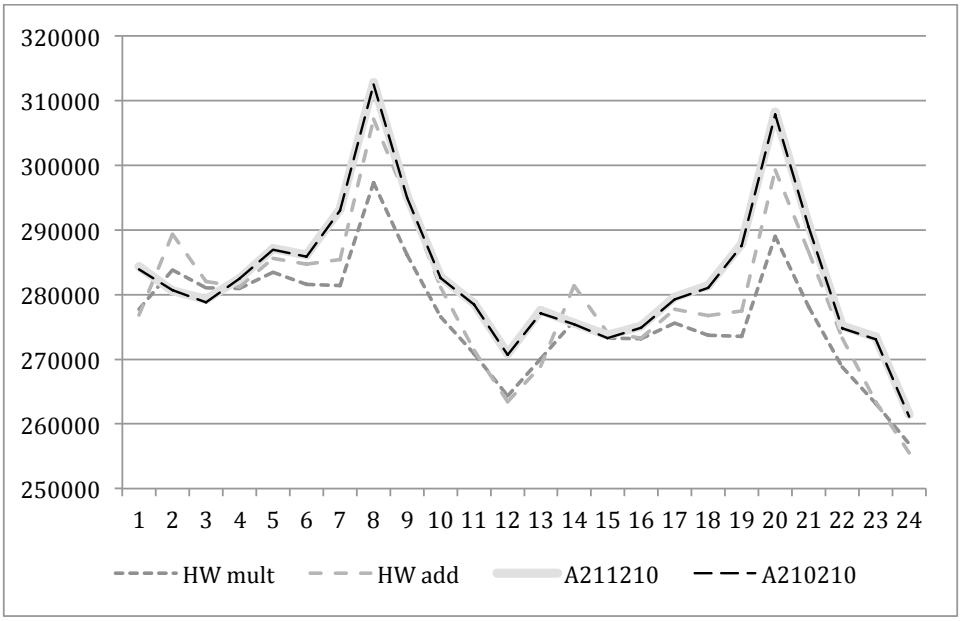
ARIMA-modellerna för tidsserien är även en prognos framtagen utifrån denna modell motiverad. Prognosen utifrån denna modell återfinns i tabell 4.3.




Figur 4.3 Prognos med ARIMA(2,1,0)(2,1,0)s

### 4.2.3. Utvärdering av prognoserna

I figur 4.4 kan man se samtliga prognoser för tidsperioden december 2010 - november 2012. Det som blir tydligast är att de båda ARIMA-modellerna gör nästan identiska prognoser, och man kan knappt skönja skillnaden. ARIMA-modellerna ger något högre skattning än de som fås fram i Holt-Winter-modellerna, särskilt vid säsongstopparna. I tabell 4.3 kan man också se att ARIMA-modellerna har betydligt bredare konfidensintervall än Holt-Winter-modellerna, och att denna skillnad ökar över tid. ARIMA(2,1,0)(2,1,0)s (ARIMA1 i figur 4.4) har dock ett kortare konfidensintervall än ARIMA(2,1,1)(2,1,0)s (ARIMA2 i figur 4.4). Samtliga modeller påvisar dock liknande mönster vad gäller såväl trend som säsong, och tidsseriens karaktäristika kan således anses klarlagd.




**Figur 4.4 Jämförelse mellan prognoser för multiplikativ och additiv Holt-Winter, samt ARIMA(2,1,0)(2,1,0)s och ARIMA(2,1,1)(2,1,0)s**

## *4.3 Modell för den Bayesianska regressionen*

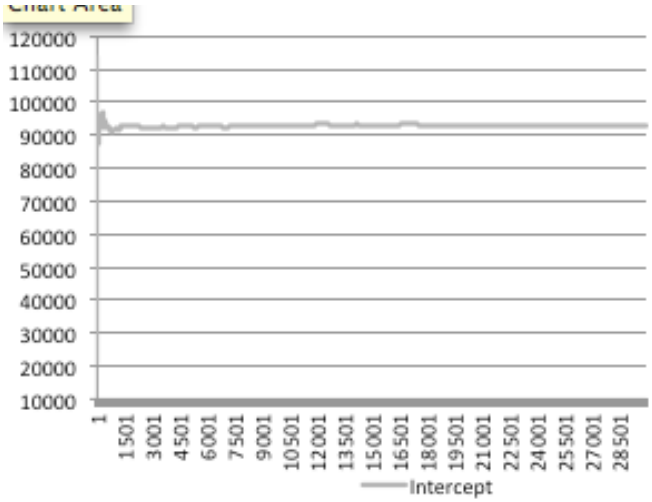
För den Bayesianska regressionen har data från tidsperioden januari 2002 - december 2009 använts. Den beroende variabeln är samma variabel som använts i tidsserieanalysen, det vill säga försäljning av tidningen Expressen. Modellen som först utvärderas är

$$E(y_i|\boldsymbol{\beta}, X) = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 D_{i1} + \beta_7 D_{i2} + \beta_8 D_{i3}$$

där $x_{i1}$ antas vara 1 för alla observationer i, och utgör därför interceptet i regressionsmodellen. $x_{i2}$ är besökare på expressen.se vid tidpunkt i, $x_{i3}$ är försäljning av tidningen Aftonbladet vid tidpunkt i, $x_{i4}$ är pris på tidningen Expressen vid tidpunkt i och $x_{i5}$ är en variabel för tiden. $D_{i1}$ är en dummyvariabel för kvartal två, $D_{i2}$är en dummyvariabel för kvartal tre och $D_{i3}$ är en dummyvariabel för kvartal fyra.

För samtliga beräkningar har programvaran R 2.14 använts.

I den Bayesianska regressionsanalysen dras värden på $\boldsymbol{\beta}$ ett upprepat antal gånger för att på så sätt ta fram fördelningen för varje parameter. Först utförs en undersökning av hur många dragningar som behövs för att medelvärdet för respektive parameter ska stabiliseras kring det verkliga medelvärdet. Vi drog 30 000 parametervärden vilket visade att medelvärdet för respektive parameter planade ut/konvergerade efter ca 15 000 dragningar, se figur 4.5. De 30 000 dragningarna används för att ta fram den marginella posteriorfördelningen för varje parameter för att på så sätt kunna undersöka om parametrarna är signifikanta eller inte. Figurer över konvergensen för de övriga parametrarnas medelvärden återfinns i Appendix 5.



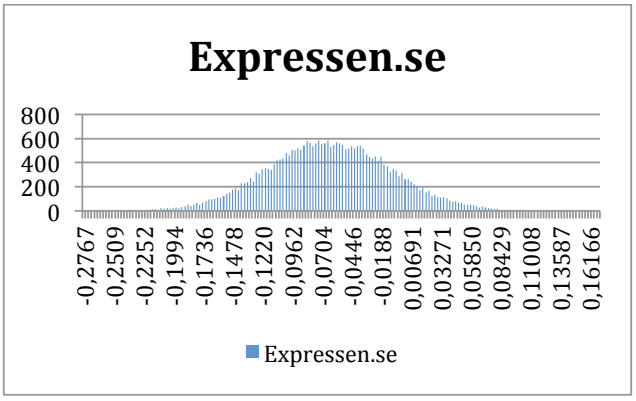

Figur 4.5 Konvergens kring medelvärden för parametrarna för interceptet

### 4.3.1 Resultat för den Bayesianska regressionen

Den första regressionsanpassningen gav följande medelvärde, $\overline{\boldsymbol{\beta}}$, på respektive parameter,

$$\overline{\boldsymbol{\beta}} = \begin{pmatrix} \overline{\beta}_1 = 93268{,}855 \\ \overline{\beta}_2 = -0{,}065 \\ \overline{\beta}_3 = 0{,}758 \\ \overline{\beta}_4 = -10893{,}223 \\ \overline{\beta}_5 = 724{,}573 \\ \overline{\beta}_6 = 2636{,}115 \\ \overline{\beta}_7 = 4349{,}768 \\ \overline{\beta}_8 = 4570{,}576 \end{pmatrix}$$

Vid en undersökning av fördelningarna för dessa variabler konstateras att vi endast med 88,5 % sannolikhet kan säga att parametern för besökare på expressen.se, $\overline{\beta}_2$, är negativ, eftersom endast 88,5 % av observationerna är lägre än 0, se figur 4.6. Därmed kan vi sluta oss till att 0 ingår i ett ensidigt 90 % -igt kredibilitetsintervall och att variabeln därmed inte riktigt lever upp till kraven för att vara signifikant.
Dummyvariabeln för kvartal 2 visar sig vara positiv med 83,7 % sannolikhet – variabeln är således klart icke-signifikant.
Samtliga övriga variabler visade sig vara signifikanta på 10 % signifikansnivå.




**Figur 4.6 Fördelning för parametern för expressen.se**

Tvåsidiga kredibilitetsintervall på 95 % för parametrarna blir

$$\text{Expressen. se}, \bar{\beta}_2: [-0{,}170; 0{,}041]$$
$$\text{Aftonbladet}, \bar{\beta}_3: [0{,}615; 0{,}901]$$
$$\text{Pris}, \bar{\beta}_4: [-15942{,}133; -5904{,}453]$$
$$\text{Tid}, \bar{\beta}_5: [319{,}901; 1122{,}598]$$
$$\text{Kvartal 2}, \bar{\beta}_6: [-2653{,}084; 7912{,}613]$$
$$\text{Kvartal 3}, \bar{\beta}_7: [-2183{,}719; 11029{,}129]$$
$$\text{Kvartal 4}, \bar{\beta}_8: [-597{,}990; 9766{,}219]$$

Några av variablerna, Expressen.se, och dummyvariablerna för kvartalen blir inte riktigt signifikanta med 95 % sannolikhet. Att döma av kredibilitetsintervallen är det variablerna för pris och expressen.se som har tydligast negativ effekt. Det är också mycket tydligt att försäljningen av Aftonbladet har en positiv korrelation med försäljningen av Expressen.

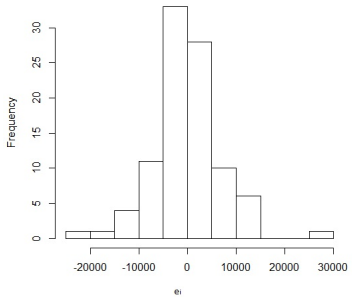
### 4.3.2 Modellutvärdering med residualanalys

Residualanalysen utfördes med 300 slumpmässigt utvalda parameterdragningarna från de ursprungliga 30 000 dragningarna. Dessa användes för att skatta medelvärden för feltermerna, $\bar{\varepsilon}_{i.}$, gör respektive observation *i*. Detta gjordes genom att ta det observerade värdet på den beroende variabeln minus medelvärdet för de skattade värdena från modellen för den beroende variabeln.

$$\bar{\varepsilon}_{i.} = y_i - \bar{\hat{y}}_{i.},$$

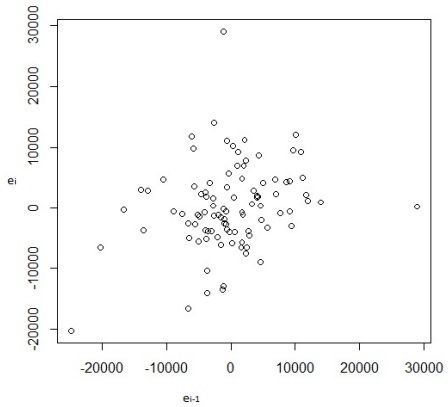
där $i = (1,2,\ldots,96)$

I residualanalysen visar figur 4.7 att antagandet om normalfördelade feltermer verkar vara uppfyllt.



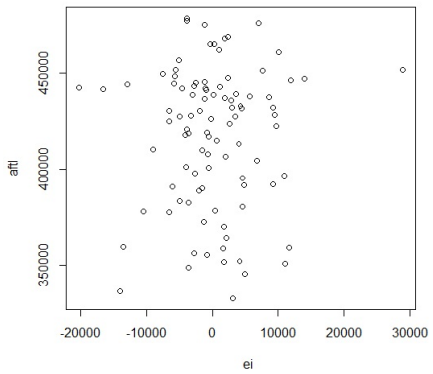

**Figur 4.7 Histogram över feltermerna**

En plot över $\varepsilon_t$ och $\varepsilon_{t-1}$ visar att det inte heller verkar finnas någon korrelation mellan feltermerna, se figur 4.8.




**Figur 4.8 Plot över $\varepsilon_t$ och $\varepsilon_{t-1}$**

Dessutom kan man anta från figur 4.9 att variabeln Aftonbladet, som visat sig vara en viktig förklaringsvariabel, är oberoende av feltermen.




**Figur 4.9 Plot över feltermen och variabeln för försäljning av Aftonbladet**

Residualanalysen visar således att vår modell verkar vara korrekt specificerad, och vi erhåller optimala skattningar. Det går alltså att göra en korrekt inferens utifrån resultaten i regressionen.

# 5. Slutsats

## *5.1 Sammanfattande slutsats och huvudupptäckter*

I detta arbete har vi visat hur försäljningen av kvällstidningen Expressen, som varit den tidning som fått tjäna som exempel för kvällstidningar i allmänhet, har förändrats över tid. Det har konstaterats att det finns en negativ trend för tidningsförsäljningen och att det också finns säsongsvariationer.

För tidsserien för försäljning av Expressen har flera olika modeller utvärderats. Såväl metoder för exponentiell utjämning och ARIMA-modeller har testats. Utvärderingen av metoderna visade att den multiplikativa Holt-Winter-modellen var den bästa modellen, eftersom den påvisade bäst anpassning till datamaterialet. Den hade också de lägsta värdena på informationskriterierna som inkluderar straff för överidentifiering samtidigt som den hade ett högt justerat $R^2$-värde. Prognoserna för denna modell visar att försäljningen kommer att minska över tid. Det sista kända värdet som finns för tidsserien är 270 900 exemplar för november 2010. Det predikterade värdet för november 2011 och 2012 är 264 369 respektive 256 886 exemplar. En av ARIMA-modellerna, ARIMA(2,1,0)(2,1,0)s, klarade Ljung-Box-testet för autokorrelation i residualerna och hade samtidigt resultat för AIK och SIK som är jämförbara med Holt-Winter-modellen, men anses inte vara lika bra.

Variabeln expressen.se visade sig ha en relativt liten betydelse för nedgången då man bara med ca 88,5 % sannolikhet kan säga att den har en negativ inverkan på försäljningen. Inverkan är också relativt låg – ett ytterligare besök på expressen.se per dag leder till 0,07 färre sålda tidningar per dag i snitt. Detta skulle kunna bero på att det finns såväl komplementära som ersättande element som tar ut varandra. Förklaringen kan också tänkas vara att den traditionella Expressen-läsaren vänder sig till andra nyhetskällor på nätet, så som andra nyhetssajter eller communitys.

Regressionsanalysen påvisar en positiv korrelation mellan försäljningen av Expressen och dess huvudkonkurrent Aftonbladet. Om försäljningen av Aftonbladet ökar med 1 tidning per dag i snitt kommer snittet för försäljningen av Expressen att öka med ca 0,76 tidningar per dag, givet att övriga variabler hålls konstanta.

Priset visade sig ha en klart negativ betydelse för försäljningen, vilket är rimligt då tidningar troligen är en så kallad ”vanlig” vara där en prisökning leder till minskad försäljning. En realprisökning med en krona (i januari 2002-års priser) leder enligt regressionsanalysen till en försäljningsminskning på i snitt ca 10 893 tidningar per dag, givet att övriga variabler hålls konstanta.

Utifrån parametrarna för dummyvariablerna kan man dra slutsatsen att det förväntas säljas minst antal tidningar under det första kvartalet. Under det tredje kvartalet säljs i snitt ca 4 350 fler tidningar per dag jämfört med första kvartalet, vilket stärker uppfattningen från tidsserieanalysen att säsong spelar stor roll för försäljningen av tidningen.

## 5.2 Jämförelse med tidigare resultat

Den utförda tidsserieanalysen visar på liknande resultat som andra studier inom området: försäljningen av papperstidningar minskar. Till skillnad från andra studier där endast årsdata används kan man i detta arbete också se att det finns en säsongskomponent i tidsserien, och att benägenheten att köpa tidningar är högre på sommaren, kanske framförallt under industrisemestern. Till skillnad från andra undersökningar på området inkluderar denna en undersökning om huruvida priset har betydelse, och i den genomförda regressionsanalysen står det klart att det finns en sådan effekt som är värd att ta i beaktande.

Det finns en positiv korrelation mellan Expressen och Aftonbladet, där båda tidningar ökar och minskar i försäljning samtidigt. Detta resultat antyder att det är andra faktorer som förklarar nedgången i försäljningen av Expressen och att den pågående nedgången drabbar alla aktörer i samma kategori. Resultatet betyder alltså att det väsentliga beslutet hos de potentiella tidningsköparna är *om* man ska köpa en kvällstidning överhuvudtaget och inte *vilken* tidning man väljer.

Sammantaget kan man säga att de resultat som framkommit i arbetet framstår som trovärdiga med bakgrund av tidigare studier. Papperstidningarna tappar i andelar enligt den genomförda tidsserieanalysen, precis som övriga studier indikerar. Resultaten från regressionsanalysen verkar också rimliga. Att ökad användning av internet (här mätt med antalet besökare på expressen.se) och prisökningar har en negativ effekt på försäljning är inte oväntat. Att det finns en positiv korrelation mellan Aftonbladet och Expressen kan först tyckas förvånande, men resultatet innebär att båda tidningarna påverkas av andra faktorer än den inbördes konkurrensen, och det är andra variabler som driver på den nedåtgående trenden för pappersupplagan av kvällstidningarna.

## 5.3 Kritik mot de egna slutsatserna

I tidsserieanalysen har metoden där man gör en ”out of sample”-analys använts för att bedöma de olika modellernas lämplighet. Då tidsserien delvis ändrar mönster runt år 2002 och våra observationer går tillbaka relativt långt i tiden finns det en risk att detta förfarande har lett till att den här föreslagna modellen inte är den modell som skulle ge absolut bäst prognoser om hela datamaterialet hade använts för modellvalet. Men om man undersöker de fyra bästa modellerna i tabell 4.3, där hela tidsserien används kan man trots denna invändning se att det är de exponentiella modellerna som har lägst varians, vilket antyder att dessa är mest lämpliga. Samma tabell ger dock också indikationer på att det skulle kunna vara så att det är den additiva och inte den multiplikativa modellen som har den bästa anpassningen till datan, om hela datamaterialet används för anpassningen. Denna metod för modellval är vald utifrån den arbetsgång som förordas i Montgomery m fl. (2008), och invändningarna mot den har vi inte ansett vara tillräckligt allvarliga för att vi skulle avvika från metoden. Särskilt inte eftersom de sista sex-sju åren av den period som modellen anpassas för följer samma synliga mönster som utvärderingsperioden.

## *5.4 Förslag till fortsatt forskning*

Det finns andra variabler som också kan inkluderas i en regressionsanalys. Exempelvis kan det tänkas vara intressant att inkludera morgontidningar, besöken på andra tidningars hemsidor etc. Något som vore önskvärt är att inkludera en variabel för att beskriva internets framväxt, så som bredbandstillgång eller motsvarande. På så vis skulle man också kunna skatta andra nya sätt att ta till sig nyheter, så som sociala medier och communitys. En annan viktig aspekt i undersökandet av tidningars relation till internet är en kvalitativ analys av ämnet. Likheter och skillnader i kvalitativa egenskaper för nät- och pappersupplaga skulle kunna vara en viktig förklaring i hur upplagan förändras över tiden.

I detta arbete föreslås inga reformer eller tänkbara förändringar för upplagans utveckling baserade på våra resultat. Exempelvis potentiella fördelar med att driva kategorin (kvällstidningar som helhet) jämfört med att driva den egna produkten (Expressen).

Trenden just nu är tydligt negativ, men det kan vara intressant att närmre undersöka om det finns skäl att tro att detta kommer att fortsätta, eller om det finns en kritisk gräns vid vilken nedgången planar ut.

I detta arbete utförs ingen analys av den marginella likelihood-funktionen, vilket skulle kunna vara av intresse. Att undersöka denna skulle kunna vara en möjlighet till fortsatt forskning på området.

## Referenser


Bergström, A. och Wadbring, I. (2010). Nya tidningsformer – konkurrenter eller komplement? i Holmberg, Sören och Weibull, Lennart (Red) Nordiskt ljus. 37 kapitel om politik, medier och samhälle. SOM institutet, Göteborgs universitet

Dutta-Bergman, M. (2004). Complementarity in consumption of news types across traditional and new media. Journal of Broadcasting and Electronic Media, 48(1), 41–61.

Färdigh, M. A. & Westlund, O (2011). Kvällspress i gamla och nya tappningar i Sören Holmberg, Lennart Weibull & Henrik Oscarsson (red) Lycksalighetens ö. Göteborg: SOM-institutet, Göteborgs universitet.

Gelman, A., Carlin, J.B., Stern, H.S., och Rubin D.B (2004). Bayesian data analysis, 2nd ed. Chapman & Hall/CRC.

Ghersetti, M. (2008). Mera film - men inte på bio, Publicerad i : Sören Holmberg och Lennart Weibull (red): Skilda världar. Trettioåtta kapitel om politik, medier och samhälle. SOM-institutet, s. 285-298

Ghersetti, M. (2011). Olika men ändå lika. Rapporteringen om riksdagsvalet 2010 i fem stora pappers- och webbtidningar. Göteborg: University of Gothenburg.

Jaynes, E. T. (1976). ”Confidence intervals vs. Bayesian intervals” i Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, (W.L. Harper och C. A. Hooker, eds.) Dordrecht D. Reidel

KTH (2011), Kapitel 1, Bayesianska metoder, Stockholm: KTH Matematik, avd. Matematisk Statistik; http://www.math.kth.se/matstat/gru/godis/bayes.pdf

Montgomery D.C,. Jennings C.L., Kulahci, M (2008). Introduction to Time Series Analysis and Forecasting. Hoboken, New Jersey: Wiley.

Nordicom (2011). Sveriges Internetbarometer 2010. MedieNotiser: 2-2011. Nordicom

Pindyck R.S och Rubinfeld D.L (1991). Econometric Models & Economic Forecasts, 3rd ed. New York: McGraw-Hill.

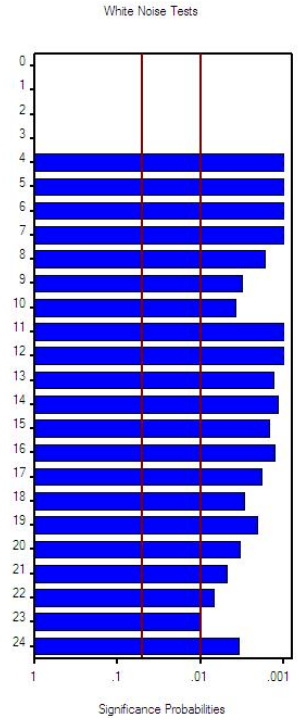
Robert C.P. (2001).The Bayesian Choice, 2nd ed. New York: Springer Verlag.

SAS Institute Inc. (2011). SAS/STAT 9.2 Users Guide, Second Edition. SAS Campus Drive, Cary, North Carolina

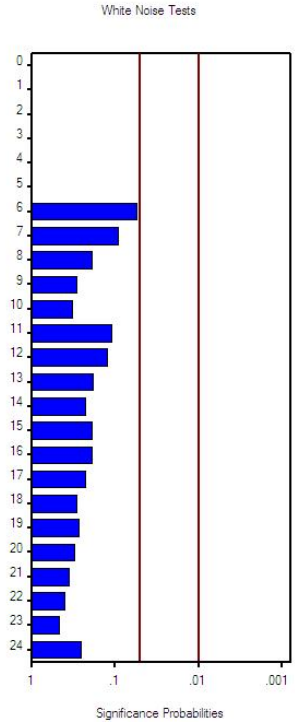
Tidningsstatistik AB (2010). Månadsupplagor Aftonbladet 2009-12

# Appendix

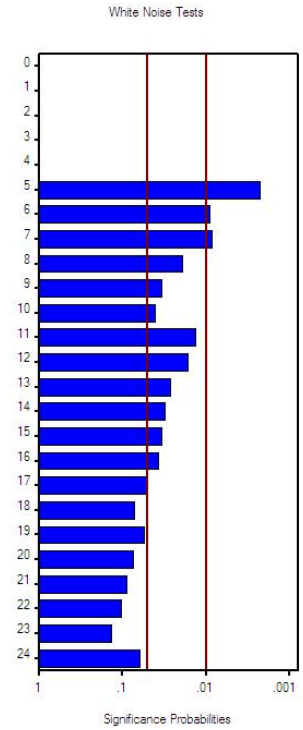
## Appendix 1, Ljung-Box test




Ljung-Box 1 ARIMA(1,1,0)(2,1,0)s

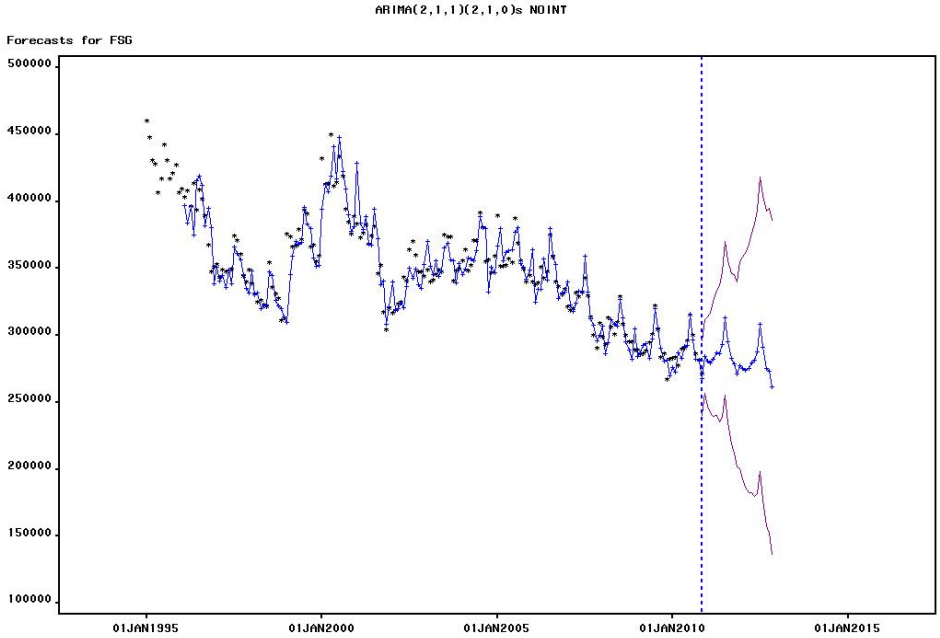



Ljung-Box 3 ARIMA(2,1,1)(2,1,0)s

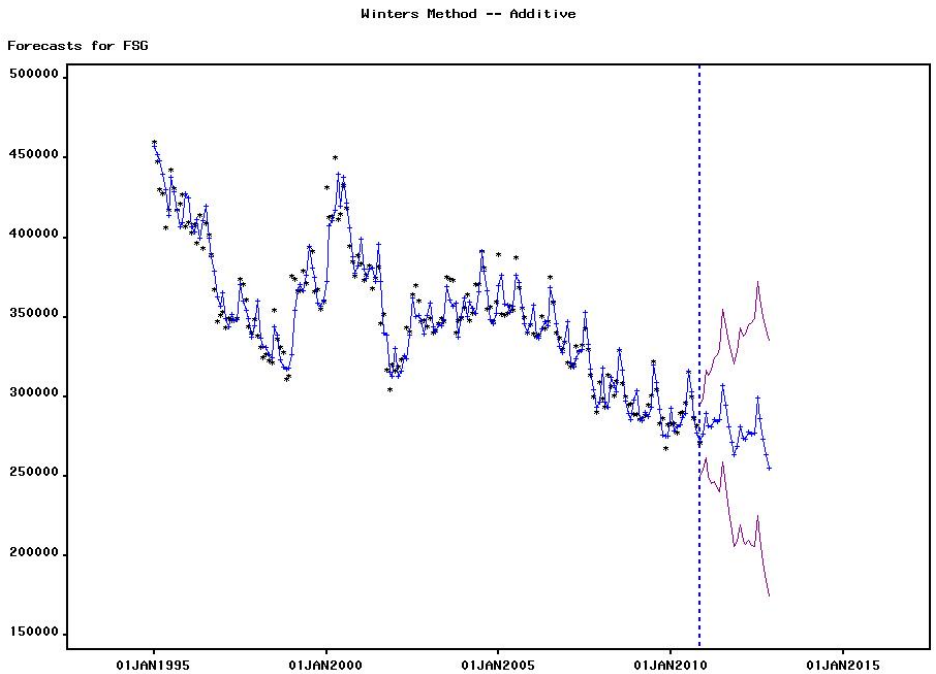



Ljung-Box 2 ARIMA(2,1,0)(2,1,0)s

## Appendix 2, Prognoser

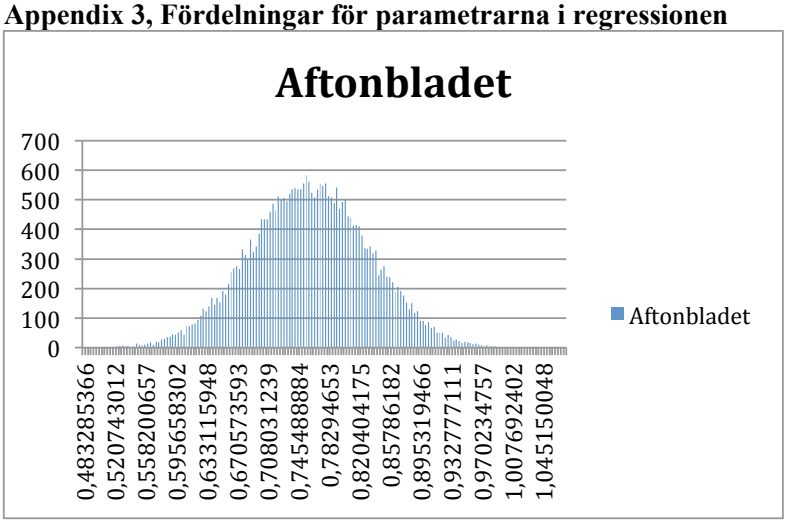



Prognos 1, ARIMA(2,1,1)(2,1,0)s

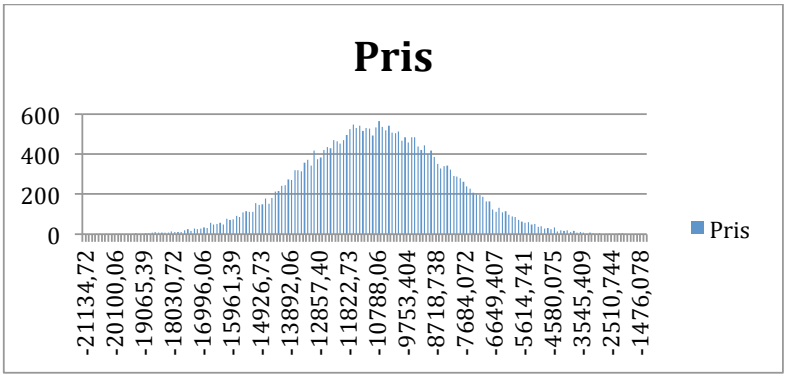



Prognos 2, Additiv Holt-Winter

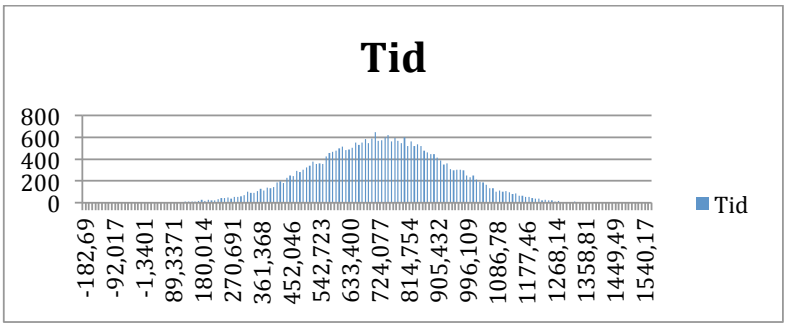
**Appendix 3, Fördelningar för parametrarna i regressionen**



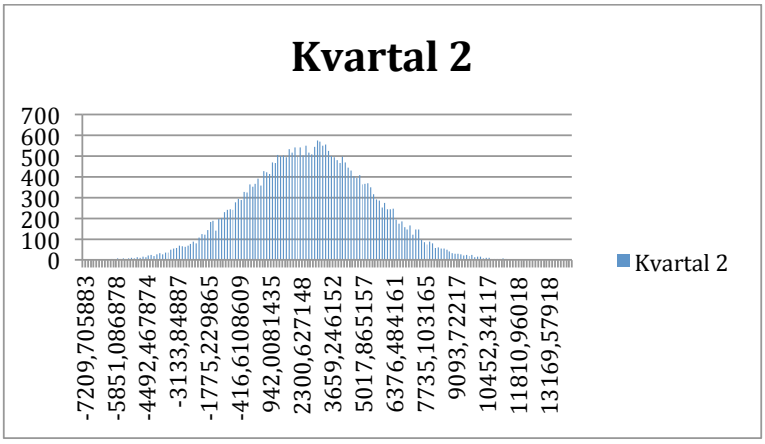

Fördelning 1. Parametern för försäljning av Aftonbladet



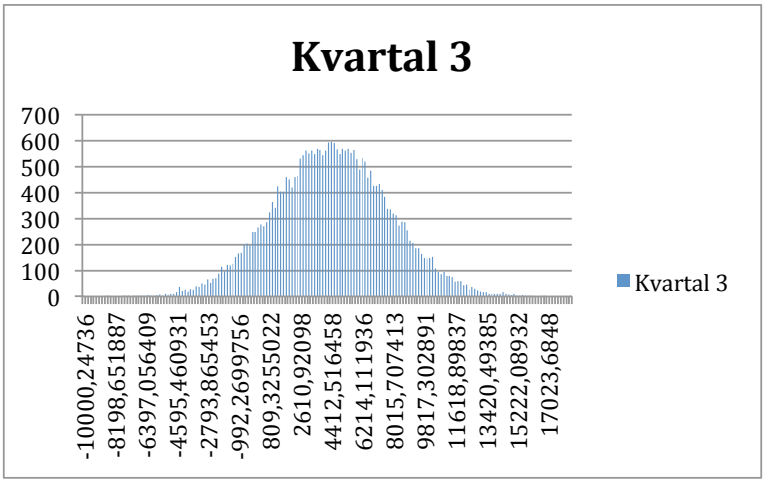

Fördelning 2. Parametern för priset på Expressen uttryckt i 2002 års priser



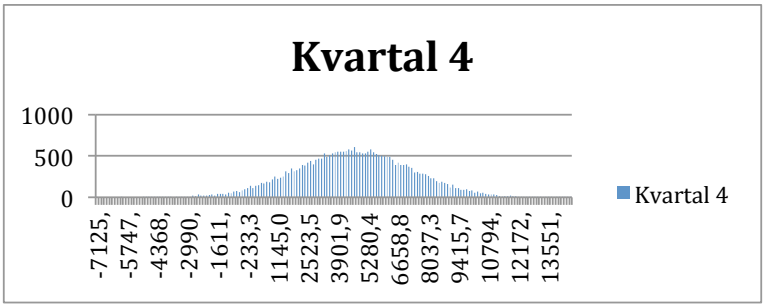

Fördelning 3. Parametern för effekten av tid




Fördelning 4. Parametern för effekten av kvartal 2 (jmf kvartal 1)




Fördelning 5. Parametern för effekten av kvartal 3 (jmf kvartal 1)

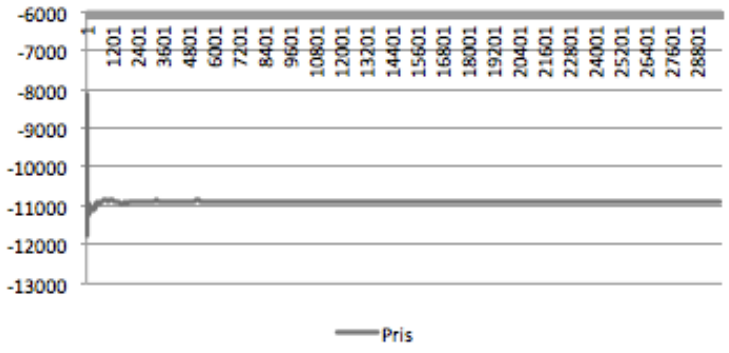


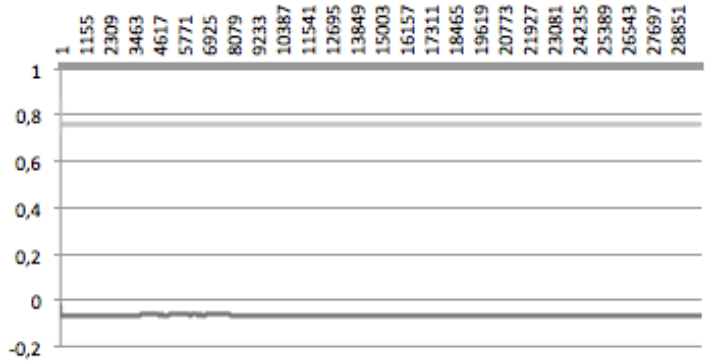

Fördelning 6. Parametern för effekten av kvartal 4 (jmf kvartal 1)

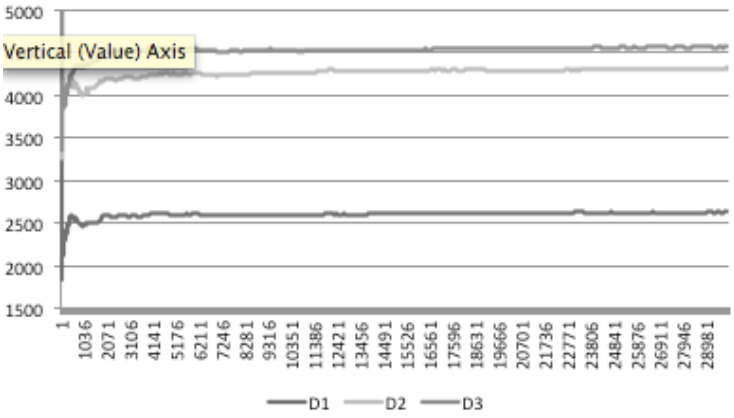
**Appendix 4. Prognos för 24 månader framåt, samt konfidensintervallsbredd för de fyra valda modellerna.**

| | Holt-Winter multip. | | Holt-Winter add | | ARIMA(2,1,1)(2,1,0)s | | ARIMA(2,1,0)(2,1,0)s | |
|---|---|---|---|---|---|---|---|---|
| | Predikt. | k.i./2 | Predikt. | k.i./2 | Predikt. | k.i./2 | Predikt. | k.i./2 |
| dec-10 | 277742 | ±23203 | 276846 | ±22273 | 284317 | ±27614 | 283921 | ±26546 |
| jan-11 | 283799 | ±28670 | 289325 | ±27378 | 280590 | ±33507 | 280677 | ±31689 |
| feb-11 | 281028 | ±32933 | 281989 | ±31679 | 279280 | ±37301 | 278807 | ±35467 |
| mar-11 | 280953 | ±36870 | 281213 | ±35470 | 282498 | ±43218 | 282445 | ±39776 |
| apr-11 | 283463 | ±40696 | 285623 | ±38898 | 287134 | ±46825 | 286951 | ±43464 |
| maj-11 | 281555 | ±43777 | 284670 | ±42054 | 286170 | ±50668 | 285893 | ±46792 |
| jun-11 | 281417 | ±46857 | 285397 | ±44996 | 293209 | ±54321 | 293039 | ±49951 |
| jul-11 | 297353 | ±52002 | 307156 | ±47761 | 312768 | ±57513 | 312538 | ±52914 |
| aug-11 | 286139 | ±52999 | 294527 | ±50381 | 295223 | ±60702 | 294994 | ±55715 |
| sep-11 | 276619 | ±54114 | 281168 | ±52875 | 282719 | ±63670 | 282523 | ±58385 |
| okt-11 | 270824 | ±55722 | 271354 | ±55262 | 278669 | ±66502 | 278450 | ±60937 |
| nov-11 | 264369 | ±57090 | 263422 | ±57553 | 270882 | ±69242 | 270710 | ±63387 |
| dec-11 | 270079 | ±61176 | 268933 | ±59764 | 277539 | ±77191 | 277097 | ±70889 |
| jan-12 | 275951 | ±64702 | 281412 | ±61898 | 275595 | ±82631 | 275416 | ±75773 |
| feb-12 | 273239 | ±66379 | 274076 | ±63973 | 273737 | ±87276 | 273251 | ±80118 |
| mar-12 | 273148 | ±68562 | 273301 | ±65970 | 275149 | ±92623 | 274894 | ±84591 |
| apr-12 | 275570 | ±71264 | 277710 | ±67921 | 279589 | ±97048 | 279247 | ±88758 |
| maj-12 | 273697 | ±72909 | 276757 | ±69820 | 281445 | ±101486 | 281070 | ±92711 |
| jun-12 | 273545 | ±74925 | 277484 | ±71673 | 287866 | ±105774 | 287547 | ±96523 |
| jul-12 | 289015 | ±80860 | 299243 | ±73483 | 308200 | ±109793 | 307854 | ±100188 |
| aug-12 | 278097 | ±79920 | 286615 | ±75252 | 290760 | ±113744 | 290411 | ±103720 |
| sep-12 | 268826 | ±79363 | 273255 | ±76985 | 275142 | ±117535 | 274789 | ±107138 |
| okt-12 | 263177 | ±79744 | 263442 | ±78682 | 273461 | ±121205 | 273115 | ±110450 |
| nov-12 | 256886 | ±79897 | 255509 | ±80347 | 261469 | ±124778 | 261123 | ±113664 |

**Appendix 5. Figurer över konvergens kring medelvärden.**




1
1155
2309
3463
4617
5771
6925
8079
9233
10387
11541
12695
13849
15003
16157
17311
18465
19619
20773
21927
23081
24235
25389
26543
27697
28851
1
0,8
0,6
0,4
0,2
0
-0,2
Expressen.se
Aftonbladet


5000
Vertical (Value) Axis
4000
3500
3000
2500
2000
1500
1
1036
2071
3106
4141
5176
6211
7246
8281
9316
10351
11386
12421
13456
14491
15526
16561
17596
18631
19666
20701
21736
22771
23806
24841
25876
26911
27946
28981
D1
D2
D3