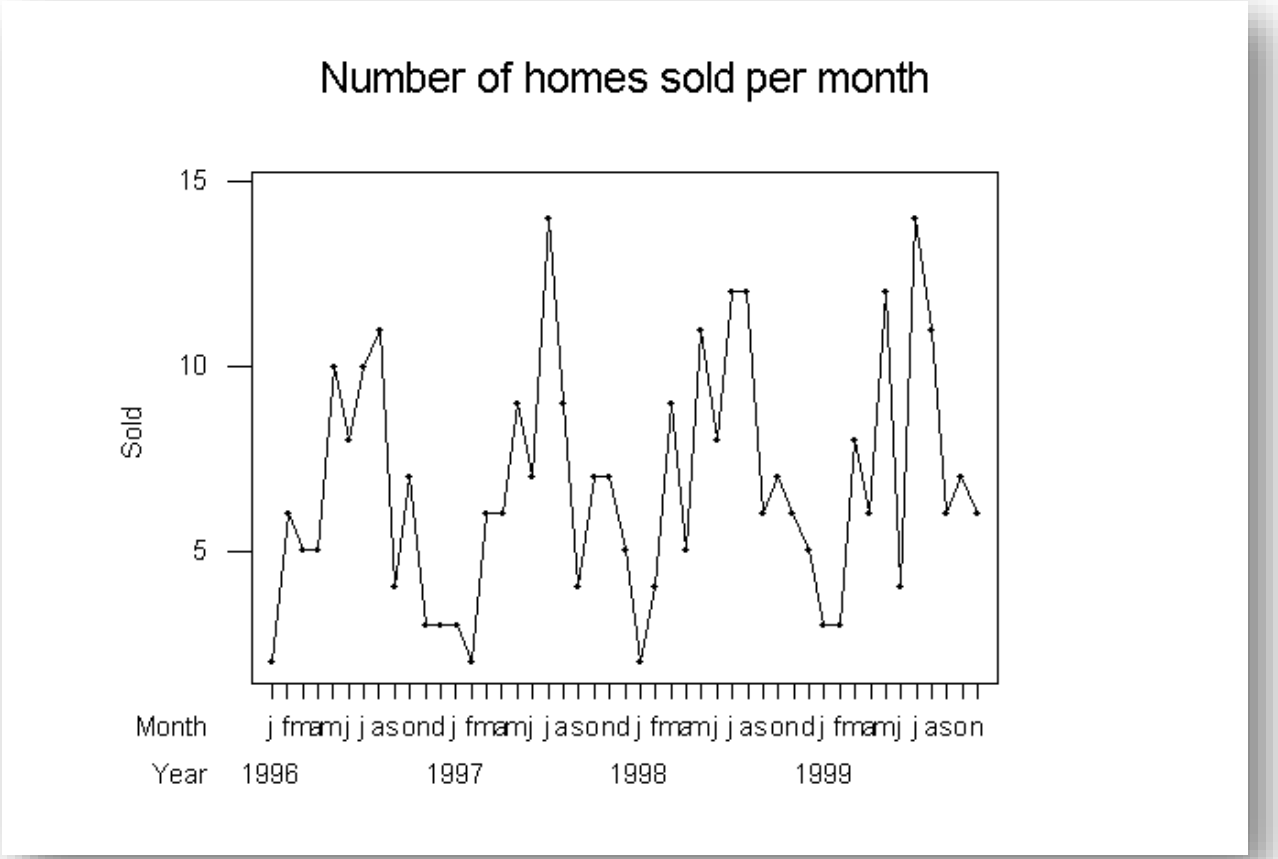
Föreläsning 3
Tidsserie-regression mm

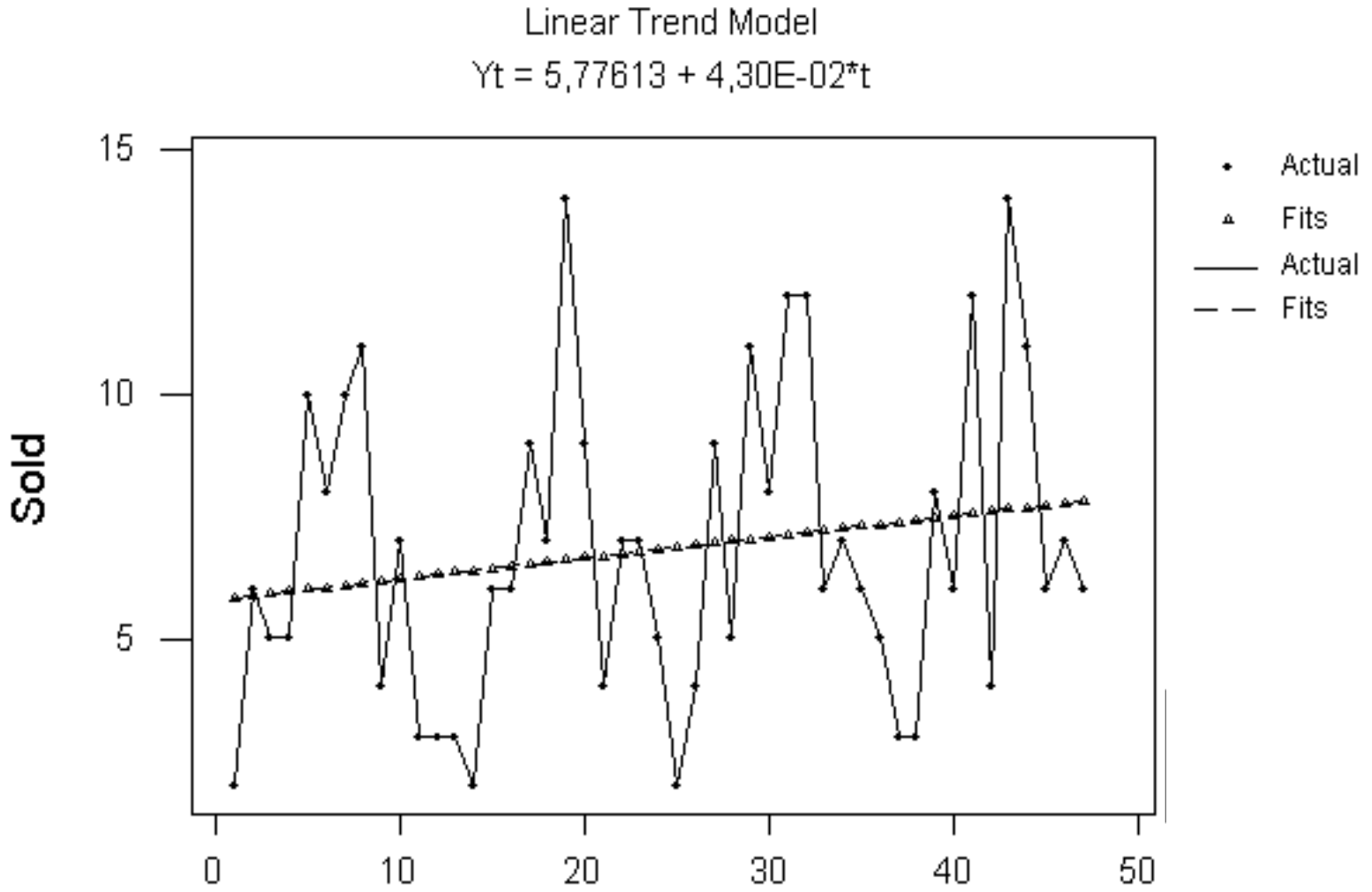
Kapital 3

Åter till ex med Sold, antal sålda hus över tiden.

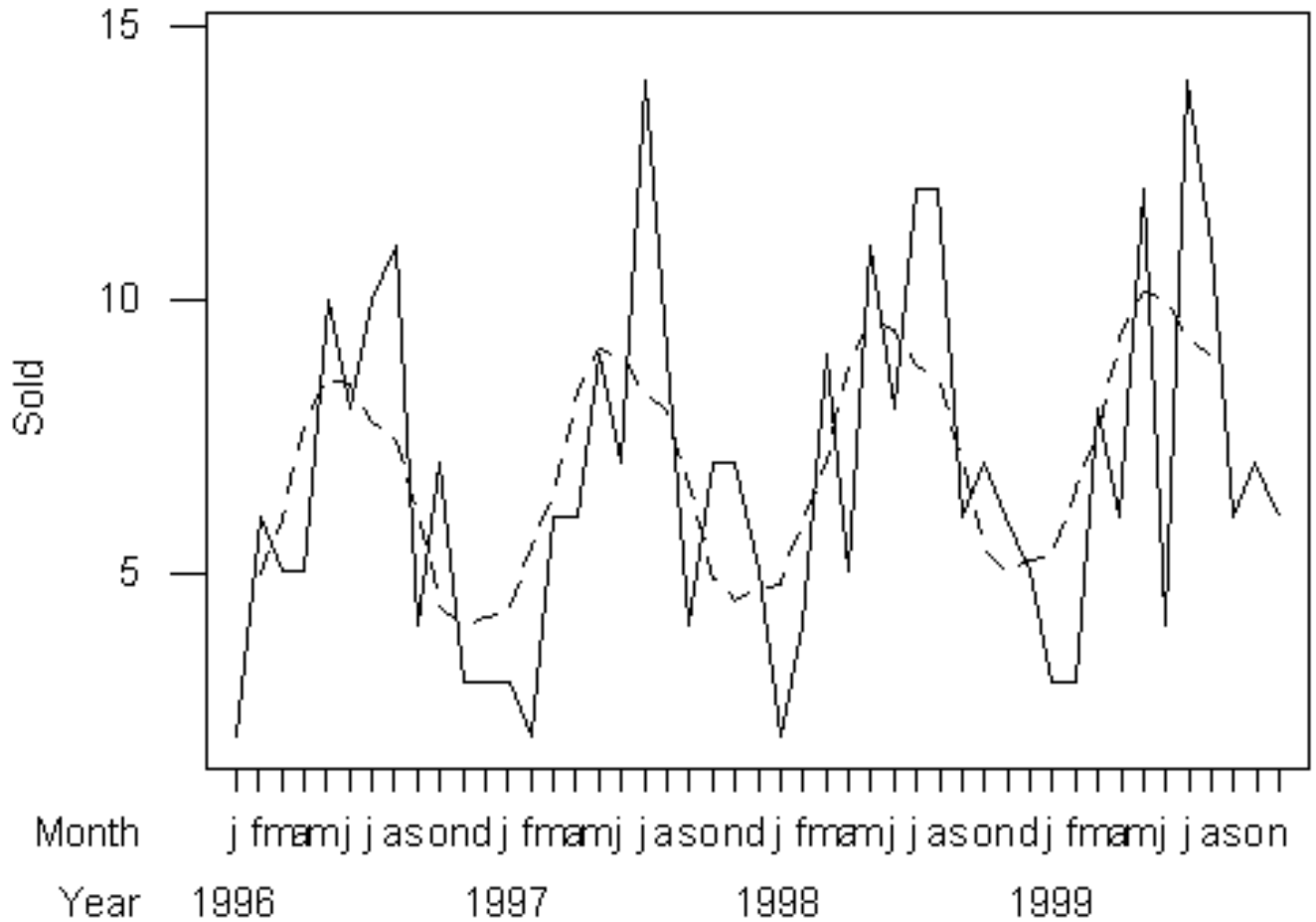
- Typ av trend
- Typ av säsongsvariation
- Additivt eller multiplikativt mönster
- Statiskt eller dynamiskt mönster
- Tolka serien eller beräkna prognos



Finns det en linjär trend i data?

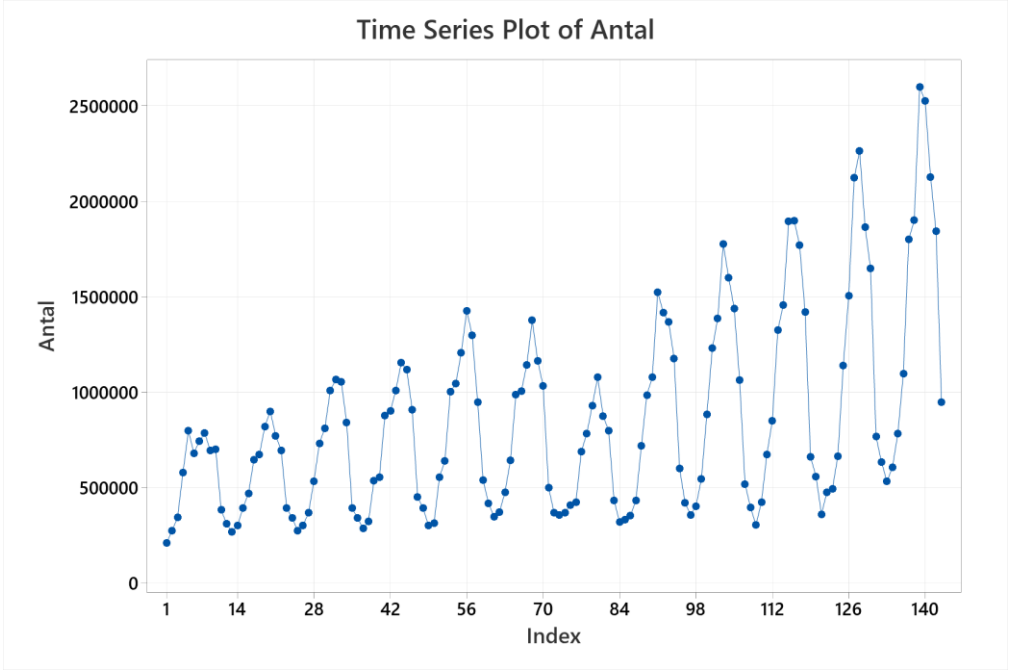


Finns det säsongvariation i data?



Transformera om nödvändigt

- Om transformation av tidsserien behövs så måste det göras först av allt
- Så var inte fallet vid komponentuppdelning. Då valdes multiplikativ modell istället



Trendanalys

Gäller både tidsseriergression och komponentuppdelning

- Oavsett om tidsserien har säsongsvariation eller annan variation så kan vi göra en analys av trenden. Beteckna trendfunktonen med TR

Data med endast linjär trend :

$$y_t = TR_t + \varepsilon_t = \beta_0 + \beta_1 \cdot t + \varepsilon_t$$

Data med kvadratisk trend :

$$y_t = TR_t + \varepsilon_t = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \varepsilon_t$$

Data med kubisk trend :

$$y_t = TR_t + \varepsilon_t = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot t^2 + \beta_3 \cdot t^3 + \varepsilon_t$$

Modellering av säsongsvariation och trend i tidsserieregression

- Låt TR vara trendkomponent och låt SN vara säsongskomponent

$$y_t = TR_t + SN_t + \varepsilon_t$$

- Vi har tittat på att TR kan modelleras med en linjär, kvadratisk eller kubisk trend.

För enkelhets skull låter vi trenden vara linjär just nu:

$$y_t = TR_t + SN_t + \varepsilon_t = \beta_0 + \beta_1 \cdot t + \beta_{s1} \cdot x_{1,t} + \beta_{s2} \cdot x_{2,t} + \ldots + \beta_{s11} \cdot x_{11,t} + \varepsilon_t$$

där $x_{1,t} = 1$ om säsong till tidpunkt t är 1; och 0 annars,

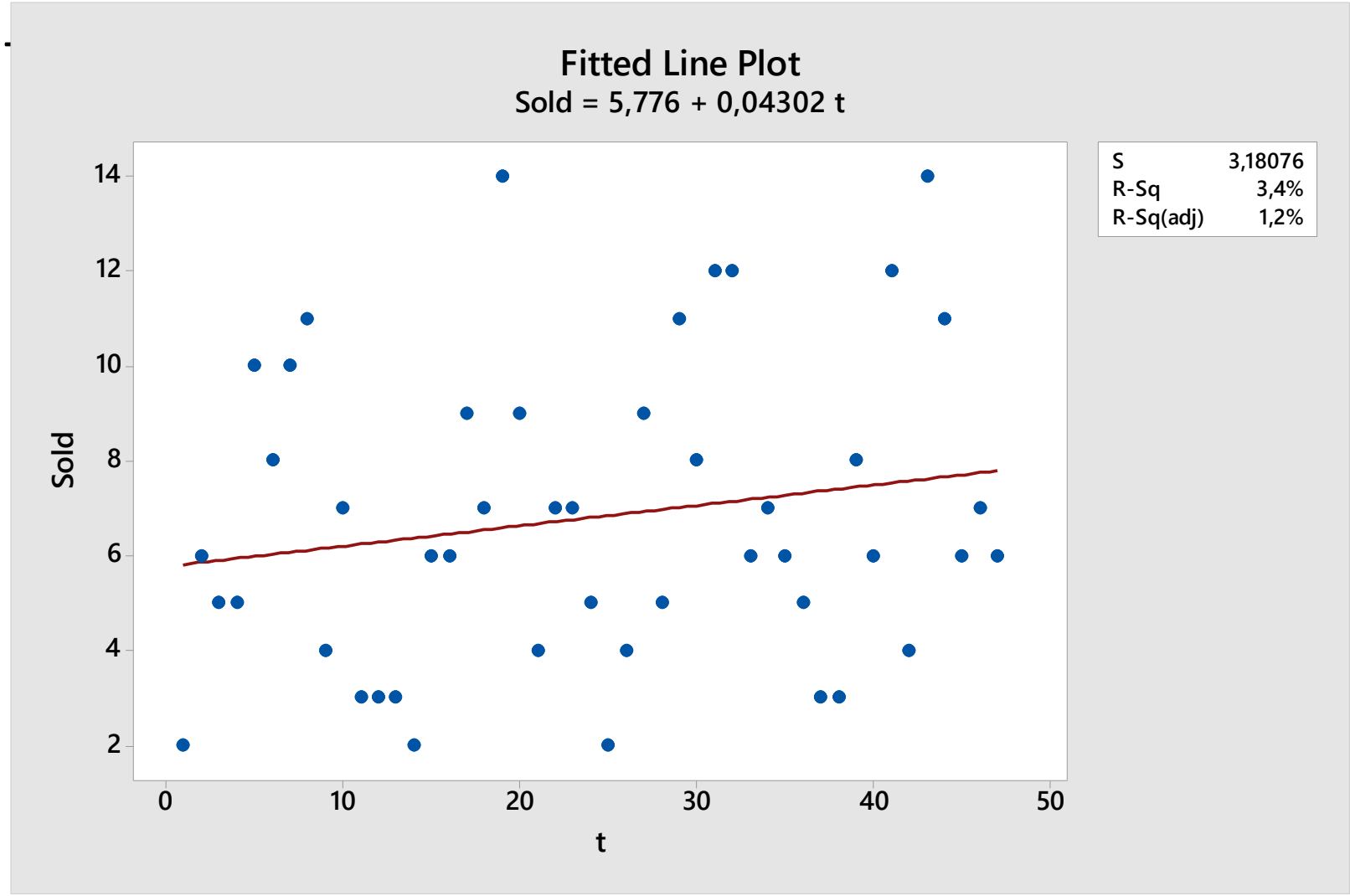
$x_{2,t} = 1$ om säsong till tidpunkt t är 2; och 0 annars,...

osv

s.k. "säsongsdummies"

sold	time	month	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}
2	1	1	1	0	0	0	0	0	0	0	0	0	0
6	2	2	0	1	0	0	0	0	0	0	0	0	0
5	3	3	0	0	1	0	0	0	0	0	0	0	0
5	4	4	0	0	0	1	0	0	0	0	0	0	0
10	5	5	0	0	0	0	1	0	0	0	0	0	0
8	6	6	0	0	0	0	0	1	0	0	0	0	0
.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.
7	46	10	0	0	0	0	0	0	0	0	0	1	0
6	47	11	0	0	0	0	0	0	0	0	0	0	1
3	48	12	0	0	0	0	0	0	0	0	0	0	0

Tidsserieregression med enbart en linjär



Regression Analysis: Sold versus t

The regression equation is
Sold = 5,776 + 0,04302 t

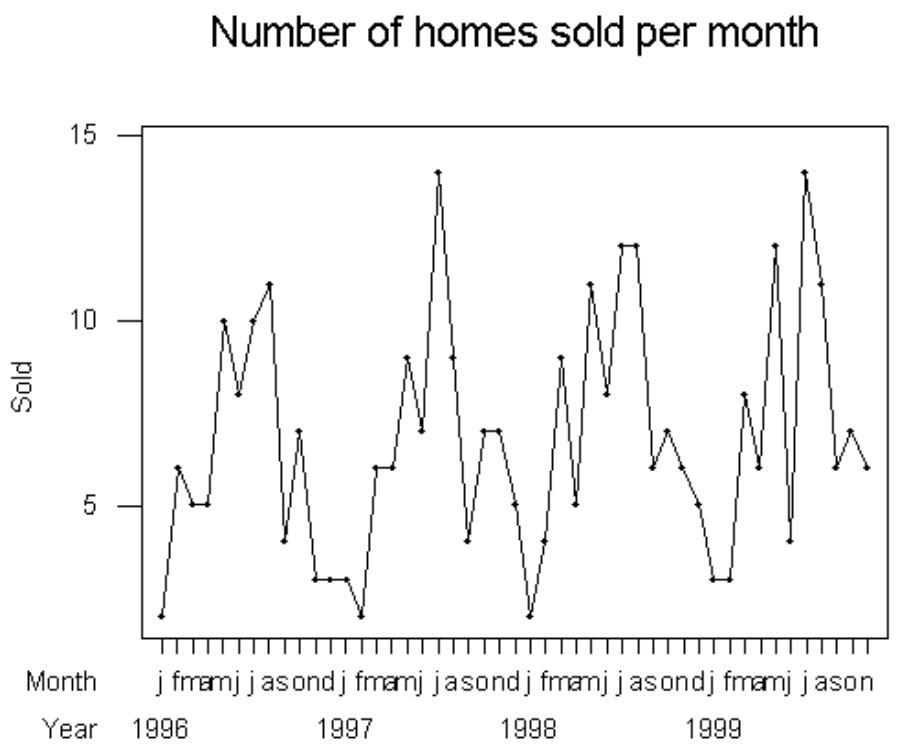
Model Summary

S	R-sq	R-sq(adj)
3,18076	3,40%	1,25%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	16,002	16,0019	1,58	0,215
Error	45	455,275	10,1172		
Total	46	471,277			

Fitted Line: Sold versus t



Tidsserieregression fungerar statistiskt som vanlig regression.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
1,34214	87,00%	82,42%	75,02%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3,649	0,853	4,28	0,000	
t	0,0285	0,0148	1,92	0,063	1,05
Månad_1	-1,69	1,03	-1,65	0,109	2,15
Månad_2	-0,47	1,03	-0,46	0,651	2,14
Månad_3	2,75	1,03	2,68	0,011	2,14
Månad_4	1,22	1,03	1,19	0,241	2,14
Månad_5	6,20	1,03	6,04	0,000	2,14
Månad_6	2,42	1,03	2,36	0,024	2,13
Månad_7	8,14	1,03	7,94	0,000	2,14
Månad_8	6,36	1,03	6,20	0,000	2,14
Månad_9	0,58	1,03	0,57	0,575	2,14
Månad_10	2,55	1,03	2,49	0,018	2,14
Månad_11	1,02	1,03	1,00	0,326	2,15

Regression Equation

Sold = 3,649 + 0,0285 t - 1,69 Månad_1 - 0,47 Månad_2 + 2,75 Månad_3 + 1,22 Månad_4 + 6,20 Månad_5 + 2,42 Månad_6 + 8,14 Månad_7 + 6,36 Månad_8 + 0,58 Månad_9 + 2,55 Månad_10 + 1,02 Månad_11

Coefficients

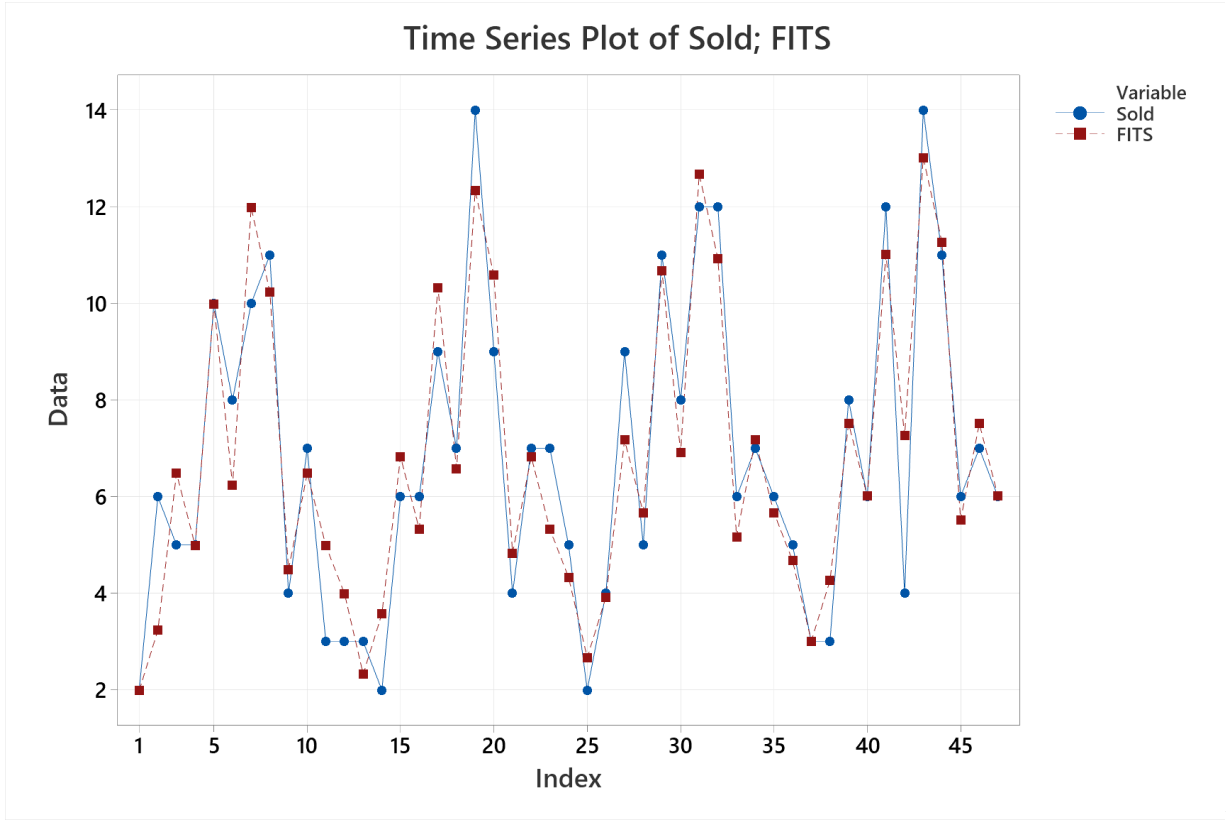
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	3,649	0,853	4,28	0,000	
t	0,0285	0,0148	1,92	0,063	1,05
Månad_1	-1,69	1,03	-1,65	0,109	2,15
Månad_2	-0,47	1,03	-0,46	0,651	2,14
Månad_3	2,75	1,03	2,68	0,011	2,14
Månad_4	1,22	1,03	1,19	0,241	2,14
Månad_5	6,20	1,03	6,04	0,000	2,14
Månad_6	2,42	1,03	2,36	0,024	2,13
Månad_7	8,14	1,03	7,94	0,000	2,14
Månad_8	6,36	1,03	6,20	0,000	2,14
Månad_9	0,58	1,03	0,57	0,575	2,14
Månad_10	2,55	1,03	2,49	0,018	2,14
Månad_11	1,02	1,03	1,00	0,326	2,15

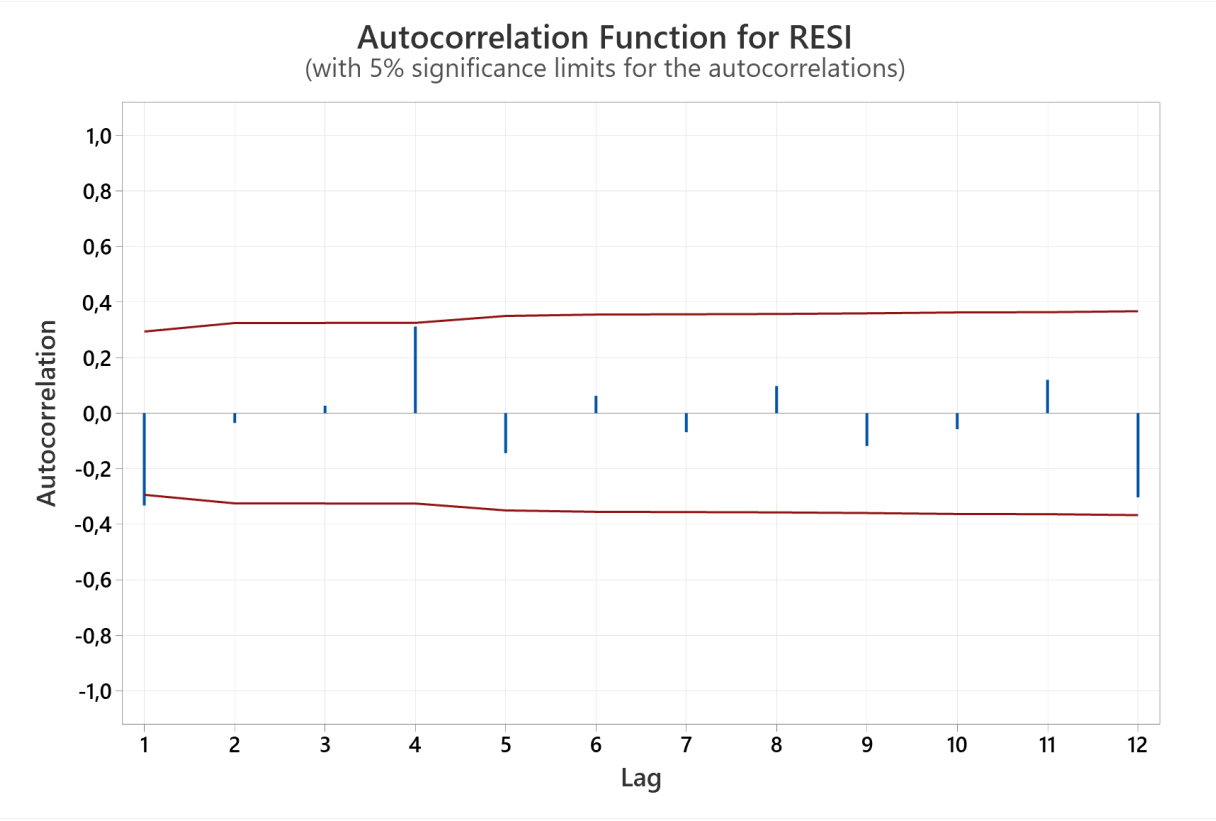
Tolkning av parametrar:

- Antal sålda hus ökar i genomsnitt med 0,0285 enheter per tidsenhet (månad) (alla andra variabler hålls fixa)
- I januari säljs det färre hus (-1.69 hus) jämfört med december, i mars säljs det fler hus (+ 2.75)....

(observera att december är basperioden (referensmånad), eftersom dummy- variabeln för december inte finns med – decembernivån är alltså inbakad i konstanten)

SAC på residualer





Trend- och säsongsanalys

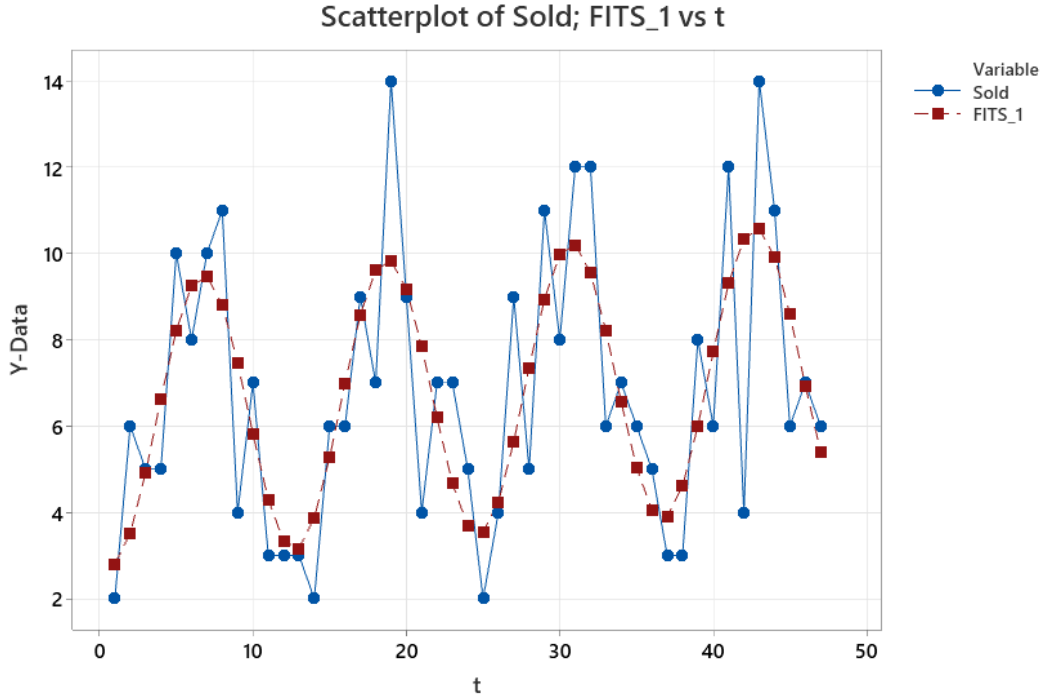
$$y_t = TR_t + SN_t + \varepsilon_t$$

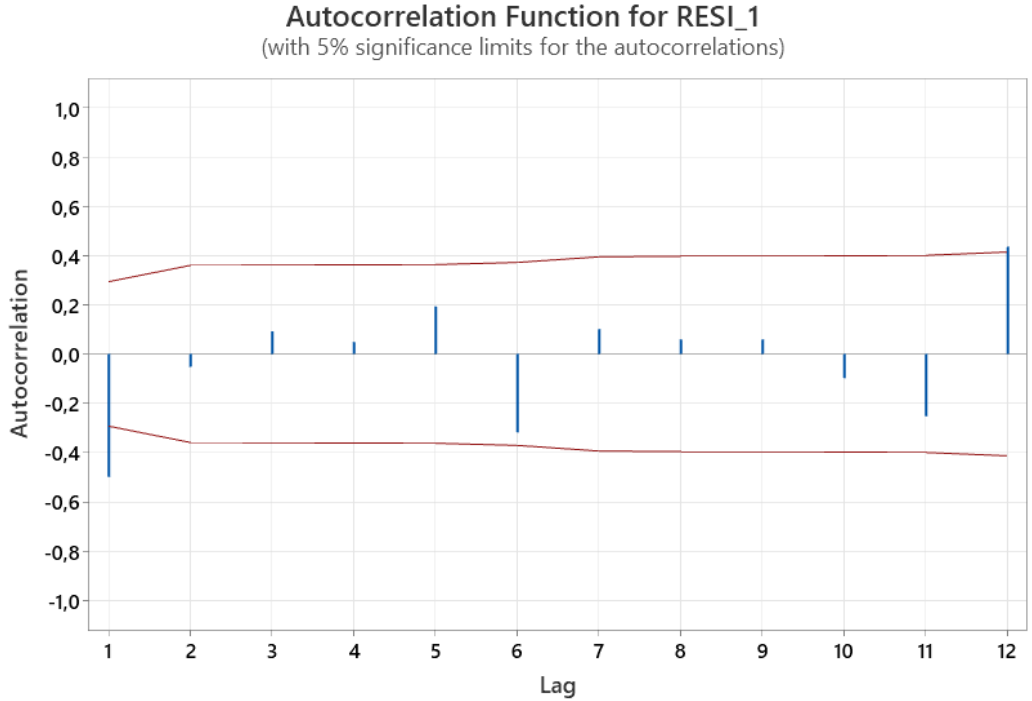
$$y_t = \beta_0 + \beta_1 t + \beta_2 sin\left(\frac{2\pi}{d} t\right) + \beta_3 cos\left(\frac{2\pi}{d} t\right) + \varepsilon_t$$

där d är säsongslängden
Ex $d = 12$ vid månadsdata

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	6,008	0,662	9,07	0,000	
t	0,0306	0,0241	1,27	0,211	1,04
sin	-1,189	0,458	-2,60	0,013	1,04
cos	-3,056	0,459	-6,66	0,000	1,00


Scatterplot of Sold; FITS_1 vs t
Variable
Sold
FITS_1
Y-Data
t
14
12
10
8
6
4
2
0
10
20
30
40
50


Autocorrelation Function for RESI_1
(with 5% significance limits for the autocorrelations)
Autocorrelation
Lag
1,0
0,8
0,6
0,4
0,2
0,0
-0,2
-0,4
-0,6
-0,8
-1,0
1
2
3
4
5
6
7
8
9
10
11
12

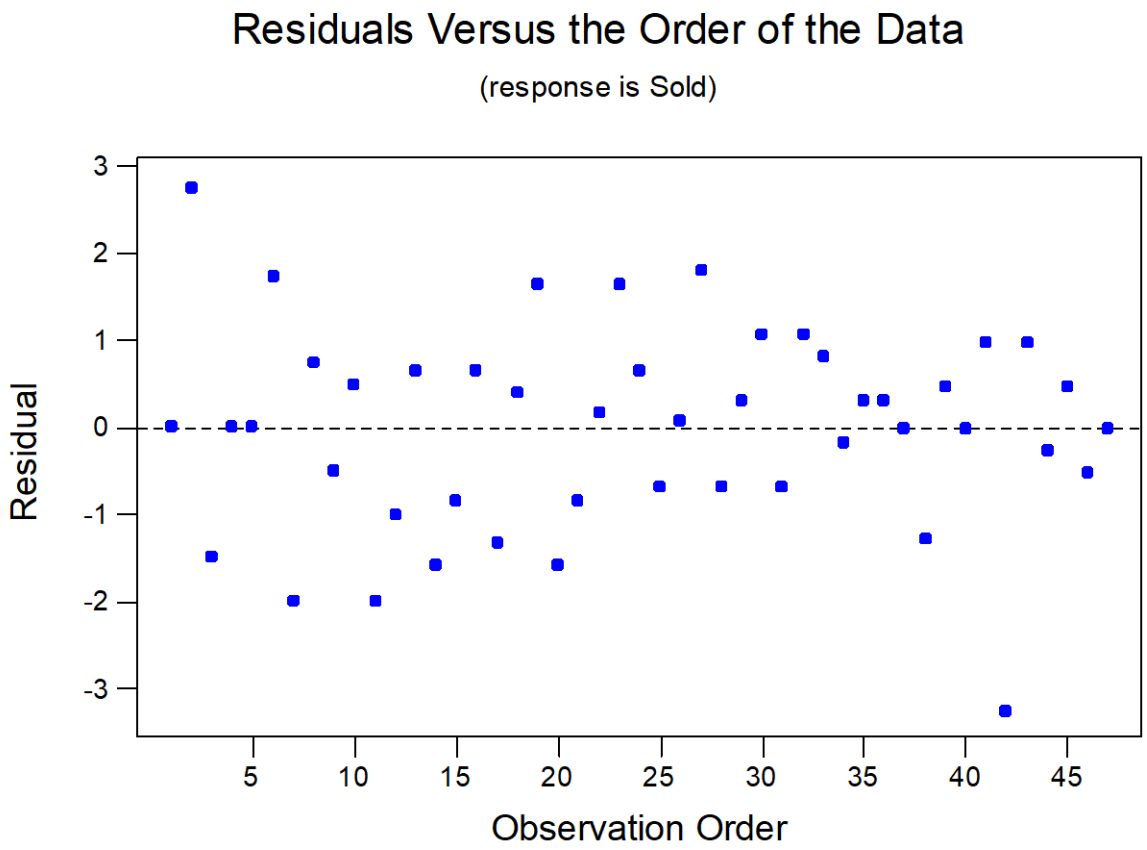
Autokorrelation i tidsserien

Inferens, som konfidensintervall, prognosintervall, t-test, F-test, partiellt F-test... kan göras på samma sätt som i vanlig regressionsanalys.

Residualanalys ska göras för att kontrollera om villkoren för regression är uppfyllt:

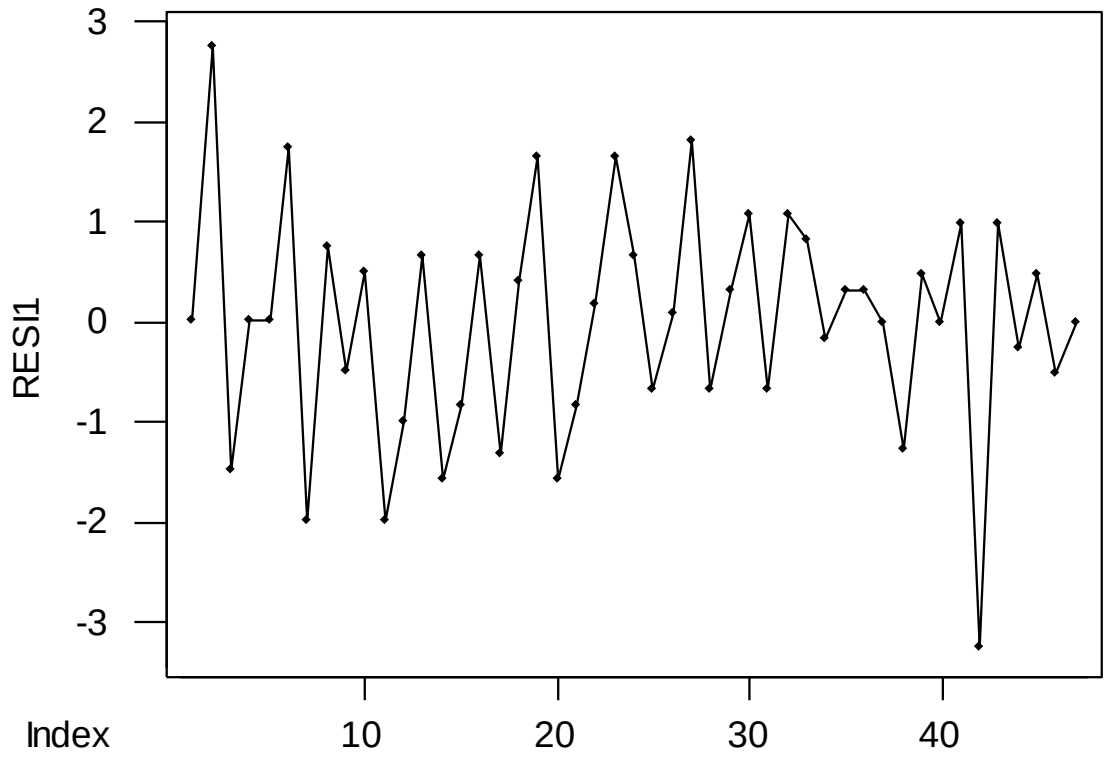
- Oberoende residualer
- Normalfördelade residualer
- Residualer med konstant varians

Antagandet om oberoende residualer är ofta inte uppfyllt när det gäller tidsseriedata. Det kan också vara svårt att kontrollera detta antagande visuellt.



Enklare att se om autokorrelation finns om observationerna är sammanbundna. Här ser man tydligt att en negativ residual vanligtvis följs av en positiv residual och tvärtom.

Detta är ett tecken på negativ autokorrelation.

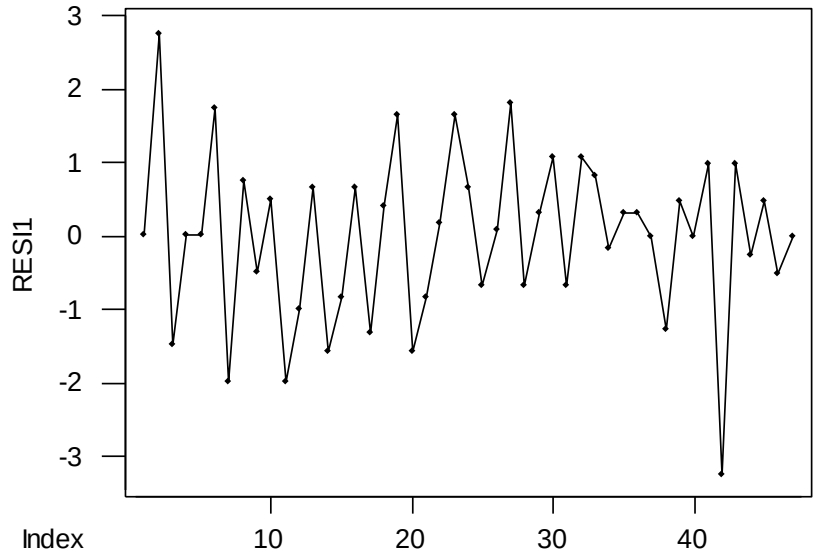


Statistiskt test för att kontrollera om residualerna är oberoende: Durbin-Watson-test

- Durbin-Watson-testet bedömer om *autokorrelation* (eller *seriell korrelation)* förekommer bland residualerna:

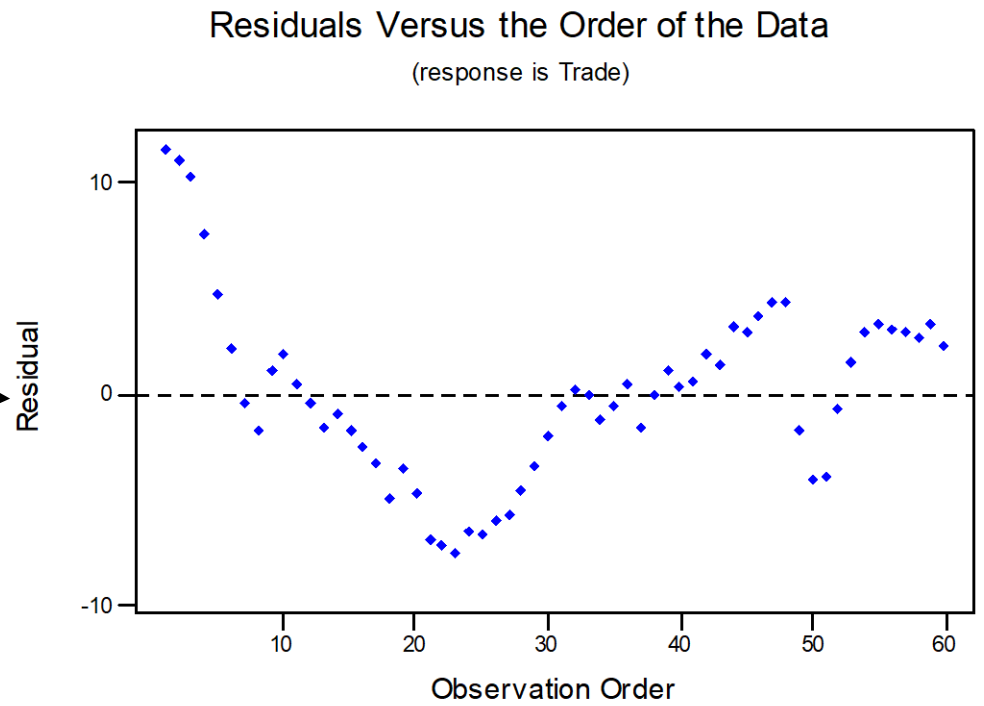
$$Corr(e_t, e_{t-1})$$

- Vi skiljer mellan positiv autokorrelation och negativ autokorrelation



Negativ autokorrelation

Positiv autokorrelation



Durbin-Watson-test

testvariabeln: $$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

Regression→Fit regression model, Result

Regression Analysis: Sold versus t; Månad_1; Månad_2; ... 10; Månad_11

Regression Equation

Sold = 3,649 + 0,0285 t - 1,69 Månad_1 - 0,47 Månad_2 + 2,75 Månad_3 + 1,22 Månad_4
+ 6,20 Månad_5 + 2,42 Månad_6 + 8,14 Månad_7 + 6,36 Månad_8 + 0,58 Månad_9
+ 2,55 Månad_10 + 1,02 Månad_11

Durbin-Watson Statistic

Durbin-Watson Statistic = 2,66414

Om det inte finns någon autokorrelation i residualerna så kommer d att ligga nära 2.

En approximativ kontroll kan göras genom att se om

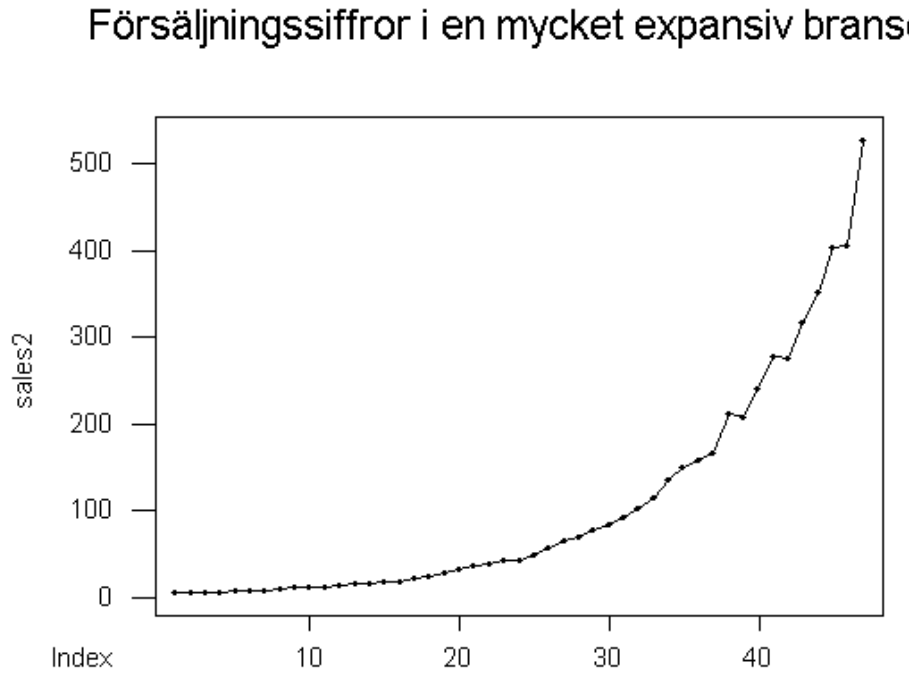
d är mindre än 1 eller större än 3

Då finns autokorrelation i residualerna.

Vi har verktyget SAC som ger mycket mer information

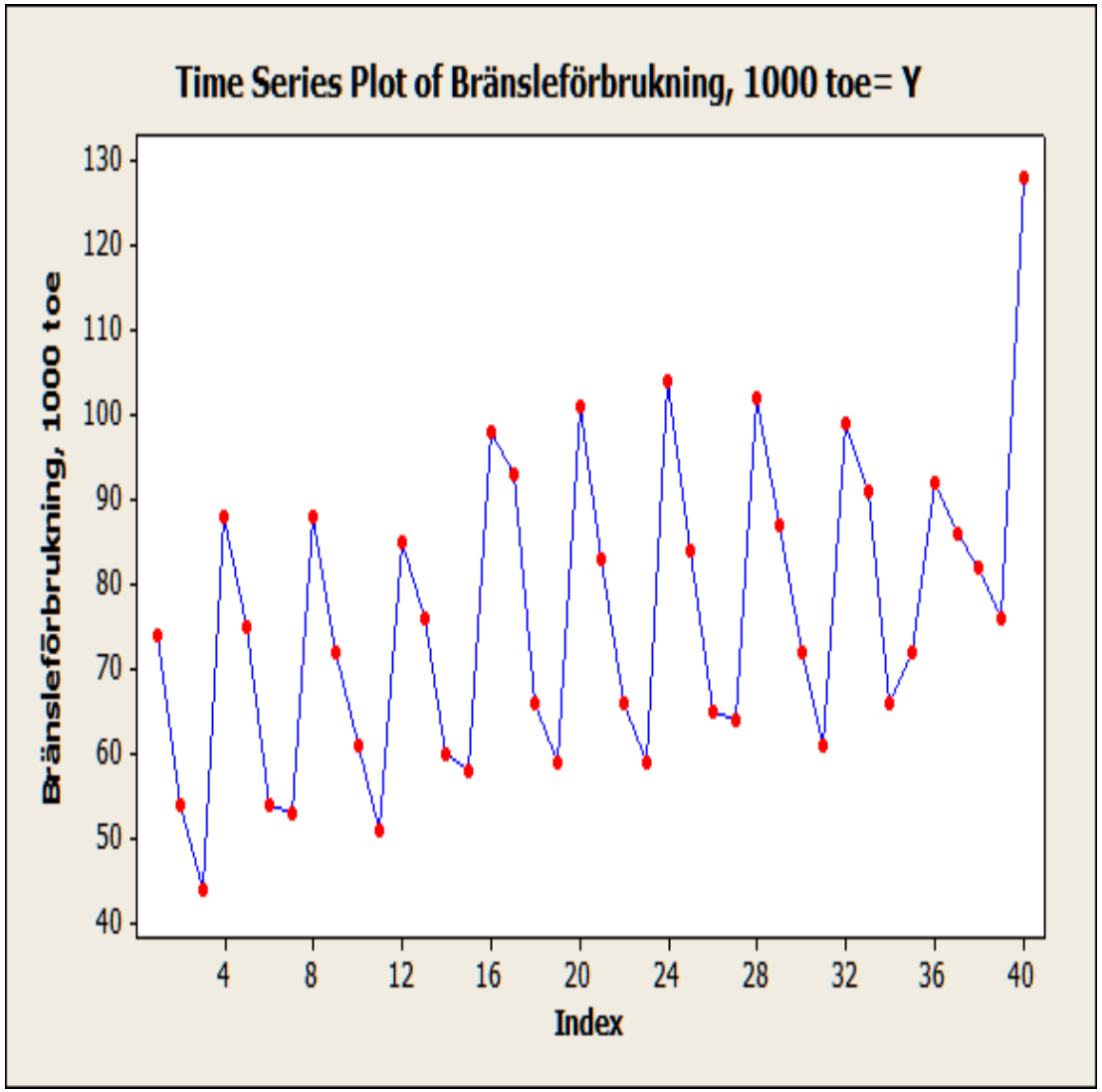
Exponentiell modell. Growth curve model

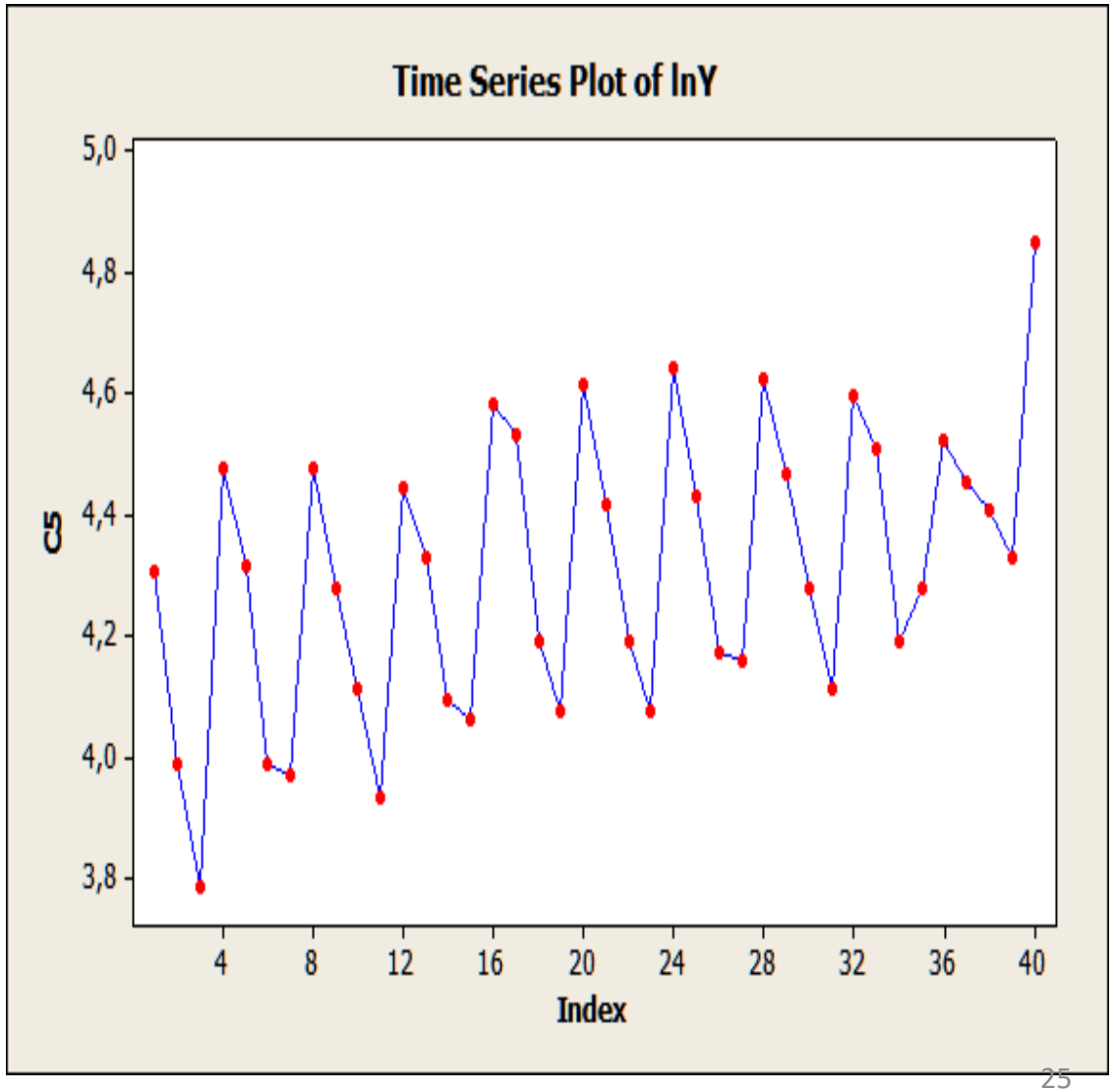
- Vissa tidsserier har en så kallad *exponentiell* trend:



$$y_t = \beta_0 \cdot \beta_1^t \cdot \delta_t$$

$Y =$ bränsleförbrukning av naturgas





Tidsserieregression för lnY

- Grafen över data med lnY ser något bättre ut, dvs amplituden ser mer jämn ut, ej ökad varians. Vi kommer dessutom få en tolkning av parametrarna i procent vilket kan vara fördelaktigt.
- Vi väljer linjär trend i modellen för lnY. Skapa en tidsvariabel.
- Skapa en kvartalsvariabel
- Bilda indikatorvariabler för kvartal med hjälp av kvartalvariabeln och välj ut tre dummies
- Kontrollera med DW om residualerna är autokorrelerade
- Finns autokorrelation med andra tidsförskjutningar?

Arbetsblad i minitab

tid	Y	lnY	t	kvartal	kv1	kv2	kv3	kv4
1992K1	74	4,30407	1	1	1	0	0	0
1992K2	54	3,98898	2	2	0	1	0	0
1992K3	44	3,78419	3	3	0	0	1	0
1992K4	88	4,47734	4	4	0	0	0	1
1993K1	75	4,31749	5	1	1	0	0	0
1993K2	54	3,98898	6	2	0	1	0	0
…..								
2000K1	91	4,51086	33	1	1	0	0	0
2000K2	66	4,18965	34	2	0	1	0	0
2000K3	72	4,27667	35	3	0	0	1	0
2000K4	92	4,52179	36	4	0	0	0	1
2001K1	86	4,45435	37	1	1	0	0	0
2001K2	82	4,40672	38	2	0	1	0	0
2001K3	76	4,33073	39	3	0	0	1	0
2001K4	128	4,85203	40	4	0	0	0	1

- Kvartal 3 är *referenskvartal* här då det ser ut att vara minst förbrukning under detta kvartal.
- Begär att få DW-statistikan
- Spara residualerna
- Begär *Four in one* för residualerna

- Analys av modellen ska göras i den linjära formen

$$lnY_t = \beta_0 + \beta_1 t + \beta_2 kv1 + \beta_3 kv2 + \beta_4 kv4 + \varepsilon_t$$

- Residualer $e_t = lny_t - \widehat{lny}_t$

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	1,99554	0,49888	100,20	0,000
Residual Error	35	0,17426	0,00498		
Total	39	2,16979			

Term	Coef	SE Coef	T	P
Constant	3,89557	0,03023	128,87	0,000
t	0,0086823	0,0009711	8,94	0,000
kv1	0,34370	0,03162	10,87	0,000
kv2	0,09178	0,03157	2,91	0,006
kv4	0,49699	0,03157	15,74	0,000

S = 0,0705602 R-Sq = 92,0% R-Sq(adj) = 91,1%

The regression equation is

lnY = 3,90 + 0,00868 t + 0,344 kv1 + 0,0918 kv2 + 0,497 kv4

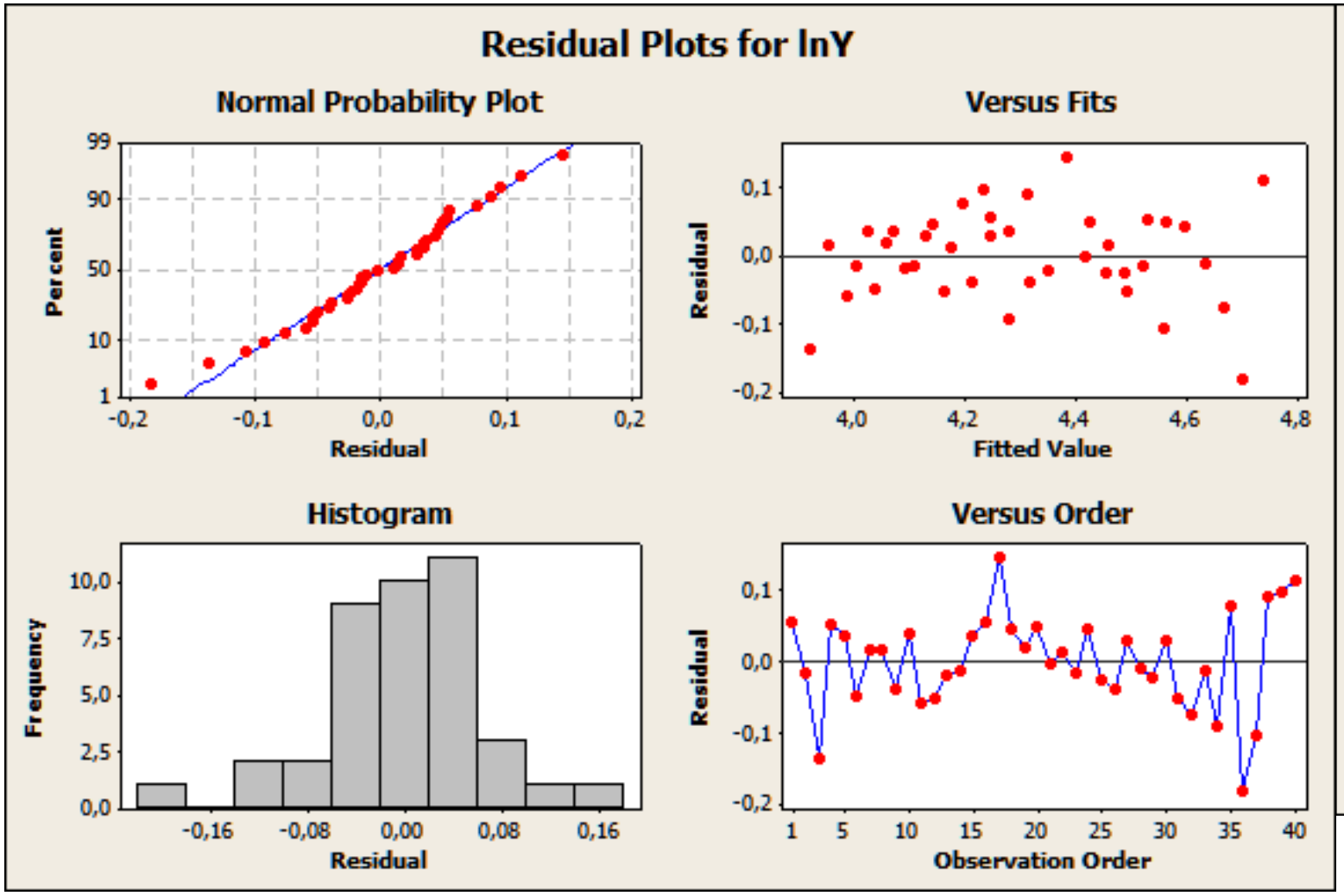
Durbin-Watson statistic = 1,66045

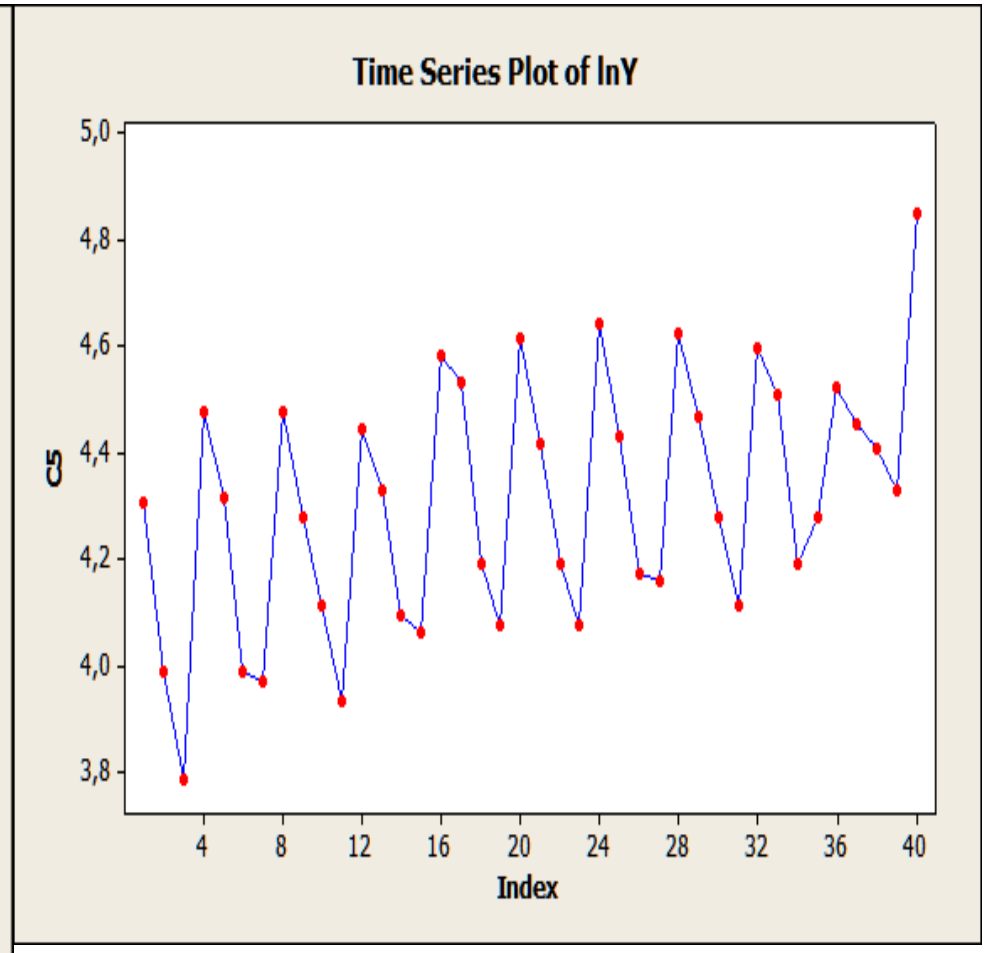
Analys av modell

- Anpassad modell (fitted model)

$$\widehat{lny}_t = 3{,}89557 + 0{,}0086823 \cdot t + 0{,}34370 \cdot kv1 + 0{,}09178 \cdot kv2 + 0{,}49699 \cdot kv4$$

- Signifikant regression. $F = 100{,}20$
- Alla regressionskoefficienter är signifikanta
- Förklaringsgraden är hög, 92%
- $DW = 1{,}66$. Residualerna verkar inte vara autokorrelerade


Residual Plots for lnY
Normal Probability Plot
Percent
99
90
50
10
1
-0,2
-0,1
0,0
0,1
0,2
Residual
Versus Fits
Residual
0,1
0,0
-0,1
-0,2
4,0
4,2
4,4
4,6
4,8
Fitted Value
Histogram
Frequency
10,0
7,5
5,0
2,5
0,0
-0,16
-0,08
0,00
0,08
0,16
Residual
Versus Order
Residual
0,1
0,0
-0,1
-0,2
1
5
10
15
20
25
30
35
40
Observation Order


Time Series Plot of lnY
C5
5,0
4,8
4,6
4,4
4,2
4,0
3,8
4
8
12
16
20
24
28
32
36
40
Index

Autokorrelation.

DW är en statistika som mäter om vi har autokorrelation vid en tidsförskjutning (lag) för residualerna e_t

$$Corr[e_t, e_{t-1}]$$

Men det kan ju finnas autokorrelation för andra tidsförskjutningar (lags).

ACF = AutoCorrelationFunction (funktion i k=antal lags)

$Corr[e_t, e_{t-2}]$ Lag 2

$Corr[e_t, e_{t-k}]$ Lag k

Vi kan i minitab skatta denna autokorrelation och få en graf

Stat→Time Series→Autocorrelation

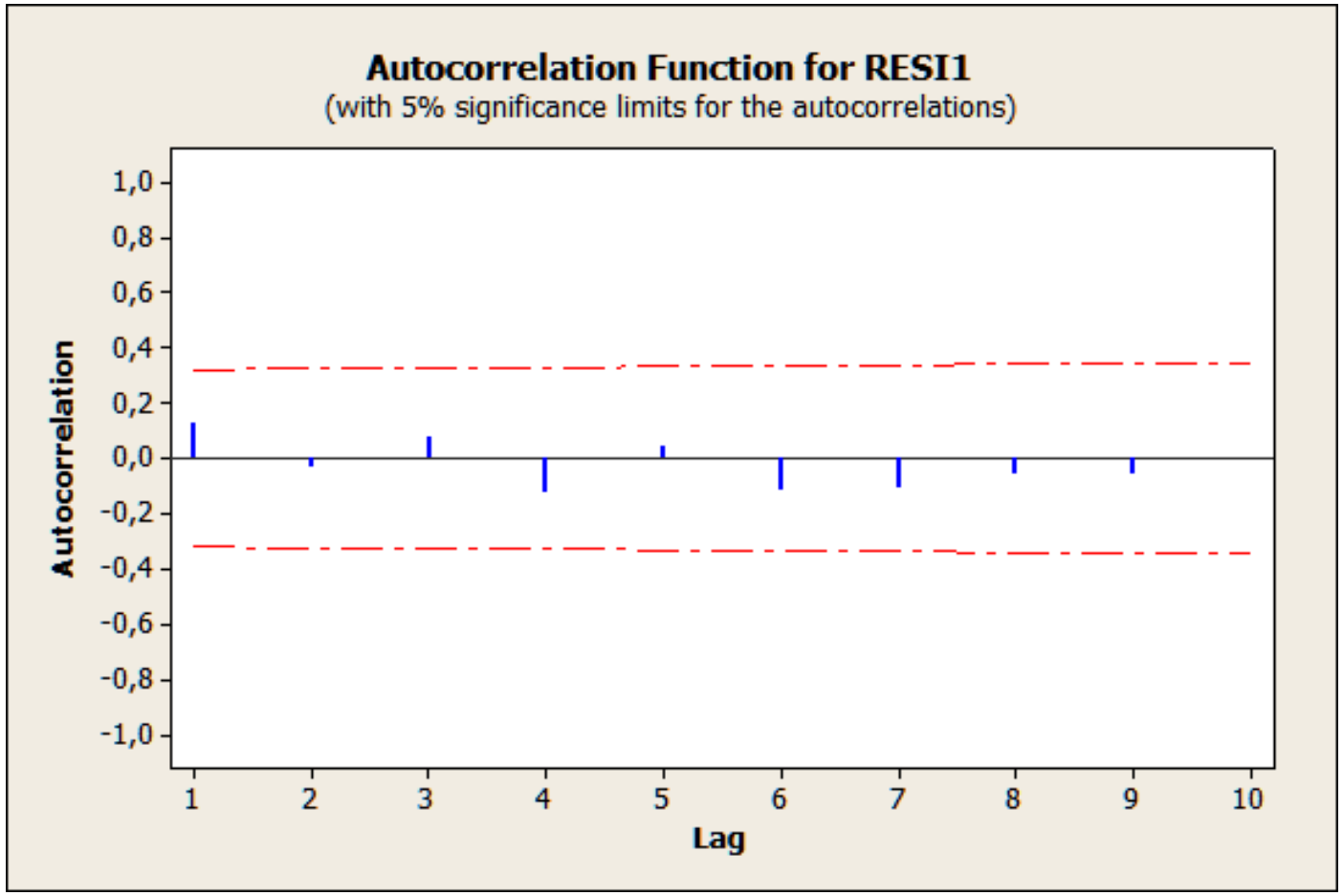
SAC för residualerna

SAC = Sample Auto Correlation

är skattad ACF

$$r_k = \frac{\sum_{t=1}^{n-k} e_t e_{t+k}}{\sum_{t=1}^{n} e_t \, e_t}$$

SAC för residualerna



Vi ser att residualerna inte är autokorrelerade för någon lag

Vi studerar utskriften igen

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	1,99554	0,49888	100,20	0,000
Residual Error	35	0,17426	0,00498		
Total	39	2,16979			

Predictor	Coef	SE Coef	T	P
Constant	3,89557	0,03023	128,87	0,000
t	0,0086823	0,0009711	8,94	0,000
kv1	0,34370	0,03162	10,87	0,000
kv2	0,09178	0,03157	2,91	0,006
kv4	0,49699	0,03157	15,74	0,000

S = 0,0705602 R-Sq = 92,0% R-Sq(adj) = 91,1%

The regression equation is

lnY = 3,90 + 0,00868 t + 0,344 kv1 + 0,0918 kv2 + 0,497 kv4

Durbin-Watson statistic = 1,66045

Modell uttryckt i Y

$$lnY_t = \beta_0 + \beta_1 t + \beta_2 kv1 + \beta_3 kv2 + \beta_4 kv4 + \varepsilon_t$$

$$Y_t = e^{(\beta_0 + \beta_1 t + \beta_2 kv1 + \beta_3 kv2 + \beta_4 kv4 + \varepsilon_t)}$$

$$Y_t = \beta_0' \cdot \beta_1'^{\,t} \cdot \beta_2'^{\,kv_1} \cdot \beta_3'^{\,kv_2} \cdot \beta_4'^{\,kv_4} \cdot \varepsilon_t'$$

Skattade regressionskoefficienter:

$$b_0' = e^{3{,}89557} = 49{,}1841$$
$$b_1' = e^{0{,}0086823} = 1{,}0087$$
$$b_2' = e^{0{,}34370} = 1{,}4102$$
$$b_3' = e^{0{,}09178} = 1{,}0961$$
$$b_4' = e^{0{,}49699} = 1{,}6438$$

Anpassad modell och tolkning

Något avrundade värden på koefficienterna

$$\hat{y}_t = 49{,}18 \cdot 1{,}0087^{\,t} \cdot 1{,}41^{kv1_t} \cdot 1{,}096^{kv2_t} \cdot 1{,}64^{kv4_t}$$

Ex: $1{,}0087$; För varje kvartal (tidsenhet) så ökar förbränningen av naturgas med 0,87% enligt modellen.

$e^{0{,}0086823 \cdot 4} = e^{0{,}0347292} = 1{,}0353$. För varje **år** så ökar förbränningen med 3,5%

1,64; I kvartal 4 är förbränningen av naturgas 64% högre jämfört med kvartal 3.

Prognosintervall för Y

Vid tidsserier är det lämpligare att säga Prognos (forecast) än Prediktion.

Bilda först, på vanligt sätt, prognosintervall för lny.

Därefter antiloggas intervallgränserna.

Beräkna ett prognosintervall *Y* för kvartal 1 och 2 år 2002, dvs för $Y_{41}\ och\ \ Y_{42}$.

```
Predicted Values for New Observations

New Obs      Fit   SE Fit         95% CI              95% PI
      1   4,5952   0,0309   (4,5325; 4,6580)   (4,4389; 4,7516)

Values of Predictors for New Observations
New Obs      t    kv1         kv2         kv4
      1   41,0   1,00   0,000000   0,000000
```

95% prognosintervall för Y_{41}:

(4,4389; 4,7516) → $\left(e^{4,4389}; e^{4,7516}\right)$ →

(84,68; 115,77) Prognos=99,01

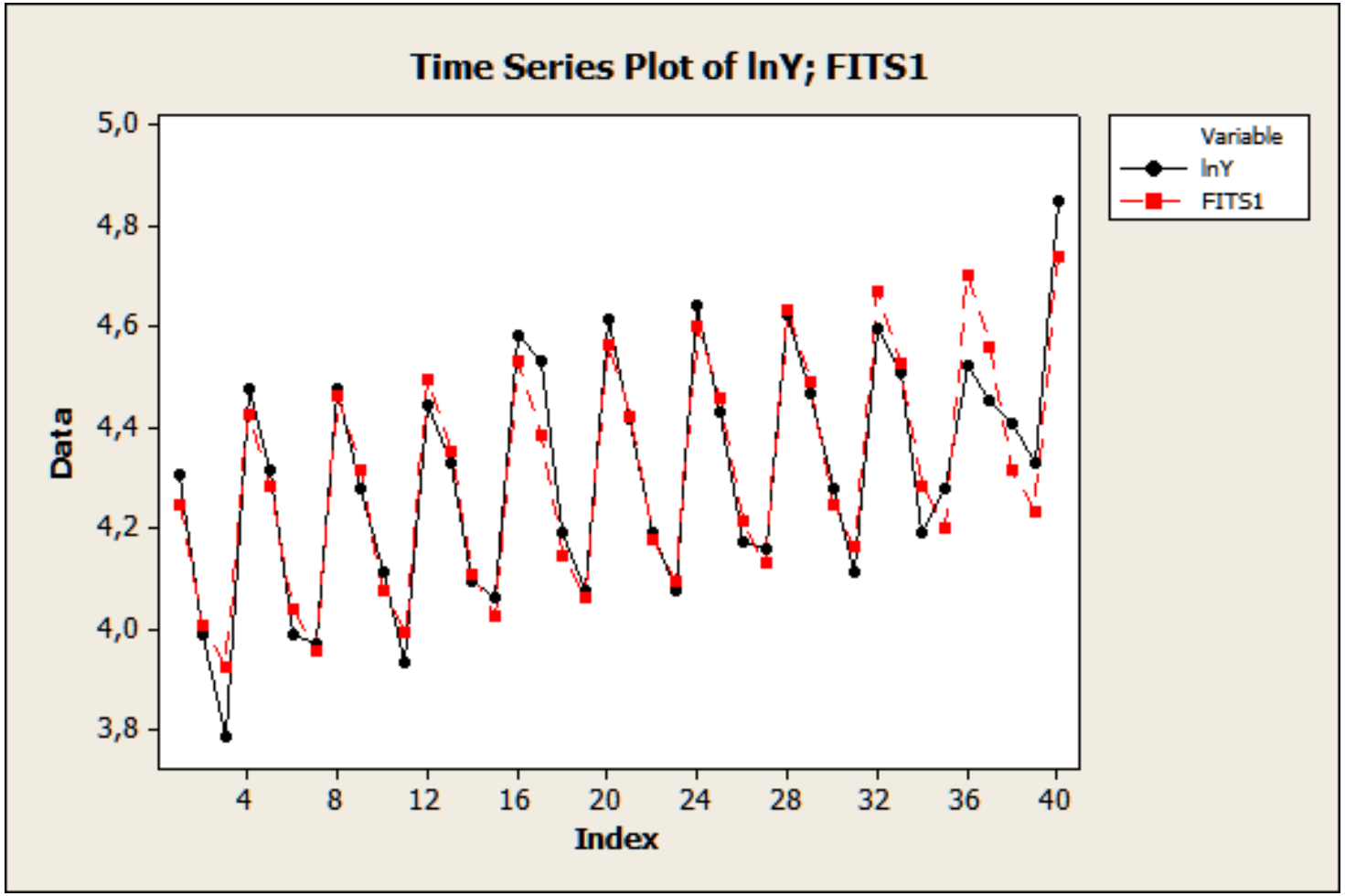
```
Predicted Values for New Observations
New Obs      Fit   SE Fit        95% CI              95% PI
      1   4,3520   0,0309   (4,2893; 4,4147)   (4,1956; 4,5084)

Values of Predictors for New Observations
New Obs      t         kv1    kv2         kv4
      1   42,0   0,000000   1,00   0,000000
```

95% prognosintervall för Y_{42}:

(66,39; 90,78) Prognos=77,63

Graf med lny och anpassade värden



Prognoser i tidsserieregressionen

Term	Coef
Constant	3.6491
time	0.02851
jan	-1.691
Feb	-0.469
mar	2.752
apr	1.224
maj	6.195
jun	2.417
jul	8.138
aug	6.360
sep	0.581
okt	2.553
nov	1.024

Prognos för december 1999

tid : 48; jan-nov: 0

$$\hat{y} = 3.649 + 0.0285 \cdot 48 = 5.017$$

Prognos för januari 2000

tid : 49; jan: 1; feb-nov: 0

$$\hat{y} = 3.649 + 0.0285 \cdot 49 + \left(-1.69\right) = 3.36$$

Autoregressiva regressionsmodeller

Kapitel 3,8 3,9 ytligt

Låt y_t vara den tidsserie som ska analyseras. Denna serie förklaras till viss del av en annan tidsserie x_t. y_t kan då modelleras så att den förklaras av sig själv i tidigare tidsenheter och av den andra tidsserien

$$y_t = \beta_0 + \beta_1 t + \beta_2 y_{t-1} + \beta_3 x_t + \beta_4 x_{t-1} + \varepsilon_t$$

Detta är ett enkelt exempel på hur y_t kan modelleras.

Även ε_t kan modelleras vidare

Table B.6 Global Mean Surface Air Temperature Anomaly and Global CO2 Concentration

Kan CO2 förklara Temperaturdifferensen?

År 1880 till 2004

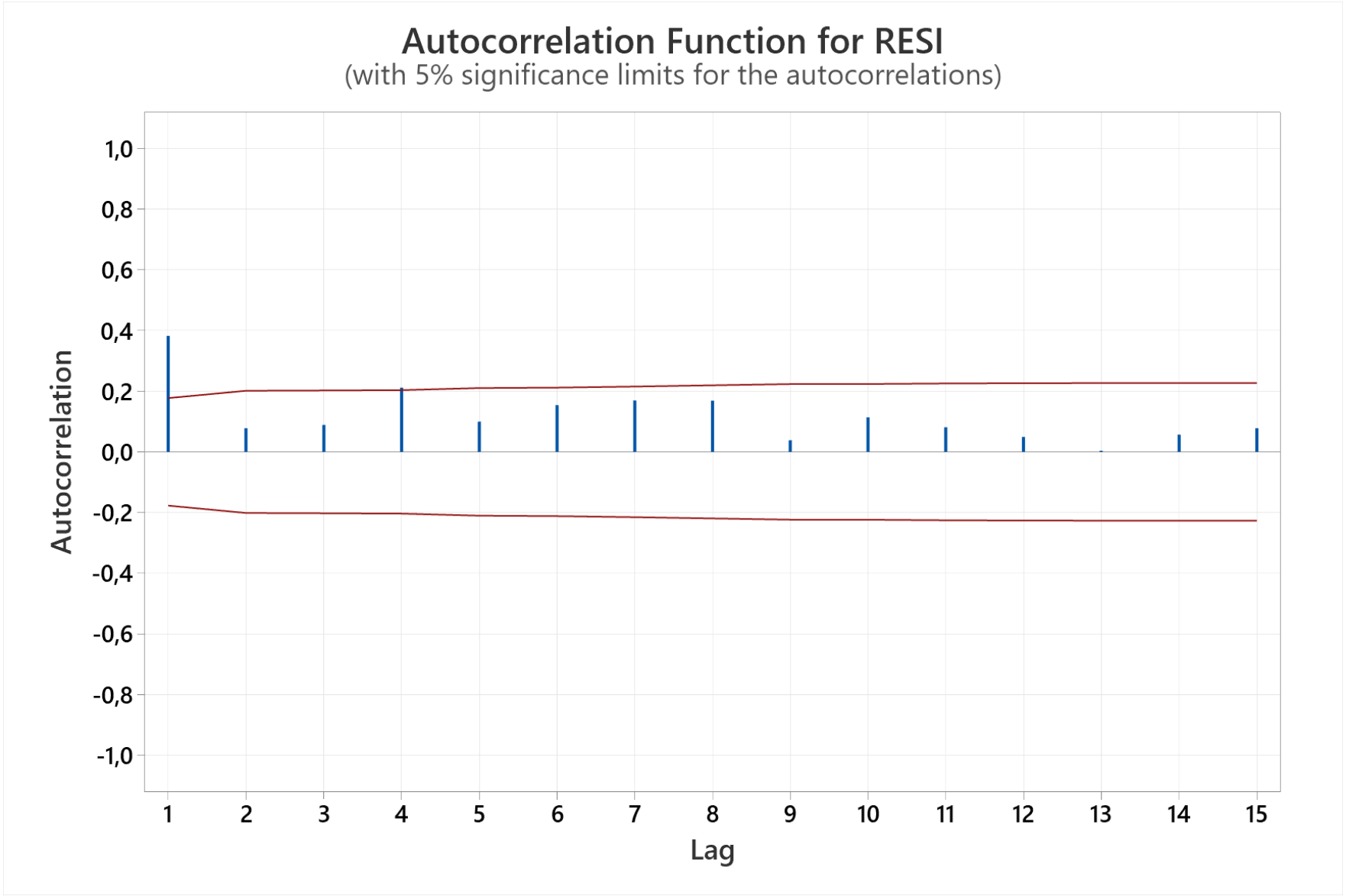
Regression Equation

Temp_diff = -3,030 + 0,009531 CO2

Coefficients

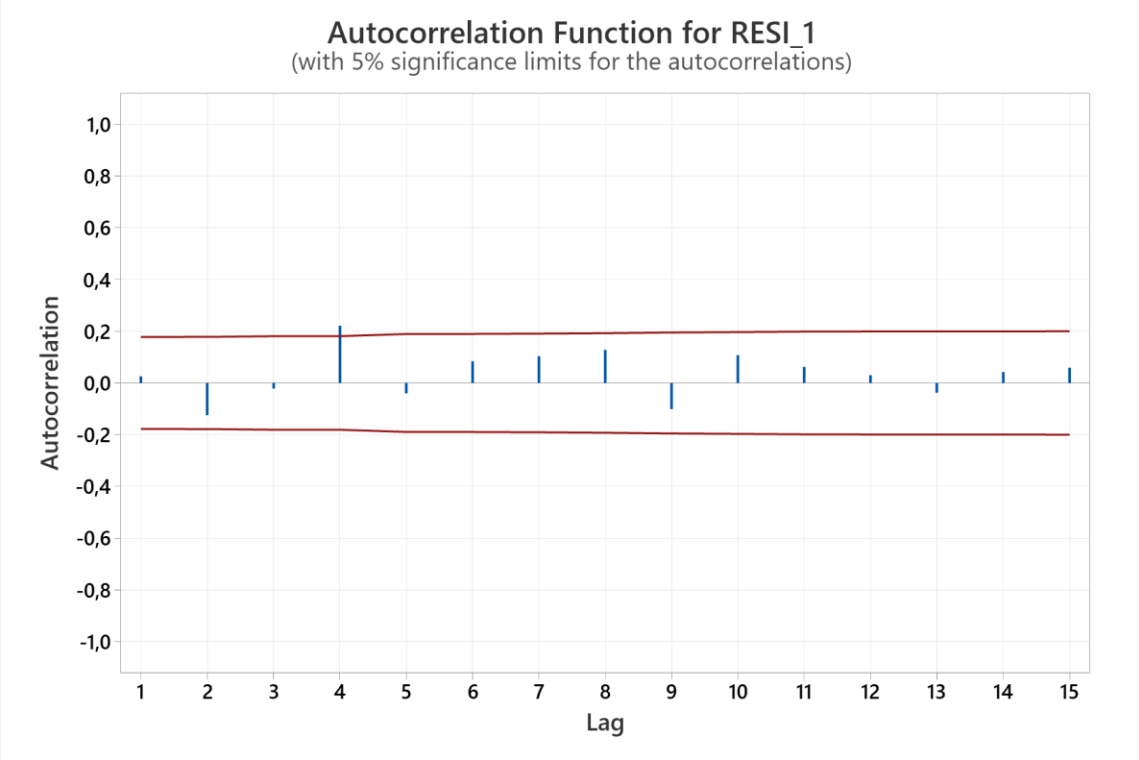
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-3,030	0,168	-18,07	0,000	
CO2	0,009531	0,000527	18,07	0,000	1,00

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$$


Autocorrelation Function for RESI
(with 5% significance limits for the autocorrelations)
Autocorrelation
1,0
0,8
0,6
0,4
0,2
0,0
-0,2
-0,4
-0,6
-0,8
-1,0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Lag

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1,934	0,291	-6,64	0,000	
Temp_diff(t-1)	0,3784	0,0832	4,55	0,000	3,46
CO2	0,006089	0,000914	6,66	0,000	3,46



$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \varepsilon_t$$

Detta var genomgång av:

- Kapitel 3

3,1-3,6 är repetition

3,7 tar Josef senare

3,8-3,9 ytligt