

Regressions- och tidsserieanalys

Föreläsning 10 - Autokorrelation. Autoregressiva modeller.

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet



 mattiasvillani.com

 [@matvil](https://twitter.com/matvil)



 [mattiasvillani](https://github.com/mattiasvillani)

- Autokorrelation
- Autoregressiva modeller
- Tidsserieregression
- Prognosutvärderingsmått

Repetition - Korrelation

- **Kovarians** mellan två variabler

$$s_{xy} = \text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

- **Korrelation** mellan två variabler:

$$r_{xy} = \text{corr}(x, y) = \frac{s_{xy}}{s_x s_y}$$

där

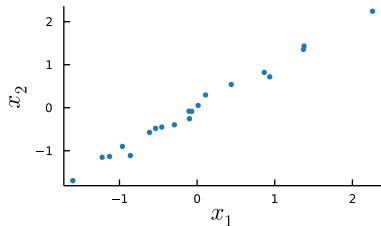
$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

- Samma formel som i F2, men med andra symboler:

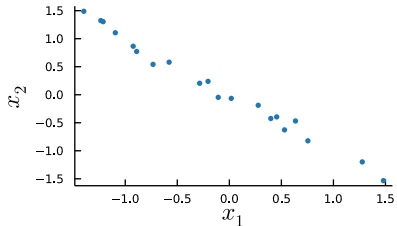
$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x}_i)(y_i - \bar{y}_i)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_i)^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y}_i)^2}}$$

Repetition - Korrelation

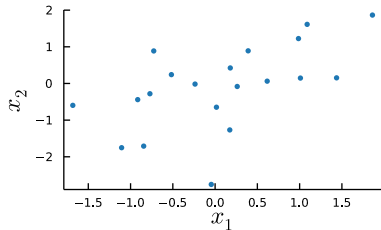
$r = 0.994$



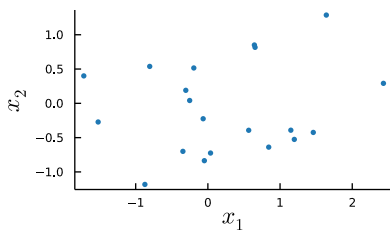
$r = -0.994$



$r = 0.563$



$r = 0.165$



Autokorrelation av ordning 1

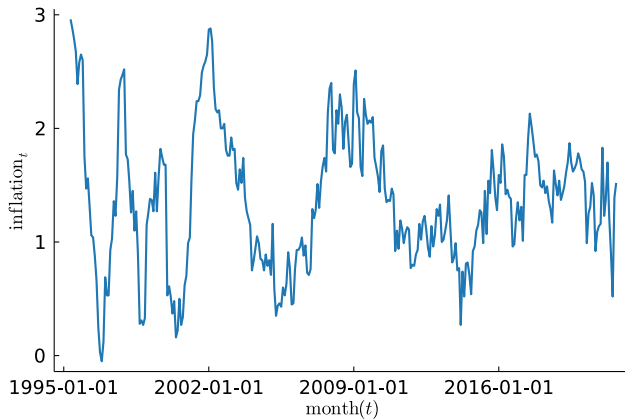
- Observationerna i en **tidsserie** y_t är ofta beroende/**korrelerade**.

- **Autokorrelation** av **ordning 1**:

$$r_1 = \text{corr}(y_t, y_{t-1})$$

- “Korrelation mellan dagens värde och gårdagens värde.”
- “Korrelation mellan denna månad och förra månaden”.
- “Första laggen”: y_{t-1} .

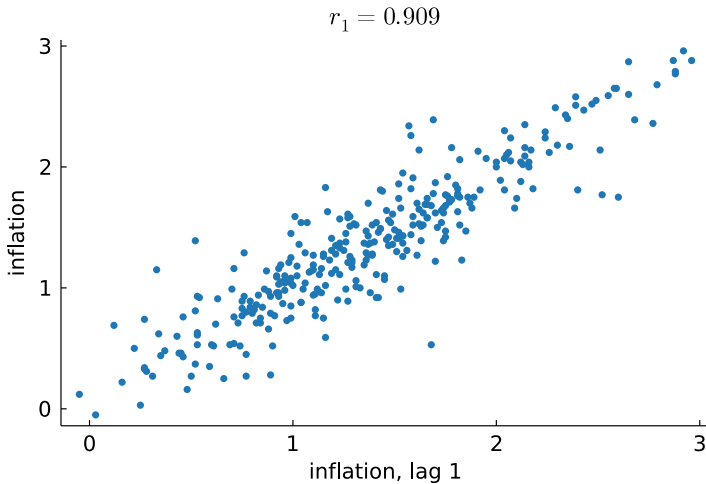
Inflation



Laggade variabler - inflation

	A	B	C	D	E	F
1	Månad	Inflation(t)	Inflation(t-1)	Inflation(t-2)	Inflation(t-3)	Inflation(t-4)
2	1995-05-01	2.96				
3	1995-06-01	2.88	2.96			
4	1995-07-01	2.79	2.88	2.96		
5	1995-08-01	2.68	2.79	2.88	2.96	
6	1995-09-01	2.39	2.68	2.79	2.88	2.96
7	1995-10-01	2.58	2.39	2.68	2.79	2.88
8	1995-11-01	2.65	2.58	2.39	2.68	2.79
9	1995-12-01	2.6	2.65	2.58	2.39	2.68
10	1996-01-01	1.75	2.6	2.65	2.58	2.39
11	1996-02-01	1.47	1.75	2.6	2.65	2.58
12	1996-03-01	1.56	1.47	1.75	2.6	2.65
13	1996-04-01	1.31	1.56	1.47	1.75	2.6
14	1996-05-01	1.06	1.31	1.56	1.47	1.75
15	1996-06-01	1.04	1.06	1.31	1.56	1.47
16	1996-07-01	0.88	1.04	1.06	1.31	1.56
17	1996-08-01	0.66	0.88	1.04	1.06	1.31
18	1996-09-01	0.25	0.66	0.88	1.04	1.06
19	1996-10-01	0.03	0.25	0.66	0.88	1.04
20	1996-11-01	-0.05	0.03	0.25	0.66	0.88
21	1996-12-01	0.12	-0.05	0.03	0.25	0.66
22	1997-01-01	0.69	0.12	-0.05	0.03	0.25

Inflation - autokorrelation lag 1



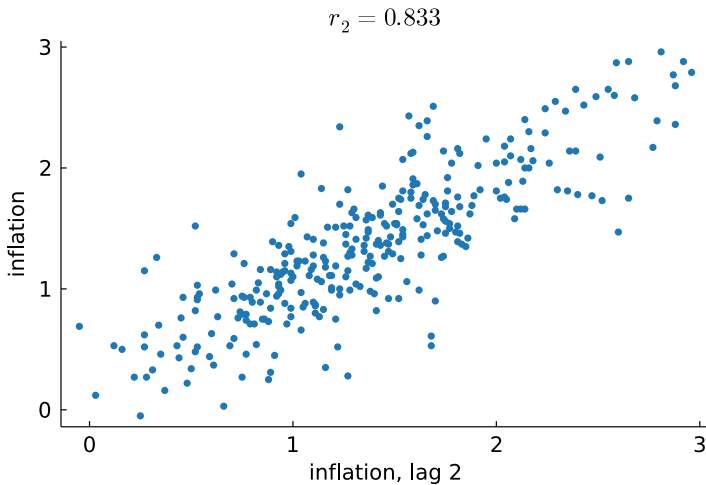
Autokorrelation av ordning 2

- Autokorrelation av ordning 2:

$$r_2 = \text{corr}(y_t, y_{t-2})$$

- “Korrelation mellan dagens värde och förrgårs värde.”
- “Korrelation mellan denna månad och förrförra månaden”.
- “Andra laggen”: y_{t-2} .

Inflation - autokorrelation lag 2



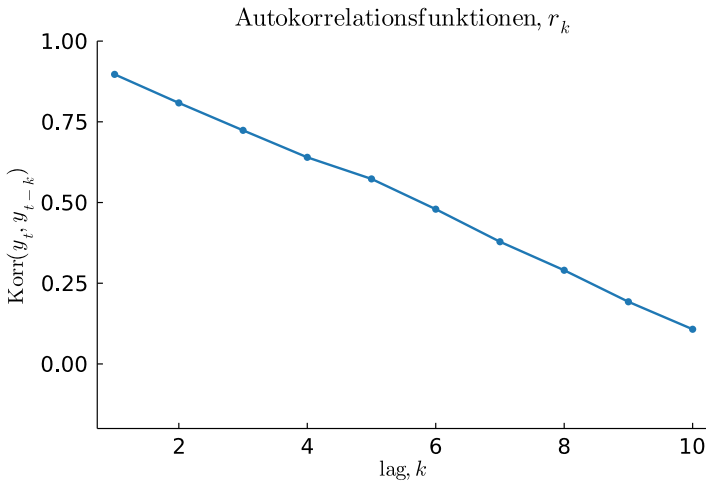
Autokorrelationsfunktioner

- Autokorrelation av ordning k

$$r_k = \text{corr}(y_t, y_{t-k})$$

- “Korrelation mellan månadens värde och k månader innan”.
- Autokorrelationsfunktionen (ACF) är r_k som en funktion av tidsavståndet k .

Inflation - autokorrelationsfunktion



Autoregressiva modeller

- Autoregressiv modell av ordning 1 (**AR(1)**)

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

- **AR(1)** är **regression med y_{t-1} som förklarande variabel!**
- Skattas med **minstakvadrat-metoden**

$$y_t = a + by_{t-1}$$

- Autoregressiv modell av ordning p (**AR(p)**)

$$y_t = \alpha + \beta_1 y_{t-1} + \dots + \beta_p y_{t-p} + \varepsilon_t$$

- **AR(p)** är en **multipl regression** med de p förklarande variablerna y_{t-1}, \dots, y_{t-p} .

AR(1) för inflation - SAS

AR(1)

The AUTOREG Procedure

Dependent Variable y

Ordinary Least Squares Estimates			
SSE	19.7318994	DFE	318
MSE	0.06205	Root MSE	0.24910
SBC	28.1102919	AIC	20.5736499
MAE	0.18252339	AICC	20.6115048
MAPE	22.9090928	HQC	23.5831741
Durbin-Watson	2.0578	Total R-Square	0.8352

Parameter Estimates					
Variable	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	0.1253	0.0343	3.65	0.0003
lag1_y	1	0.9069	0.0226	40.14	<.0001

```
title "AR(1)";  
proc autoreg data=work.inflrepo;  
model y=lag1_y;  
run;
```

AR(4) för inflation - SAS

AR(4)

The AUTOREG Procedure

Dependent Variable y

Ordinary Least Squares Estimates

SSE	19.5434434	DFE	312
MSE	0.06264	Root MSE	0.25028
SBC	45.1565103	AIC	26.3620015
MAE	0.18244114	AICC	26.5549275
MAPE	23.1092492	HQC	33.8694694
Durbin-Watson	1.9989	Total R-Square	0.8273

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > t
Intercept	1	0.1375	0.0367	3.75	0.0002
lag1_y	1	0.8734	0.0565	15.45	<.0001
lag2_y	1	0.0565	0.0755	0.75	0.4551
lag3_y	1	0.006678	0.0769	0.09	0.9309
lag4_y	1	-0.0396	0.0572	-0.69	0.4893

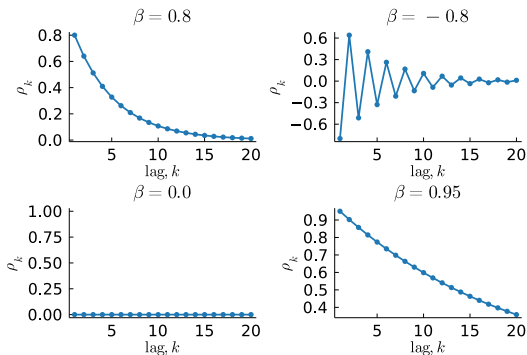
Autoregressiva modeller

■ AR(1)

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

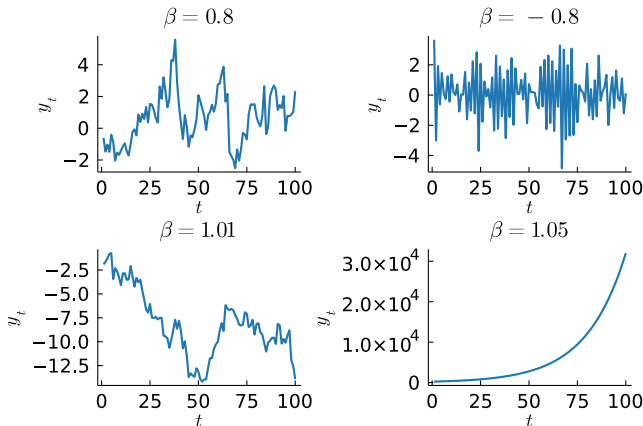
■ Autokorrelationsfunktion för AR(1) i populationen:

$$\rho_k = \beta^k, \text{ för } k = 1, 2, \dots$$



Autoregressiva modeller - stationäritet

- AR(1) är **stationär** (icke-explosiv) modell om $-1 < \beta < 1$.



Autoregressiva modeller - prognoser

- Skattad AR(1)-modell

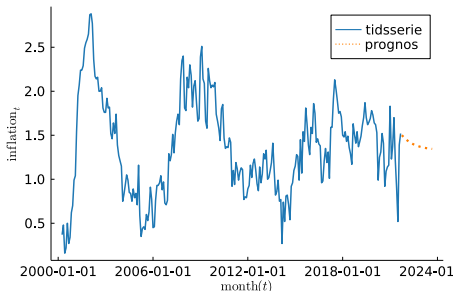
$$y_t = a + b \cdot y_{t-1}$$

- Vid tidpunkt T , **prognos för nästa månad $T + 1$**

$$\hat{y}_{T+1} = a + b \cdot y_T$$

- **Prognos för $T + 2$**

$$\hat{y}_{T+2} = a + b \cdot \hat{y}_{T+1}$$



Regression för tidsserier

■ Regression

$$y = \alpha + \beta_1 x_1 + \dots + \beta_k x_k + \varepsilon$$

där **feltermerna** ε antas bara **oberoende** från $N(0, \sigma_\varepsilon^2)$.

■ **Oberoende = korrelerade** för **normalfördelade** variabler.

■ Regressionen skattas med

$$y = a + b_1 x_1 + \dots + b_k x_k$$

och vi får residualer

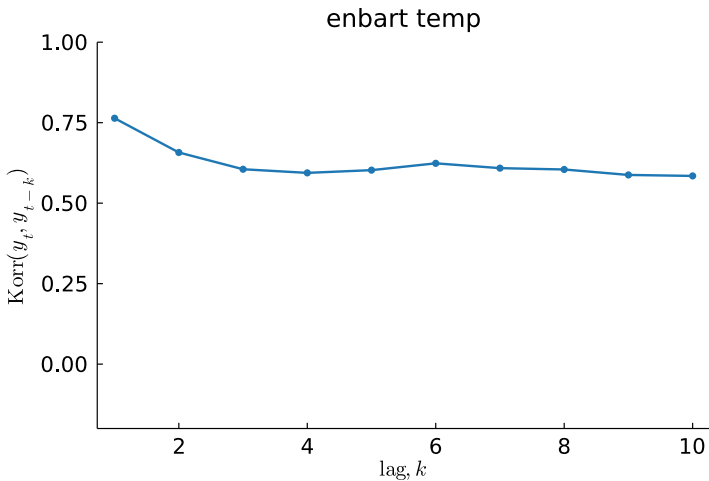
$$e_t = y_t - \hat{y}_t.$$

■ Vi kan undersöka om **residualerna är okorrelerade**.

■ Två metoder:

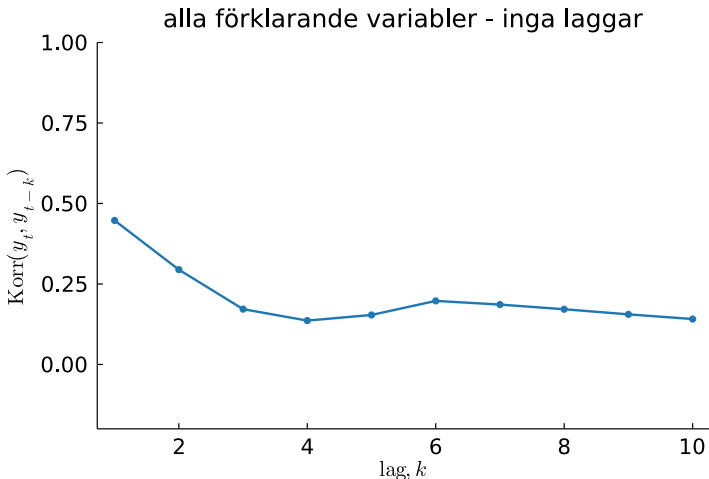
- ▶ Visuellt genom att **plotta autokorrelationsfunktionen för** e_t
- ▶ **Durbin-Watson test**

ACF residualer - temp



ACF residualer - alla variabler

- Regression med alla förklarande variabler:
temp,hum,windspeed,holiday,workingday,säsong,yr.



Regression för tidsserier

■ Regressionsmodeller för tidsserier

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

får ofta korrelerade residualer. 🙄

■ Kombinera enkel regression och AR(1) 😊

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

■ Kombinera multipel regression och AR(p) 😍

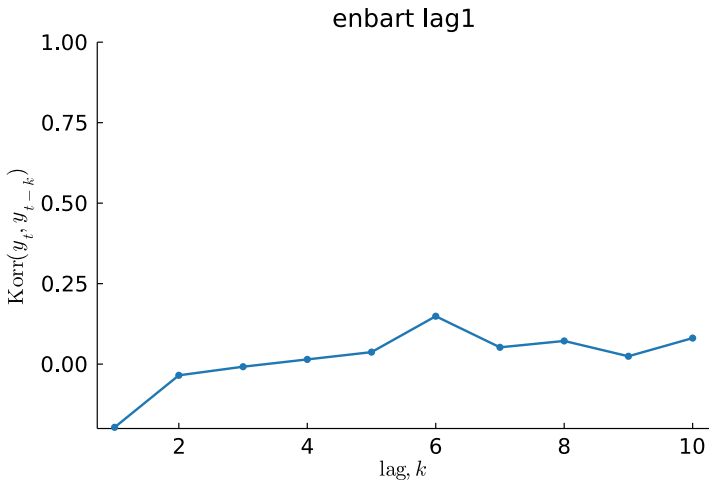
$$y_t = \alpha + \beta_1 x_t + \dots + \beta_k x_{kt} + \beta_{k+1} y_{t-1} + \dots + \beta_{k+p} y_{t-p} + \varepsilon_t$$

■ Cykeluthyrning:

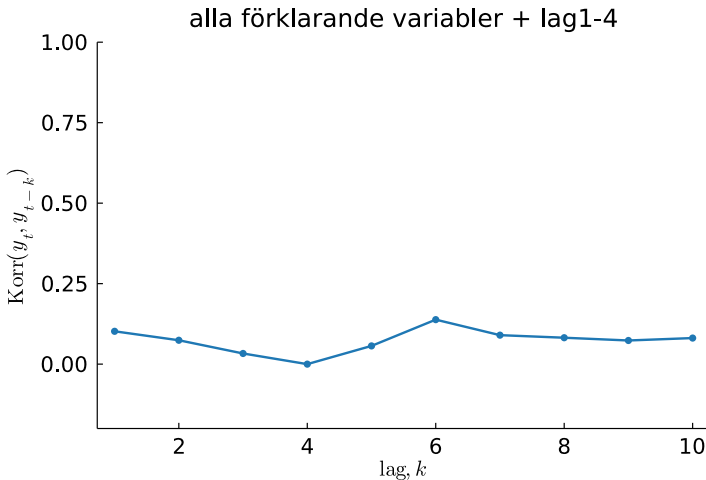
$$\text{AntalUthyr}_{\text{idag}} = a + b_1 \cdot \text{temp}_{\text{idag}} + b_2 \cdot \text{AntalUthyr}_{\text{igar}}$$

■ Standardfel och hypotestest måste korrigeras om laggar av y_t används som förklarande variabel. SAS proc autoreg gör detta.

ACF residualer - enbart lag 1



ACF residualer - alla variabler + lag 1-4



Durbin-Watson test

- Test för autokorrelation (i feltermen).

- Teststatistika

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

- Durbin-Watson **testar första autokorrelationen** (AJÅ)

$$d \approx 2(1 - r_1)$$

- Teststatistikan uppfyller

$$0 \leq d \leq 4$$

- Grova **kritiska gränser**:

d nära 2 \implies ej signifikant

$d < 1$ \implies signifikant positiv autokorrelation

$d > 1$ \implies signifikant negativ autokorrelation

- Durbin-Watson test kan inte användas när man har laggar av målvariabeln (y_{t-1} etc) som förklarande variabler.

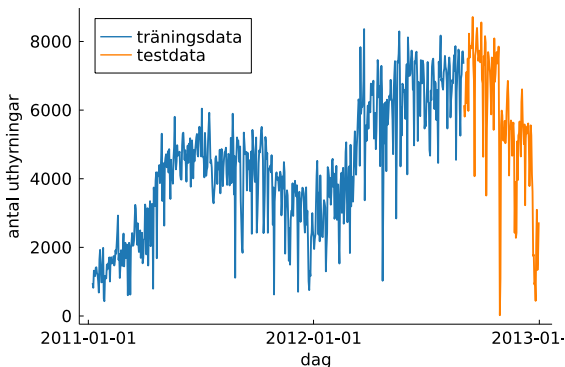
Durbin-Watson test - cykeluthyrning

Förklarande variabler	R^2	$r_1^{(\text{res})}$	d	p -värde
temp	0.385	0.764	0.471***	< 1e-93
temp,hum,windspeed,holiday,workingday,säsong,yr	0.795	0.447	1.104***	< 1e-33

Cykeluthyrningar - utvärdera prognosförmåga

- **Träningsdata:** Jan 1, 2011 - Aug 31, 2012.
- **Testdata:** Sept 1, 2012 - Dec 31, 2012.
- **Prediktionsmått RMSE**

$$\text{RMSE}_{\text{test}} = \sqrt{\frac{1}{n_{\text{test}}} \sum_{t \in \text{Testdata}} (y_t - \hat{y}_t)^2}$$



Cykeluthyrningar

- Training data: Jan 1, 2011 - Aug 31, 2012.
- Test data: Sept 1, 2012 - Dec 31, 2012.

Förklarande variabler	R^2	RMSE _{test}
temp	0.385	2346.60
temp,hum,windspeed,holiday,workingday,säsong,yr	0.795	1292.07
lag1	0.714	1274.32
lag1,lag2	0.730	1279.30
lag1-lag4	0.746	1267.84
lag1-lag6	0.764	1262.10
temp,hum,windspeed,holiday,workingday,säsong,yr,lag1	0.825	1127.63
temp,hum,windspeed,holiday,workingday,säsong,yr,lag1-lag4	0.827	1118.83
temp,hum,windspeed,holiday,workingday,säsong,yr,lag1-lag6	0.830	1117.63
temp,hum,windspeed,holiday,workingday,säsong,yr,lag1-lag6,Lasso	NA	1118.34