Regressions- och tidsserieanalys

Föreläsning 9 - Autokorrelation. Autoregressiva modeller.

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet

mattiasvillani.com @matvil mattiasvillani

Översikt

- Autokorrelation
- Autoregressiva modeller
- Prognosutvärderingsmått

Repetition - Korrelation

- **Kovarians** mellan två variabler

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

- **Korrelation** mellan två variabler:

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

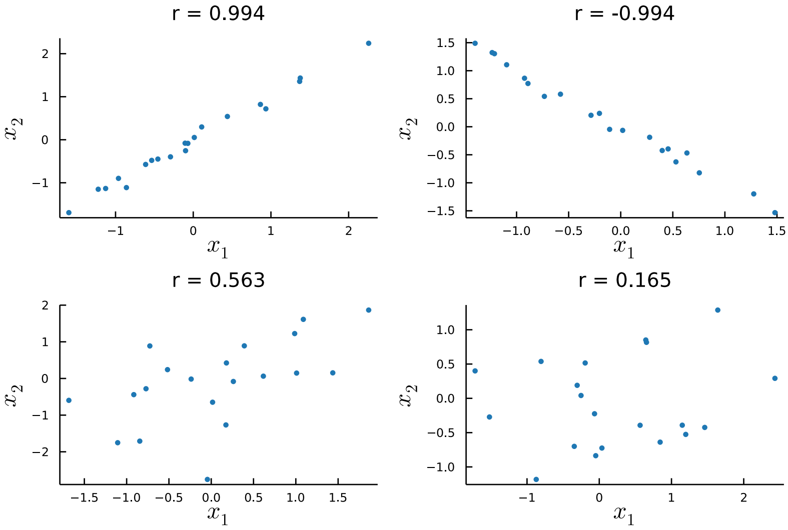
där

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

- 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



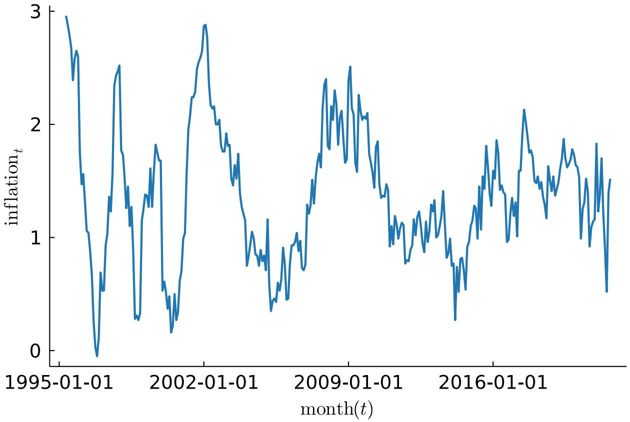
Autokorrelation av ordning 1

- 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

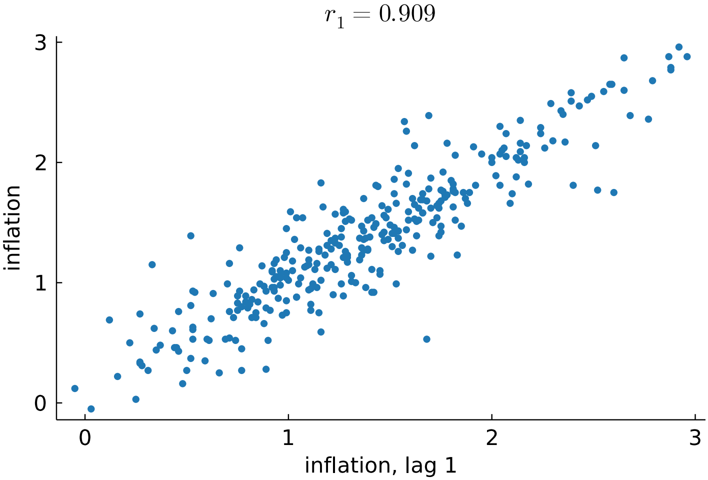


3
2
1
0
inflation_t
1995-01-01
2002-01-01
2009-01-01
2016-01-01
month(t)

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



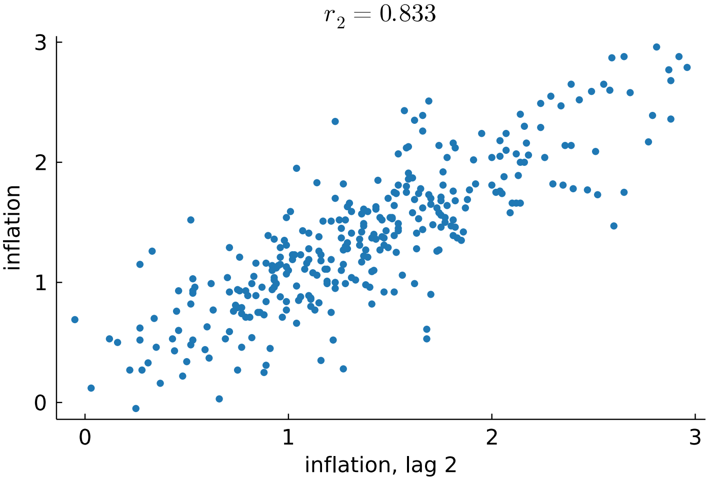
Autokorrelation av ordning 2

- **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



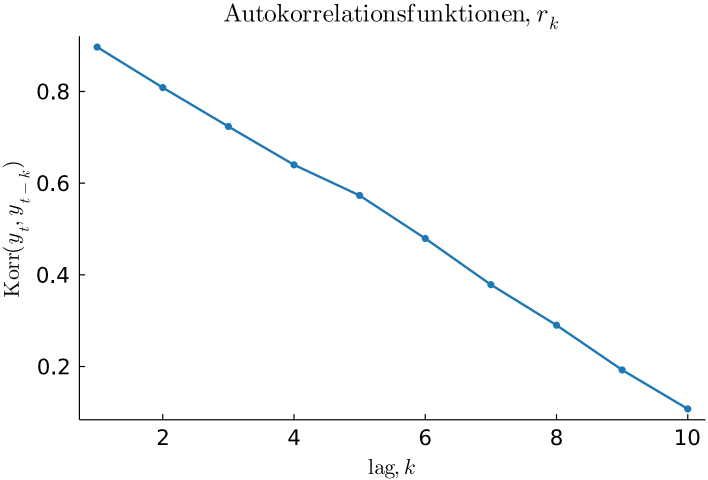
Autokorrelationsfunktionen

- **Autokorrelation** av **ordning** k

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

- "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 ($\mathrm{AR}(1)$)

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

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

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

- Autoregressiv modell av ordning p ($\mathrm{AR}(p)$)

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

- $\mathrm{AR}(p)$ är en **multipel regression** med de p förklarande variablerna $y_{t-1}, \ldots, y_{t-p}$.

AR(1) för inflation - R

```
> library(SUdatasets)
> arimafit = arima(swedinfl$KPIF, order = c(1,0,0))
> arima_coef_summary(arimafit)

Parameter estimates
--------------------------------------------------------------
     Estimate Std. Error z-ratio Pr(>|z|)   2.5 %  97.5 %
ar1   0.91801   0.022383 41.0135        0 0.87414 0.96188
mean  1.43624   0.165006  8.7042        0 1.11282 1.75965
>
> 
```

AR(4) för inflation - R

```
> library(SUdatasets)
> arimafit = arima(swedinfl$KPIF, order = c(4,0,0))
> arima_coef_summary(arimafit)

Parameter estimates
--------------------------------------------------------------
        Estimate Std. Error   z-ratio Pr(>|z|)     2.5 %   97.5 %
ar1    0.8900015   0.055640 15.995742  0.00000  0.780947 0.999056
ar2    0.0586250   0.075101  0.780619  0.43503 -0.088572 0.205822
ar3    0.0062025   0.076370  0.081216  0.93527 -0.143483 0.155888
ar4   -0.0405666   0.057249 -0.708605  0.47857 -0.152774 0.071641
mean   1.4334525   0.158225  9.059583  0.00000  1.123331 1.743573
```

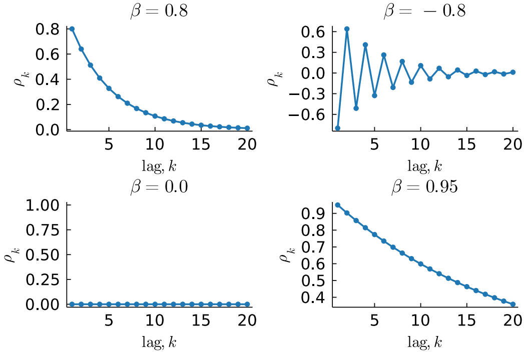
Autokorrelationsfunktion AR(1)

- AR(1)

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

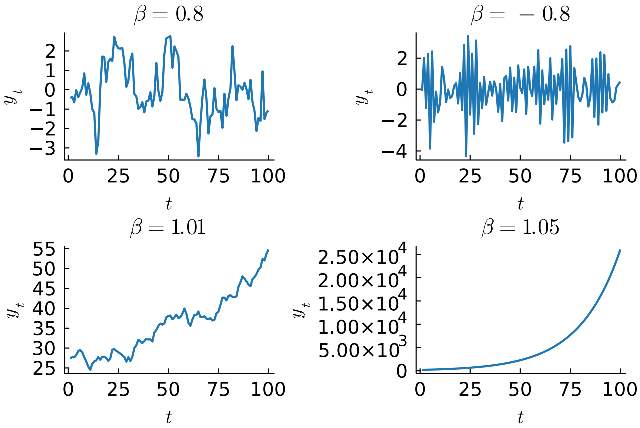
- **Autokorrelationsfunktion** (**ACF**) för AR(1) i populationen:

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



Autoregressiva modeller - stationäritet

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



Prognoser med AR(1) modell

- Skattad AR(1)-modell

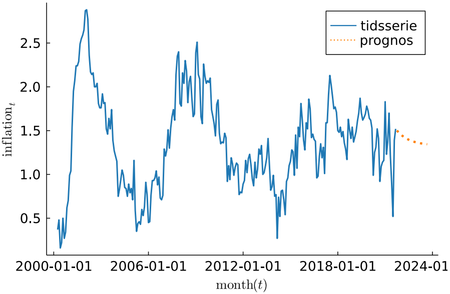
$$y_t = a + b \cdot y_{t-1}$$

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

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

- **Prognos för $T+2$**

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



Prognoser med AR(2) modell

- Skattad AR(2)-modell

$$y_t = a + b_1 \cdot y_{t-1} + b_2 \cdot y_{t-2}$$

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

$$\hat{y}_{T+1} = a + b_1 \cdot y_T + b_2 \cdot y_{T-1}$$

- **Prognos för $T+2$**

$$\hat{y}_{T+2} = a + b_1 \cdot \hat{y}_{T+1} + b_2 \cdot y_T$$

- **Prognos för $T+3$**

$$\hat{y}_{T+3} = a + b_1 \cdot \hat{y}_{T+2} + b_2 \cdot \hat{y}_{T+1}$$

Mått på prognosförmåga

- Genomsnittliga **kvadrerade prognosfel**

$$\mathrm{MSE} = \frac{\sum_{t=1}^{n} \left(y_t - \hat{y}_t\right)^2}{n}$$

- Genomsnittliga **absoluta prognosfel**

$$\mathrm{MAE} = \frac{\sum_{t=1}^{n} \left|y_t - \hat{y}_t\right|}{n}$$

- Genomsnittlig **procentuella absoluta prognosfel**

$$\mathrm{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \frac{\left|y_t - \hat{y}_t\right|}{y_t} \cdot 100$$