

# Regressions- och tidsserieanalys

## Föreläsning 10 - ARMA modeller och Enkel logistisk regression.

**Mattias Villani** 🧑

Statistiska institutionen  
Stockholms universitet

Institutionen för datavetenskap  
Linköpings universitet



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



@matvil



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

- ARMA modeller
- Regression för tidsserier
- Odds och logodds
- Enkel logistisk regression

# Autokorrelationsfunktion - AR(1)

- AR(1) som populationsmodell:

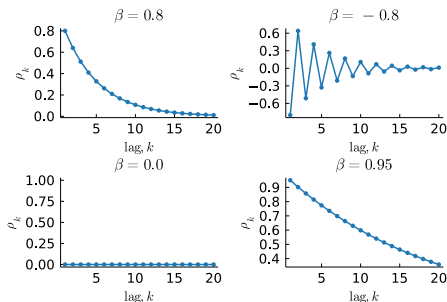
$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

- Autokorrelationsfunktion (ACF)

$$\rho_k = \text{corr}(y_t, y_{t-k}), \text{ för } k = 1, 2, \dots$$

- ACF för AR(1)

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



# Partiell autokorrelationsfunktion - AR(1)

- AR(1)

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

- Partiell ACF (PACF) multipel regressions-variant av ACF

$$\rho_k^* = \text{corr}(y_t, y_{t-k} | y_{t-1}, \dots, y_{t-k-1}), \text{ för } k = 1, 2, \dots$$

- för AR(1) i populationen:

$$\rho_1^* = \beta$$

$$\rho_k^* = 0, \text{ för } k = 2, 3, \dots$$

# Partiell autokorrelationsfunktion - AR(2)

## ■ AR(2)

$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

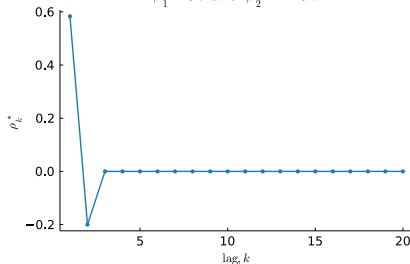
## ■ Partiell ACF för AR(2) i populationen:

$$\rho_1^* = \frac{\beta_1}{1 - \beta_2}$$

$$\rho_2^* = \beta_2$$

$$\rho_k^* = 0, \text{ för } k = 3, 4, \dots$$

$$\beta_1 = 0.7 \text{ and } \beta_2 = -0.2$$



# ARMA modeller

- AR(1) modell beror på laggad tidsserie  $y_{t-1}$

$$y_t = \alpha + \phi_1 y_{t-1} + \varepsilon_t$$

- MA(1) modell beror på laggad felterm  $\varepsilon_{t-1}$

$$y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

- ACF för MA(1)

$$\rho_1 = \theta_1$$

$$\rho_k = 0, \text{ för } k = 2, 3, \dots$$

- Box-Jenkins identifiering av AR och MA ordning:

- ▶ ACF noll efter  $q$  laggar  $\iff$  MA( $q$ ) modell.
- ▶ PACF noll efter  $p$  laggar  $\iff$  AR( $p$ ) modell.

# Kombinera AR och MA: ARMA modeller

- ARMA(1, 1) modell

$$y_t = \alpha + \phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

- ARMA( $p, q$ ) modell

$$y_t = \alpha + \phi_1 y_{t-1} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \dots + \theta_p \varepsilon_{t-p}$$

- ARMA( $p, q$ ) modeller är svårare att välja från ACF och PACF. ARMA( $p, q$ ) har inte tydliga nollor i ACF och PACF.
- Ibland måste man analysera **skillnaden mellan två tidsperioder** för att få tidsserien **stationär**:

$$\Delta y_t = y_t - y_{t-1}$$

- Vi har **differentierat** tidsserien. Kan “diffa” flera gånger.
- En **ARIMA( $p, d, q$ )** modell är en ARMA( $p, q$ ) modell för en tidsserie  $y_t$  som vi diffat  $d$  gånger.

# Estimation av en ARMA(2,2) modell

```
> library(SUdatasets)
> arimafit = arima(swedinfl$KPIF, order = c(2,0,2))
> arimasumm = arima_coef_summary(arimafit)
```

Parameter estimates

```
-----
      Estimate Std. Error  z-ratio Pr(>|z|)    2.5 %  97.5 %
ar1    0.023018   0.043356   0.53091  0.59548 -0.06196 0.10800
ar2    0.836117   0.037591  22.24263  0.00000  0.76244 0.90979
ma1    0.898033   0.065537  13.70271  0.00000  0.76958 1.02648
ma2   -0.071553   0.060119  -1.19019  0.23397 -0.18939 0.04628
mean    1.437798   0.172772   8.32195  0.00000  1.09916 1.77643
```



# Regression för tidsserier

## ■ Regression

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

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

## ■ **Oberoende = okorrelerade** för **normalfördelade** variabler.

## ■ Regressionen skattas med

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

och vi får residualer

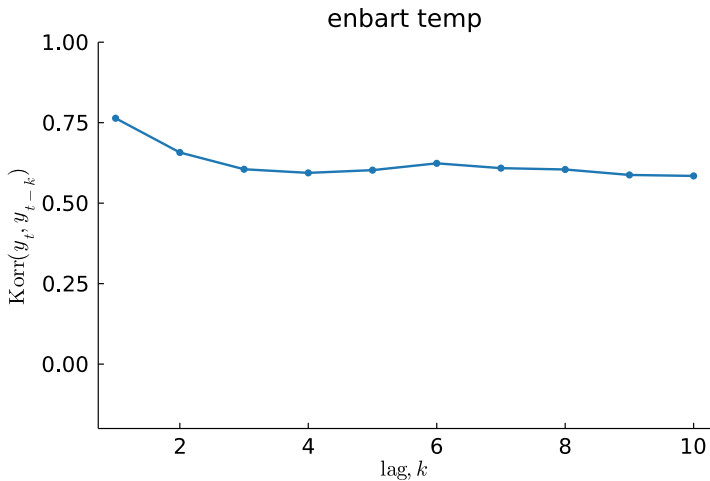
$$e_t = y_t - \hat{y}_t.$$

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

## ■ Två metoder:

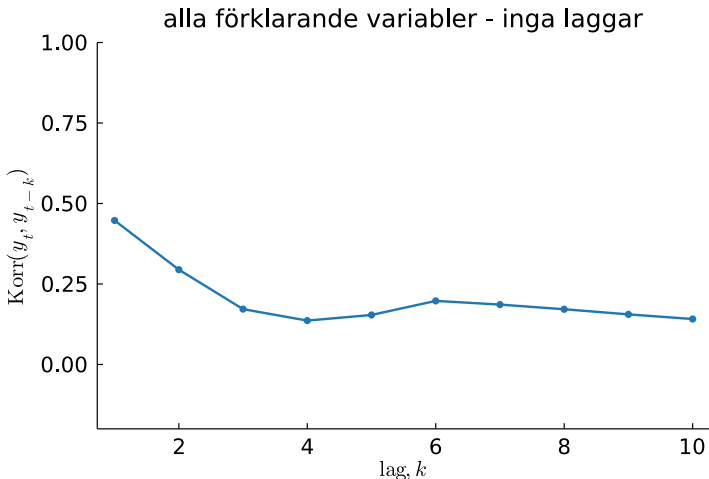
- ▶ Visuellt genom att **plotta autokorrelationsfunktionen för**  $e_t$
- ▶ **Durbin-Watson test**

# 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) 😍

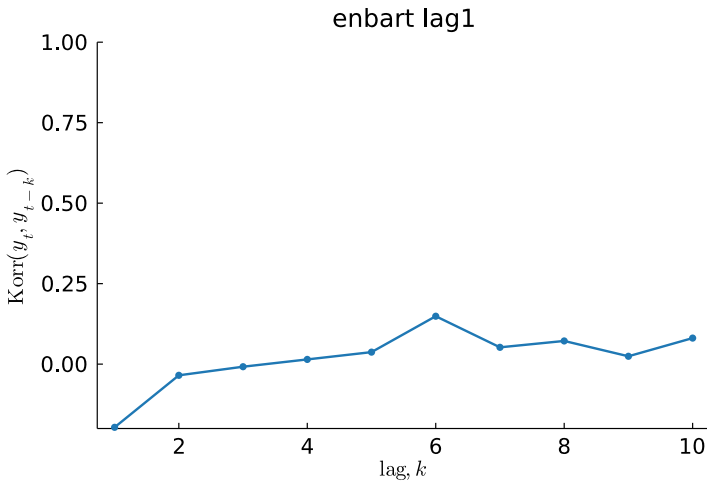
$$y_t = \alpha + \beta_1 x_t + \dots + \beta_k x_{kt} + \beta_{k+1} y_{t-1} + \dots + \beta_{k+p} y_{t-p} + \varepsilon_t$$

## ■ Cykeluthyrning:

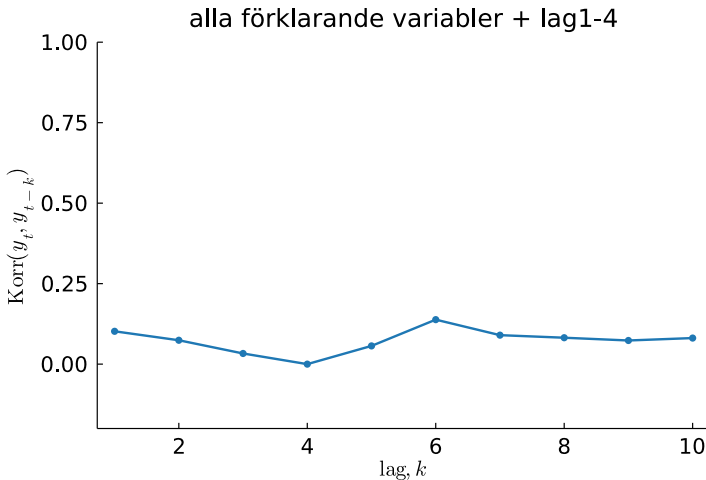
$$\text{AntalUthyr}_{\text{idag}} = a + b_1 \cdot \text{temp}_{\text{idag}} + b_2 \cdot \text{AntalUthyr}_{\text{igar}}$$

## ■ Standardfel och hypotestest måste korrigeras om laggar av $y_t$ används som förklarande variabel.

# 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

```
> library(car)
> lmfit = lm(nRides ~ temp , data = bike)
> durbinWatsonTest(lmfit)
lag Autocorrelation D-W Statistic p-value
1      0.7641582      0.4678707      0
Alternative hypothesis: rho != 0

> lmfit = lm(nRides ~ temp + hum + windspeed + holiday + workingday + as.factor(season) + yr, data = bike)
> durbinWatsonTest(lmfit)
lag Autocorrelation D-W Statistic p-value
1      0.4472755      1.104221      0
Alternative hypothesis: rho != 0
```

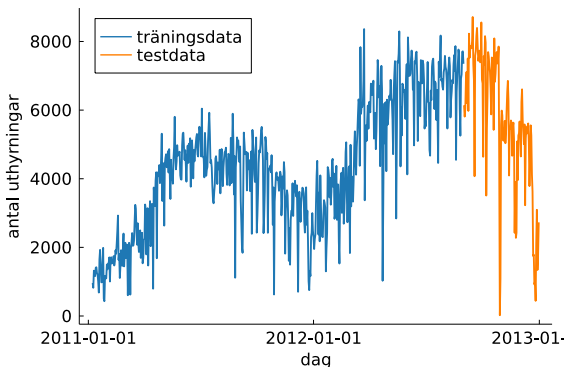
Förklarande variabler	$R^2$	$r_1^{(\text{res})}$	$d$	$p$ -värde
temp	0.385	0.764	0.471***	< 1e-93
temp,hum,windspeed,holiday,workingday,säsong,yr	0.795	0.447	1.104***	< 1e-33



# 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

■ Träningsdata: Jan 1, 2011 - Aug 31, 2012.

■ Testdata: Sept 1, 2012 - Dec 31, 2012.

Förklarande variabler	$R^2$	RMSE <sub>test</sub>
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

# Odds och logodds

- Låt  $P(A)$  vara sannolikheten för en händelse  $A$ .

$$P(A) = \frac{\text{antal fall där } A \text{ inträffar}}{\text{antal möjliga fall}}$$

- Odds

$$\text{Odds}(A) = \frac{\text{antal fall där } A \text{ inträffar}}{\text{antal fall där } A \text{ inte inträffar}}$$

$$\text{Odds}(A) = \frac{P(A)}{1 - P(A)}$$

- Exempel: Sannolikheten att slå en 6:a med en vanlig tärning:

- ▶ Sannolikhet  $P(A) = 1/6$
- ▶ Odds

$$\text{Odds}(A) = \frac{1/6}{5/6} = \frac{1}{5}$$

Oddset är 1 : 5 ("1mot 5").

# Exponentialfunktioner

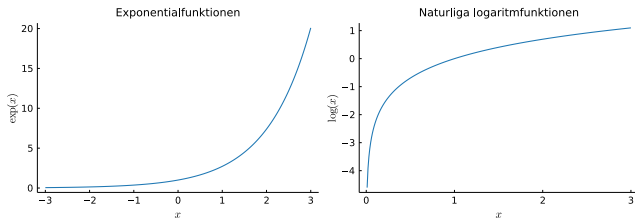
## ■ Exponentialfunktioner

$$\exp(x) = e^x$$

där  $e \approx 2.71828$  är **Eulers tal** som **bas**, istället för basen 10.

■ **Naturliga logaritmen**  $\ln(x)$  är inversa funktionen till  $\exp(x)$ .

$$\ln(e^x) = x$$

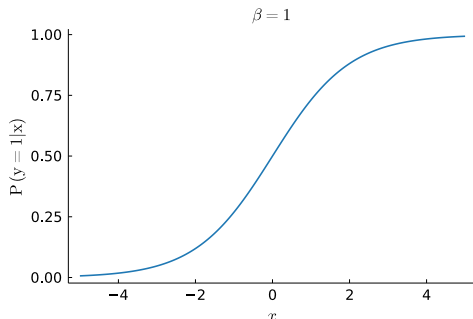


# Logistisk regression - sannolikhet för $y = 1$

- Binär responsvariabel:  $y = 0$  och  $y = 1$ .
- Logistisk regression

$$P(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$P(y = 0|x) = 1 - P(y = 1|x) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x)}$$



# Logistisk regression - oddskvot

## ■ Logistisk regression

$$P(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

$$P(y = 0|x) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x)}$$

## ■ Odds

$$\text{Odds}(y = 1|x) = \frac{P(y = 1|x)}{P(y = 0|x)} = \exp(\beta_0 + \beta_1 x)$$

- Odds i logistisk regression är **multiplikativa effekter**  
[ $\exp(a + b) = \exp(a) \exp(b)$ ]

$$\text{Odds}(y = 1|x) = \exp(\beta_0) \cdot \exp(\beta_1 x)$$

- Tolkning intercept  $\beta_0$  [ $\exp(0) = 1$ ]

$$\text{Odds}(y = 1|x = 0) = \exp(\beta_0)$$

# Logistisk regression - oddskvot

## ■ Odds

$$\text{Odds}(y = 1|x) = \exp(\beta_0 + \beta_1 x)$$

## ■ För $x = 1$

$$\text{Odds}(y = 1|x = 1) = \exp(\beta_0 + \beta_1) = \exp(\beta_0) \exp(\beta_1)$$

## ■ För $x = 2$ [ $\exp(ab) = \exp(a)^b$ ]

$$\text{Odds}(y = 1|x = 2) = \exp(\beta_0) \exp(2\beta_1) = \exp(\beta_0) \exp(\beta_1)^2$$

## ■ Tolkning $\beta_1$ : $x$ ökar med en enhet, oddset multipliceras med $\exp(\beta_1)$

## ■ **Oddskvot** för att tolka $\beta_1$

$$\text{OR}(x) = \frac{\text{Odds}(y = 1|x + 1)}{\text{Odds}(y = 1|x)} = \exp(\beta_1)$$

## ■ Bevis:

$$\text{OR}(x) = \frac{\text{Odds}(y = 1|x + 1)}{\text{Odds}(y = 1|x)} = \frac{\exp(\beta_0 + \beta_1 x + \beta_1)}{\exp(\beta_0 + \beta_1 x)} = \frac{\exp(\beta_0 + \beta_1 x) \exp(\beta_1)}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1)$$

## Oddskvot - exempel

- Sannolikhet för cancer ( $y = 1$ ) bestäms av personens ålder  $x$

$$P(y = 1|x) = \frac{\exp(-4.6 + 0.04 \cdot x)}{1 + \exp(-4.6 + 0.04 \cdot x)}$$

- Dvs  $\beta_0 = -4.6$  och  $\beta_1 = 0.04$ .
- Oddskvoten: ökning av odds med ca 4% per levnadsår:

$$\exp(\beta_1) = \exp(0.04) = 1.040811$$

- Oddset för en nyfödd är ca 1 : 100

$$\text{Odds}(y = 1|x = 0) = \exp(-4.6) = 0.01005184$$

- Oddset för en 1-åring

$$\exp(\beta_0) \exp(\beta_1) = 0.01005184 \cdot 1.040811 = 0.01046207$$

- Oddset för en 2-åring

$$\text{Odds}(y = 1|x = 1) \exp(\beta_1) = 0.01046207 \cdot 1.040811$$

- Oddset för en 100-åring:  $\exp(\beta_0 + 100\beta_1) = 0.548811$



# Logistisk regression - log-odds

- Repetition: Logaritm med bas 10:

$$\log(10^a) = a$$

- **Naturlig logaritm** (bas  $e \approx 2.7183$ )

$$\ln(\exp(a)) = \ln e^a = a$$

- Logistisk regression

$$P(y = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

- Odds

$$\text{Odds}(y = 1|x) = \exp(\beta_0 + \beta_1 x)$$

- **Log-odds**

$$\text{LogOdds}(y = 1|x) = \beta_0 + \beta_1 x$$

- Logistisk regression är en **linjär modell** för log-oddset.

# Vilka överlevde Titanic? Enkel logistisk regression

- $n = 891$  personer på Titanic, varav 342 överlevande.
- Responsvariabel:  $y = 1$  om överlevde, annars  $y = 0$ .
- Förklarande variabel: age

```
> library(regkurs)
> fit <- glm(survived ~ age, data = titanic, family = binomial)
> logisticregsummary(fit)
```

Parameter estimates

```
-----
                Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.2091888   0.1594937 -1.3116 0.189662
age          -0.0087744   0.0049474 -1.7735 0.076139
```

Odds ratio estimates

```
-----
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.81124      1.1729 -1.3116 0.189662
age           0.99126      1.0050 -1.7735 0.076139
```

~ |

# Vilka överlevde Titanic? Enkel logistisk regression

Parameter estimates  
-----

	Estimate
(Intercept)	-0.2091888
age	-0.0087744

Odds ratio estimates  
-----

	Estimate
(Intercept)	0.81124
age	0.99126

- Oddset för att överleva för en nyfödd ( $\text{age}=0$ ) är  $\exp(-0.2091888) = 0.81124$ .
- Sannolikheten för att överleva för en nyfödd ( $\text{age}=0$ ) är  $0.81124 / (1 + 0.81124) = 0.4478927$ .
- Oddset för att överleva för en 1-åring:  $0.81124 \cdot 0.99126 = 0.8041498$
- ... vilket är en minskning med  $(1 - 0.99126) \cdot 100 = 0.874\%$ .
- Varje extra levnadsår minskar oddset med 0.874%.