Regressions- och tidsserieanalys

Föreläsning 10 - AR modeller och Enkel logistisk regression.

Mattias Villani 👨

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet

mattiasvillani.com   @matvil   mattiasvillani

Översikt

- AR modeller och partiell autokorrelation
- Regression för tidsserier
- Odds och logodds
- Enkel logistisk regression

Autokorrelationsfunktion - AR(1)

- $\mathrm{AR}(1)$ som populationsmodell:

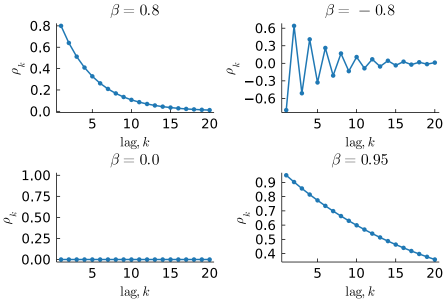
$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

- Autokorrelationsfunktion (ACF)

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

- ACF för $\mathrm{AR}(1)$

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



Partiell autokorrelationsfunktion - PACF

- $\mathrm{AR}(p)$

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

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

$$\rho_k^\star = \mathrm{corr}\left(y_t, y_{t-k} \middle| \underbrace{y_{t-1}, \ldots, y_{t-k-1}}_{\text{laggar mellan } t \text{ och } t-k} \right), \text{ för } k = 1, 2, \ldots$$

- PACF för lag k kontrollerar för laggarna mellan t och $t-k$.

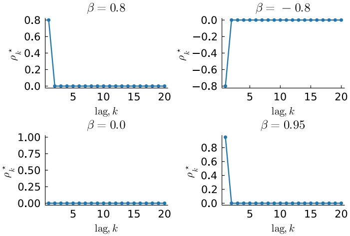
Partiell autokorrelationsfunktion - AR(1)

- $\mathrm{AR}(1)$

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

- PACF för $\mathrm{AR}(1)$ i populationen:

$$\rho_1^\star = \beta$$
$$\rho_k^\star = 0, \text{ för } k = 2, 3, \ldots$$



- Identifiera AR-modell:
  - skatta PACF från data.
  - bara första laggen signifikant $\Rightarrow$ AR(1) bra modell.

Partiell autokorrelationsfunktion - AR(2)

- AR(2)

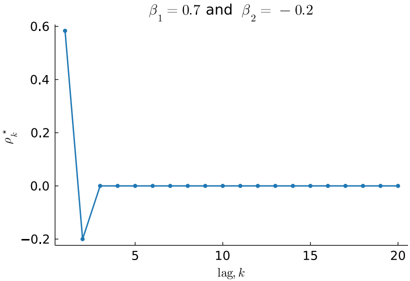
$$y_t = \alpha + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_\varepsilon^2)$$

- **Partiell ACF** för AR(2) i populationen:

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

$$\rho_2^\star = \beta_2$$

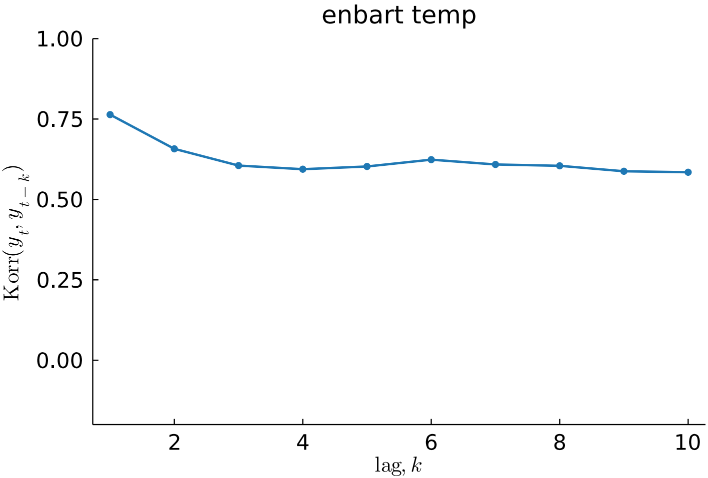
$$\rho_k^\star = 0, \text{ för } k = 3, 4, \ldots$$



Regression för tidsserier

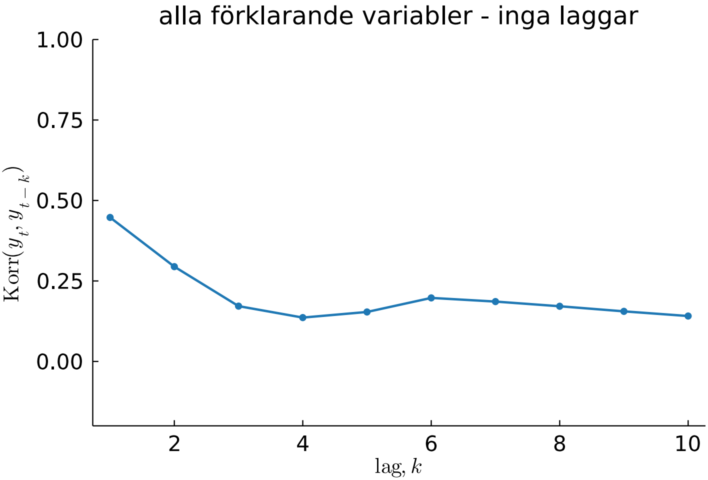
- Regression
$$y = \alpha + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon$$
där **feltermerna ε** antas bara **oberoende** från $N(0, \sigma_\varepsilon^2)$.
- **Oberoende = okorrelerade** för **normalfördelade** variabler.
- Regressionen skattas med
$$y = a + b_1 x_1 + \ldots + b_k x_k$$
och vi får 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** $\mathrm{AR}(1)$ 😁

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

- Kombinera **multipel regression och** $\mathrm{AR}(p)$ 😍

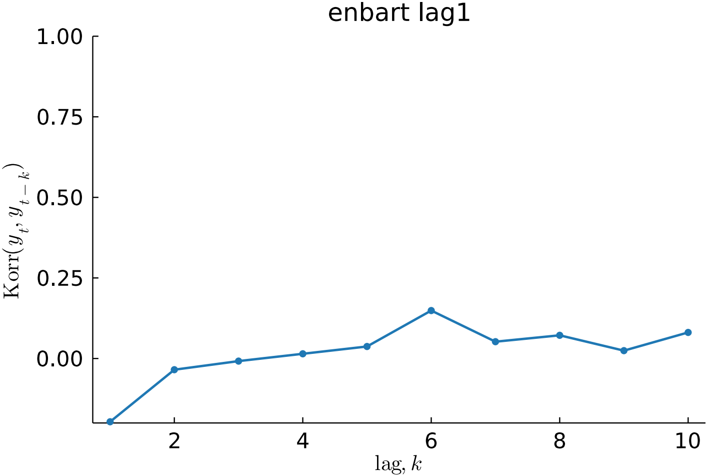
$$y_t = \alpha + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + \beta_{k+1} y_{t-1} + \ldots + \beta_{k+p} y_{t-p} + \varepsilon_t$$

- Cykeluthyrning:

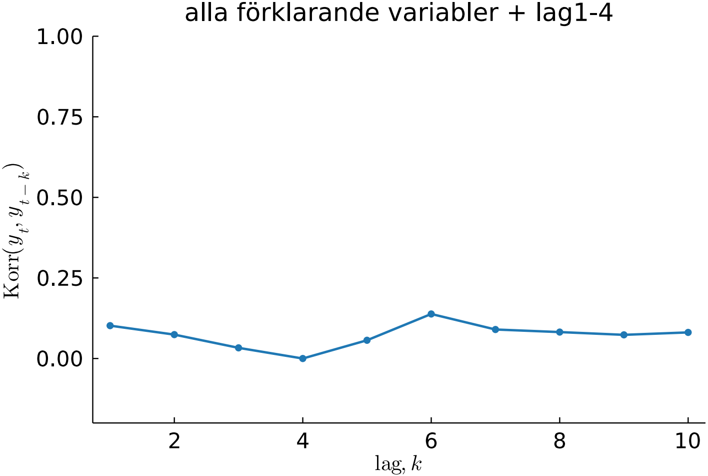
$$\texttt{AntalUthyr}_{\mathrm{idag}} = a + b_1 \cdot \texttt{temp}_{\mathrm{idag}} + b_2 \cdot \texttt{AntalUthyr}_{\mathrm{igar}}$$

- 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 feltermer).
- 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:
$$\begin{aligned} d \text{ nära } 2 &\implies \text{ej signifikant} \\ d < 1 &\implies \text{signifikant positiv autokorrelation} \\ d > 1 &\implies \text{signifikant negativ autokorrelation} \end{aligned}$$
- 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
```

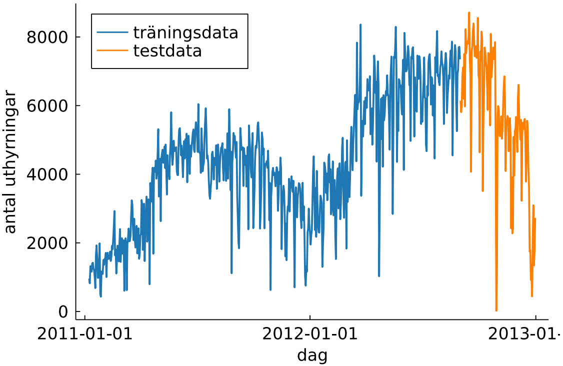
```
> lmfit = lm(nRides ~ temp + hum + windspeed + holiday + workingday + as.factor(season) + yr, data = bike)
> durbinWatsonTest(lmfit)
 lag Autocorrelation D-W Statistic p-value
   1       0.4472755      1.104221       0
 Alternative hypothesis: rho != 0
```

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

Cykeluthyrningar - utvärdera prognosförmåga

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

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



Cykeluthyrningar

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

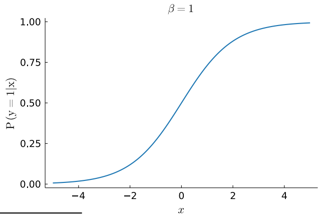
Förklarande variabler	R^2	$\mathrm{RMSE}_{\mathrm{test}}$
temp	0.385	2346.60
temp,hum,windspeed,holiday,workingday,säsong,yr	0.795	1292.07
lag1	0.714	1274.32
lag1,lag2	0.730	1279.30
lag1-lag4	0.746	1267.84
lag1-lag6	0.764	1262.10
temp,hum,windspeed,holiday,workingday,säsong,yr,lag1	0.825	1127.63
temp,hum,windspeed,holiday,workingday,säsong,yr,lag1-lag4	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

Logistisk regression för binär responsvariabel[1]

- **Binär responsvariabel**: $y = 0$ och $y = 1$.
  - företag i konkurs ($y_i = 1$) eller ej konkurs ($y_i = 0$).
  - handskriven 1:a ($y_i = 1$) eller 0:a ($y_i = 0$).
  - mynt på eBay sålt ($y_i = 1$) eller icke sålt ($y_i = 0$).
- **Logistisk regression** är modell för sannolikheten för 1:a

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

- Vilka faktorer (x-variabler) ökar/minskar $\mathrm{P}(\text{konkurs})$?



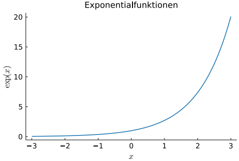
[1]Videon med matterepetition **PotenserLogaritmer.mp4** finns i Videos mappen på Athena. Se även slides PotensOchLogaritm.pdf.

Exponentialfunktionen

- **Exponentialfunktionen**

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

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



- Som för alla potenstal:

$$e^x \cdot e^y = e^{x+y}$$

och

$$(e^x)^y = e^{xy}$$

- Uttryckt i $\exp(x)$ **notation**

$$\exp(x)\exp(y) = \exp(x+y)$$

Odds och logodds

- Låt $\mathrm{P}(A)$ vara **sannolikheten** för en händelse A.

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

- **Odds**

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

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

- Exempel: Sannolikheten att slå en 6:a med en vanlig tärning:
  - Sannolikhet $\mathrm{P}(A) = 1/6$
  - Odds

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

Oddset är 1 : 5 (“1mot 5”).

Logistisk regression - oddskvot

- Logistisk regression

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

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

- **Odds**

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

- Odds i logistisk regression är **multiplikativa effekter**

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

- Jämför med **exponentiell tillväxt**: $y = a \cdot b^x$.
- Tolkning intercept β_0

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

Logistisk regression - oddskvot

- **Odds**

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

- För $x = 1$

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

- För $x = 2$

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

- Tolkning β_1: x ökar med en enhet, oddset multipliceras med

$$\exp(\beta_1)$$

- **Oddskvot** för att tolka β_1

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

Oddskvot - exempel

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

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

- 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

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

- Oddset för en 1-åring

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

- Oddset för en 2-åring

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

- 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

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

- Odds

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

- **Log-odds**

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

- 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 (age=0) är $\exp(-0.2091888) = 0.81124$.
- 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 lite mindre än 1%.