Regressions- och tidsserieanalys

Föreläsning 8 - Tidsserieanalys. Komponenter. Säsongsrensning med glidande medelvärden

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet

mattiasvillani.com   @matvil   mattiasvillani

Översikt

- Saknade förklarande variabler i regression
- Tidsserier
- Trendskattning - parametriska modeller
- Trendskattning - glidande medelvärden
- Säsongsrensning med glidande medelvärden
- Komponentsuppdelning av tidsserie.

Felspecifikation - saknade förklarande variabler

- **Population**:

$$y = \alpha + \beta_1 x_1 + \beta_2 x^2 + \varepsilon$$

- Skattad modell **korrekt specificerad**. **Väntevärderiktiga**:

$$\mathbb{E}(a) = \alpha,\ \mathbb{E}(b_1) = \beta_1 \text{ och } \mathbb{E}(b_2) = \beta_2$$

- **Skattad modell** missar att ta med x_2

$$y = a + b_1 x_1 + \varepsilon$$

- Bias

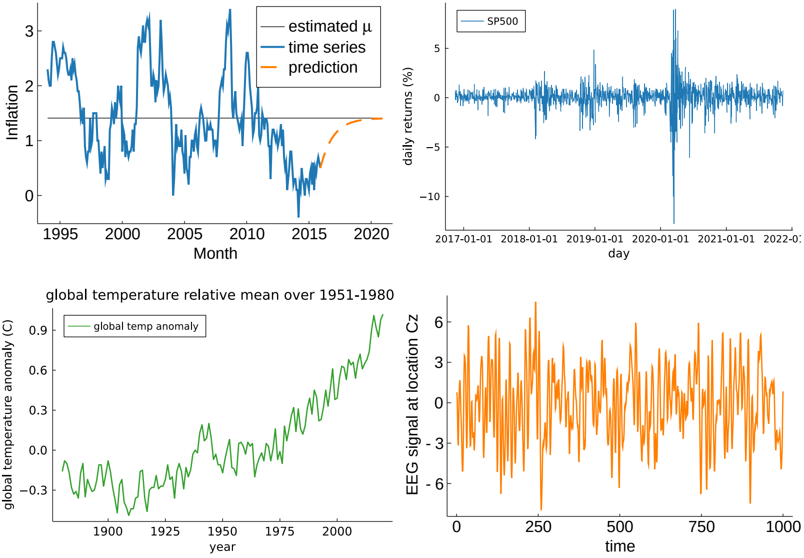
$$E(b_1) \neq \beta_1$$

- Storleken på biasen beror på **korrelationen mellan x_1 och x_2**.
- x_1 plockar upp variation i y som egentligen förklaras av x_2.

Tidsserier

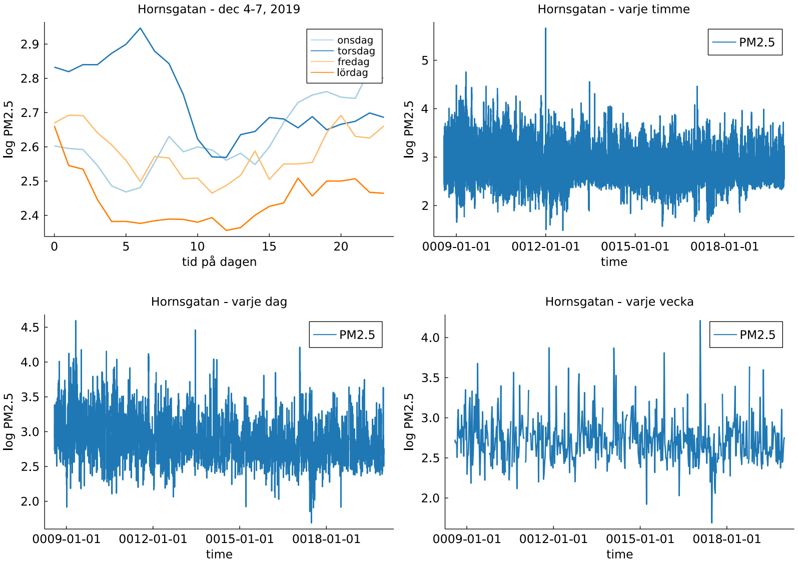
- **Tvärsnittsdata** data uppmätta vid en tidpunkt. Regression.
- **Tidsseriedata:** data uppmätta över **tid**. y_t, $t = 1, 2, \ldots$
- Mäts ofta vid tidpunkter med **likstora avstånd** (varje månad).
- Tidsserier är speciella:
  - **Trender**, **säsong**.
  - **Beroende observationer** över tid. Värdet igår y_{t-1} kan användas för att prediktera dagens värde y_t. **Autokorrelation**.
  - Kräver **speciella modeller** som tar hänsyn till beroenden.

Tidsserier

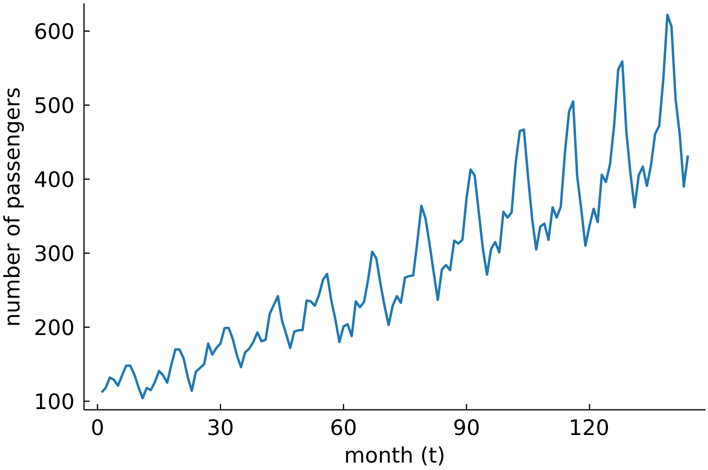


estimated μ
time series
prediction
Inflation
Month
SP500
daily returns (%)
day
2017-01-01
2018-01-01
2019-01-01
2020-01-01
2021-01-01
global temperature relative mean over 1951-1980
global temp anomaly
global temperature anomaly (C)
year
EEG signal at location Cz
time

Miljöskadliga partiklar i luften på Hornsgatan



Airline passenger data



Airline passenger data - linjär trend

- **Linjär trend**

$$y = a + b \cdot t$$

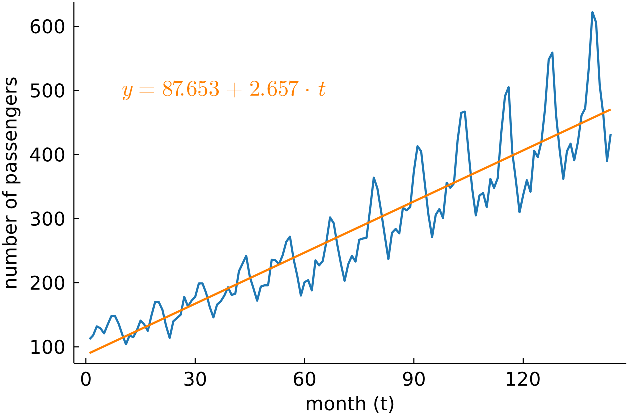
- **Minsta kvadrat**

```
passengers ~ 1 + time

Coefficients:
```

	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	87.6528	7.71635	11.36	<1e-20	72.399	102.907
time	2.65718	0.0923325	28.78	<1e-60	2.47466	2.83971

Airline passenger data - linjär trend



- $R^2 = 0.853$.

Airline passenger data - kvadratisk trend

- **Kvadratisk trend**

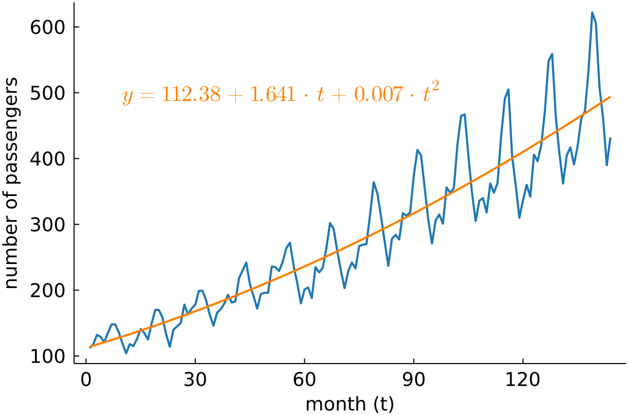
$$y = a + b_1 \cdot t + b_2 \cdot t^2$$

- Minsta kvadrat

```
passengers ~ 1 + time + :(time ^ 2)

Coefficients:
─────────────────────────────────────────────────────────────────────────
                  Coef.    Std. Error     t  Pr(>|t|)   Lower 95%   Upper 95%
─────────────────────────────────────────────────────────────────────────
(Intercept)  112.38       11.3841       9.87    <1e-17  89.8744     134.886
time           1.641       0.362473     4.53    <1e-04   0.92441      2.35758
time ^ 2       0.0070082   0.00242149   2.89    0.0044   0.00222108   0.0117953
─────────────────────────────────────────────────────────────────────────
```

Airline passenger data - kvadratisk trend



- $R^2 = 0.862$.

Airline passenger data - exponentiell trend

- **Exponentiell trend**

$$y = a \cdot b^t$$

- Skattas med minsta kvadrat genom att **logaritmera data**

$$\underbrace{\log y}_{\tilde{y}} = \underbrace{\log a}_{\tilde{a}} + \underbrace{\log b}_{\tilde{b}} \cdot t$$

$$\tilde{y} = \tilde{a} + \tilde{b} \cdot t$$

$$\tilde{a} = \log a$$

$$\tilde{b} = \log b$$

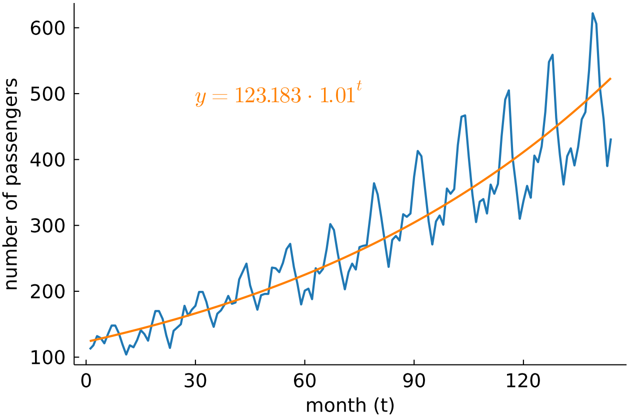
```
logpassengers ~ 1 + time

Coefficients:
```

	Coef.	Std. Error	t	Pr(>|t|)	Lower 95%	Upper 95%
(Intercept)	2.09055	0.0101165	206.65	<1e-99	2.07055	2.11055
time	0.00436396	0.000121052	36.05	<1e-72	0.00412466	0.00460325

- $a = 10^{\tilde{a}} = 10^{2.09055} \approx 123.183$
- $b = 10^{\tilde{b}} = 10^{0.00436396} \approx 1.010.$

Airline passenger data - exponentiell trend



- $R^2 = 0.902$ för logarimerade data. Kan inte jämföras med tidigare modeller!

Airline passenger data - exponentiell trend

```
logpassengers ~ 1 + time

Coefficients:
─────────────────────────────────────────────────────────────────────────────
                  Coef.    Std. Error       t  Pr(>|t|)   Lower 95%   Upper 95%
─────────────────────────────────────────────────────────────────────────────
(Intercept)  2.09055     0.0101165     206.65    <1e-99  2.07055     2.11055
time         0.00436396  0.000121052    36.05    <1e-72  0.00412466  0.00460325
─────────────────────────────────────────────────────────────────────────────
```

- Approximativt (n=144) **95% konfidensintervall för $\tilde{b}$**

  $$0.00436396 \pm 1.96 \cdot 0.0001211052 = (0.004126594, 0.00460133)$$

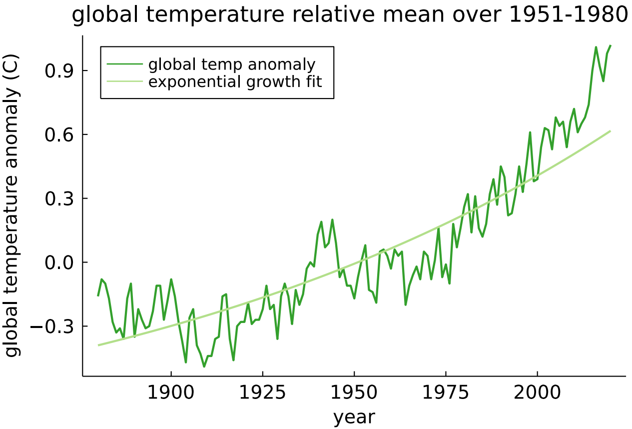
- Approximativt (n=144) **95% konfidensintervall för b** genom att anti-logga gränserna

  $$(10^{0.004126594}, 10^{0.00460133}) \approx (1.0095, 1.0107)$$

  dvs mellan 0.95% och 1.07% ökning per månad.

- 1.07% ökning per månad blir $1.0107^{12} \approx 1.1362$, dvs ca 13.62% ökning per år.

Global temperatur - exponentiell trend



- $R^2 = 0.764$ för logarimerade data.

Trendskattning genom glidande medelvärden

- 3-punkts (centrerat) **glidande medelvärde** med **lika vikter**:

$$M_t = (y_{t-1} + y_t + y_{t+1})\,/3$$

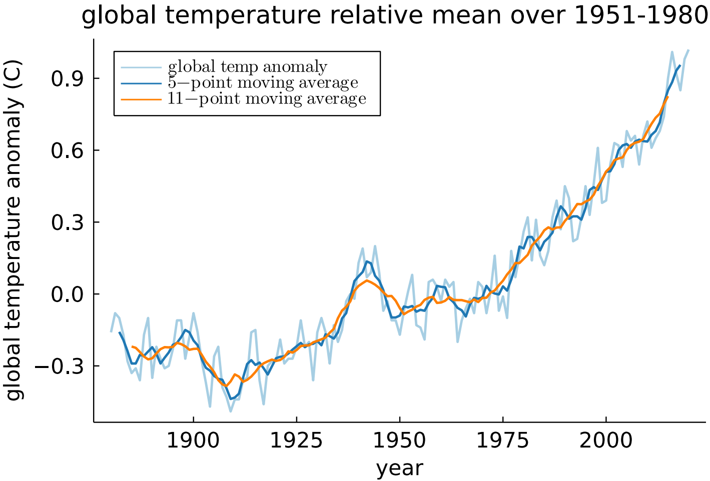
- 3-punkts **glidande medelvärde** med **olika vikter**:

$$M_t = w_{-1}y_{t-1} + w_0 y_t + w_1 y_{t+1}$$

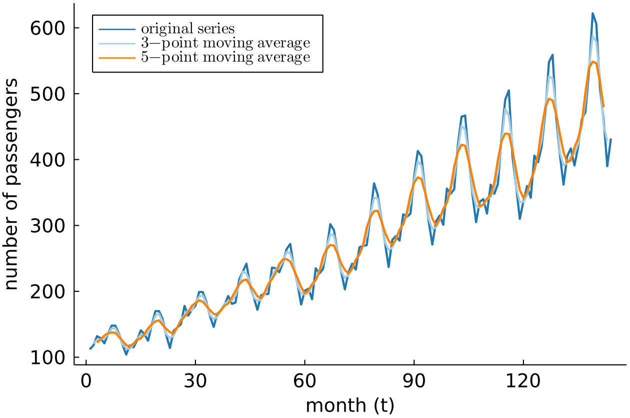
- Notera att vikterna måste summera till 1.
- ***r*-punkts glidande medelvärde**

$$M_t = \sum_{s-r}^{r} w_s y_{t+s}$$

Trendskattning genom glidande medelvärden



Airline passenger data - glidande medelvärden



Trendskattning - glidande medelvärden - säsong

- Kvartalsdata (ex: $t = \text{Kvartal3}$):

$$M_t^{(5)} = \left(\underbrace{y_{t-2}}_{\text{Kv2}} + 2\underbrace{y_{t-1}}_{\text{Kv3}} + 2\underbrace{y_t}_{\text{Kv3}} + 2\underbrace{y_{t+1}}_{\text{Kv4}} + \underbrace{y_{t+2}}_{\text{Kv1}} \right) /8$$

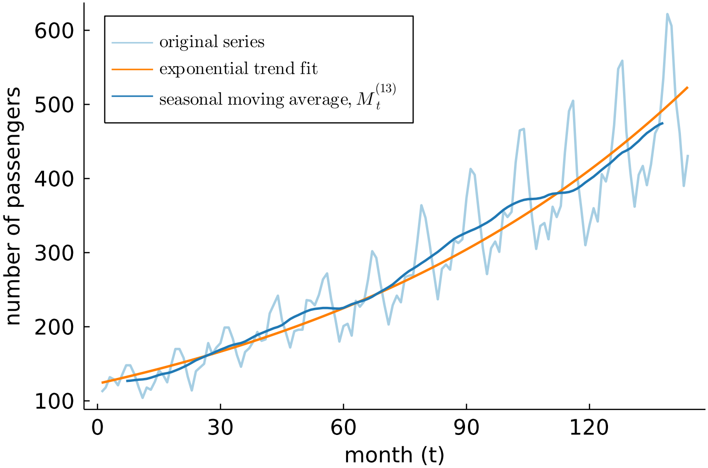
- Månadsdata (ex: $t = \text{juni}$):

$$M_t^{(13)} = \left(\underbrace{y_{t-6}}_{\text{dec}} + 2\underbrace{y_{t-5}}_{\text{jan}} + \ldots + 2\underbrace{y_t}_{\text{juni}} + \ldots + 2\underbrace{y_{t+5}}_{\text{nov}} + \underbrace{y_{t+6}}_{\text{dec}} \right) /24$$

´

Trendskattning - glidande medelvärden - säsong

´



Komponentsuppdelning

- En tidsserie kan delas upp i komponenter:
  - **Trend variation** (T)
  - **Cyklisk variation** (C)
  - **Säsongsvariation** (S)
  - **Slumpkomponent** (E)
- **Additiv modell**

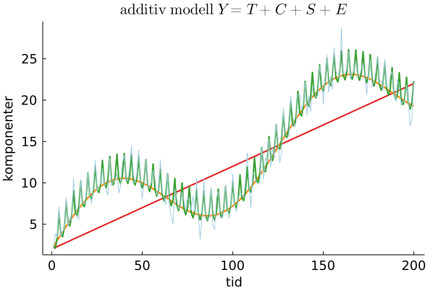
$$y_t = T_t + C_t + S_t + E_t$$

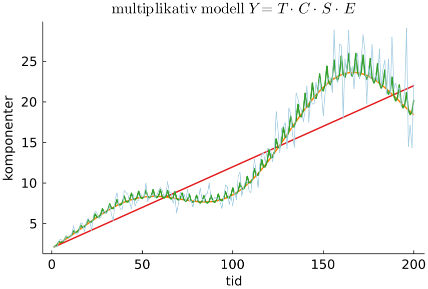
- Säsongseffekten är **visst värde över/under trend**, t ex decemberförsäljningen är 200 tkr högre i december.
- **Multiplikativ modell**

$$y_t = T_t \cdot C_t \cdot S_t \cdot E_t$$

- Säsongseffekten är **visst procent över/under trend**, t ex decemberförsäljningen är 18% högre i december.

Additiv vs multiplikativ uppdelning





Komponentsuppdelning - additiv modell

- Additiv model utan cyklisk komponent:

$$y_t = T_t + S_t + E_t$$

- Steg 1: **Bedöm modelltypen** genom att plotta tidsserien: **additiv** eller **multiplikativ**? Vilken **trendmodell**?
- Steg 2: Skatta **trendkomponenten** $\hat{T}_t$.
  T ex parametrisk modell eller glidande medelvärde.
- Steg 3: **Rensa bort trenden**: $y_t - \hat{T}_t \approx S_t + E_t$
- Steg 4: Skatta säsongskomponenten genom att beräkna medelvärden av $y_t - \hat{T}_t$ för varje säsong separat.

Skattning av säsongskomponenten

- Steg 4: **Skatta säsongskomponenten**. Ex kvartalsdata:

$$\bar{S}_1 = \frac{\sum_{\text{alla t som är kvartal 1}} (y_t - \hat{T}_t)}{\text{antal kvartal 1 observationer}}$$
$$\bar{S}_2 = \frac{\sum_{\text{alla t som är kvartal 2}} (y_t - \hat{T}_t)}{\text{antal kvartal 2 observationer}}$$
$$\bar{S}_3 = \frac{\sum_{\text{alla t som är kvartal 3}} (y_t - \hat{T}_t)}{\text{antal kvartal 3 observationer}}$$
$$\bar{S}_4 = \frac{\sum_{\text{alla t som är kvartal 4}} (y_t - \hat{T}_t)}{\text{antal kvartal 4 observationer}}$$

- Steg 5: **Korrigera säsongen** så summan av säsongskomponenterna är noll:

$$S_i^+ = \bar{S}_i - \frac{\bar{S}_1 + \bar{S}_2 + \bar{S}_3 + \bar{S}_4}{4}$$

Skattning av säsongskomponenten

- Steg 6: **Rensa bort säsongen** genom att:
  - dra av S_1^+ från alla observationer i kvartal 1
  - dra av S_2^+ från alla observationer i kvartal 2, osv

  $$y_t - \hat{T}_t - S_{i_t}^+ \approx E_t$$

  där i_t är säsongen vid tidpunkt t. T ex $i_7 = 2$ om tidpunkt $t = 7$ är i kvartal 2.

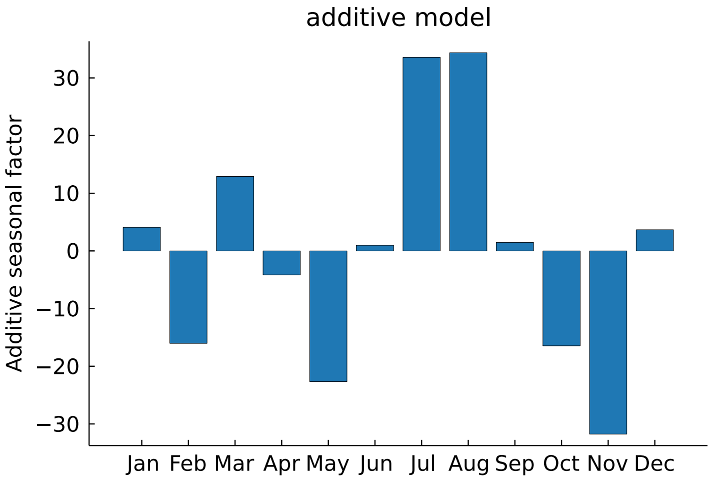
- **Multiplikativ modell - Variant 1**: logga för göra additiv

  $$\log y_t = \log T_t + \log C_t + \log S_t + \log E_t = \tilde{T}_t + \tilde{C}_t + \tilde{S}_t + \tilde{E}_t$$

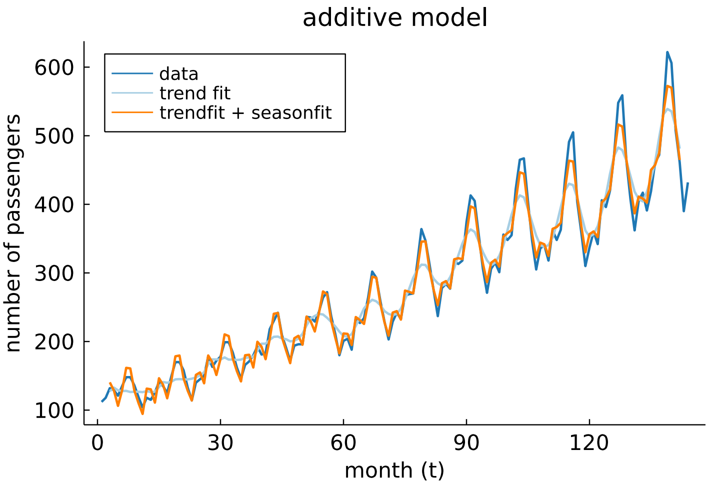
- **Multiplikativ modell - Variant 2**: uppdelning på orginalskala. Dividera istället för subtrahera för att rensa, ex:

  $$\frac{y_t}{\hat{T}_t} \approx S_t \cdot E_t$$

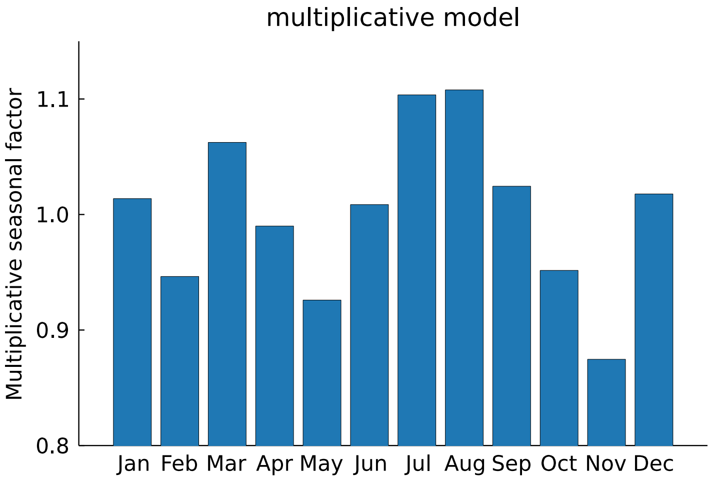
Airline passenger data - säsongskomponent S_i^+



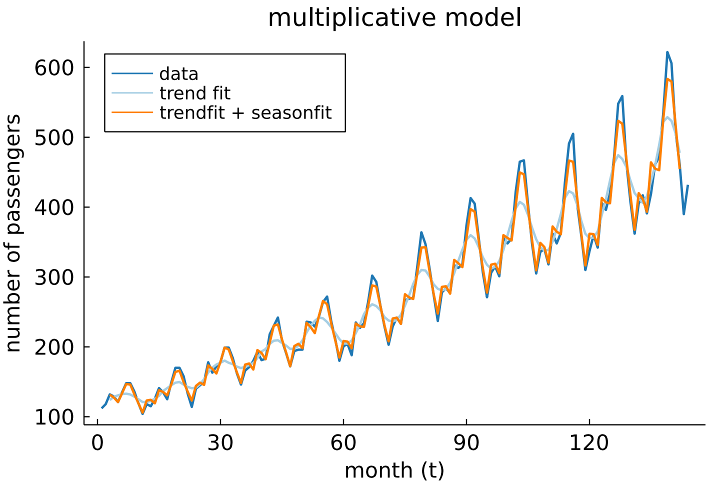
Airline passenger data - komponentanpassning



Airline passenger data - säsongskomponent S_i^+



Airline passenger data - komponentanpassning



Airline passenger data - komponentanpassning