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

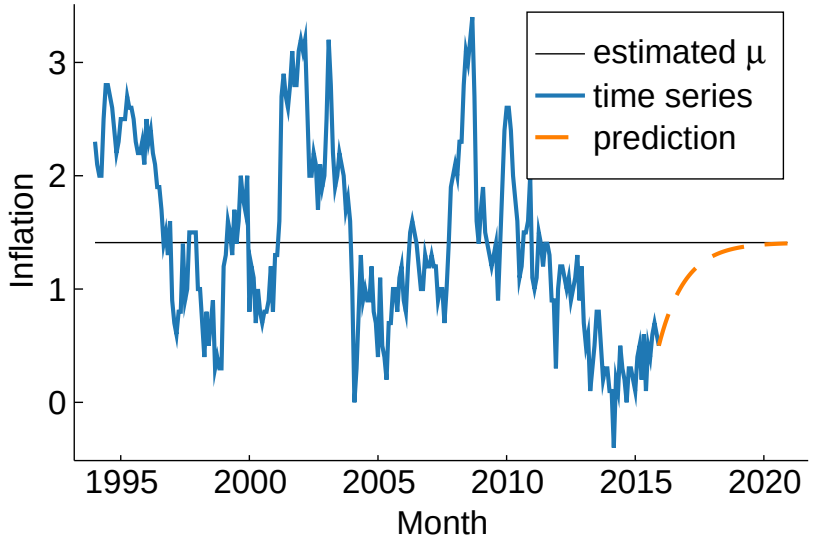
$$E(b_1) \neq \beta_1$$

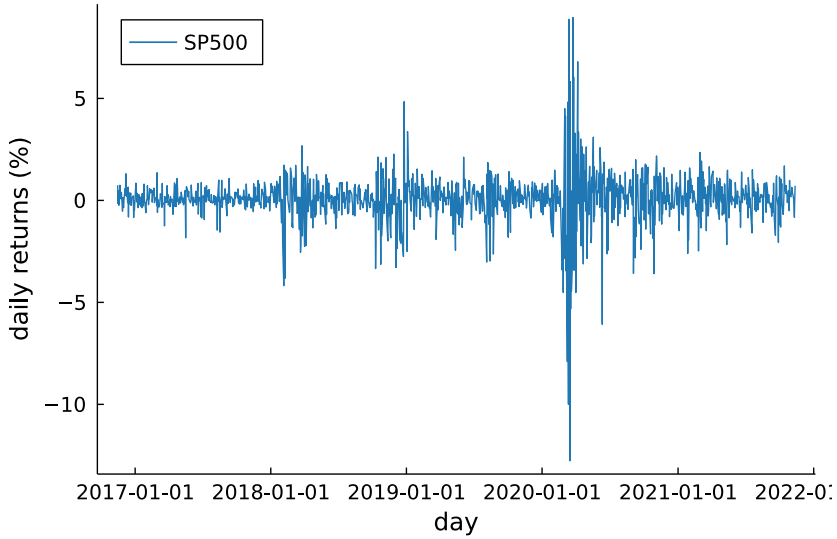
- Storleken på biasen beror på **korrelationen mellan x_1 och x_2**.
- x_1 plockar upp variation i y som egentligen förklaras av x_2.

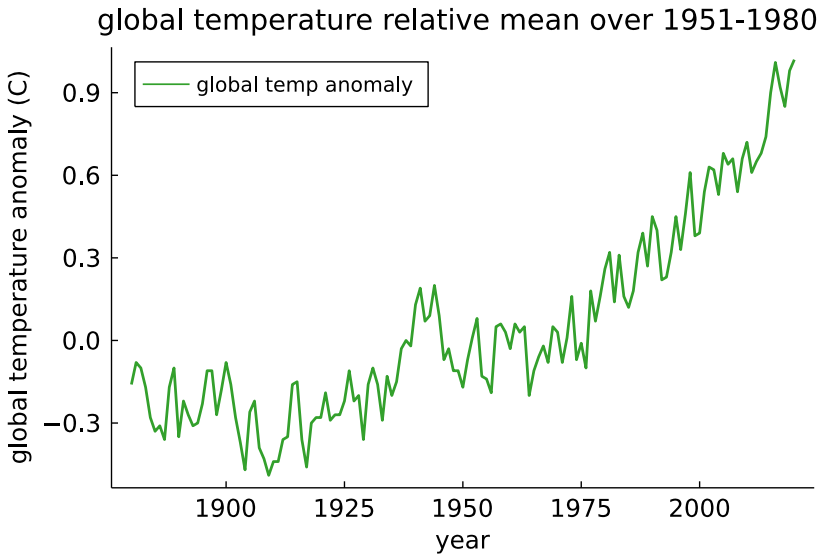
Tidsserier

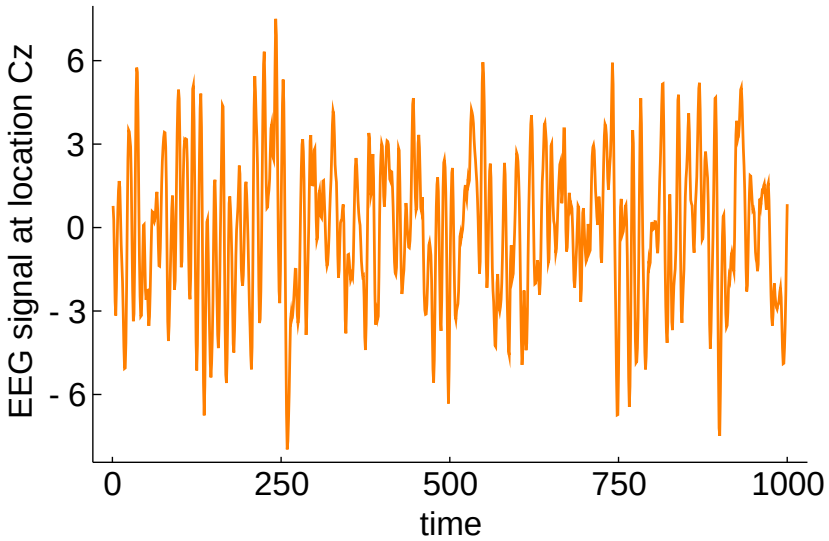
- **Tvärsnittsdata** data uppmätta vid en tidpunkt. Regression.
- **Tidsseriedata:** data uppmätta över **tid**. y_t, $t = 1, 2, ...$
- Mäts ofta vid tidpunkter med **likstora avstånd** (varje månad).
- Tidsserier är speciella:
  - **Trender**, **säsong**.
  - **Beroende observationer** över tid. Värdet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

Tidsserier

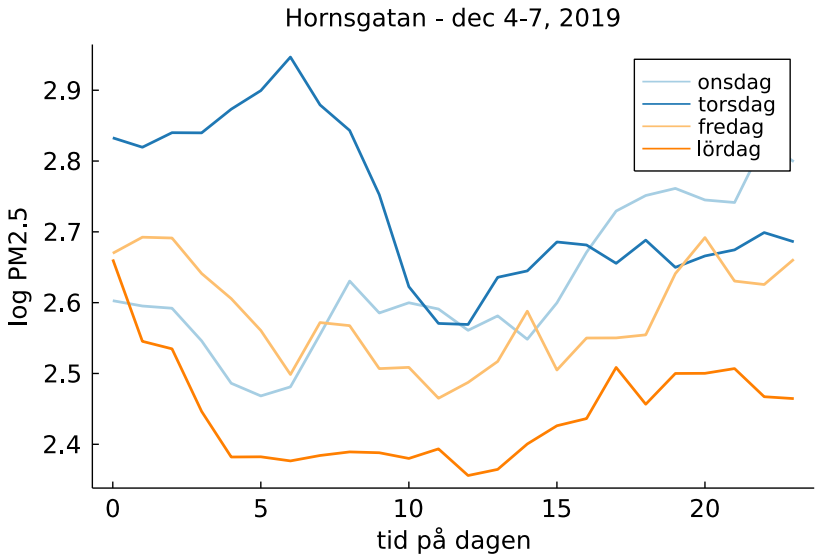

estimated μ
time series
prediction
Inflation
Month
1995
2000
2005
2010
2015
2020

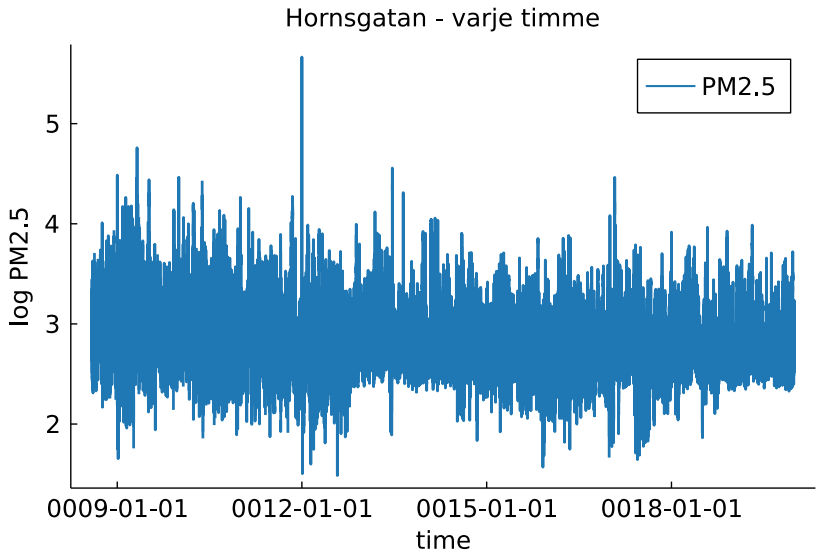

SP500
daily returns (%)
day
2017-01-01
2018-01-01
2019-01-01
2020-01-01
2021-01-01

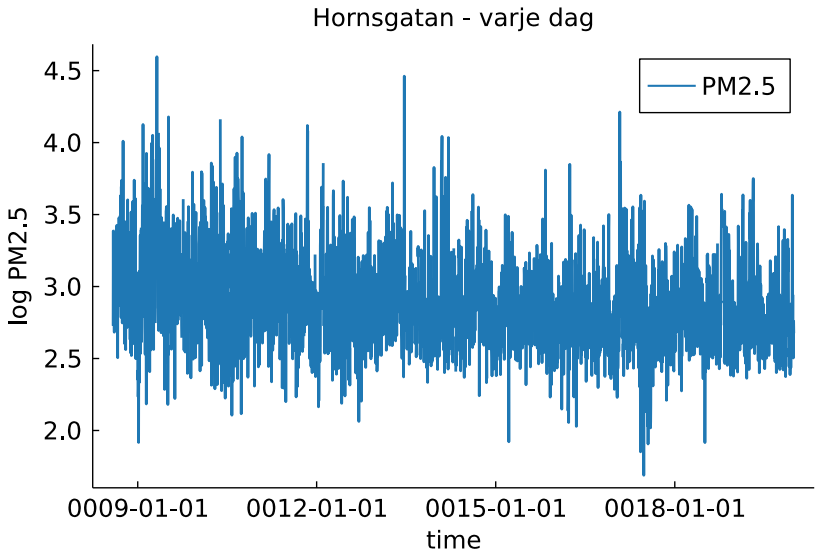

global temperature relative mean over 1951-1980
global temp anomaly
global temperature anomaly (C)
year
1900
1925
1950
1975
2000

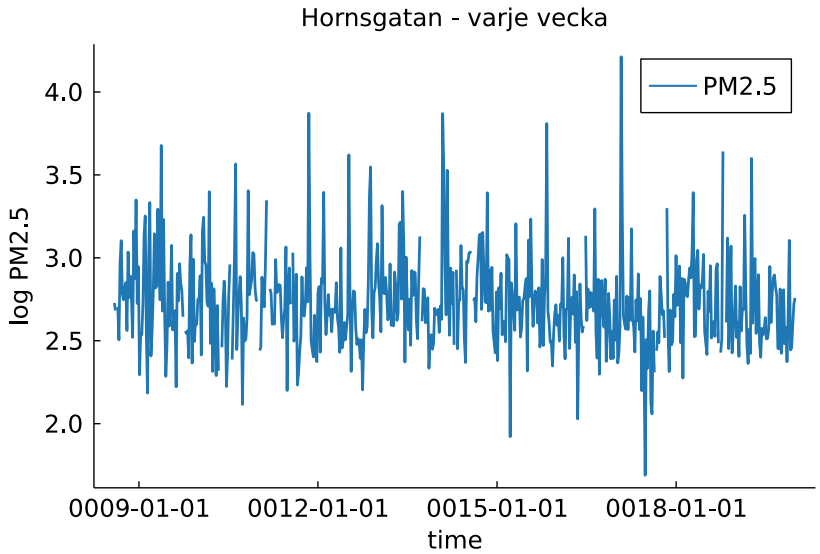

EEG signal at location Cz
time
0
250
500
750
1000

Miljöskadliga partiklar i luften på Hornsgatan





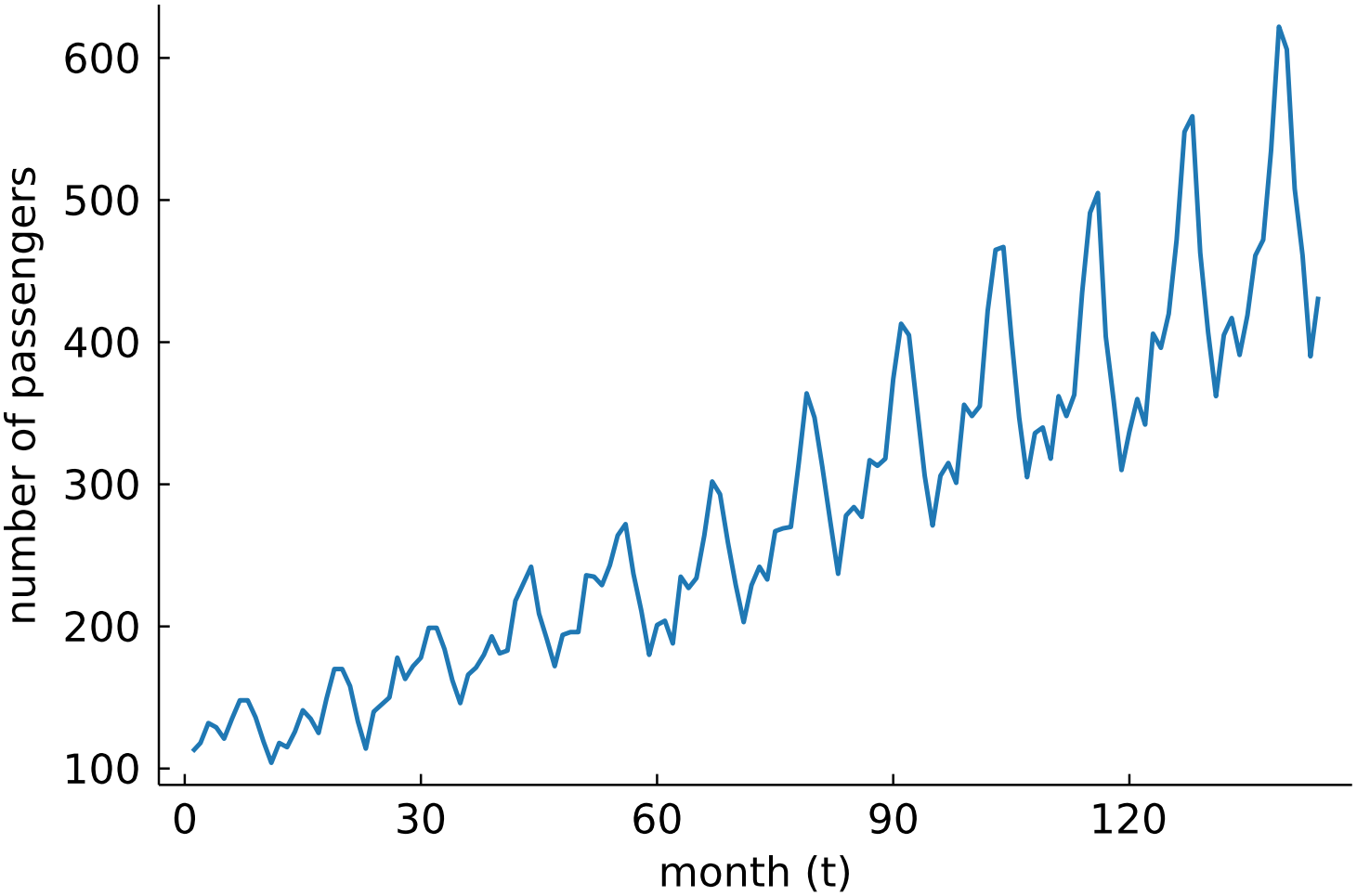




Airline passenger data

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
1960	417	391	419	461	472	535	622	606	508	461	390	432

Airline passenger data



Airline passenger data - linjär trend

- **Linjär trend**

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

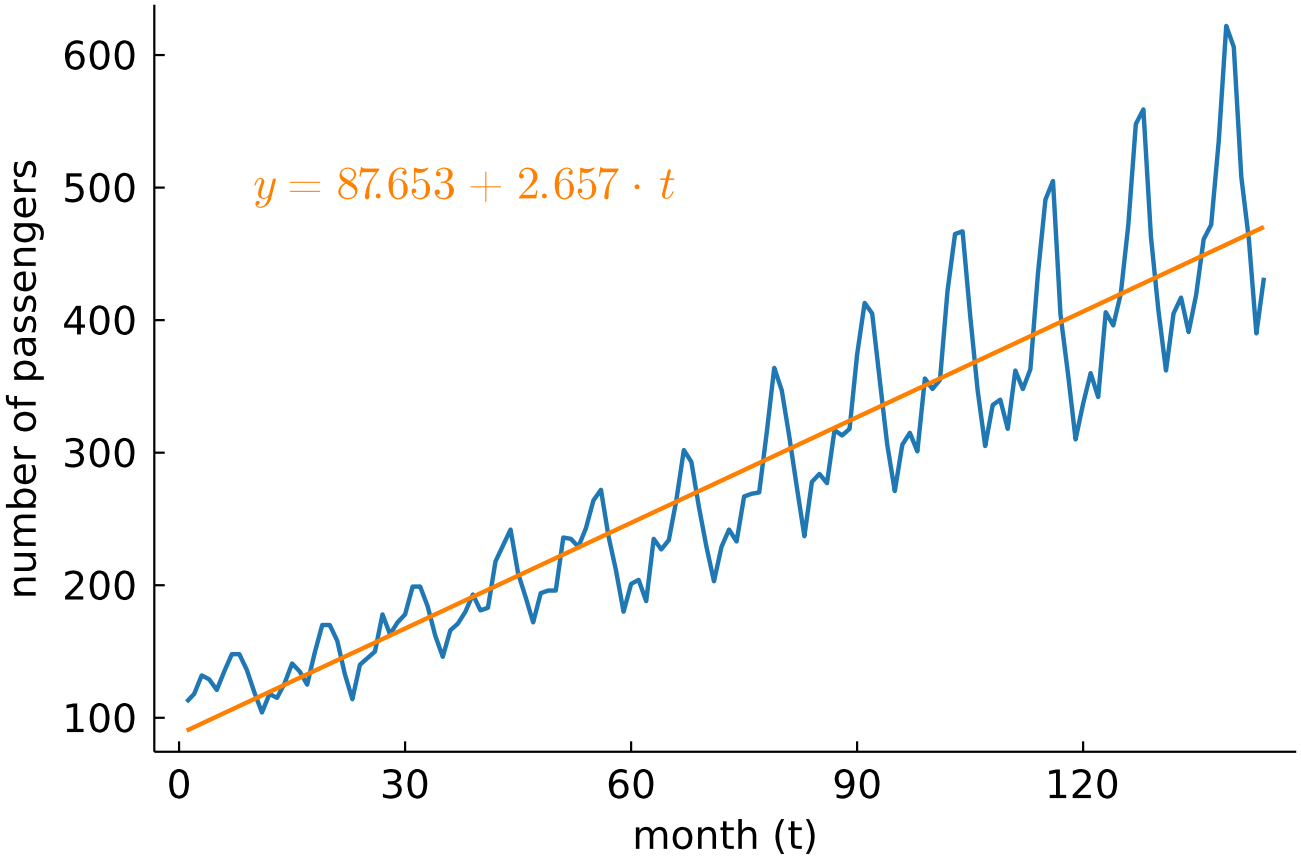
- **Minsta kvadrat**

```
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**

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

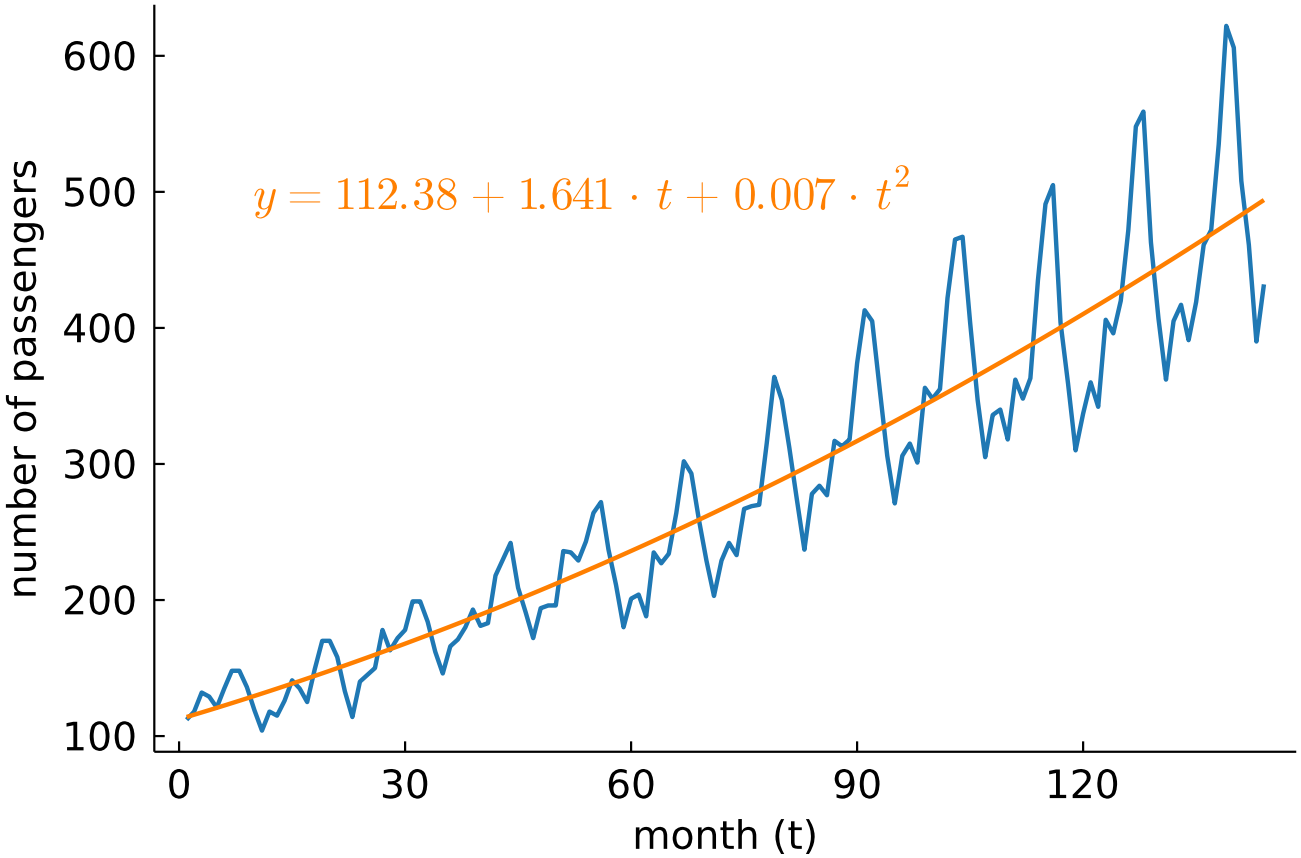
- Minsta kvadrat

```
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**

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

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

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

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

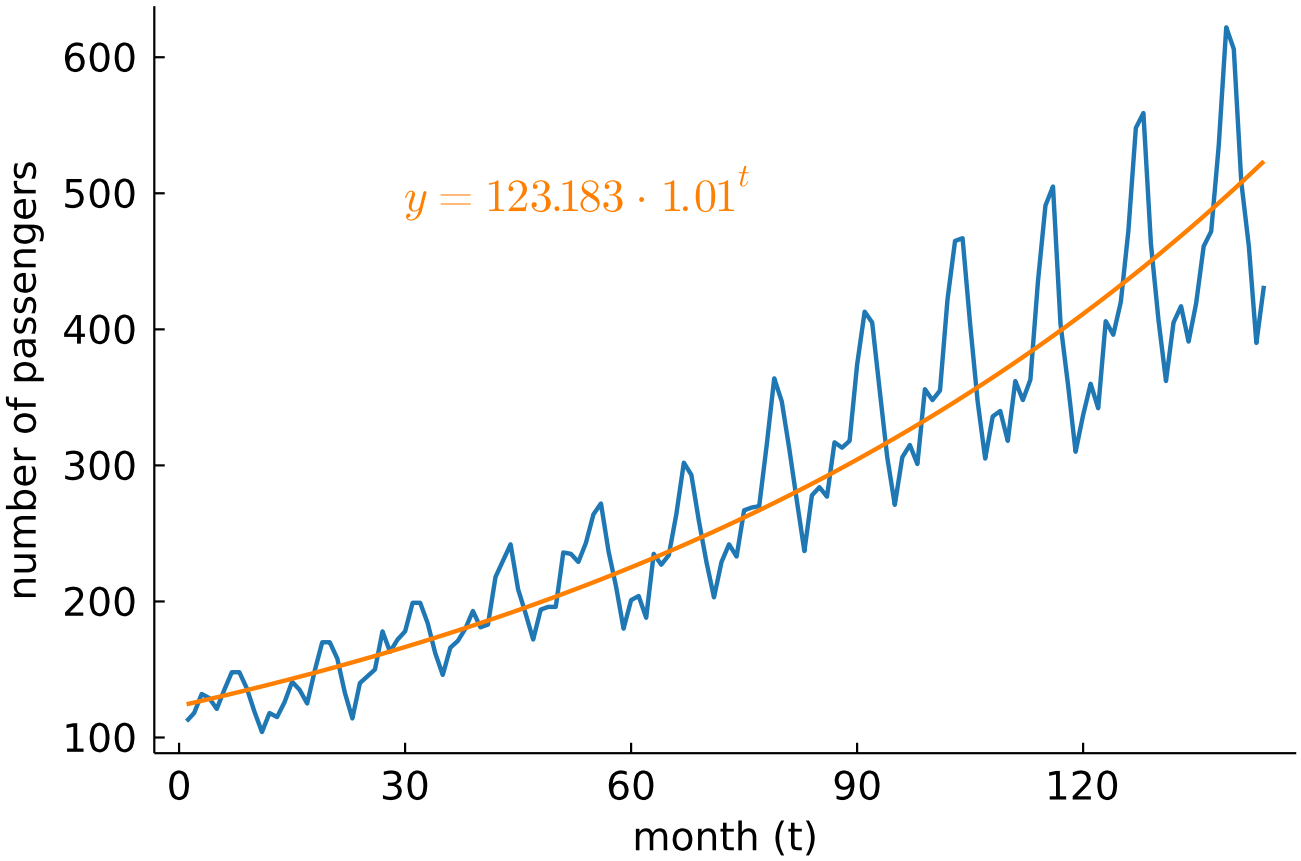
```
logpassengers ~ 1 + time

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 $\tilde{b}$

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

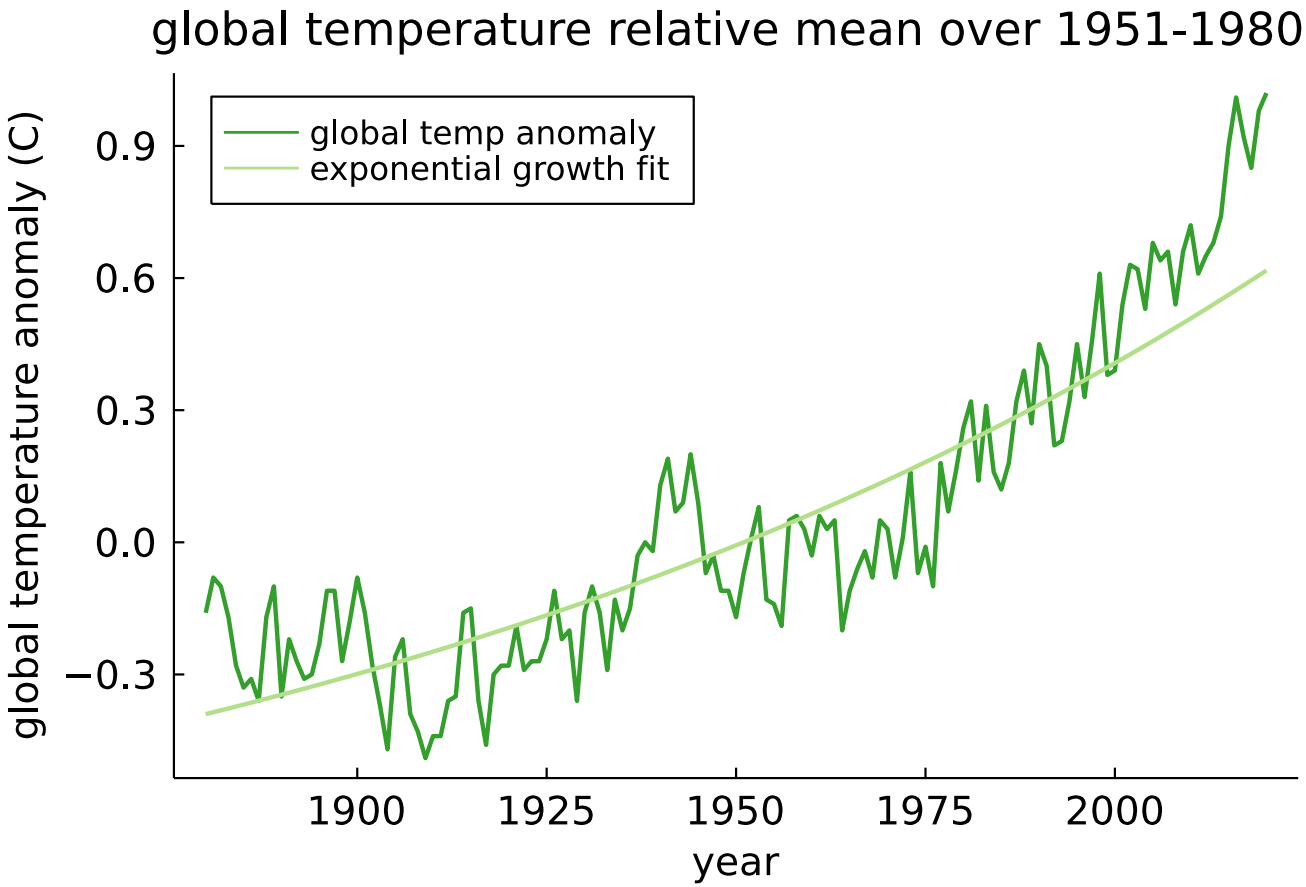
- Approximativt (n=144) 95% konfidensintervall för b genom att anti-logga gränserna

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

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**:

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

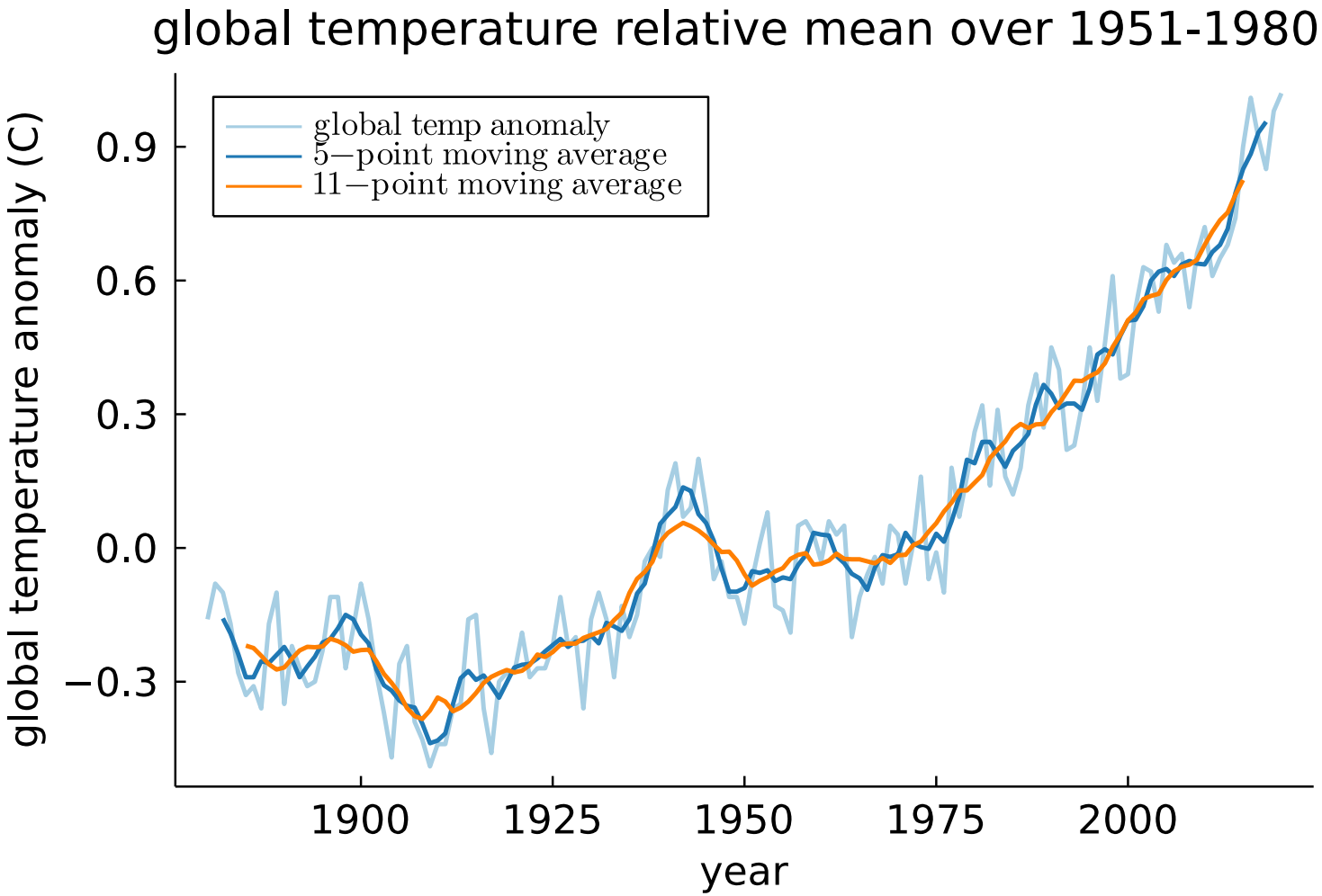
- 3-punkts **glidande medelvärde** med **olika vikter**:

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

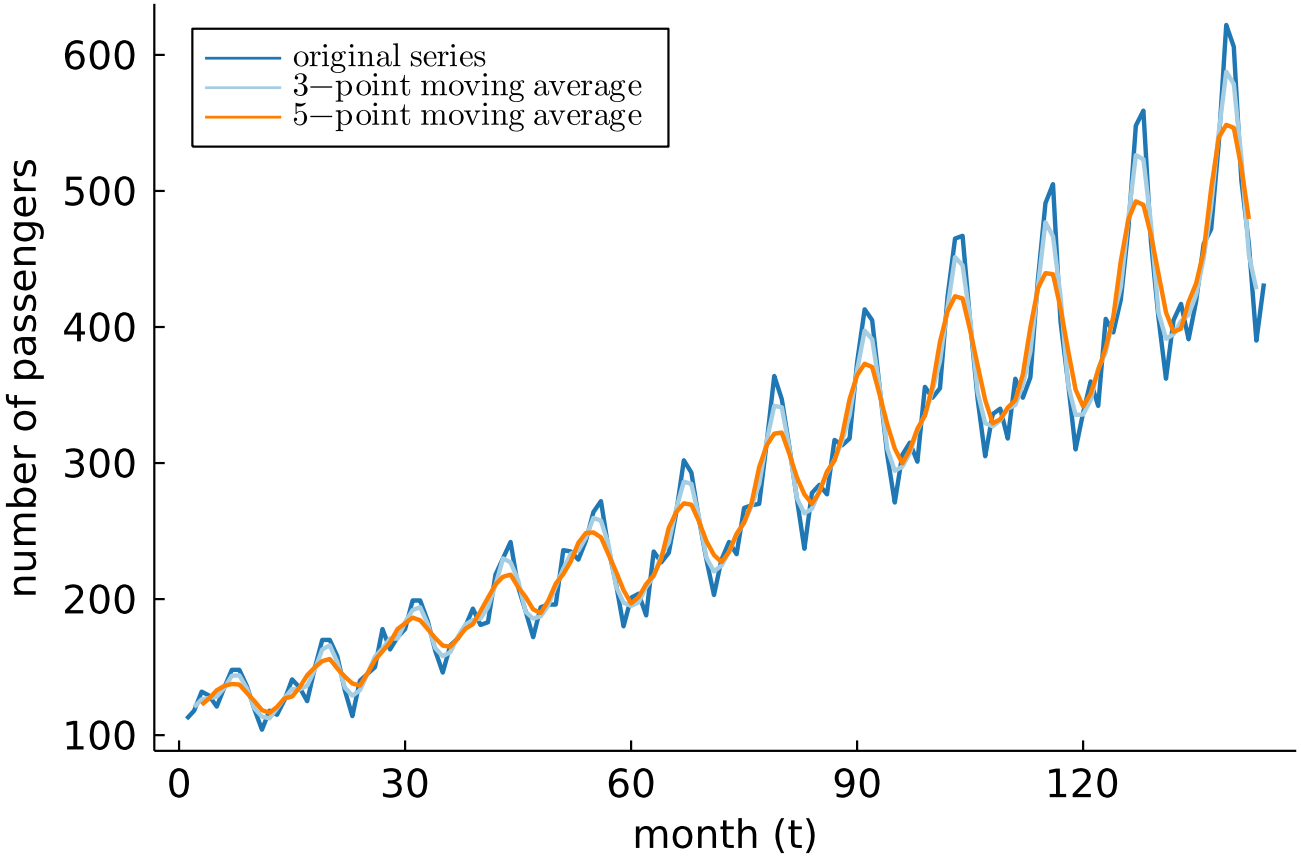
- Notera att vikterna måste summera till 1.
- ***r*-punkts glidande medelvärde**

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

Trendskattning genom glidande medelvärden



Airline passenger data - glidande medelvärden



Trendskattning - glidande medelvärden - säsong

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

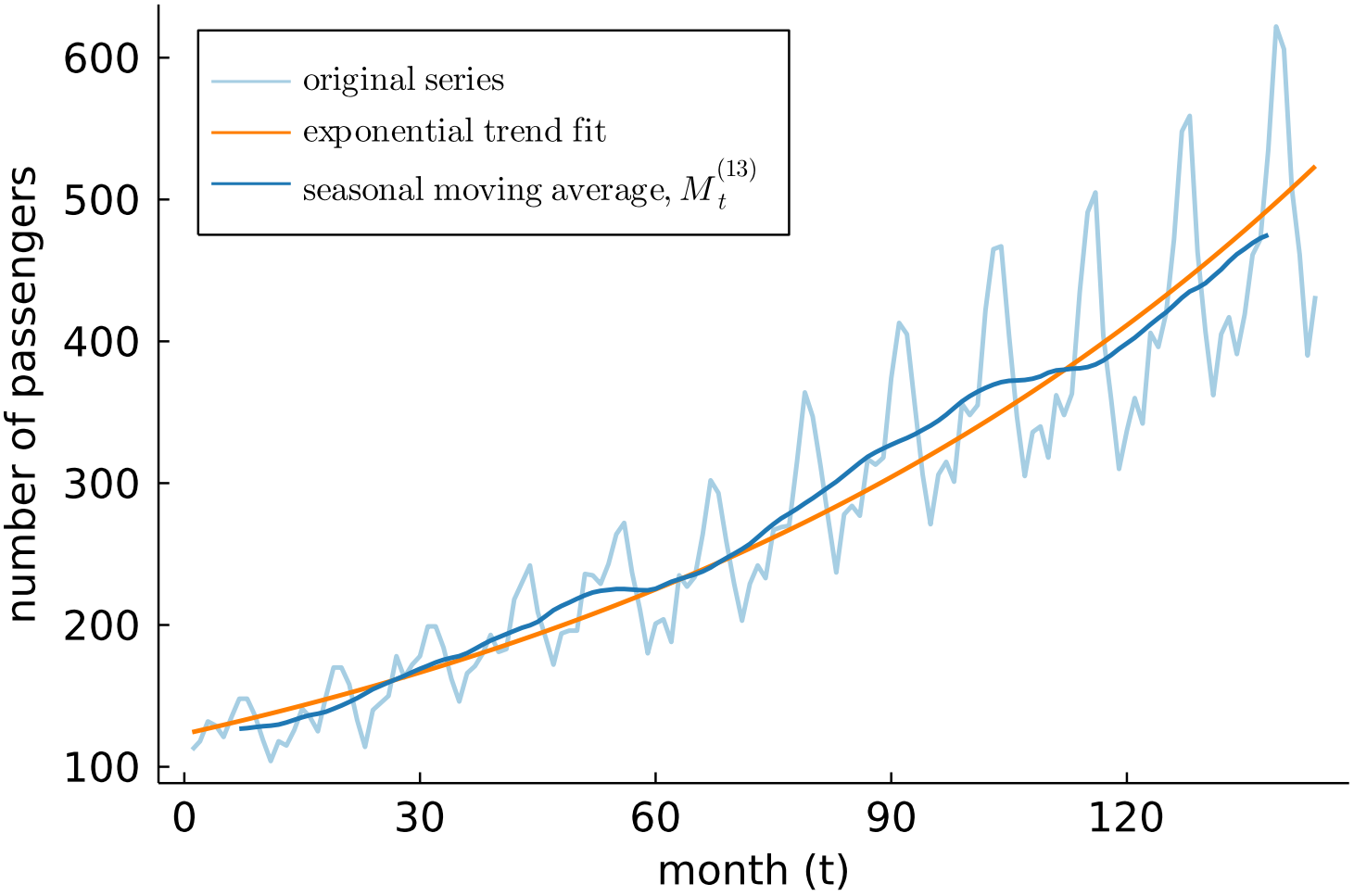
$$M_t^{(5)} = \left(\underbrace{y_{t-2}}_{\text{Kv2}} + 2\underbrace{y_{t-1}}_{\text{Kv3}} + 2\underbrace{y_t}_{\text{Kv3}} + 2\underbrace{y_{t+1}}_{\text{Kv4}} + \underbrace{y_{t+2}}_{\text{Kv1}} \right) /8$$

- 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**

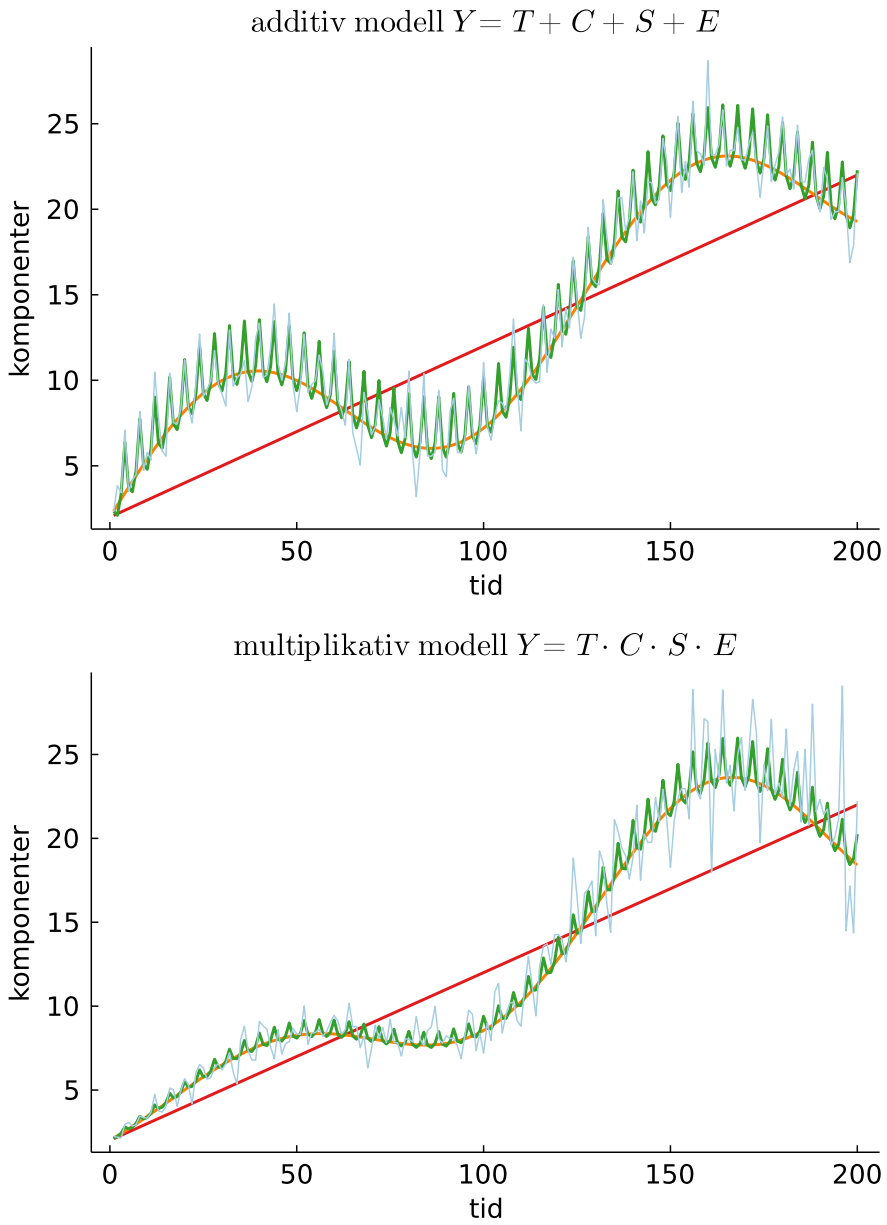
$$y_t = T_t + C_t + S_t + E_t$$

- 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.

Airline - additiv med 5-punkts glidande medel

månad	tidsserie y_t	trend $\hat{T}_t$	grov säsong $\bar{S} = y_t - \hat{T}_t$	säsong S^+	säsongsjust. $y_t - S^+$
1949-01-01	112	·	·	4.092	107.908
1949-02-01	118	·	·	-16.035	134.035
1949-03-01	132	122.4	9.6	12.910	119.090
1949-04-01	129	127.0	2.0	-4.156	133.156
1949-05-01	121	133.0	-12.0	-22.673	143.673
1949-06-01	135	136.2	-1.2	0.977	134.023
1949-07-01	148	137.6	10.4	33.577	114.423
1949-08-01	148	137.2	10.8	34.377	113.623
1949-09-01	136	131.0	5.0	1.477	134.523
1949-10-01	119	125.0	-6.0	-16.456	135.456
1949-11-01	104	118.4	-14.4	-31.763	135.763
1949-12-01	118	116.4	1.6	3.674	114.326
1950-01-01	115	120.8	-5.8	4.092	110.908
1950-02-01	126	127.0	-1.0	-16.035	142.035
1950-03-01	141	128.4	12.6	12.910	128.090
1950-04-01	135	135.2	-0.2	-4.156	139.156
1950-05-01	125	144.0	-19.0	-22.673	147.673
1950-06-01	149	149.8	-0.8	0.977	148.023
1950-07-01	170	154.4	15.6	33.577	136.423
1950-08-01	170	156.0	14.0	34.377	135.623
1950-09-01	158	149.0	9.0	1.477	156.523
1950-10-01	133	143.0	-10.0	-16.456	149.456
1950-11-01	114	138.0	-24.0	-31.763	145.763
1950-12-01	140	136.4	3.6	3.674	136.326
1951-01-01	145	145.4	-0.4	4.092	140.908
1951-02-01	150	155.2	-5.2	-16.035	166.035
1951-03-01	178	161.6	16.4	12.910	165.090
⋮	⋮	⋮	⋮	⋮	⋮

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.

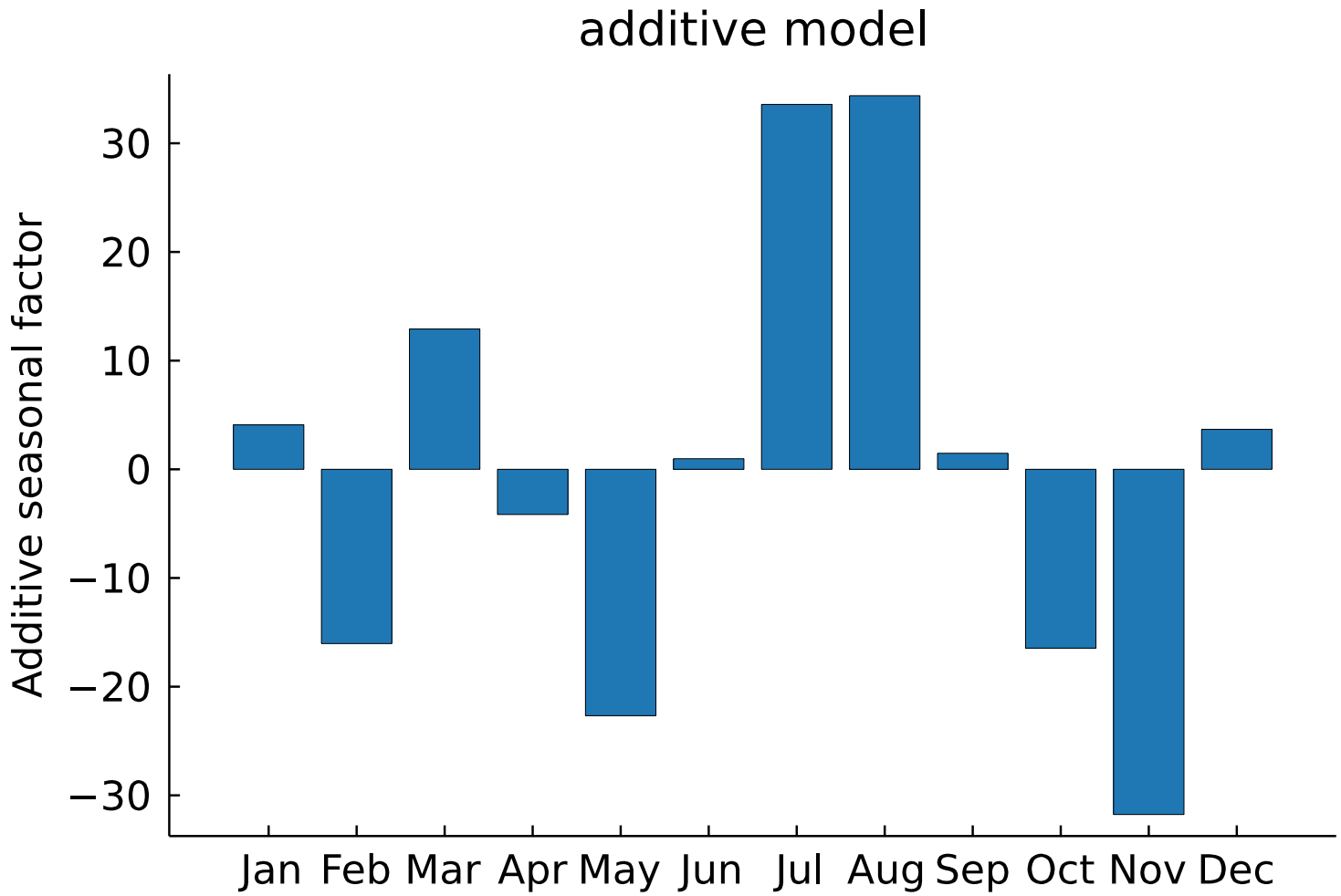
- **Multiplikativ modell - Variant 1**: logga för göra additiv

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

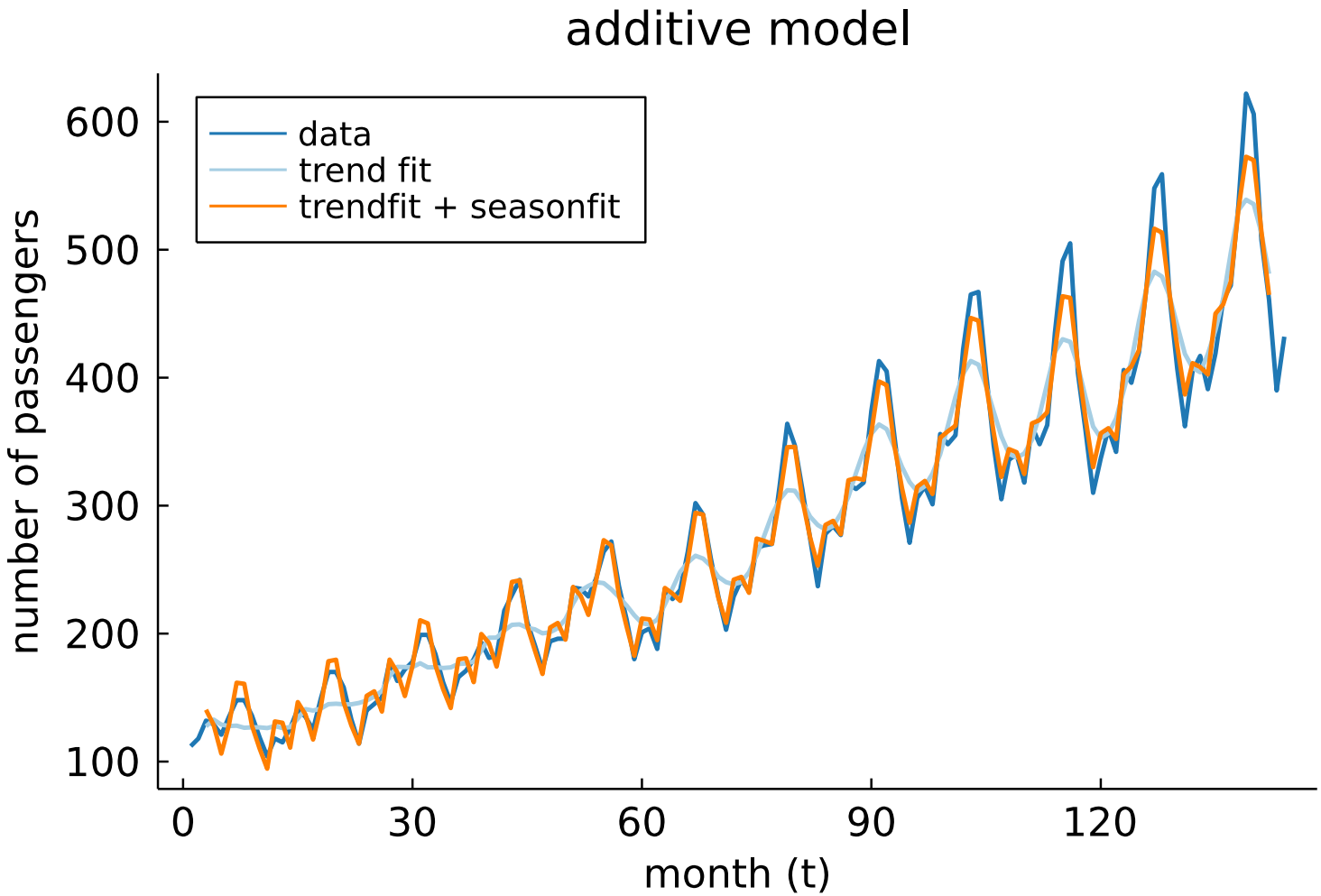
- **Multiplikativ modell - Variant 2**: uppdelning på orginalskala. Dividera istället för subtrahera för att rensa, ex:

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

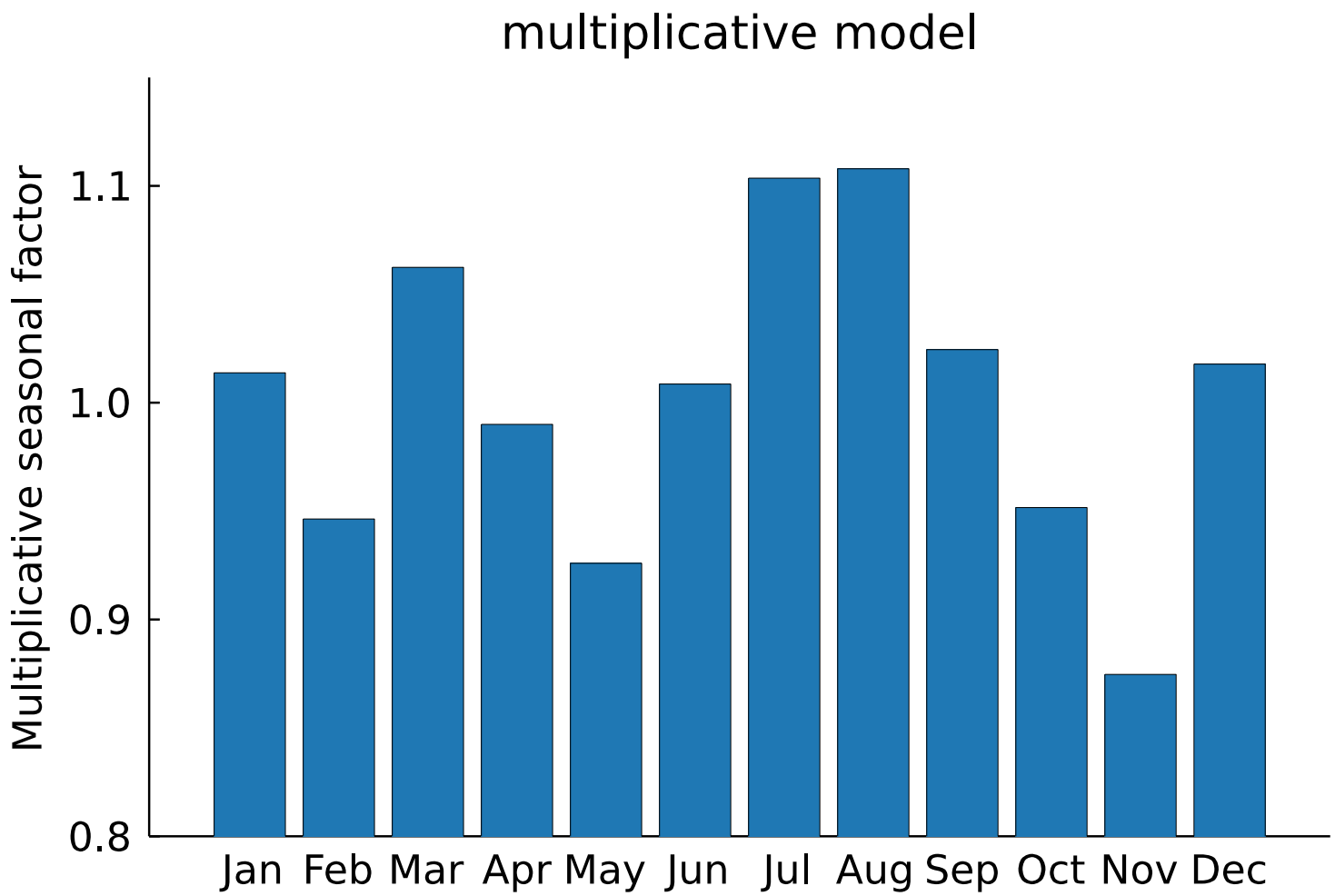
Airline passenger data - säsongskomponent S_i^+



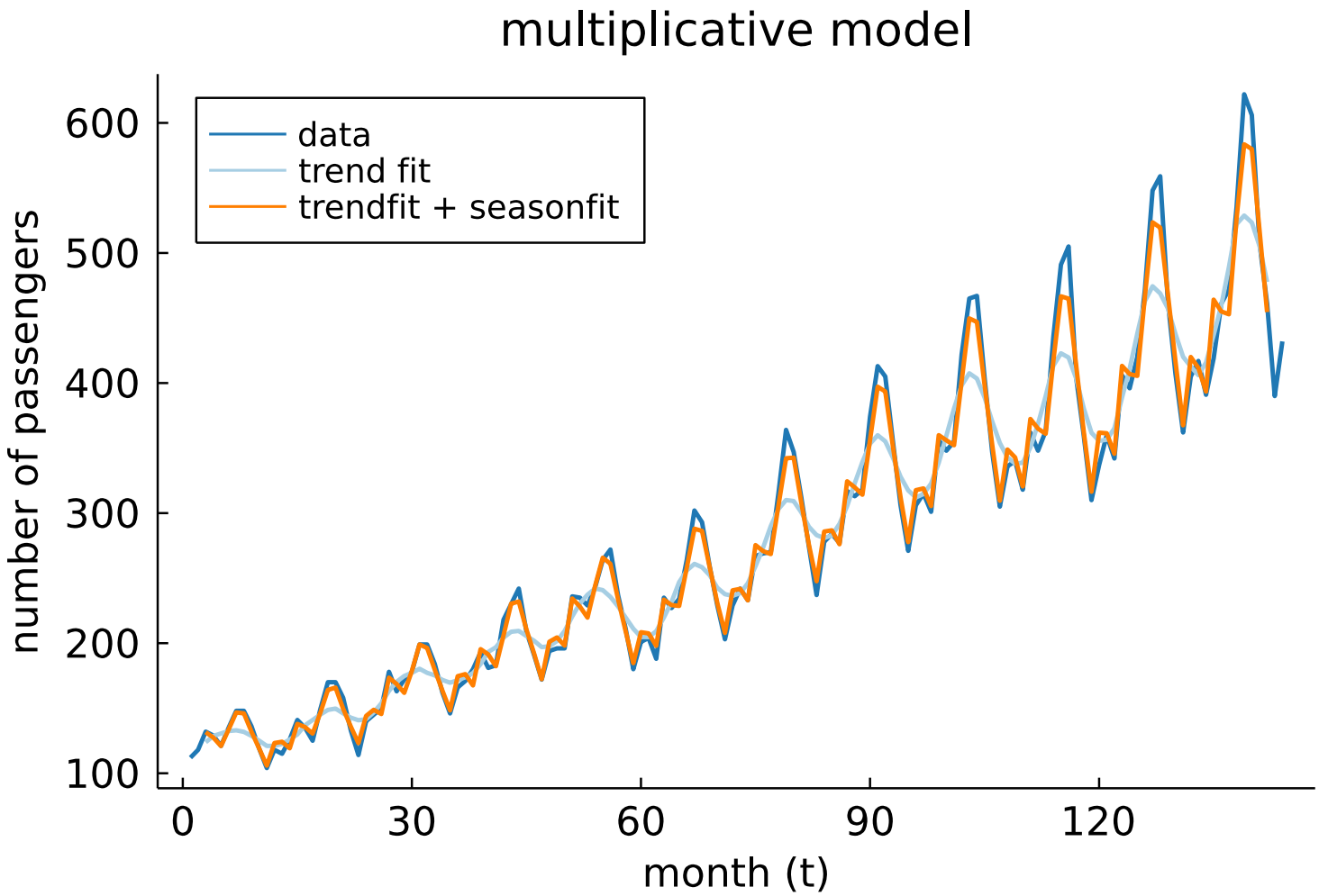
Airline passenger data - komponentanpassning



Airline passenger data - säsongskomponent S_i^+



Airline passenger data - komponentanpassning



Airline passenger data - komponentanpassning