

Regressions- och tidsserieanalys

Föreläsning 8 - Säsongrensning med regression

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet



mattiasvillani.com



@matvil



mattiasvillani

- Repetition kategoriska förklarande variabler
- Säsongrensning med regression
- Saknade förklarande variabler i regression

Binära förklarande variabler

- **Binära (dummy) variabler** som bara kan anta två värden. Ex:

$$\text{holiday} = \begin{cases} 1 & \text{om röd dag} \\ 0 & \text{annars} \end{cases}$$

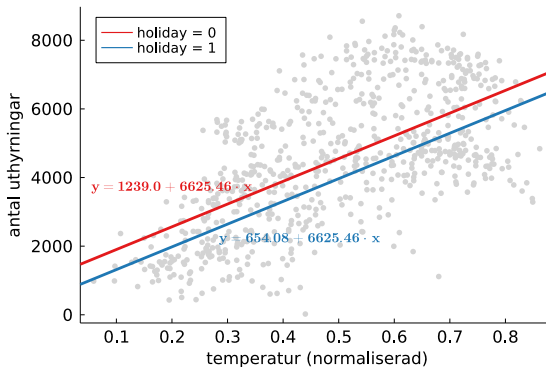
- Regressionsmodell med binär förklarande variabel:

$$y = \alpha + \beta_1 \cdot \text{temp} + \beta_2 \cdot \text{holiday} + \varepsilon$$

innebär att vi får **två parallella regressionlinjer**

$$y = \begin{cases} \alpha + \beta_1 \cdot \text{temp} + \varepsilon & \text{om holiday} = 0 \\ (\alpha + \beta_2) + \beta_1 \cdot \text{temp} + \varepsilon & \text{om holiday} = 1 \end{cases}$$

Binära förklarande variabler



Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	1	1239.00123	161.53482	7.67	<.0001	921.87157	1556.13090
temp	1	6625.45780	304.88015	21.73	<.0001	6026.90857	7224.00704
holiday	1	-584.91862	333.87396	-1.75	0.0802	-1240.38930	70.55207

Säsongrensning med regression - kvartalsdata

- Skatta **regression** med **trendmodell + kvartalsdummies**:

$$y_t = a + b \cdot t + c_1 \cdot D_{1t} + c_2 \cdot D_{2t} + c_3 \cdot D_{3t}$$

$$D_{1t} = \begin{cases} 1 & \text{om observationen vid tidpunkt } t \text{ är kvartal 1} \\ 0 & \text{annars} \end{cases}$$

$$D_{2t} = \begin{cases} 1 & \text{om observationen vid tidpunkt } t \text{ är kvartal 2} \\ 0 & \text{annars} \end{cases}$$

$$D_{3t} = \begin{cases} 1 & \text{om observationen vid tidpunkt } t \text{ är kvartal 3} \\ 0 & \text{annars} \end{cases}$$

Säsongrensning med regression - månadsdata

- Skatta **regression** med **trendmodell + månadsdummies**:

$$y_t = a + b \cdot t + c_1 \cdot D_{1t} + c_2 \cdot D_{2t} + \dots + c_{11} \cdot D_{11t}$$

$$D_{kt} = \begin{cases} 1 & \text{om observationen vid tidpunkt } t \text{ är månad nr } k \\ 0 & \text{annars} \end{cases}$$

- Multiplikativ modell: logga y_t .

Airline passenger data - säsongsdummies

Year	nPass	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949-01-01	112	1	0	0	0	0	0	0	0	0	0	0	0
1949-02-01	118	0	1	0	0	0	0	0	0	0	0	0	0
1949-03-01	132	0	0	1	0	0	0	0	0	0	0	0	0
1949-04-01	129	0	0	0	1	0	0	0	0	0	0	0	0
1949-05-01	121	0	0	0	0	1	0	0	0	0	0	0	0
1949-06-01	135	0	0	0	0	0	1	0	0	0	0	0	0
1949-07-01	148	0	0	0	0	0	0	1	0	0	0	0	0
1949-08-01	148	0	0	0	0	0	0	0	1	0	0	0	0
1949-09-01	136	0	0	0	0	0	0	0	0	1	0	0	0
1949-10-01	119	0	0	0	0	0	0	0	0	0	1	0	0
1949-11-01	104	0	0	0	0	0	0	0	0	0	0	1	0
1949-12-01	118	0	0	0	0	0	0	0	0	0	0	0	1
1950-01-01	115	1	0	0	0	0	0	0	0	0	0	0	0
1950-02-01	126	0	1	0	0	0	0	0	0	0	0	0	0
.
.
1960-11-01	390	0	0	0	0	0	0	0	0	0	0	1	0
1960-12-01	432	0	0	0	0	0	0	0	0	0	0	0	1

Säsongrensning med regression

- Skatta **regression** med **trendmodell + kvartalsdummies**:

$$y_t = a + b \cdot t + c_1 \cdot D_{1t} + c_2 \cdot D_{2t} + c_3 \cdot D_{3t}$$

- Uttrycket $a + b \cdot t$ är **inte** trendkomponenten.
Lutningen b är korrekt, men a är interceptet under vintern.
- Den **genomsnittliga trenden**:

$$y_t = a_0 + b \cdot t$$

- Vi vet att:

$$\bar{y} = a_0 + b \cdot \bar{t}$$

$$a_0 = \bar{y} - b \cdot \bar{t}$$

- **Säsongen** som avvikelse från trenden

$$S_1 = a + c_1 - a_0 \text{ (vår)}$$

$$S_2 = a + c_2 - a_0 \text{ (sommar)}$$

$$S_3 = a + c_3 - a_0 \text{ (höst)}$$

$$S_4 = a - a_0 \text{ (vinter)}$$

Airline - regression med säsongsdummies

	Coef.	Std. Error	t	Pr(> t)	Lower 95%	Upper 95%
(Intercept)	54.3277	8.65118	6.28	<1e-08	37.2135	71.4418
time	2.66033	0.0529682	50.23	<1e-86	2.55555	2.76511
Jan	9.18029	10.7651	0.85	0.3953	-12.1156	30.4761
Feb	-0.230041	10.7623	-0.02	0.9830	-21.5205	21.0604
Mar	32.2763	10.7598	3.00	0.0032	10.9908	53.5618
Apr	26.5326	10.7576	2.47	0.0149	5.25147	47.8138
May	28.6223	10.7557	2.66	0.0088	7.34501	49.8996
Jun	65.7953	10.754	6.12	<1e-07	44.5214	87.0692
Jul	102.802	10.7525	9.56	<1e-16	81.5305	124.073
Aug	99.8913	10.7514	9.29	<1e-15	78.6225	121.16
Sep	48.5643	10.7505	4.52	<1e-04	27.2974	69.8313
Oct	10.0707	10.7498	0.94	0.3506	-11.195	31.3363
Nov	-26.3397	10.7494	-2.45	0.0156	-47.6046	-5.07477

■ $a_0 = \bar{y} - b \cdot \bar{t} = 280.299 - 2.66033 \cdot 72.5 \approx 87.425$. **Trend:**

$$87.425 + 2.660 \cdot t$$

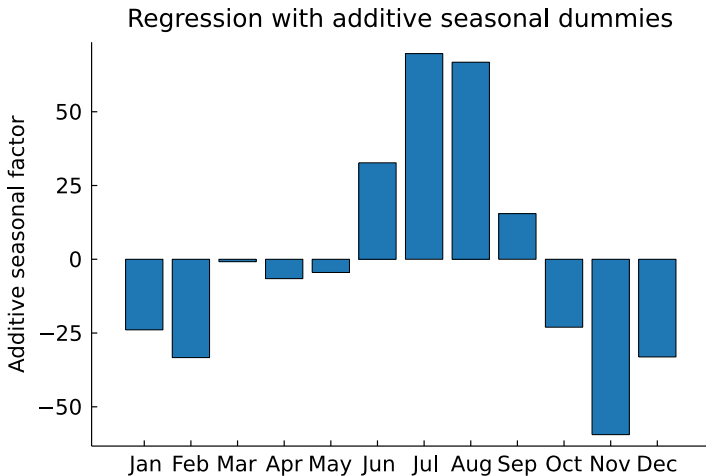
■ **Säsong januari:**

$$S_1 = a + c_1 - a_0 = 54.3277 + 9.18029 - 87.425 \approx -23.917$$

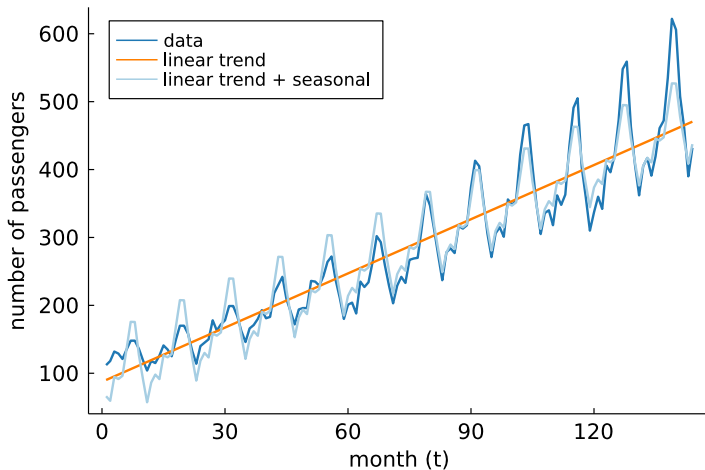
■ **Säsong december:**

$$S_{12} = a - a_0 = 54.3277 - 87.425 \approx 33.097$$

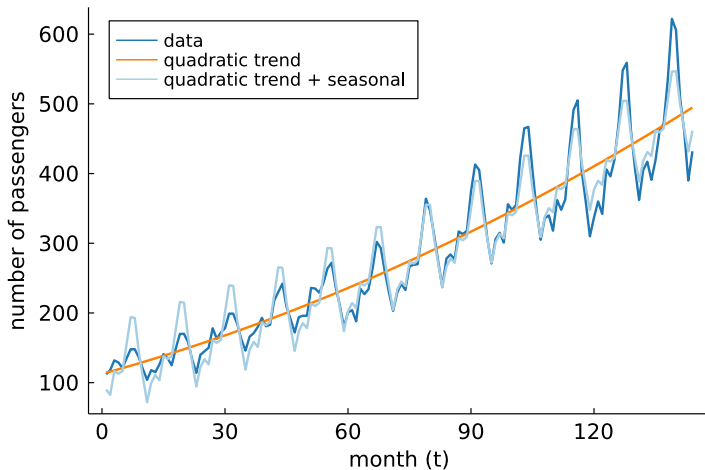
Airline - regression med säsongsdummies



Airline - regression med säsongsdummies



Airline - regression med säsongsdummies



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 .