# Regressions- och tidsserieanalys

## Föreläsning 9 - Säsongsrensning med regression

**Mattias Villani**

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet



mattiasvillani.com


@matvil


mattiasvillani

# Översikt

- Repetition kategoriska förklarande variabler
- **Säsongsrensning med regression**
- Saknade förklarande variabler i regression
- **Interaktioner** i regression

# Binära förklarande variabler

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

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

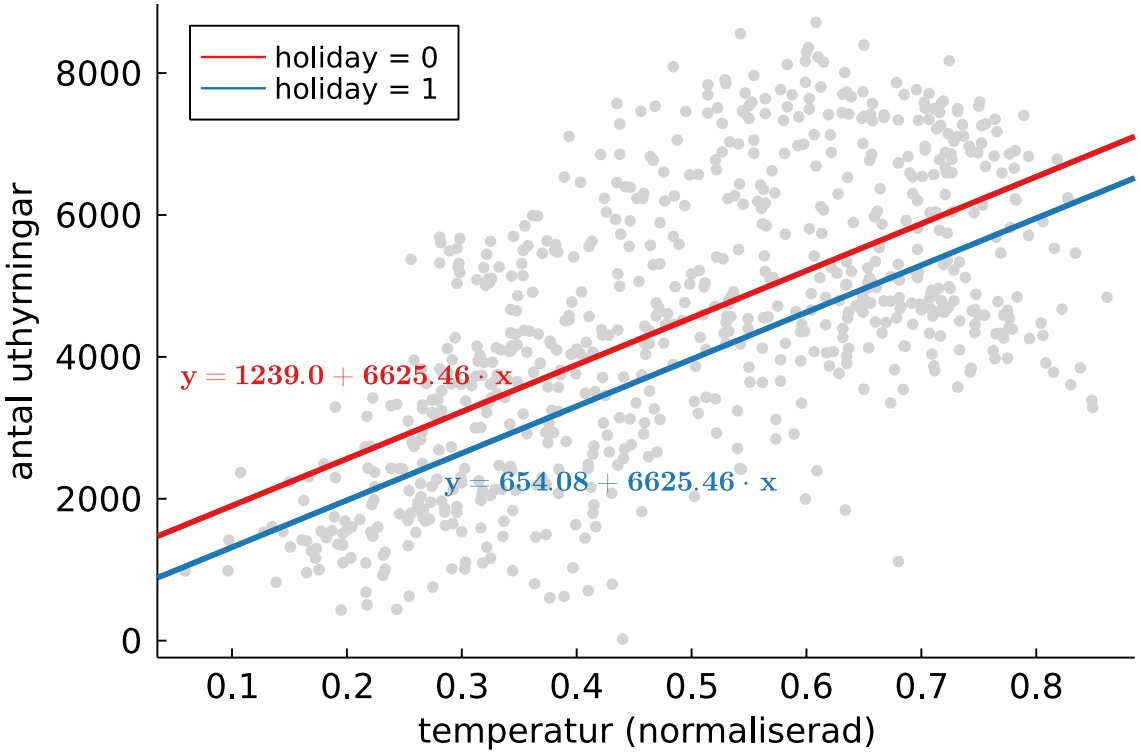
- Varianter av **kodning**: (0,1) eller $(-1,1)$, eller (`true`,`false`).
- Regressionsmodell med binär förklarande variabel:

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

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 |

# Säsongsrensning med regression - kvartalsdata

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

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

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

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

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

# Säsongsrensning med regression - månadsdata

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

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

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

- Multiplikativ modell: logga $y_t$.

# Airline passenger data - säsongsdummies

| Year | nPass | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | D |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 |

# Säsongsrensning 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**:

$$87.425 + 2.660 \cdot t$$

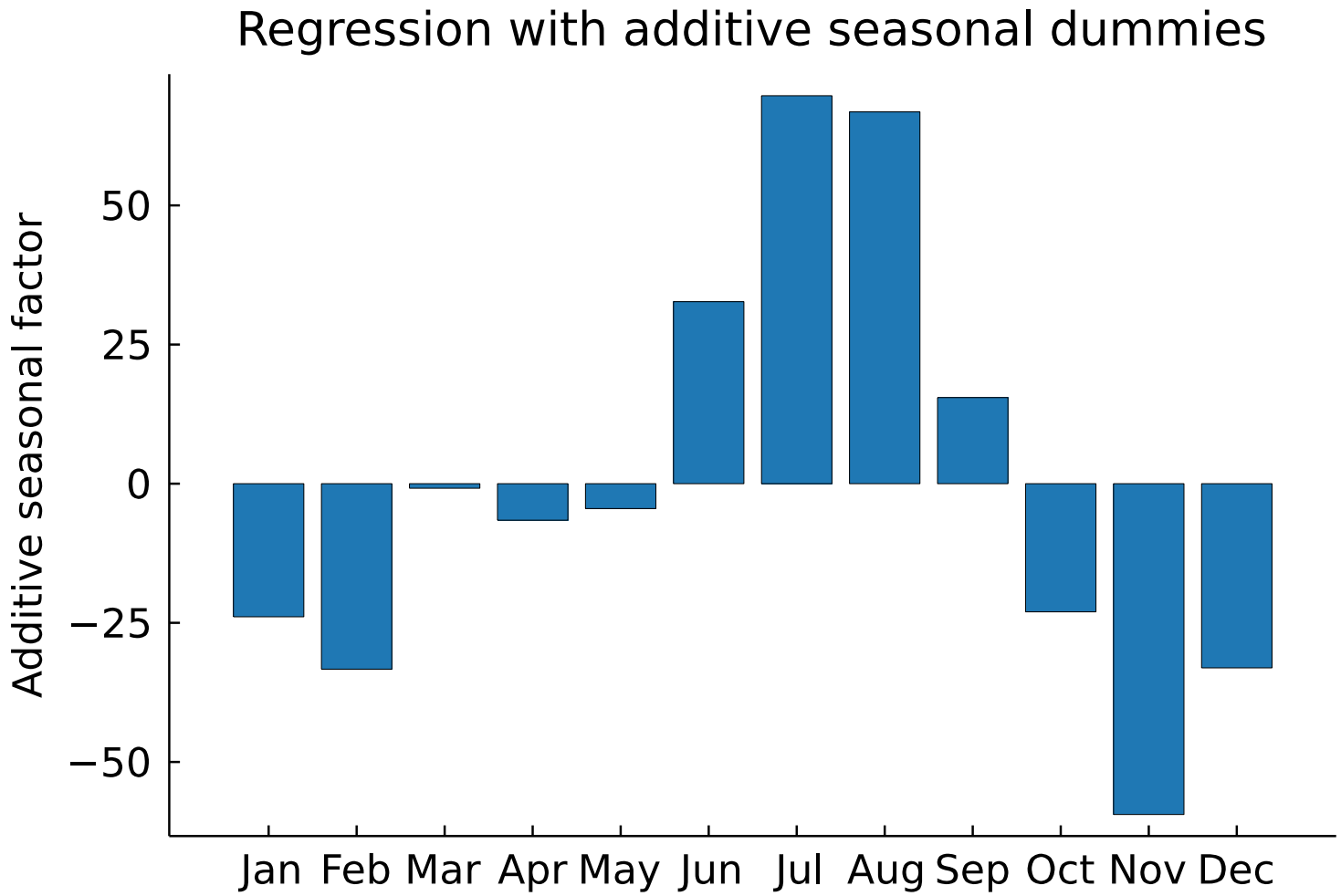
- **Säsong januari**:

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

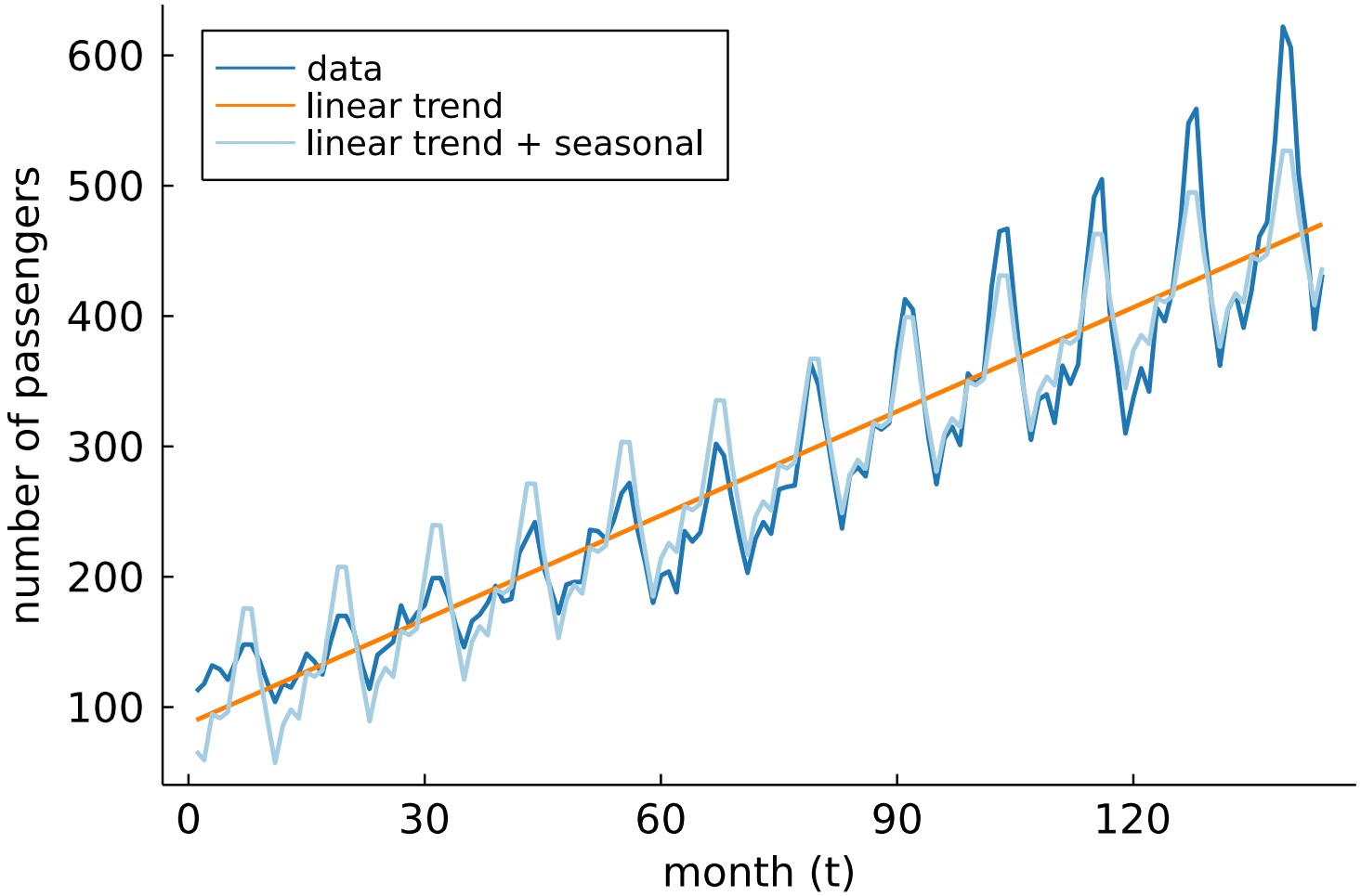
- **Säsong december**:

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

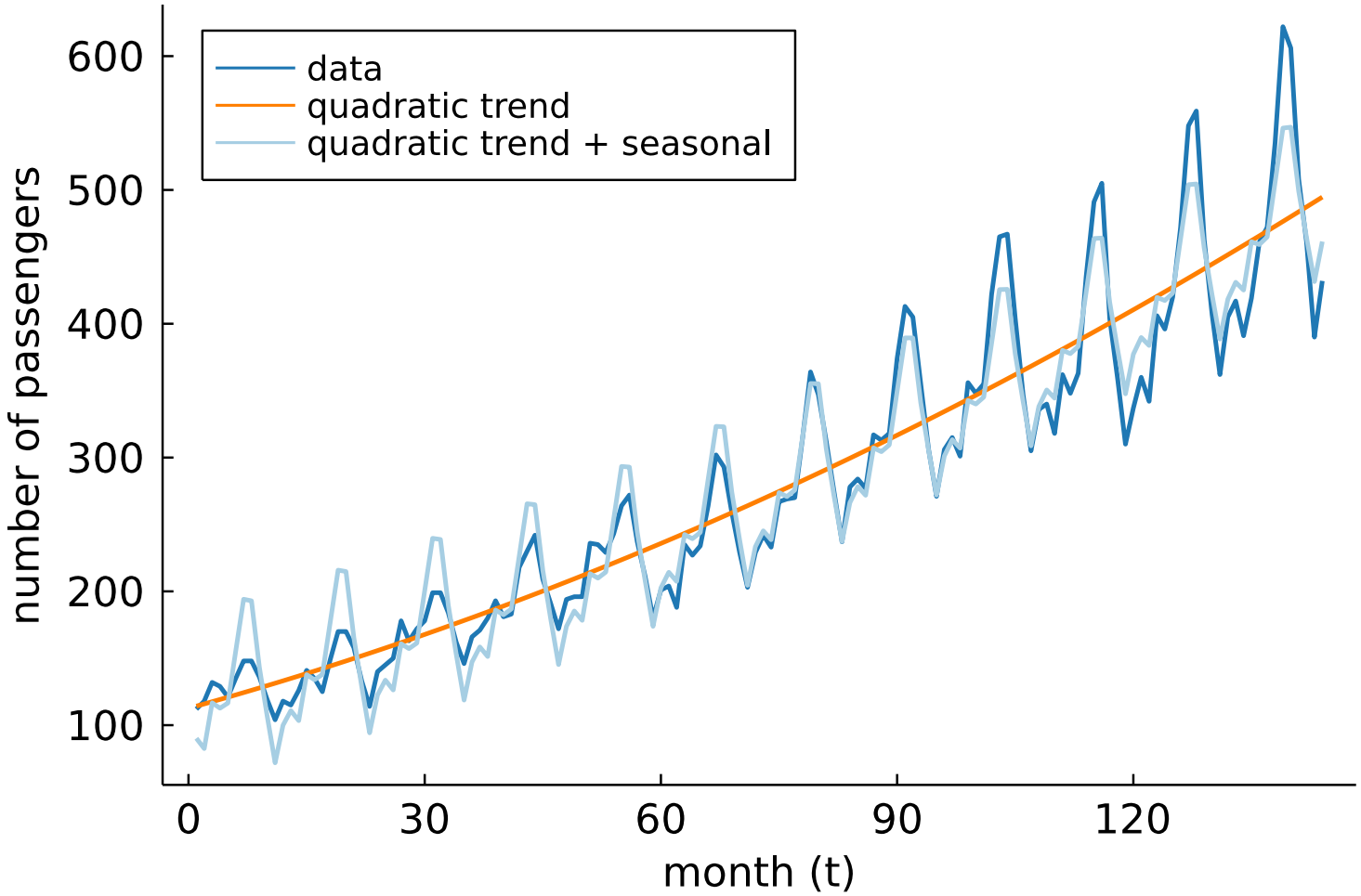
# Airline - regression med säsongsdummies



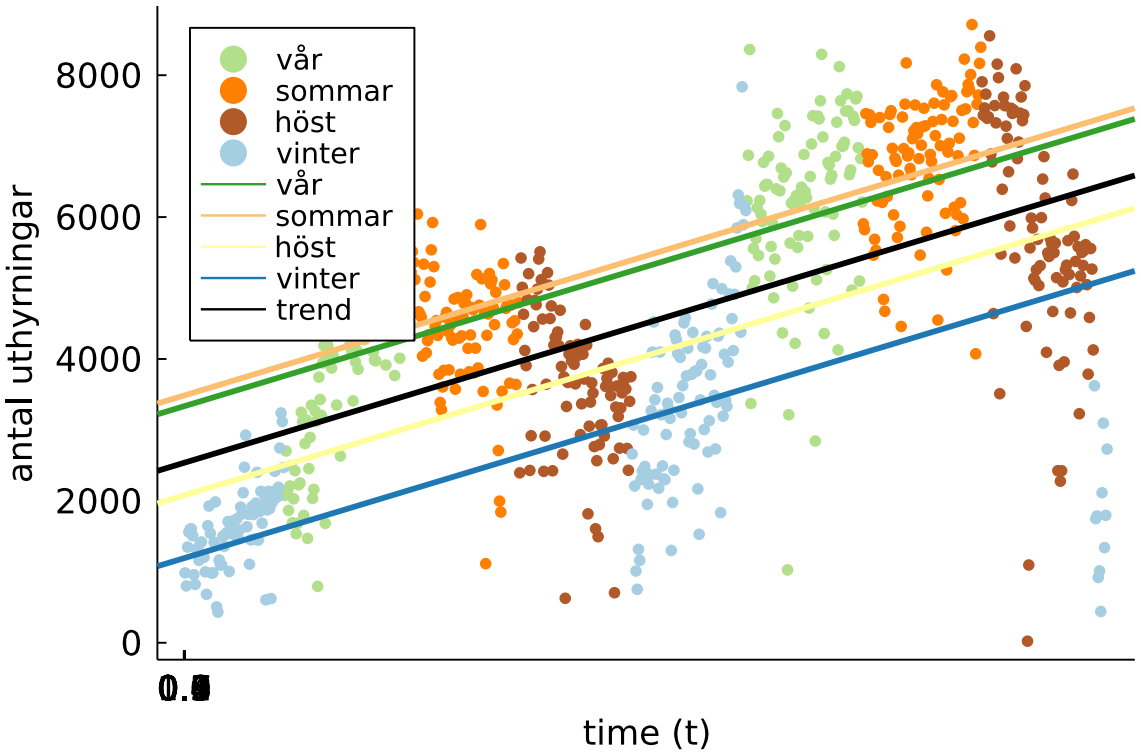
# Airline - regression med säsongsdummies



# Airline - regression med säsongsdummies



# Cykeluthyrning säsongsdummies




```
nRides ~ 1 + time + spring + summer + fall

Coefficients:
```

| | Coef. | Std. Error | t | Pr(>\|t\|) | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| (Intercept) | 1190.84 | 106.096 | 11.22 | <1e-26 | 982.543 | 1399.13 |
| time | 5.3802 | 0.227421 | 23.66 | <1e-91 | 4.93372 | 5.82668 |
| spring | 2141.71 | 123.928 | 17.28 | <1e-55 | 1898.4 | 2385.01 |
| summer | 2293.32 | 126.828 | 18.08 | <1e-59 | 2044.33 | 2542.31 |
| fall | 884.891 | 135.086 | 6.55 | <1e-09 | 619.686 | 1150.1 |

# 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 $x_2$

$$y = a + b_1 x_1$$

- 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 $x_2$.

# Interaktioner

- **Additiv modell**:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

- Modell med **interaktion**:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2 + \varepsilon$$

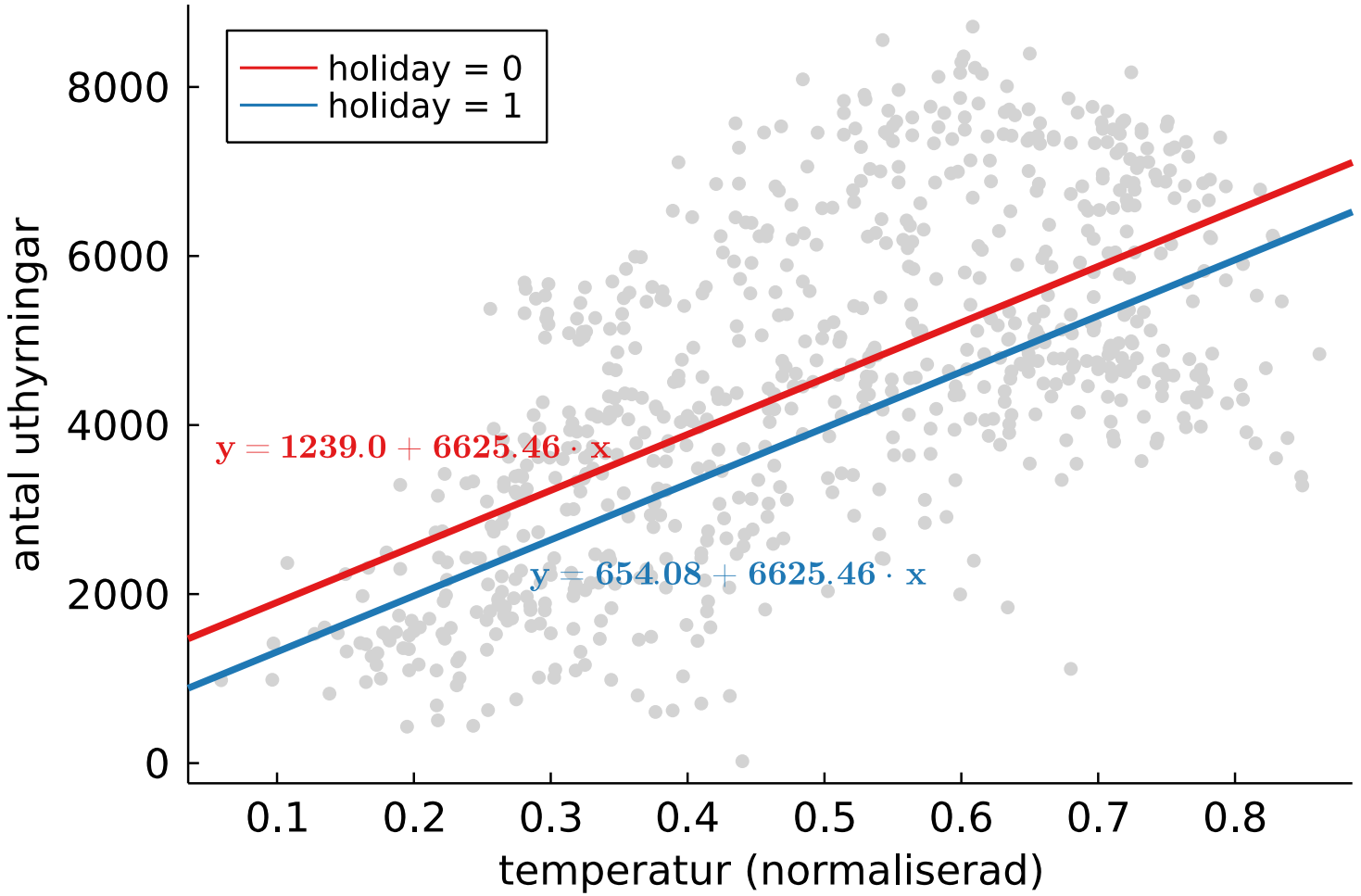
- Modell med **interaktion med dummy variabel** $(d)$

$$y = \alpha + \beta_1 x + \beta_2 d + \beta_3 x \cdot d + \varepsilon$$

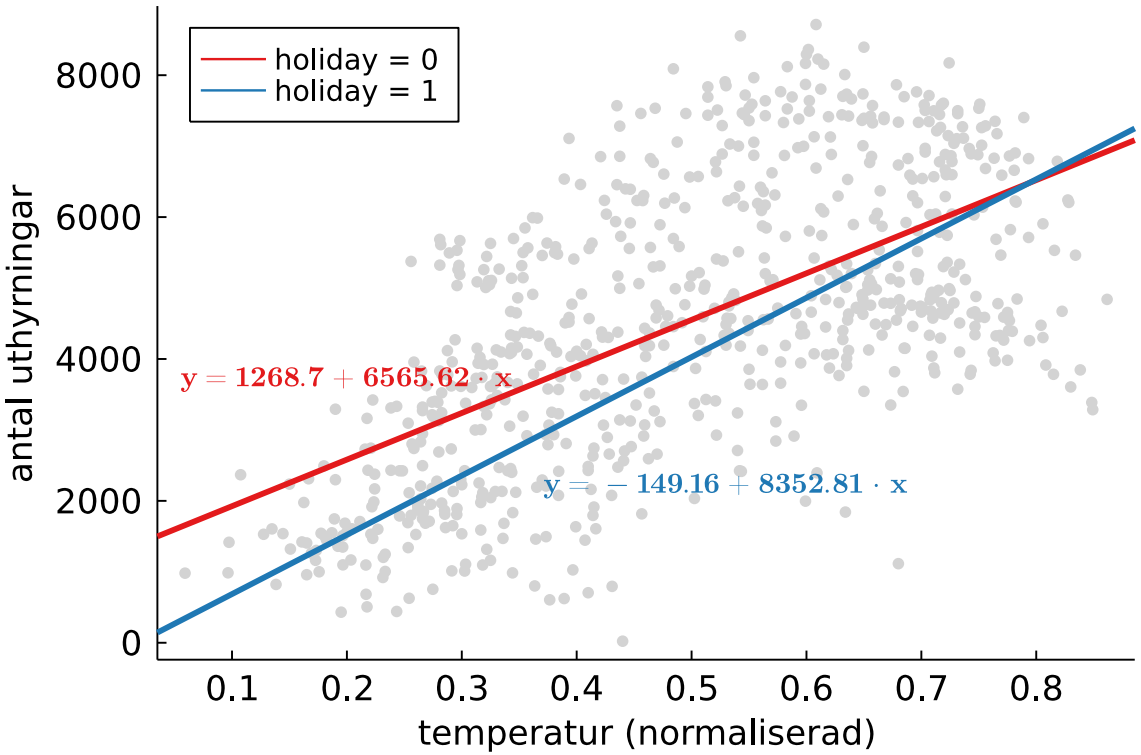
- Regressionslinjerna är inte längre parallella:

$$\begin{aligned} d = 0: &\quad y = \alpha + \beta_1 x + \varepsilon \\ d = 1: &\quad y = (\alpha + \beta_2) + (\beta_1 + \beta_3) \cdot x + \varepsilon \end{aligned}$$

# cykelhyrningar - dummy utan interaktion



# cykelhyrningar - dummy med interaktion




```
nRides ~ 1 + temp + holiday + temp & holiday

Coefficients:
```

| | Coef. | Std. Error | t | Pr(>\|t\|) | Lower 95% | Upper 95% |
|---|---|---|---|---|---|---|
| (Intercept) | 1268.7 | 163.959 | 7.74 | <1e-13 | 946.809 | 1590.59 |
| temp | 6565.62 | 310.092 | 21.17 | <1e-77 | 5956.84 | 7174.4 |
| holiday | -1417.85 | 857.484 | -1.65 | 0.0987 | -3101.29 | 265.585 |
| temp & holiday | 1787.19 | 1694.69 | 1.05 | 0.2920 | -1539.88 | 5114.26 |