Linjär regression

Bengt Carlsson

Regressionsmodeller

Modellstruktur:

$$\hat{y}(t) = \varphi_1(t)\theta_1 + \varphi_2(t)\theta_2 + \ldots + \varphi_n(t)\theta_n = \varphi^T(t)\theta$$

där $\hat{y}(t)$ är utsignalen från modellen, $\varphi(t)$ är en n-dimensionell vektor av *kända* storheter; "regressor", och θ är en n-dimensionell vektor av *okända* parametrar. T betecknar transponat. $t = 1,\ 2,\ 3\ldots$ är heltal.

Problemet är att givet mätningar/observationer (betecknas $y(t)$) hitta ett "vettigt" värde på θ.

Exempel på regressionsmodeller $\hat{y} = \varphi^T(t)\theta$

- Polynomtrend:

$$\hat{y}(t) = a_o + a_1 t + \ldots + a_n t^n$$

kan skrivas som $\hat{y}(t) = \varphi(t)^T\theta$ med

$$\begin{aligned} \varphi(t) &= (1 \;\; t \ldots t^n)^T \\ \theta &= (a_o \;\; a_1 \ldots a_n)^T \end{aligned}$$

- FIR-modell (Finite Impulse Response)

$$\begin{aligned} \hat{y}(t) &= h_o u(t) + h_1 u(t-1) + \ldots + h_n u(t-n) \rightarrow \\ \varphi(t) &= (u(t) \;\; u(t-1) \ldots u(t-n))^T \\ \theta &= (h_o \;\; h_1 \ldots h_n)^T \end{aligned}$$

OBS, även flera insignaler går bra.

- Summa av exp funktioner:

$$\hat{y}(t) = b_1 e^{-k_1 t} + b_2 e^{-k_2 t} + \ldots + b_n e^{-k_n t}$$

Antag $k_1, \;\; k_2 \ldots k_n$ kända (annars olinjär regression, se senare avsnitt)

$$\begin{aligned} \varphi(t) &= (e^{-k_1 t} \;\; e^{-k_2 t} \ldots e^{-k_n t})^T \\ \theta &= (b_1 \;\; b_2 \ldots b_n)^T \end{aligned}$$

- Gaslagen

$$
\begin{aligned}
pV^{\gamma} &= C \\
p &= V^{-\gamma} C \\
\log p &= -\gamma \log V + logC
\end{aligned}
$$

Låt $\hat{y} = \log p$ (p antas mätbar) samt

$$
\begin{aligned}
\varphi(t) &= (-\log V \quad 1)^T \\
\theta &= (\gamma \quad \log C)^T
\end{aligned}
$$

- ARX-modellen (AutoRegressive model with an eXternal input)

$$
\begin{aligned}
\hat{y}(t) &= -a_1 y(t-1) - a_2 y(t-2) - \ldots - a_{na} y(t-na) \\
&\quad + b_o u(t) + b_1 u(t-1) + \ldots + b_{nb} u(t-nb)
\end{aligned}
$$

$\Rightarrow$

$$
\begin{aligned}
\varphi(t) &= (-y(t-1) \quad -y(t-2) \ldots - y(t-na) \\
&\qquad u(t) \quad u(t-1) \ldots u(t-nb))^T \\
\theta &= (a_1 \;\; a_2 \ldots a_{na} \;\; b_o \;\; b_1 \ldots b_{nb})^T
\end{aligned}
$$

Minstakvadratmetoden

Givet mätdata $\{y(t), \varphi(t)\}_{t=1,\ldots,N}$ bilda förlustfunktionen:

$$V(\theta) = \sum_{t=1}^{N} (y(t) - \hat{y}(t))^2 = \sum_{t=1}^{N} (y(t) - \varphi^T(t)\theta)^2$$

Differensen $\epsilon(t) = y(t) - \hat{y}(t)$ kallas prediktionsfel.

Antag att matrisen $\sum_{t=1}^{N} \varphi(t)\varphi^T(t)$ kan inverteras (= pos.def). Det θ som minimierar $V(\theta)$ ges av:

$$\hat{\theta} = [\sum_{t=1}^{N} \varphi(t)\varphi^T(t)]^{-1} \sum_{t=1}^{N} \varphi(t)y(t))$$

Minstakvadratmetoden matrisformulering

Definiera

$$Y = \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} \quad \Phi = \begin{bmatrix} \varphi^T(1) \\ \vdots \\ \varphi^T(N) \end{bmatrix}$$

Då kan minstakvadratskattningen skrivas

$$\hat{\theta} = [\Phi^T \Phi]^{-1} \Phi^T Y$$

Matlab:

```
>> theta_hat=Phi\Y
```

Analys -vitt brus

Antag att data genererats av ("det sanna systemet"):

$$y(t) = \varphi^T(t)\theta_o + e(t) \quad t = 1, \ldots, N$$

som kan skrivas på matrisform som

$$Y = \Phi\theta_o + e$$

där $e = [e(1);\ \ldots e(N)]^T$.

Antag att störningen ("mätbruset") $e(t)$ är *vitt brus* med medelvärde noll och varians λ.

Antag vidare att regressionsvektorn φ är *okorrelerad* med bruset $e(t)$ dvs $E\{\varphi(t)e(s)\} = 0$ för alla t and s. Detta gäller inte för ARX modellen!

Då gäller:

- Minstakvadratskattningen $\hat{\theta}$ är en väntevärdesriktig (unbiased) skattning av θ_o dvs $E\hat{\theta} = \theta_o$.
- Kovariansmatrisen för $\hat{\theta}$ ges av

$$P = E(\hat{\theta} - \theta_o)(\hat{\theta} - \theta_o)^T = \mathrm{cov}\hat{\theta} = \lambda(\Phi^T \Phi)^{-1}$$

- Brusvariansen λ kan skattas väntevärdesriktigt med $\hat{\lambda} = \frac{1}{N-n} V(\hat{\theta})$

Notera att vi antagit att det sanna systemet har samma *struktur* som vår modell!

Viktig användning av kovariansmatrisen

$$\mathrm{var}(\hat{\theta}_i) = P_{i,i} \;\; i = 1, \; ... \; n \tag{1}$$

där $P_{i,i}$ är det i'te diagonalelementet i P.
Kovariansmatrisen P kan skatta från data med

$$\hat{P} = \hat{\lambda}(\Phi^T \Phi)^{-1}$$

Analys - färgat brus, Se räkneövn L1

Antag att data genererats av:

$$Y \quad = \quad \Phi\theta_o + e$$

där $e = [e(1); \ \dots e(N))^T$ är färgat ("korrelerat") så att $Eee^T = R$. (Om vitt brus så är $R = \lambda I$, där I=enhetsmatrisen)

Då gäller:

- $E\hat{\theta} = \theta_o$ fortfarande!.
- Kovariansmatrisen för $\hat{\theta}$ ges av:

$$P = \text{cov}\hat{\theta} = (\Phi^T\Phi)^{-1}\Phi^T R\Phi(\Phi^T\Phi)^{-1}$$

- Om R är känd finns nogrannare skattning av θ (BLUE):

$$\hat{\theta} = (\Phi^T R^{-1}\Phi)^{-1}\Phi^T R^{-1}Y$$

  med kovariansmatris

$$\text{cov}\hat{\theta}_{BLUE} = (\Phi^T R^{-1}\Phi)^{-1}$$

Val av modell

1. Val av modellstruktur:
   – Fysikalisk insikt
   – Pröva olika
2. Val av modellordning n: Utvärdera förlustfunktionen. Försök hitta “knä”. Statistiska tester finns.

Mer om modellvalidering senare i kursen!

Utvärdera modell

Vanligast:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{2}$$

where

$$SS_{res} = \sum_{t=1}^{N} (y(t) - \hat{y}(t))^2 \tag{3}$$

$$SS_{tot} = \sum_{t=1}^{N} (y(t) - \bar{y})^2 \tag{4}$$

with $\bar{y} = \frac{1}{N}\sum_{t=1}^{N} y(t)$.

R^2 kallas FIT i den toolbox ni ska använda

Olinjära regressionsmodeller

Modellstruktur

$$\hat{y}(t) = g(\varphi(t), \theta)$$

där g är någon olinjär funktion (t ex exponentialfunktion). Gör pss, dvs bilda förlustfunktionen

$$V(\theta) = \sum_{t=1}^{N} [y(t) - g(\varphi(t), \theta)]^2$$

Vi väljer det värde på θ som minimerar $V(\theta)$:

$$\hat{\theta} = \arg\min_{\theta} V(\theta)$$

Notera:

- I allmänhet finns ingen analystisk lösning utan numerisk lösning (iterativ) måste användas.
- Risk för lokala minima.
- I Matlab: *fminsearch*.