Regressions- och tidsserieanalys

Föreläsning 7 - Icke-linjär regression. Polynom- och exponentiella samband

Mattias Villani

Statistiska institutionen Stockholms universitet

Institutionen för datavetenskap Linköpings universitet









Översikt

- Utvärdera och välja modeller baserat på prognosförmåga
- Polynomregression
- Exponentiella modeller

Prognosförmåga på testdata

- Dela upp observationer i två delmängder:
 - ► Träningsdata för att skatta modellens parametrar.
 - ► Testdata för att utvärdera modellens prediktioner.
- Modellen får aldrig chans att anpassa sig till testdata.
- Prediktionsmått: kvadrerade prediktionsfel på testdata

$$SSE_{test} = \sum_{j=1}^{n_{test}} (y_j - \hat{y}_j)^2$$

- Observera:
 - summan är över observationerna i testdata.
 - ightharpoonup modellen som ger \hat{y}_j är skattad enbart på träningsdata.
 - veranpassning på träningsdata ⇒ dåliga prediktioner på testdata.



Korsvalidering

- Vilka observationer ska vara i träning respektive test?
 - ▶ Tidsserier: låt de senare observationerna vara i test.
 - ▶ Regression: Korsvalidering. Dela upp data i K st folds:



Mattias Villani

ST123G

Korsvalidering



Prognosfel i testdata

$$MSE_{CV} = \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j^*)^2}{n}$$

Lättare att tolka Root MSE

$$RMSE_{CV} = \sqrt{MSE_{CV}}$$

- Prognosförmåga på testdata kan användas för modellval.
- För mer info: masterkursen Maskininlärning 7.5 hp.

Mattias Villani

Kvadratisk regression

Kvadratisk regression

$$y = a + b_1 x + b_2 x^2$$

- är multipel regression med två förklarande variabler:
 - $\rightarrow x_1 = x$
 - $x_2 = x^2$
- Populationsmodell:

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

- **Minsta-kvadratmetoden** för att beräkna a, b_1 och b_2 !
- Kvadratisk regression icke-linjär i x, men linjär i α , β_1 och β_2 .





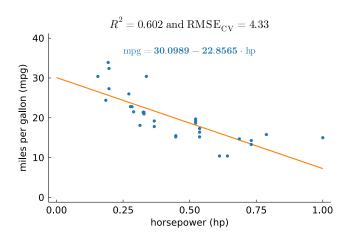
Kvadratisk regression - excel

	A	В	С	D
1		mpg (y)	hp (x)	X ²
2	Mazda RX4	21.000	0.328	0.108
3	Mazda RX4 Wag	21.000	0.328	0.108
1	Datsun 710	22.800	0.278	0.077
5	Hornet 4 Drive	21.400	0.328	0.108
5	Hornet Sportabo	18.700	0.522	0.273
7	Valiant	18.100	0.313	0.098
3	Duster 360	14.300	0.731	0.535
)	Merc 240D	24.400	0.185	0.034
0	Merc 230	22.800	0.284	0.080
1	Merc 280	19.200	0.367	0.135
2	Merc 280C	17.800	0.367	0.135
3	Merc 450SE	16.400	0.537	0.289
4	Merc 450SL	17.300	0.537	0.289
5	Merc 450SLC	15.200	0.537	0.289
6	Cadillac Fleetwo	10.400	0.612	0.374
7	Lincoln Continer	10.400	0.642	0.412
В	Chrysler Imperia	14.700	0.687	0.471
9	Fiat 128	32.400	0.197	0.039
0	Honda Civic	30.400	0.155	0.024
1	Toyota Corolla	33.900	0.194	0.038
2	Toyota Corona	21.500	0.290	0.084
3	Dodge Challeng	15.500	0.448	0.200
4	AMC Javelin	15.200	0.448	0.200
5	Camaro Z28	13.300	0.731	0.535
6	Pontiac Firebird	19.200	0.522	0.273
7	Fiat X1-9	27.300	0.197	0.039
8	Porsche 914-2	26.000	0.272	0.074
9	Lotus Europa	30.400	0.337	0.114
0	Ford Pantera L	15.800	0.788	0.621
1	Ferrari Dino	19.700	0.522	0.273
12	Maserati Bora	15.000	1.000	1.000
3	Volvo 142F	21,400	0.325	0.106

Notera att hp är normaliserad genom att dividera med max(hp) i stickprovet.

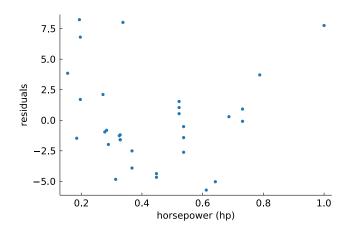
Blir numeriskt stabilare om man normaliserar så.

mtcars data - linjär regression mot hp



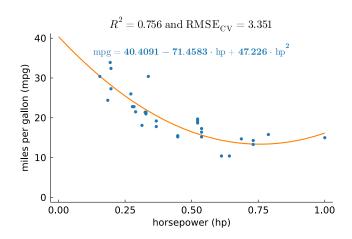
Mattias Villani

mtcars data - residualer linjär regression



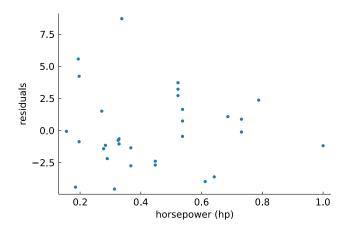
Mattias Villani

mtcars data - kvadratisk regression mot hp



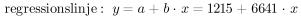
Mattias Villani

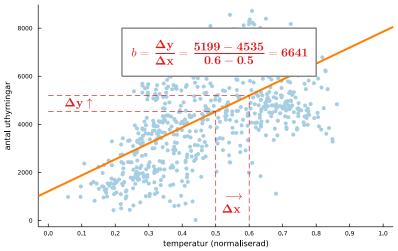
mtcars data - residualer kvadratisk regression



Mattias Villani

Linjär regression - tolkning b





Mattias Villani

Tolkningar av parametrar i kvadratisk regression

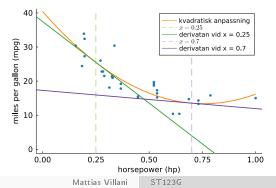
Kvadratisk regression

$$y = a + b_1 x + b_2 x^2$$

Regressionskoefficienterna tolkas som derivator:

$$\frac{dy}{dx} = b_1 + 2b_2 \cdot x$$

Effekten av en liten förändring Δx i x beror på x själv:



Polynomregression

Polynomregression

$$y = a + b_1 x + b_2 x^2 + \ldots + b_k x^k$$

- Polynomregression av ordning k är detsamma som multipel regression med k förklarande variabler:
 - $\rightarrow x_1 = x$
 - $x_2 = x^2$

 - $\rightarrow x_k = x^k$
- Populationsmodell:

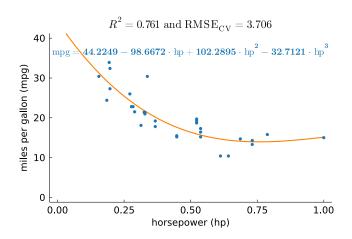
$$y = \alpha + \beta_1 x + \beta_2 x^2 + \ldots + \beta_k x^k + \varepsilon$$

- Minsta-kvadratmetoden kan användas för att beräkna $a, b_1, \ldots b_k!$
- Polynomregression är icke-linjär i x, men linjär i α , $\beta_1, \ldots \beta_k$.

Polynomregression - excel

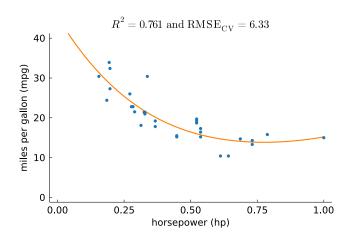
	А	В	С	D	E	F
1		mpg (y)	hp (x)	X ²	X ³	X ⁴
2	Mazda RX4	21.000	0.328	0.108	0.035	0.012
3	Mazda RX4 Wag	21.000	0.328	0.108	0.035	0.012
4	Datsun 710	22.800	0.278	0.077	0.021	0.006
5	Hornet 4 Drive	21.400	0.328	0.108	0.035	0.012
6	Hornet Sportabo	18.700	0.522	0.273	0.143	0.074
7	Valiant	18.100	0.313	0.098	0.031	0.010
8	Duster 360	14.300	0.731	0.535	0.391	0.286
9	Merc 240D	24.400	0.185	0.034	0.006	0.001
10	Merc 230	22.800	0.284	0.080	0.023	0.006
11	Merc 280	19.200	0.367	0.135	0.049	0.018
12	Merc 280C	17.800	0.367	0.135	0.049	0.018
13	Merc 450SE	16.400	0.537	0.289	0.155	0.083
14	Merc 450SL	17.300	0.537	0.289	0.155	0.083
15	Merc 450SLC	15.200	0.537	0.289	0.155	0.083
16	Cadillac Fleetwo	10.400	0.612	0.374	0.229	0.140
17	Lincoln Continer	10.400	0.642	0.412	0.264	0.170
18	Chrysler Imperia	14.700	0.687	0.471	0.324	0.222
19	Fiat 128	32.400	0.197	0.039	0.008	0.002
		Mattias	Villani ST12	3G		

mtcars data - kubisk regression mot hp

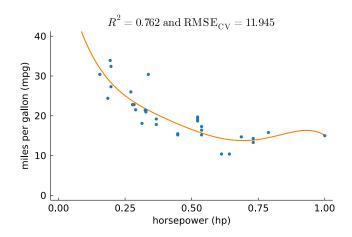


Mattias Villani

mtcars data - polynomregression ordning 4

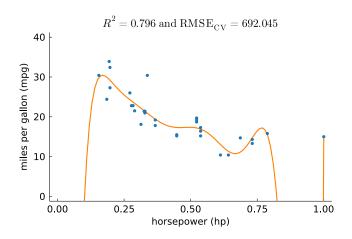


mtcars data - polynomregression ordning 5



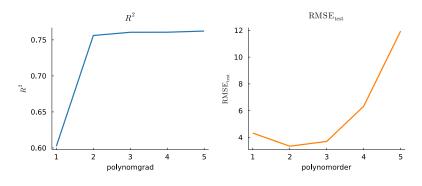
Mattias Villani

mtcars data - polynomregression ordning 10



Mattias Villani

mtcars data - R^2 och RMSE-CV(K = 4)

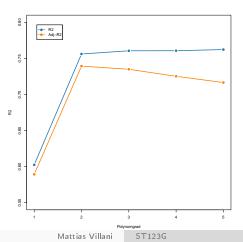


Mattias Villani

mtcars data - R^2 och R^2_{adjusted}

Justerad R^2

$$R_{\text{adjusted}}^2 = 1 - \frac{s_e^2}{s_v^2} = 1 - \frac{\text{MSE}}{\text{MST}}$$



L2-regularisering (Ridge regression)

- **För många x-variabler** \Rightarrow MK-metoden överanpassar data.
- Modellen är överparametriserad.
- Variabelselektion (t ex forward selection) är en lösning.
- L2-regularisering minimerar en straffad SSE:

$$Q_{\lambda} = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - \ldots - b_k x_{ki})^2 + \lambda \cdot \sum_{j=1}^{k} b_j^2$$

- Stort λ kommer **krympa** estimaten av b_j mot noll.
- Skattningen är nu biased, men har lägre varians.
- **■** Bias-Variance trade-off

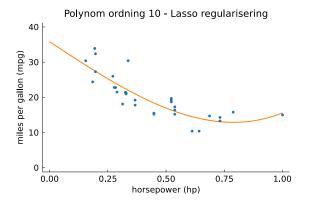
L1-regularisering (Lasso regression)

■ L1-regularisering (Lasso) straffar med absolutbelopp:

$$Q_{-} = \sum_{i=1}^{n} (y_{i} - a - b_{1}x_{1i} - \ldots - b_{k}x_{ki})^{2} + \lambda \cdot \sum_{j=1}^{k} |b_{j}|$$

- Lasso har två effekter:
 - krymper b_i mot noll (shrinkage)
 - kan sätta vissa b_j exakt till noll (selection)
- glmnet paketet i R gör både L1 och L2 regularisering och mer.
- Lasso är extremt populär. Go-to när man har väldigt många förklarande variabler.

Lasso regularisering polynom ordning 10



$$Q_{\lambda} = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - \ldots - b_k x_{ki})^2 + \lambda \cdot \sum_{j=1}^{k} |b_j|$$

- $a = 35.80, b_1 = -43.58, b_3 = 23.32.$
- $b_2 = 0$ och $b_4 = \ldots = b_{10} = 0$ (variabelselektion).

Exponentiella samband

Du sätter in 200 kr på banken till 5% årsränta. Utveckling:

1 år:
$$200 \cdot 1.05 = 210.000 \text{ kr}$$

2 år: $200 \cdot 1.05^2 = 220.500 \text{ kr}$
3 år: $200 \cdot 1.05^3 = 231.525 \text{ kr}$

- $3 \text{ ar: } 200 \cdot 1.05^{\circ} = 231.525 \text{ kr}$
- Efter x år: $200 \cdot 1.05^x$. Exponentiell tillväxt. Samma procentuella ökning varje år.
- Exponentiellt samband

$$y = a \cdot b^{x}$$

- a är det initiala beloppet eller storheten.
- b bestämmer tillväxttakten

$$b>1$$
 ökande

$$b < 1$$
 minskade

$$b = 1$$
 konstant (nolltillväxt)

Exponentiell regression

Exponentiell regression

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

■ Logaritmregler (10-logaritmer log = log₁₀)

$$\log(a \cdot b) = \log a + \log b$$
 ("log av produkten är summan av log")
 $\log b^x = x \log b$ ("exponenten hoppar ner framför")

Logaritmera båda sidor

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

$$\tilde{y} = \tilde{a} + \tilde{b}x$$

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

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

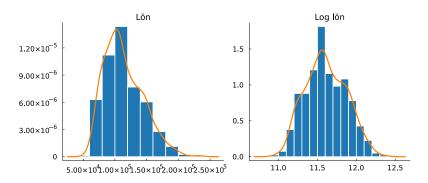
- Skatta \tilde{a} och \tilde{b} med minsta-kvadrat med $\tilde{y} = \log y!$
- Skattningar för a och b fås genom anti-logaritmering

$$a = 10^{\tilde{a}}$$
 och $b = 10^{\tilde{b}}$

Mattias Villani ST123G

Exponentiell regression

- Responsvariabler med enbart positiva värden (t ex lön):
 - Normalfördelning ofta opassande pga skevhet.
 - ► kan ge prediktioner för y som är negativa.



Exponentiell regression

Populationsmodell:

$$y = \alpha \cdot \beta^{x} \varepsilon$$

- Logaritmen av feltermen ε är normalfördelad.
- lacksquare Vi säger att feltermen arepsilon är lognormal fördelad. Innebär arepsilon>0.
- Logaritmera för att göra modellen linjär!

$$\underbrace{\log y}_{\tilde{y}} = \underbrace{\log \alpha}_{\tilde{\alpha}} + \underbrace{\log \beta}_{\tilde{\beta}} \cdot x + \underbrace{\log \varepsilon}_{\tilde{\varepsilon}}$$

$$\tilde{y} = \tilde{\alpha} + \tilde{\beta} \cdot x + \tilde{\epsilon}, \qquad \tilde{\epsilon} \sim N\left(0, \sigma_{\tilde{\epsilon}}^2\right).$$

- t-test för $H_0: \tilde{\beta} = 0$ är test för konstant tillväxt $\beta = 1$.
- Prediktion för $x = x_0$:

$$\hat{y} = a \cdot b^{x_0} = 10^{\tilde{a}} \cdot (10^{\tilde{b}})^{x_0} = 10^{\tilde{a} + \tilde{b}x_0}$$

Dvs, gör prediktion $\widehat{\log y}$ och "anti-logga" för prognosen för \hat{y} .

Kinesisk tillväxt

	A	В	С	D	E
1	year	gdp	gdpgrowth	log10(gdp)	t = year - 1999
2	2000	959.3725	9.86	2.981987265	1
3	2001	1053.1082	9.77	3.022472994	2
4	2002	1148.5083	9.06	3.060134138	3
5	2003	1288.6433	12.2	3.11013272	4
0	2004	1508.6681	17.07	3.178593708	5
7	2005	1753.4178	16.22	3.243885411	6
8	2006	2099.2294	19.72	3.3220599	7
9	2007	2693.9701	28.33	3.430392771	8
10	2008	3468.3046	28.74	3.540117232	9
11	2009	3832.2364	10.49	3.583452292	10
12	2010	4550.4531	18.74	3.658054643	11
13	2011	5618.1323	23.46	3.749591962	12
14	2012	6316.9183	12.44	3.80050526	13
15	2013	7050.6463	11.62	3.848228929	14
16	2014	7678.5995	8.91	3.885282016	15
17	2015	8066.9426	5.06	3.906708967	16
18	2016	8147.9377	1	3.9110477	17
19	2017	8879.4387	8.98	3.948385513	18
20	2018	9976.6771	12.36	3.998985916	19
21	2019	10216.6303	2.41	4.009307678	20
22	2020	10500.3956	2.78	4.021205661	21
23					

Mattias Villani

ST123G

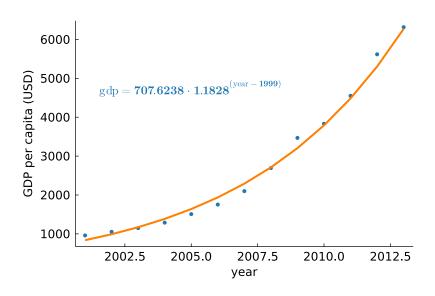
Kinesisk tillväxt 2000-2013

Coefficients:							
	Coef.	Std. Error	t	Pr(> t)	Lower 95%	Upper 95%	
(Intercept) year		0.0192341 0.00242327	148.16 30.08		2.80747 0.067567		

- $\tilde{a} = 2.8498$, så $a = 10^{\tilde{a}} = 707.62376$.
- $\tilde{b} = 0.0729005$, så $b = 10^{\tilde{b}} = 10^{0.0729005} = 1.18277$.

Mattias Villani

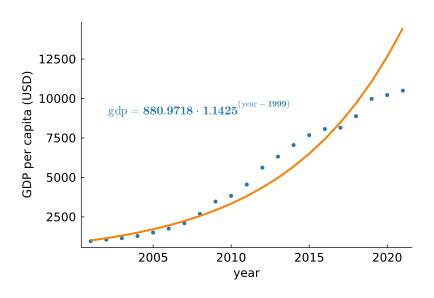
Kinesisk tillväxt 2000-2013



Mattias Villani

ST123G

Kinesisk tillväxt 2000-2021



Mattias Villani

ST123G