

Regressions- och tidsserieanalys

Föreläsning 1 - Introduktion till kursen. Motivation.

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet



mattiasvillani.com



@matvil

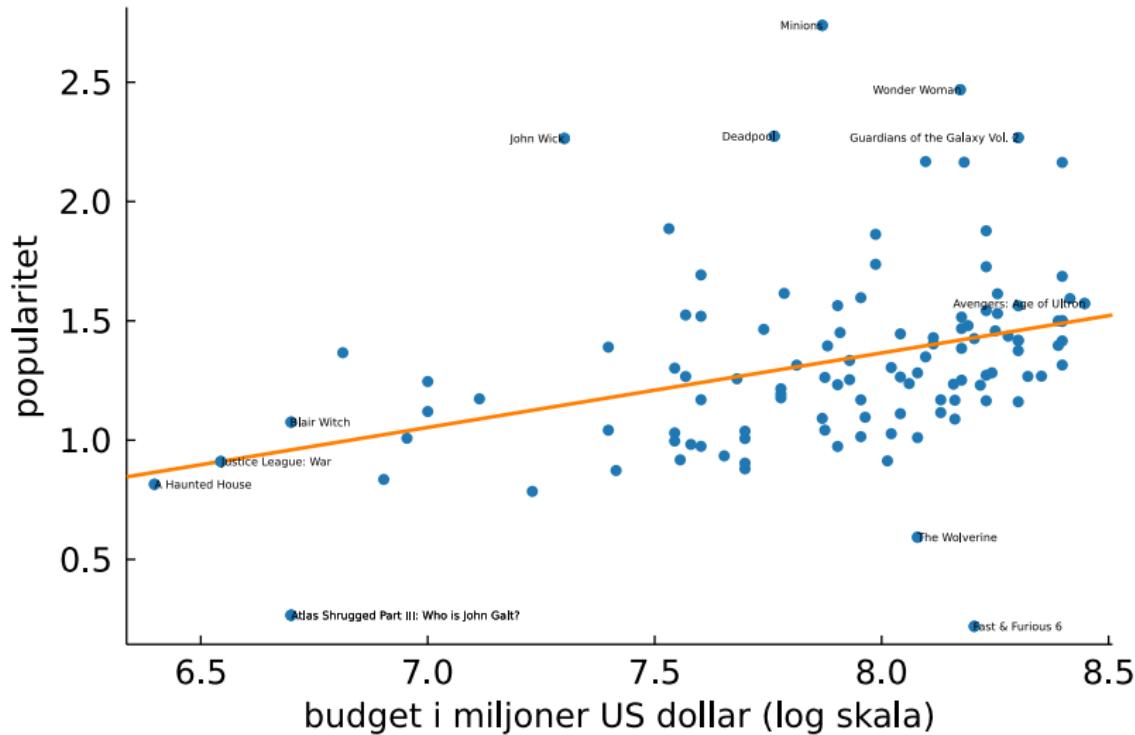


mattiasvillani

Översikt

- Kursens [kursplan](#) på Athena.
- Slides ligger på denna [webbsida](#) (länkat från Athena).
- **Kursstruktur** för Del 1 - Regressions- och tidsserieanalys:
 - ▶ Föreläsningar F1-F12 ([Mattias Villani](#))
 - ▶ Övningar Ö1-Ö6 ([Maria Anna Di Lucca](#) och Jon Lachmann)
 - ▶ Datorövningar D1-D4 ([Maria Anna Di Lucca](#) och Jon Lachmann)
- **Delar:**
 - ▶ Regressionsanalys
 - ▶ Tidsserieanalys
- **Examination**
 - ▶ Inlämningsuppgift i regressions- och tidsserieanalys, 3 hp
 - ▶ Salstentament, 4.5 hp

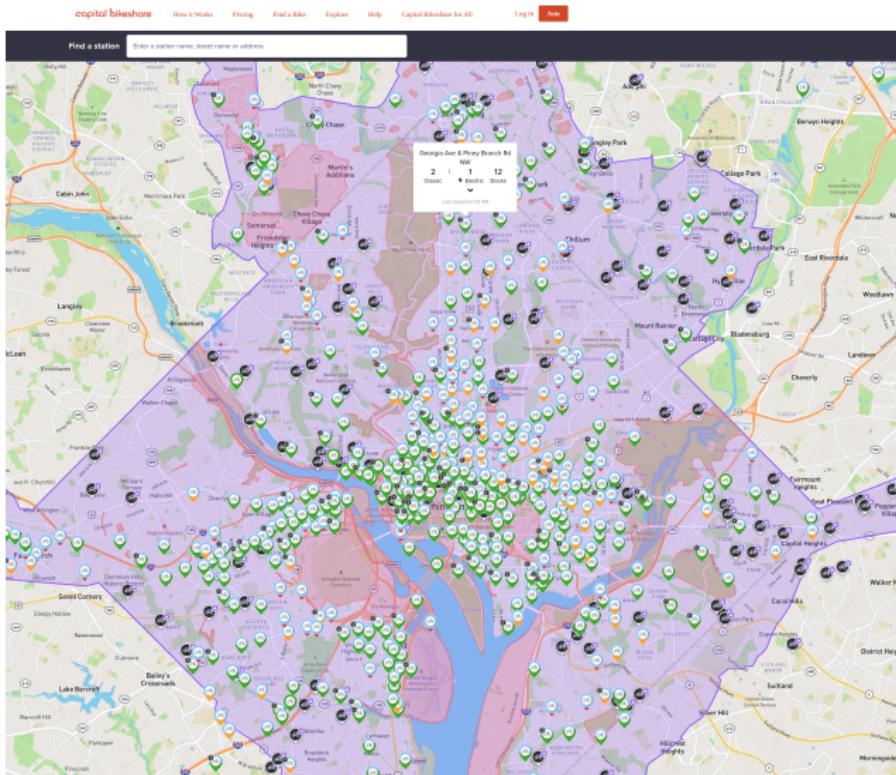
Är dyra filmer mer populära? 2013-2017



Regressionsproblem finns överallt

- Regression analyserar **samband mellan variabler**:
 - ▶ Företags försäljning (y) och deras marknadsföringsbudget (x).
 - ▶ Dos av smärtstillande (x) och upplevd smärtlindring (y).
 - ▶ Studietimmar (x) och tentaresultat (y).
 - ▶ Kvadratmeter (x) och bostadrättspris (y).
 - ▶ Ränta (x) och inflation (y).
- Regression handlar om **korrelation**. **Samvariation**.
- Samband kan utnyttjas för att göra **prediktioner**. 
- Korrelation innebär inte **kausala samband** (orsak → verkan).

Cykelpool i Washington DC



Cykelpool

- Företag som hyr ut cyklar vill
 - ▶ **förstå vilka faktorer** som påverkar användandet
 - ▶ **göra prediktioner** på kort och lång sikt över användandet
 - ▶ **fatta beslut** om hur många cyklar de ska ha i poolen.
- Data:
 - ▶ 2 års data på antalet uthyrda cyklar per dag
 - ▶ Väderinformation
 - ▶ Helger, ledigheter etc
- Originaldata inkl beskrivning finns [här](#).
- Ännu mer data, även i realtid, finns [här](#).
- Delmängd av data som jag använder finns [här](#) i CSV format.

Cykelpool - data i CSV format

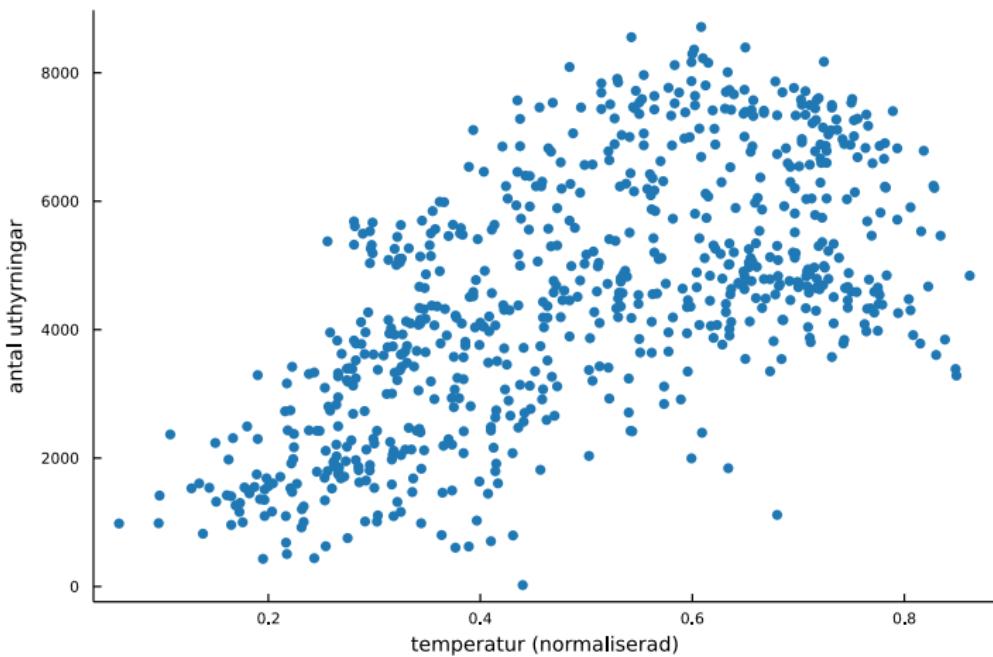
```
1 dteday,season,yr,mnth,holiday,weekday,workingday,weathersit,temp,hum,windspeed,nRides
2 2011-01-01,1,0,1,0,6,0,2,0.344167,0.805833,0.160446,985
3 2011-01-02,1,0,1,0,0,0,2,0.363478,0.696087,0.248539,801
4 2011-01-03,1,0,1,0,1,1,1,0.196364,0.437273,0.248309,1349
5 2011-01-04,1,0,1,0,2,1,1,0.2,0.590435,0.160296,1562
6 2011-01-05,1,0,1,0,3,1,1,0.226957,0.436957,0.1869,1600
7 2011-01-06,1,0,1,0,4,1,1,0.204348,0.518261,0.0895652,1606
8 2011-01-07,1,0,1,0,5,1,2,0.196522,0.498696,0.168726,1510
9 2011-01-08,1,0,1,0,6,0,2,0.165,0.535833,0.266804,959
10 2011-01-09,1,0,1,0,0,0,1,0.138333,0.434167,0.36195,822
11 2011-01-10,1,0,1,0,1,1,1,0.150833,0.482917,0.223267,1321
12 2011-01-11,1,0,1,0,2,1,2,0.169091,0.686364,0.122132,1263
13 2011-01-12,1,0,1,0,3,1,1,0.172727,0.599545,0.304627,1162
14 2011-01-13,1,0,1,0,4,1,1,0.165,0.470417,0.301,1406
15 2011-01-14,1,0,1,0,5,1,1,0.16087,0.537826,0.126548,1421
16 2011-01-15,1,0,1,0,6,0,2,0.233333,0.49875,0.157963,1248
17 2011-01-16,1,0,1,0,0,0,1,0.231667,0.48375,0.188433,1204
18 2011-01-17,1,0,1,1,1,0,2,0.175833,0.5375,0.194017,1000
19 2011-01-18,1,0,1,0,2,1,2,0.216667,0.861667,0.146775,683
20 2011-01-19,1,0,1,0,3,1,2,0.292174,0.741739,0.208317,1650
21 2011-01-20,1,0,1,0,4,1,2,0.261667,0.538333,0.195904,1927
22 2011-01-21,1,0,1,0,5,1,1,0.1775,0.457083,0.353242,1543
23 2011-01-22,1,0,1,0,6,0,1,0.0591304,0.4,0.17197,981
24 2011-01-23,1,0,1,0,0,0,1,0.0965217,0.436522,0.2466,986
25 2011-01-24,1,0,1,0,1,1,1,0.0973913,0.491739,0.15833,1416
26 2011-01-25,1,0,1,0,2,1,2,0.223478,0.616957,0.129796,1985
27 2011-01-26,1,0,1,0,3,1,3,0.2175,0.8625,0.29385,506
28 2011-01-27,1,0,1,0,4,1,1,0.195,0.6875,0.113837,431
29 2011-01-28,1,0,1,0,5,1,2,0.203478,0.703042,0.1233,1167
```

Cykelpool - data i tabellform

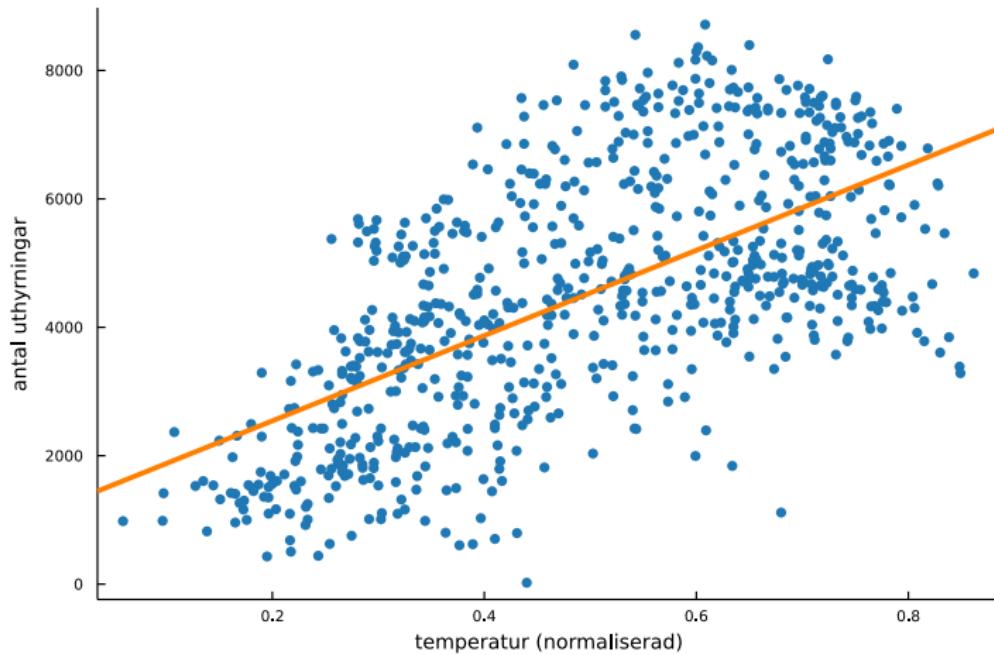
731x12 DataFrame													
Row	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	hum	windspeed	nRides	
	Date...	Int64	Int64	Int64	Int64	Int64	Int64	Int64	Float64	Float64	Float64	Int64	
1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.805833	0.160446	985	
2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.696087	0.248539	801	
3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.437273	0.248309	1349	
4	2011-01-04	1	0	1	0	2	1	1	0.2	0.590435	0.160296	1562	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
729	2012-12-29	1	1	12	0	6	0	2	0.253333	0.752917	0.124383	1341	
730	2012-12-30	1	1	12	0	0	0	1	0.255833	0.483333	0.350754	1796	
731	2012-12-31	1	1	12	0	1	1	2	0.215833	0.5775	0.154846	2729	

- dtedat: datum (dag) för observationen.
- nRides: antal uthyrningar för en given dag.
- temp: (normaliserad, 0 är kallast, 1 är varmaste).
- hum: luftfuktighet (normaliserad)
- windspeed: vindhastighet (normaliserad)
- season: (vinter = 1, vår = 2, sommar = 3, höst = 4).

Förklarande variabel: temperatur



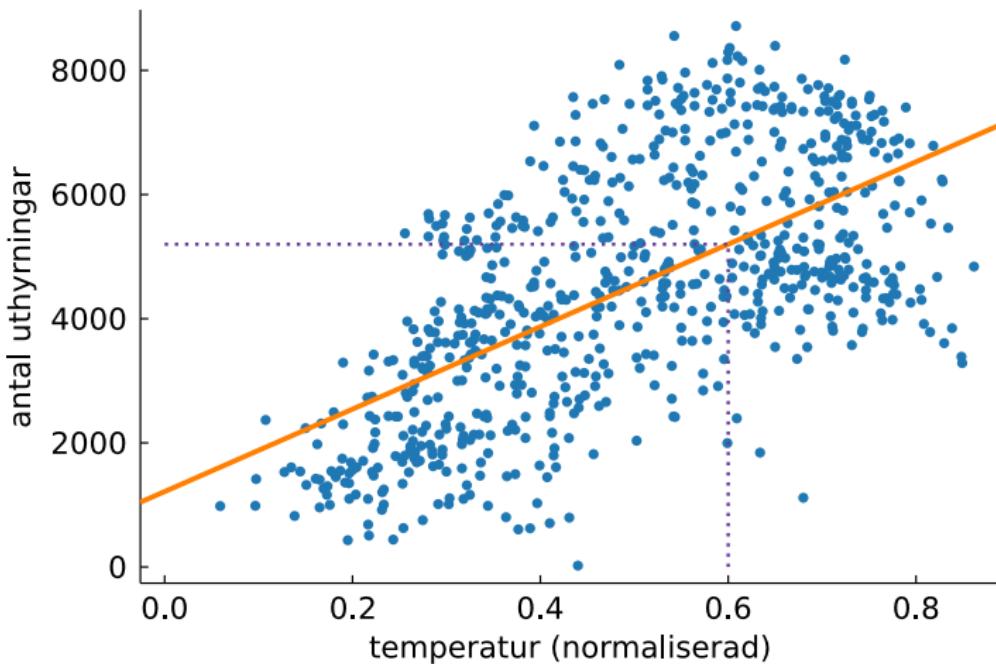
Regressionslinje



■ Regressionsekvation

$$\text{antal uthyrningar} = 1214.64 + 6640.71 \cdot \text{temperatur}$$

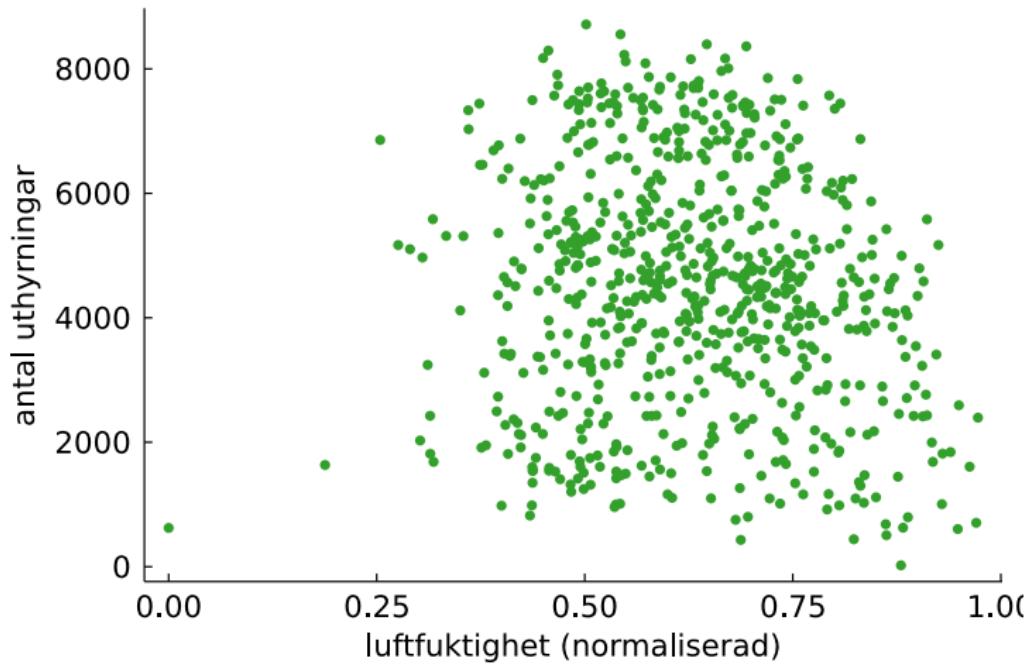
Prediktion för temp = 0.6



■ Prediktion för temperatur = 0.6 (hyfsat varm dag)

$$\text{antal uthyrningar} = 1214.64 + 6640.71 \cdot 0.6 \approx 5199 \text{ turer}$$

Förklarande variabel: luftfuktighet



Förklarande variabler: temp och luftfuktighet

■ Regressionsekvation

$$\text{antal uthyrningar} = 2657.9 + 6886.97 \cdot \text{temperatur} - 2492.85 \cdot \text{luftfuktighet}$$

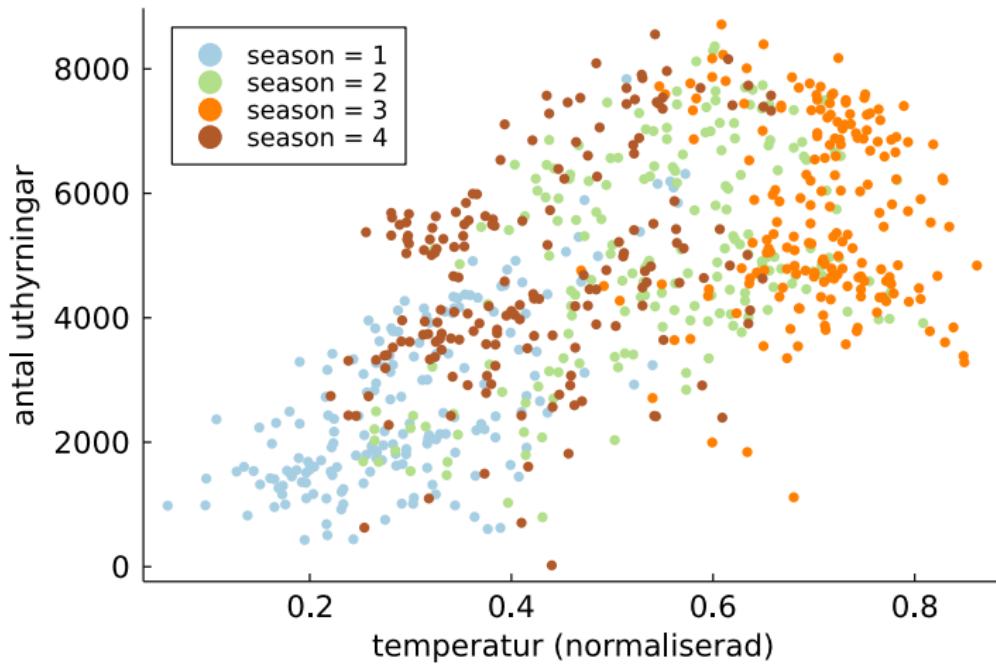
■ Prediktion för hyfsat varm, väldigt klibbig dag

$$\text{antal uthyrningar} = 2657.9 + 6886.97 \cdot 0.6 - 2492.85 \cdot 0.9 = 4546.52$$

■ Prediktion för hyfsat varm, mycket torr dag

$$\text{antal uthyrningar} = 2657.9 + 6886.97 \cdot 0.6 - 2492.85 \cdot 0.1 = 6540.80$$

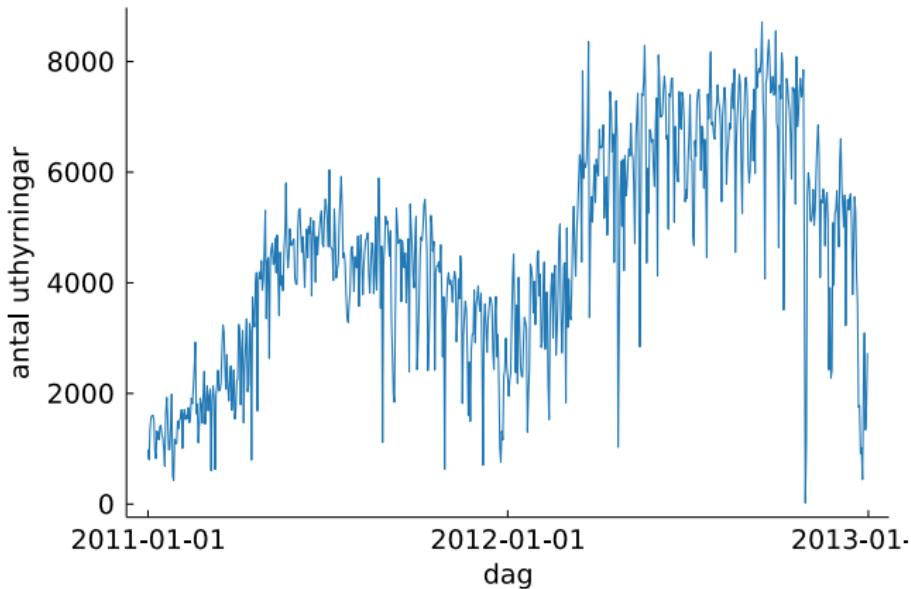
Förklarande variabler: temp och säsong



Exempel på frågor som besvaras under kursen

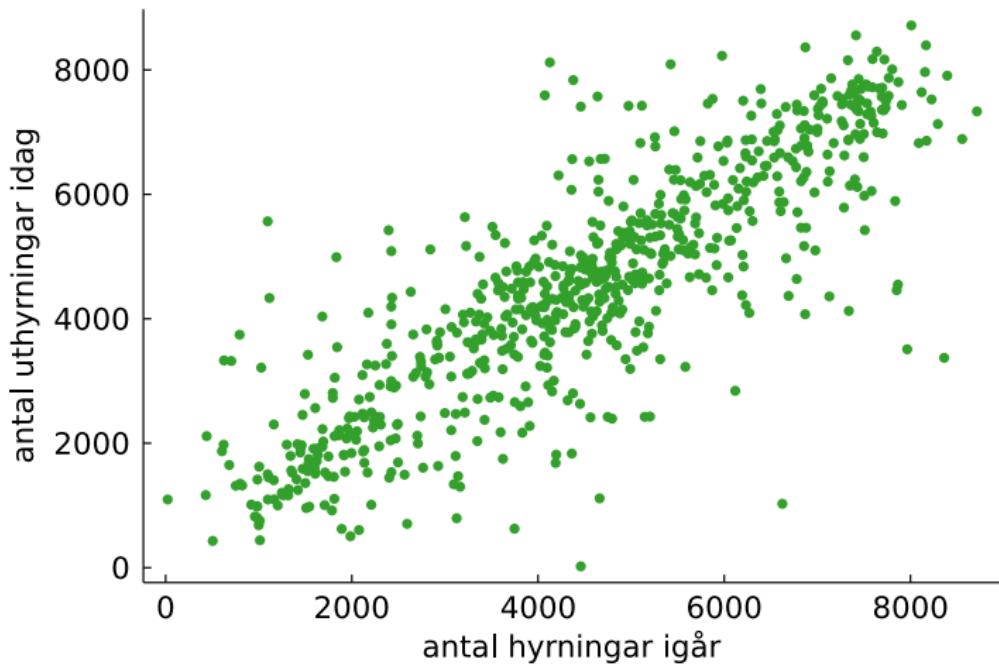
- Fler än en förklarande variabel? **Multipel regression**.
- Förklarande variabler som är **binära** ($0 = \text{vardag}$, $1 = \text{helg}$)?
- Förklarande variabler som **kategoriska** (säsong)?
- Är en förklarande variabel *verklig* korrelerad med målvariabeln? **Hypotesttest**.
- Hur **väljer** man modellens **förklarande variabler**?
- **Hur** träffsäker är en **prediktion** från en regressionsmodell?

Tidsserier - Data uppmätta över tid

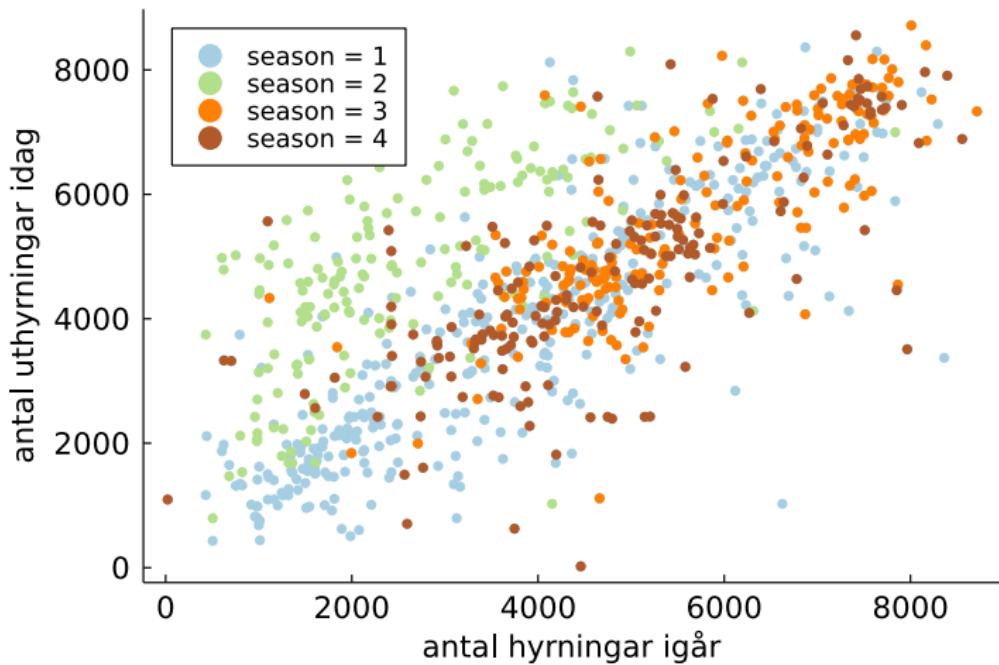


- Trend över tid?
- Säsongsvariation?

Förklarande variabel: går dagens uthyrning



Gårdagens uthyrning och säsong



Prognosticera slutpris på internetauktion

- **Mål:** statistisk modell för att förutsäga slutpriset i en auktion.
- **Data** från 1000 avslutade myntauktioner:
- **Målvariabel:** vinnande bud.
- **Förklarande variabler:**
 - ▶ värde enligt myntkatalog
 - ▶ skadad?
 - ▶ säljarens aktivitet
 - ▶ säljarens feedback score
 - ▶ obruten förpackning?

Bid History

2005 American Eagle Silver Dollar, Gem Uncirculated, 1 oz, .999 fine silver

Item number: 403251253977

Current bid: US \$6.50 (approximately 72.80 SEK)
Shipping: FREE Standard Shipping

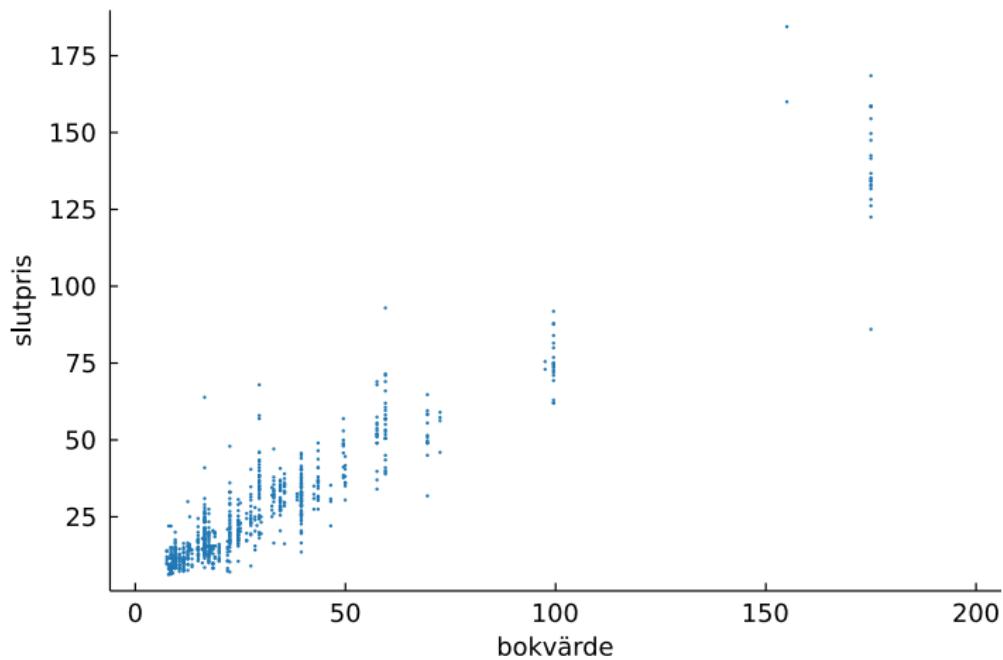
Bids: 4 Bidders: 2 Restrictions: 0 Time left: 5 days 2 hours 8 mins Duration: 7 days

(Enter US \$0.00 or more) Place bid Bid Amount

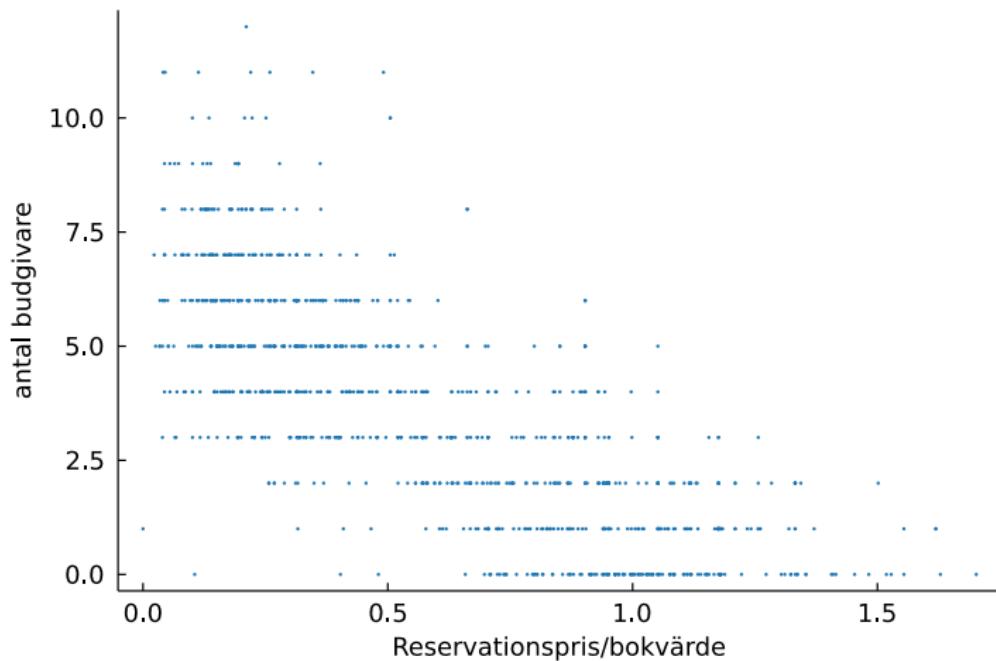
Hide automatic bids Show automatic bids [Learn more about bidding](#)

Bidder	Bid Amount	Bid Time
user1 (511)	US \$6.50	25 Oct 2021 at 4:45:25am PDT
user2 (156)	\$8.00	25 Oct 2021 at 7:04:51am PDT
user3 (156)	\$7.00	25 Oct 2021 at 6:08:23am PDT
user4 (156)	\$9.00	25 Oct 2021 at 6:08:23am PDT
Starting price	\$0.99	23 Oct 2021 at 11:31:27am PDT

Är myntkatalogens värderingar en bra prediktor?

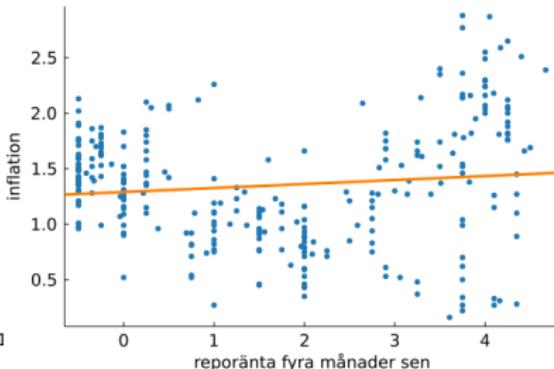
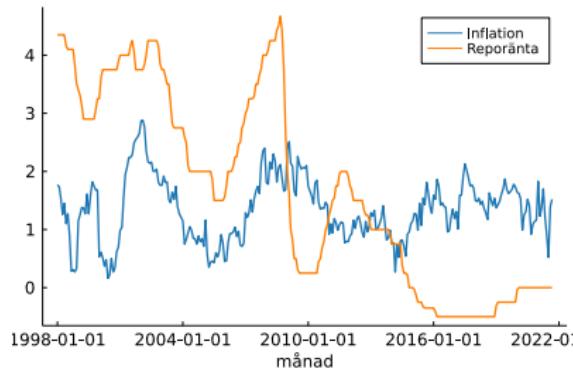


Skrämmer höga startbud bort budgivare?



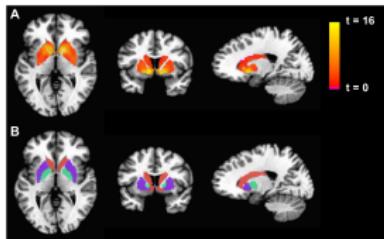
Riksbanken och styrräntan

- Riksbankens mål är att hålla inflationen nära 2% per år.
- Riksbanken bestämmer den s k reporäntan in ekonomin.
- Hur beror inflationen på räntan?
- Både inflation och reporänta är exempel på **tidsserier**.
- **Sambandet** mellan inflation och ränta: **regression**.



Var i hjärnan skapas vårt språk?

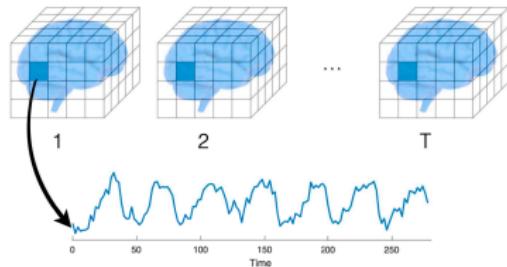
- Person i MR scanner pratar omväxlande med att knyta handen.



Lars Kruse, AU Kommunikation, CC license

[Source](#), CC license

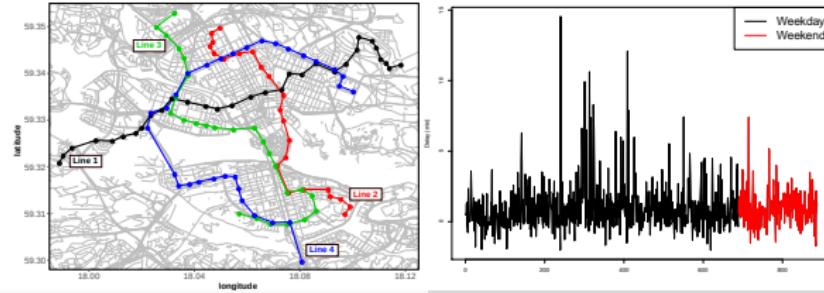
- Mäter mängden syresatt blod på tusentals ställen i hjärnan.



- Regression med förklarande variabeln pratar/knyter hand.

Förseningar i lokaltrafiken

- Mål1: **förutsäga förseningar** för stadsbussar.
- Mål2: **säkerheten** i prediktionen: **5 min, 5 min, 5 min**
- Data: alla förseningar för alla buslinjer i Sthlm under 1 år.
- Mål: förutsäga förseningen för 12.15-bussen till Tegnérsgatan.
- Förklarande variabler:
 - ▶ försening för 12.15-bussen vid hållplatser innan Tegnérsgatan.
 - ▶ förseningar för tidigare bussar vid hållplats Tegnérsgatan.
 - ▶ tid på dagen
 - ▶ rusningstid?



Artificiell intelligens och maskininlärning

- Mål: få en maskin att känna igen handskrivna siffor.
- Data: 60000 handskrivna siffror mellan 0-9.
- Förenkling: enbart skilja mellan 0:or och 1:or.
- Varje bild har 28×28 pixlar med värde mellan 0 och 255:

0 = svart.

128 = mellangrå.

255 = .



- Använd alla pixlar. Totalt $28 \cdot 28 = 784$ förklarande variabler:
 - ▶ Gråhet i pixel (1,1)
 - ▶ Gråhet i pixel (1,2)
 - ▶ ...
 - ▶ Gråhet i pixel (28,28)
- **Målvariabeln är binär:** 0 eller 1.
- **Logistisk regression:** modell för *sannolikheten* för 1:a.
- Djupa neural nätverk (deep learning) är en form av regression.

Artificiell intelligens och maskininlärning



A screenshot of a web browser's header. It shows a search bar with the URL "machinelearningmastery.com/what-is-statistics/". Below the search bar is a horizontal menu bar with several items: a blue square icon, a blue triangle icon, a green circle icon, a black square icon with "N", a black square icon with "GITHUB", a black square icon with "Editorial", a black square icon with "Julia", a blue square icon with "Proxy Liu", a blue square icon with "Proxy SU", a black square icon with "Liu", a black square icon with "SU", a black square icon with "Study", a black square icon with "Family", a black square icon with "Mail SU", a blue square icon with "Mail Liu", a yellow square icon with "Athena", and a black square icon with "M".

Statistics is Required Prerequisite

Machine learning and statistics are two tightly related fields of study. So much so that statisticians refer to machine learning as "*applied statistics*" or "*statistical learning*" rather than the computer-science-centric name.

Machine learning is almost universally presented to beginners assuming that the reader has some background in statistics. We can make this concrete with a few cherry picked examples.

Take a look at this quote from the beginning of a popular applied machine learning book titled "*Applied Predictive Modeling*":

 ... the reader should have some knowledge of basic statistics, including variance, correlation, simple linear regression, and basic hypothesis testing (e.g. p-values and test statistics).

— Page vii, [Applied Predictive Modeling](#), 2013

Några datakällor

- [Statistikdatabasen SCB](#)
- [OECD](#)
- [Gapminder](#)
- [UCI - machine learning repository](#)
- [Kaggle - machine learning](#)