Regressions- och tidsserieanalys

Föreläsning 2 - Enkel linjär regression

Mattias Villani

Statistiska institutionen
Stockholms universitet

Institutionen för datavetenskap
Linköpings universitet


mattiasvillani.com

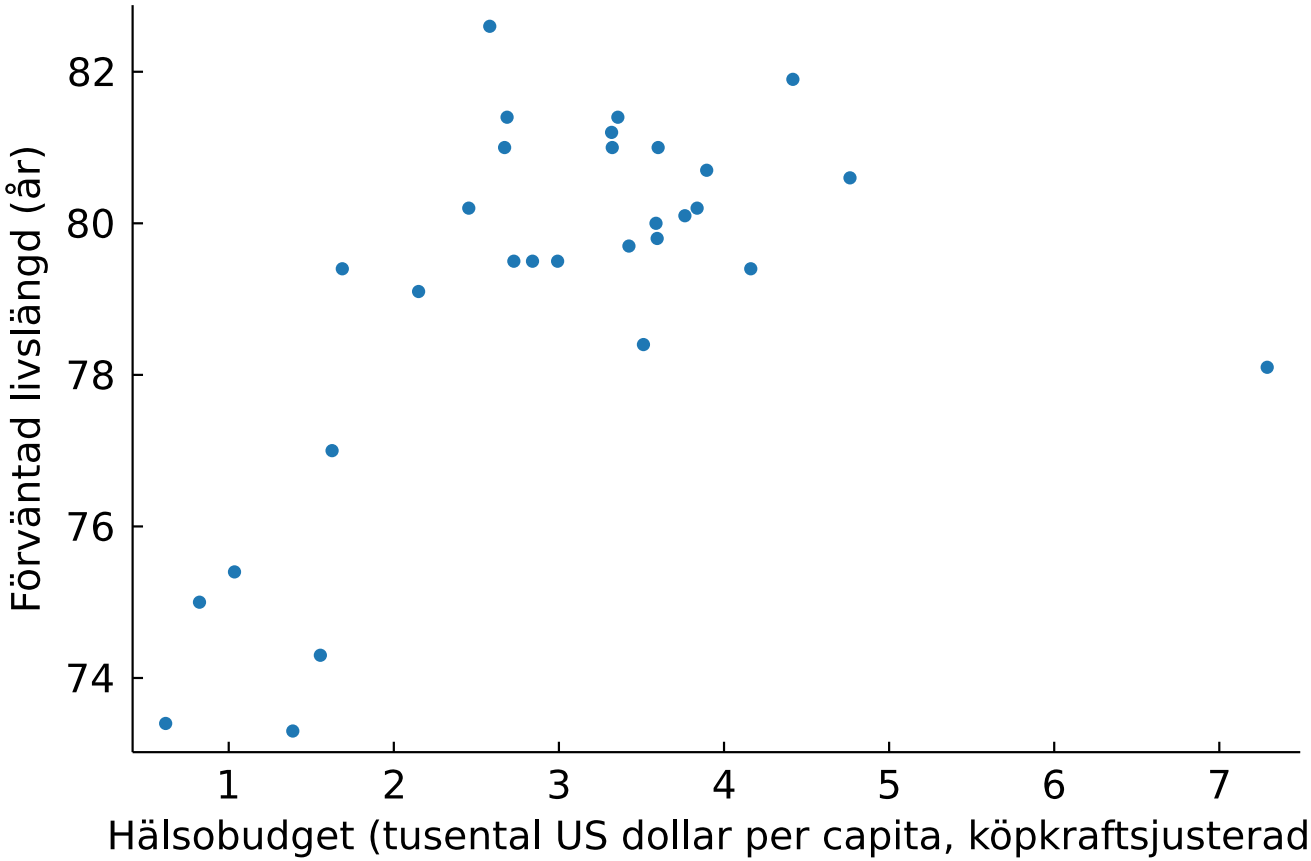

@matvil


mattiasvillani

Översikt

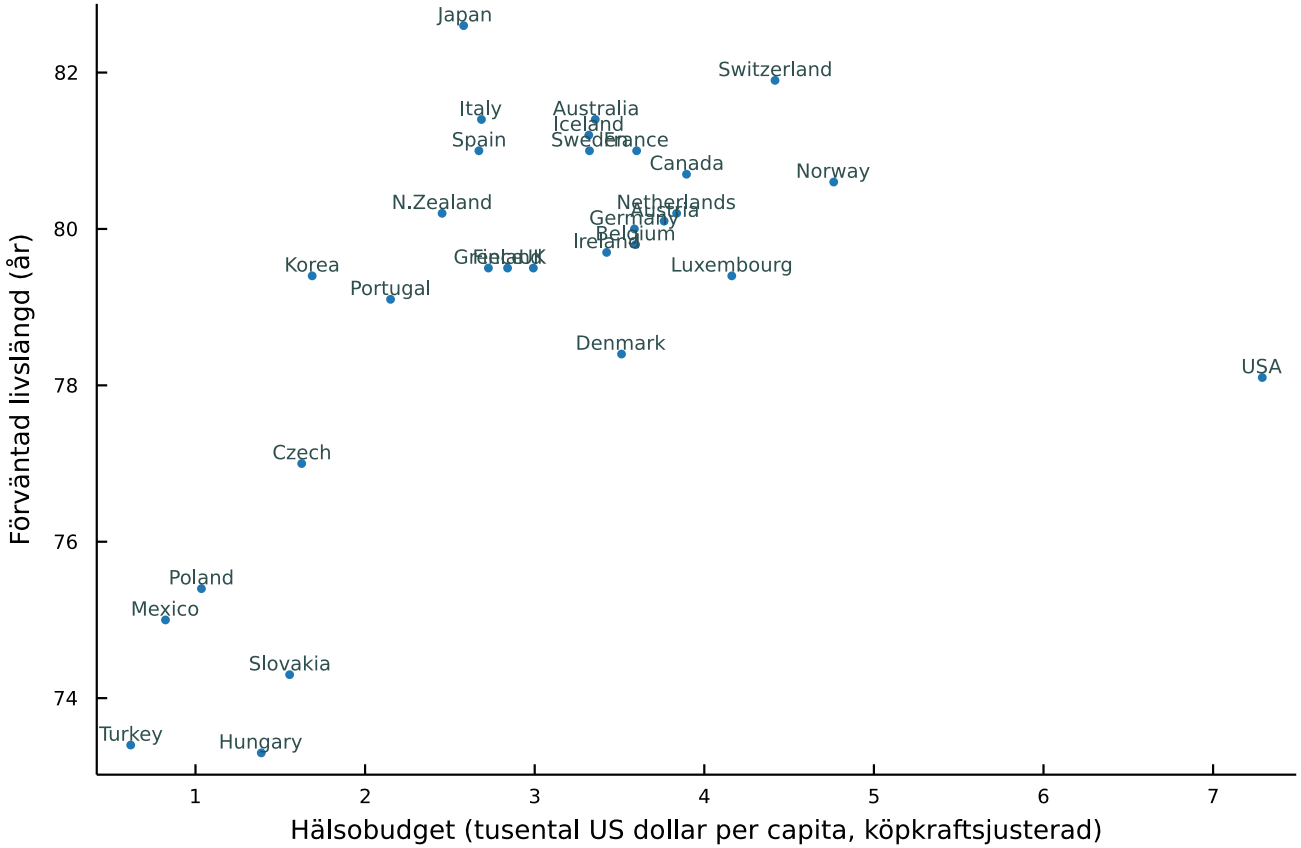
- **Enkel linjär regression**
- **Minsta-kvadratmetoden** för att skatta regression.
- **Korrelation**
- **Variansanalys**
- **Inflytelserika observationer**
- **Extrapolation**

Samband mellan hälsovårdsbudget och livslängd

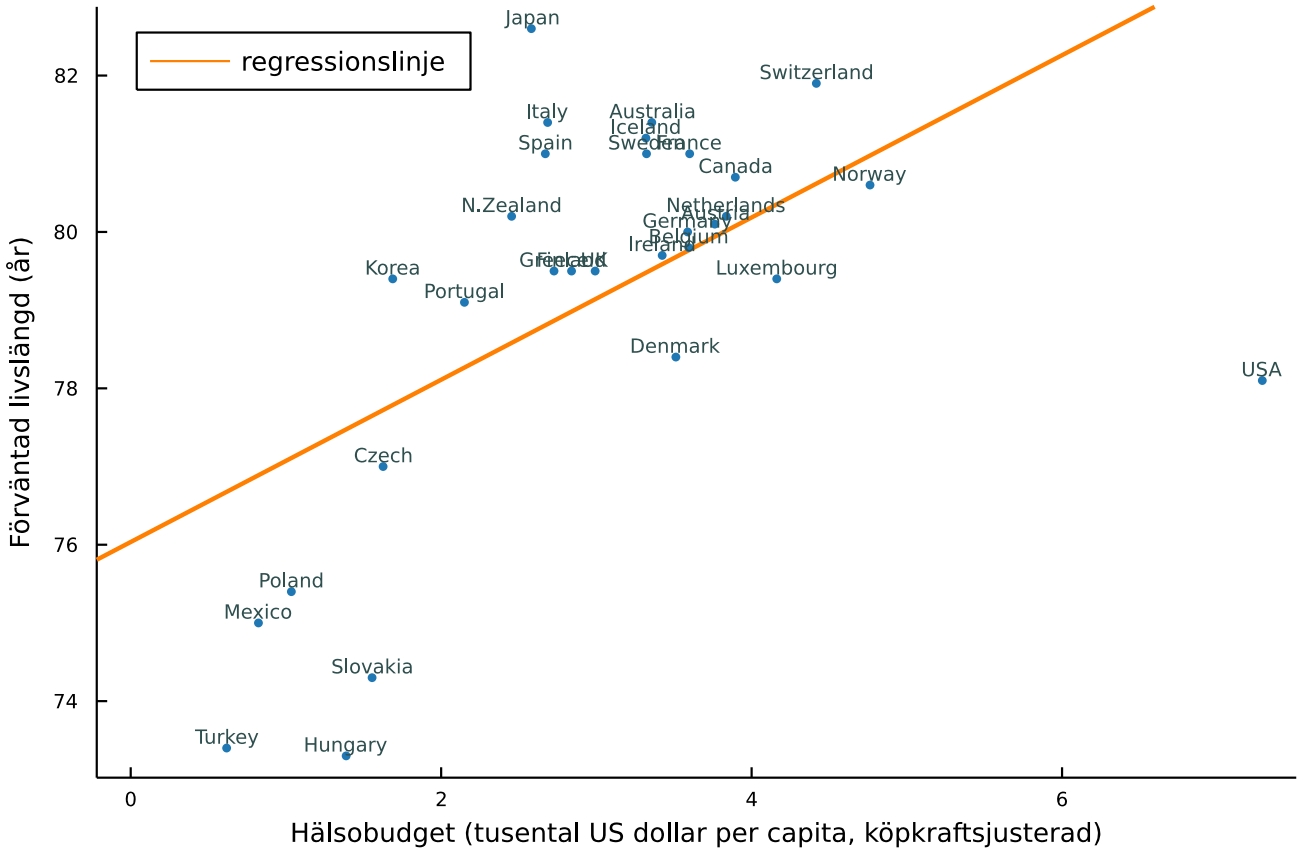


SUdatasets healthbudget

Samband mellan hälsovårdsbudget och livslängd



Samband mellan hälsovårdsbudget och livslängd



Samband mellan hälsovårdsbudget och livslängd

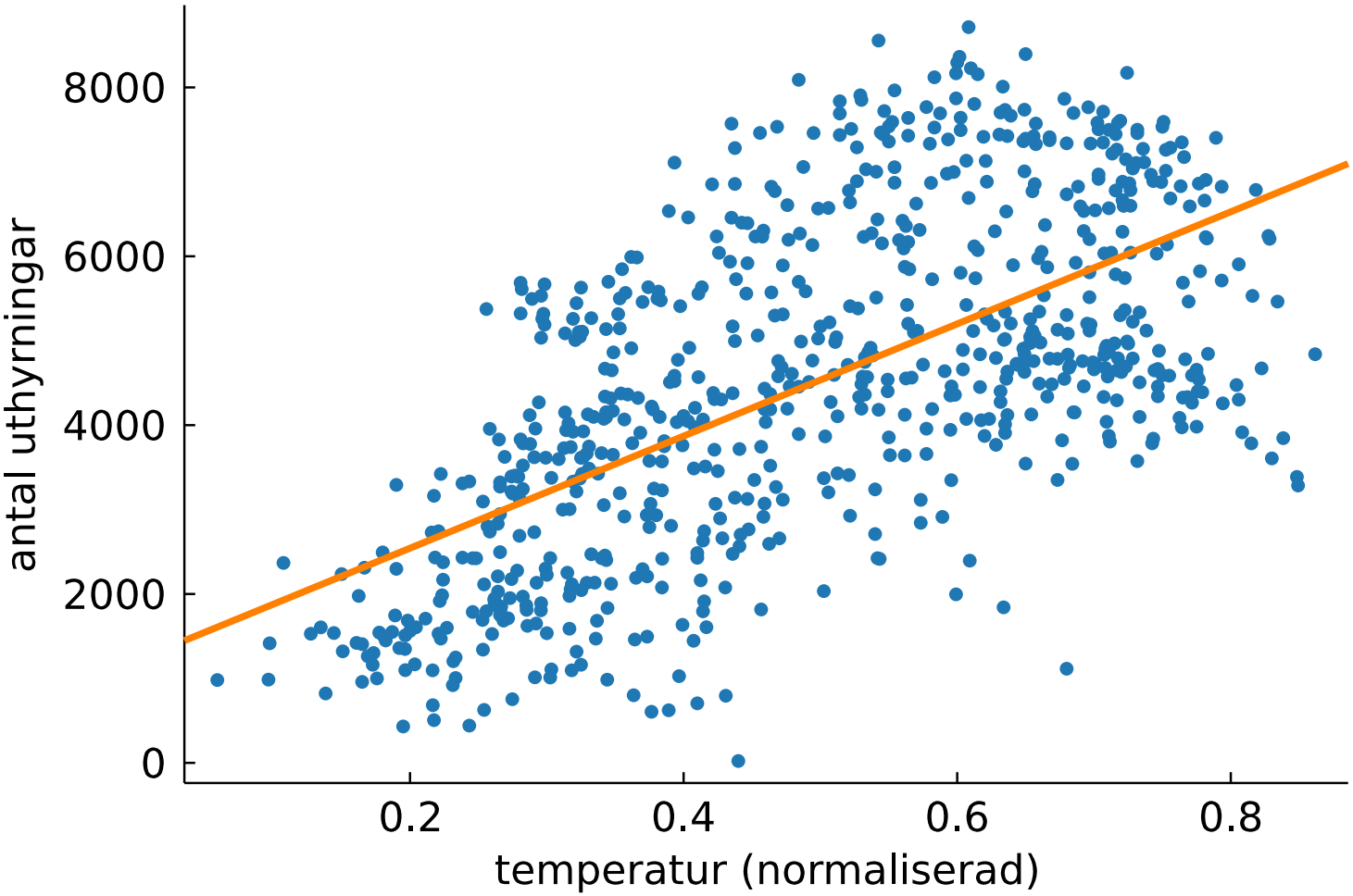
- Skattad regressionslinje hälsobudget (x) $\rightarrow$ livslängd (y)

$$\text{livslängd} = 76.035 + 1.03757 \cdot \text{hälsobudget}$$

$$y = \underbrace{76.035}_{a} + \underbrace{1.038}_{b} \cdot x$$

- Förväntade livslängden är ca 76 år om hälsobudget $= 0$.
- Livslängden ökar med 1.038 år om hälsobudgeten ökar med 1 (tusen US dollar per capita).

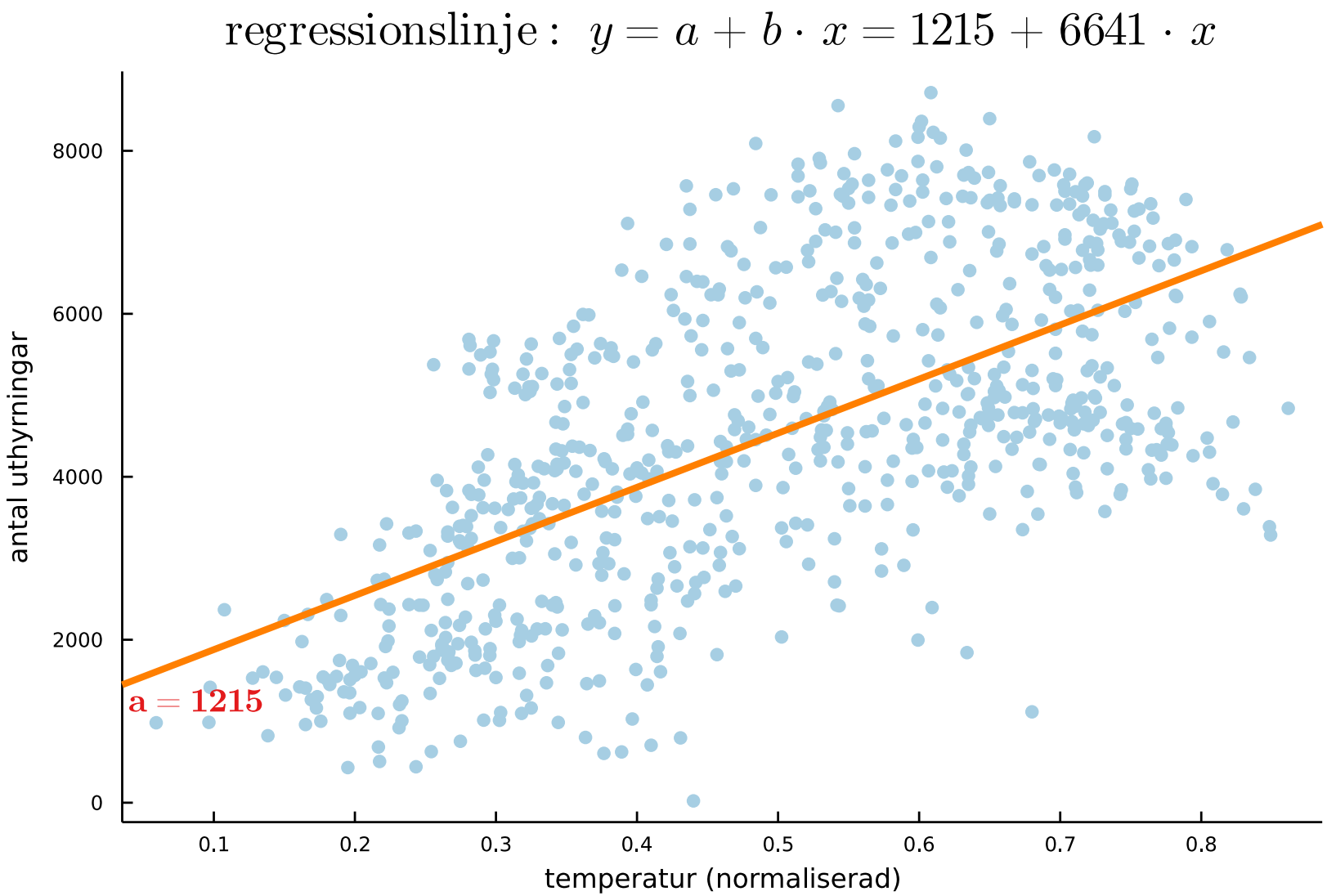
Cykeluthyrningar



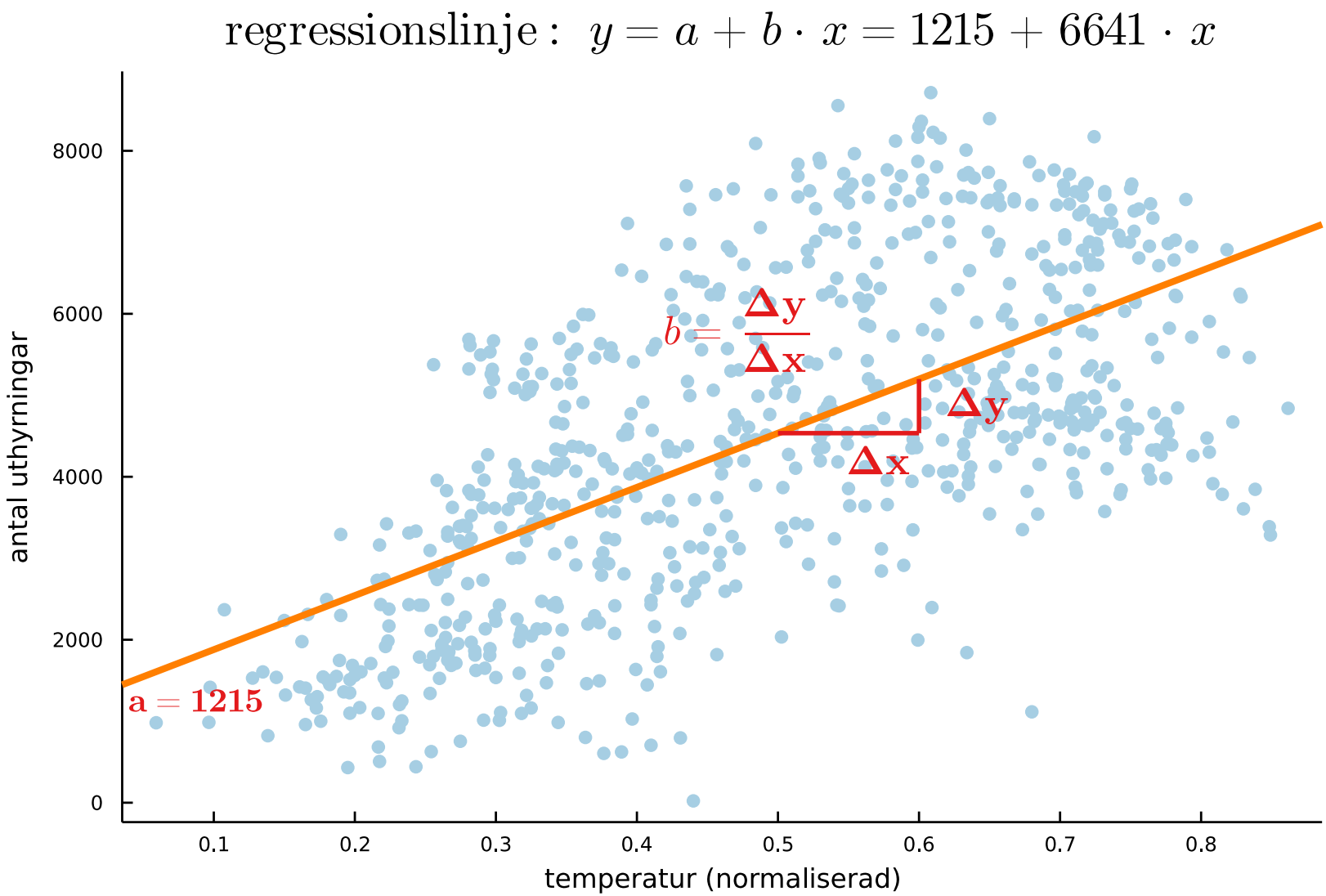
- Regressionsekvation

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

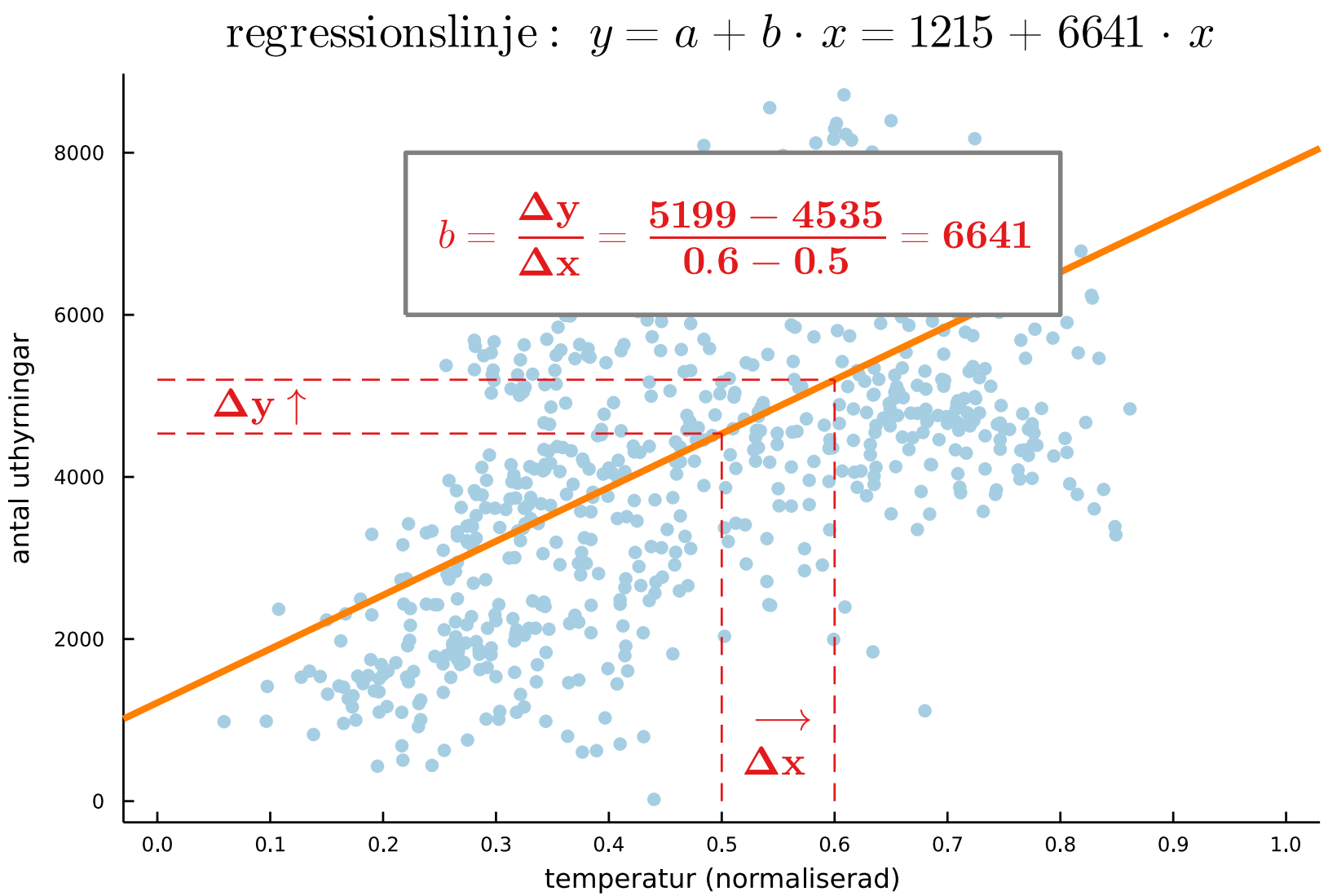
Interceptet a - värdet på y när x=0



Lutningen b - hur ändras y när x ändras en enhet?



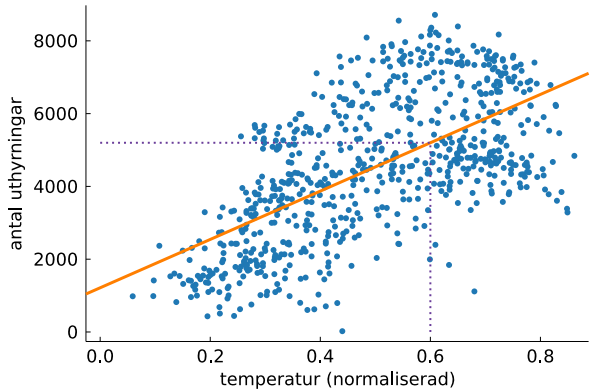
Lutningen b - hur ändras y när x ändras en enhet?



Skattning av regressionslinjen - minsta kvadrat

- **Prediktion** för den i:te observationen i stickprovet:

$$\hat{y}_i = a + b \cdot x_i$$



- **Prediktionsfel** (**residualer**)

$$e_i = y_i - \hat{y}_i$$

- Välj a och b som **minimerar residualkvadratsumman**

$$Q = \sum_{i=1}^{n} e_i^2$$

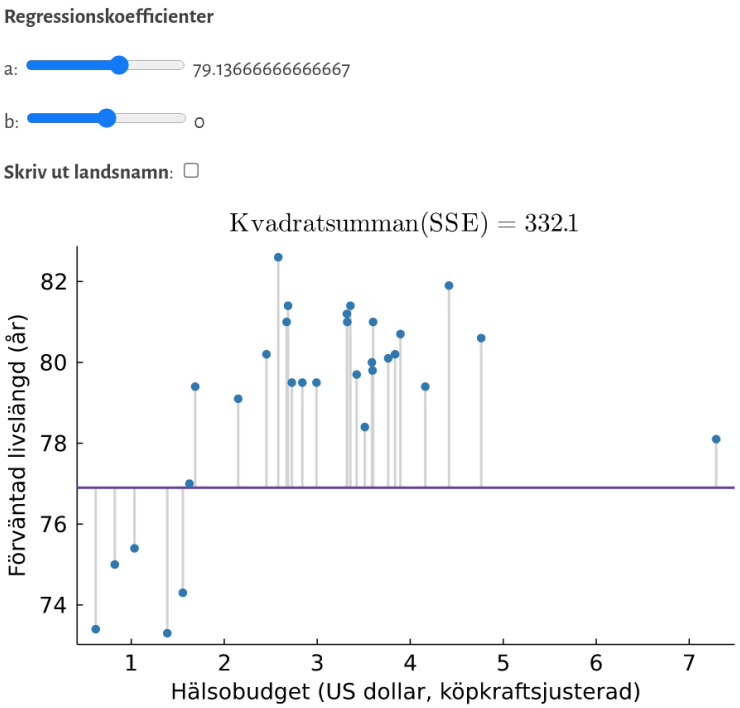
- **Sum of Squared Errors** (**SSE**).

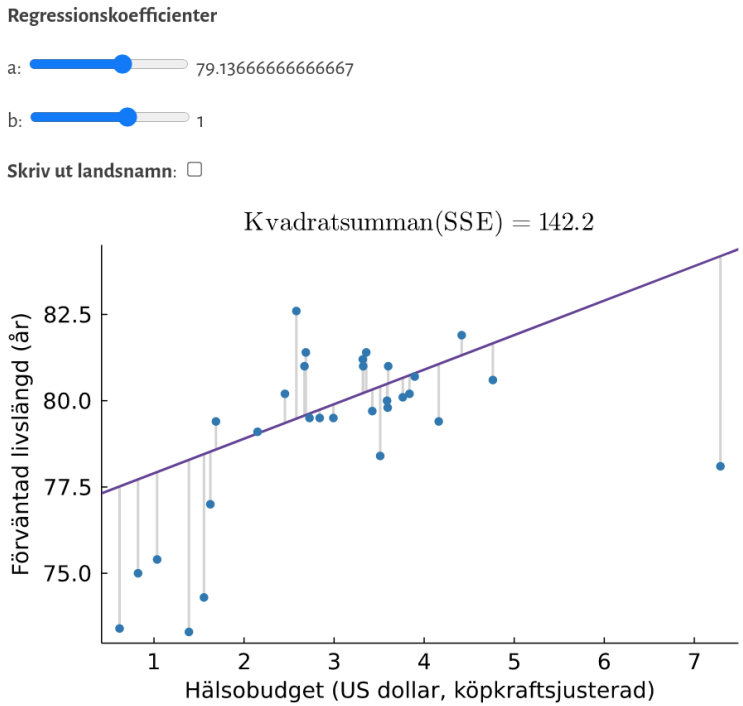
Samband mellan hälsovårdsbudget och livslängd

- Kalkylark (Excel) kan beräkna residualer och kvadrater etc.
- Orange cell är residualkvadratsumman Q för a och b i blå cell.
- Notera att t ex $\hat{y}_1 = 77 + 1 \cdot 3.357 = 80.357$.
- Se länk på kurssida till kalkylarket.

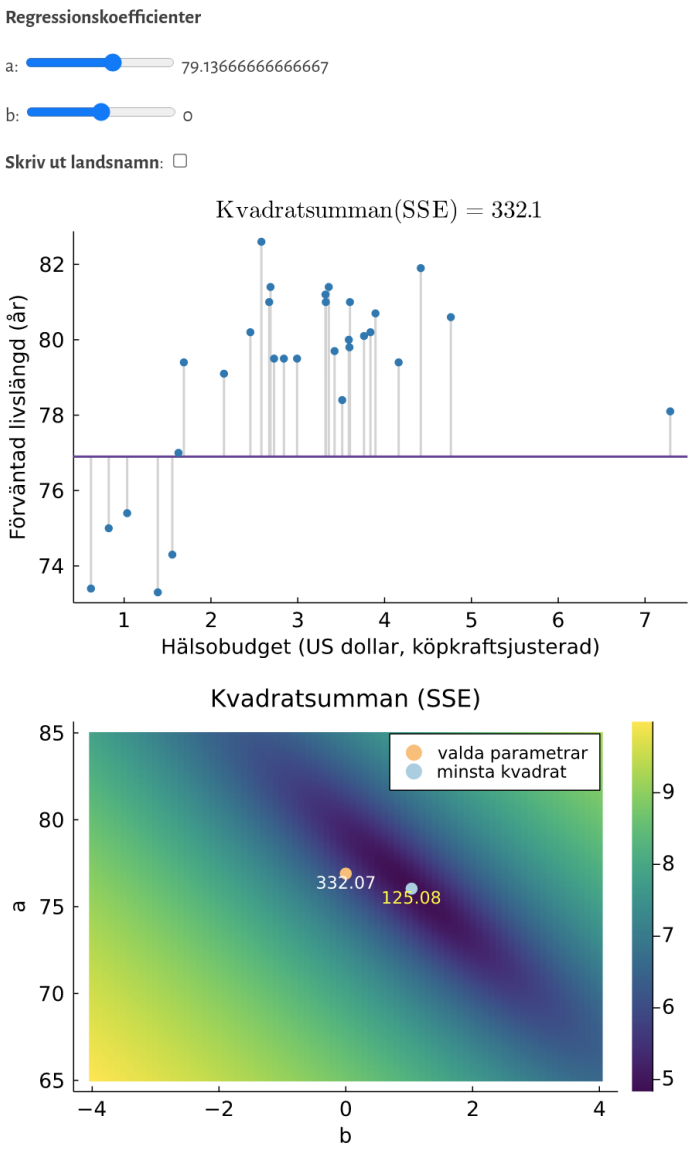
	A	B	C	D	E	F
1	country	spending (x)	lifespan (y)	yHat	e = y-yHat	e²
2	Australia	3.357	81.4	80.357	1.043	1.087849
3	Austria	3.763	80.1	80.763	-0.663	0.439569
4	Belgium	3.595	79.8	80.595	-0.795	0.632025
5	Canada	3.895	80.7	80.895	-0.195	0.038025
6	Czech	1.626	77	78.626	-1.626	2.643876
7	Denmark	3.512	78.4	80.512	-2.112	4.460544
8	Finland	2.84	79.5	79.84	-0.34	0.1156
9	France	3.601	81	80.601	0.399	0.159201
10	Germany	3.588	80	80.588	-0.588	0.345744
11	Greece	2.727	79.5	79.727	-0.227	0.051529
12	Hungary	1.388	73.3	78.388	-5.088	25.887744
13	Iceland	3.319	81.2	80.319	0.881	0.776161
14	Ireland	3.424	79.7	80.424	-0.724	0.524176
15	Italy	2.686	81.4	79.686	1.714	2.937796
16	Japan	2.581	82.6	79.581	3.019	9.114361
17	Korea	1.688	79.4	78.688	0.712	0.506944
18	Luxembourg	4.162	79.4	81.162	-1.762	3.104644
19	Mexico	0.823	75	77.823	-2.823	7.969329
20	Netherlands	3.837	80.2	80.837	-0.637	0.405769
21	N.Zealand	2.454	80.2	79.454	0.746	0.556516
22	Norway	4.763	80.6	81.763	-1.163	1.352569
23	Poland	1.035	75.4	78.035	-2.635	6.943225
24	Portugal	2.15	79.1	79.15	-0.05	0.0025
25	Slovakia	1.555	74.3	78.555	-4.255	18.105025
26	Spain	2.671	81	79.671	1.329	1.766241
27	Sweden	3.323	81	80.323	0.677	0.458329
28	Switzerland	4.417	81.9	81.417	0.483	0.233289
29	Turkey	0.618	73.4	77.618	-4.218	17.791524
30	UK	2.992	79.5	79.992	-0.492	0.242064
31	USA	7.29	78.1	84.29	-6.19	38.3161
32	Summa				-25.58	146.968268
33						
34	Regressionscoefficienter					
35	a	77				
36	b	1				

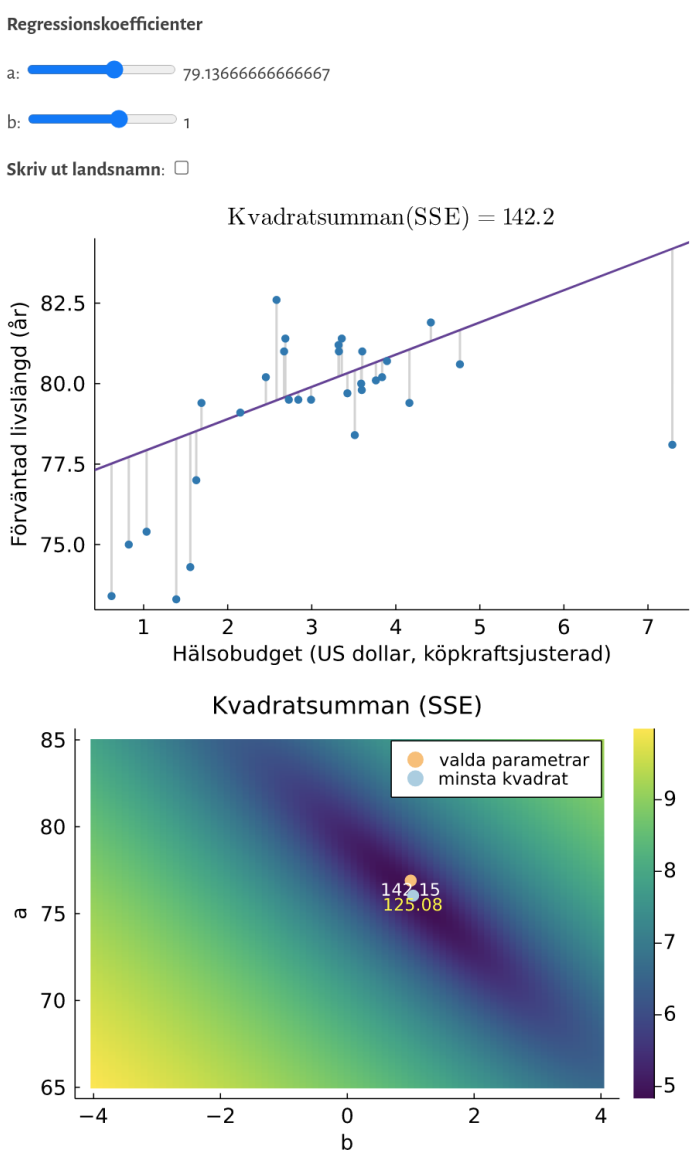
Samband mellan hälsovårdsbudget och livslängd





Samband mellan hälsovårdsbudget och livslängd





Skattning av regressionslinjen - minsta kvadrat

- Residualkvadratsumman beror på a och b:

$$Q(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - b \cdot x_i)^2$$

- $Q(a,b)$ minimeras när partialderivatorna är noll:

$$\frac{\partial Q}{\partial a} = -2 \sum_{i=1}^{n} (y_i - a - b \cdot x_i) = 0$$

$$\frac{\partial Q}{\partial b} = -2 \sum_{i=1}^{n} (y_i - a - b \cdot x_i) x_i = 0$$

- Derivatan av en summa: $\frac{d}{dx}(f(x) + g(x)) = \frac{df(x)}{dx} + \frac{dg(x)}{dx}$
- Potensregeln för derivator: $\frac{dx^p}{dx} = px^{p-1}$, t ex $\frac{dx^2}{dx} = 2x$
- Kedjeregeln för derivator (specialfall): $\frac{d}{dx} f(g(x)) = f'(g(x)) \cdot g'(x)$, där $f'(x) = \frac{d}{dx} f(x)$ är ett alternativt sätt att uttrycka derivatan.

Minsta kvadrat - alternativa formler

- Minstakvadratskattningar

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Alternativ formel för b för handberäkning:

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

- Bevis

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i^2 + \bar{x}^2 - 2\bar{x}x_i) = \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i^2 + n\bar{x}^2 - 2\bar{x}n\bar{x} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

- Skattning av intercept:

$$a = \bar{y} - b \cdot \bar{x}$$

- Hälsobudgetdata

$$b = \frac{7151.8229 - 30 \cdot 2.989333333 \cdot 79.13666667}{320.944068 - 30 \cdot 2.989333333^2} \approx 1.03757$$

Minsta kvadrat i kalkylark

	A	B	C	D	E	F	G	H
1	country	spending (x)	lifespan (y)	yHat	e = y-yHat	e^2	x^2	xy
2	Australia	3.357	81.4	79.5181466	1.881853403	3.541372231	11.269449	273.2598
3	Austria	3.763	80.1	79.93940005	0.1605999533	0.02579234501	14.160169	301.4163
4	Belgium	3.595	79.8	79.76508827	0.03491172565	0.001218828588	12.924025	286.881
5	Canada	3.895	80.7	80.0763593	0.6236407037	0.3889277273	15.171025	314.3265
6	Czech	1.626	77	77.7221128	-0.7221128002	0.5214468962	2.643876	125.202
7	Denmark	3.512	78.4	79.67896996	-1.278969958	1.635764154	12.334144	275.3408
8	Finland	2.84	79.5	78.98172287	0.5182771309	0.2686111845	8.0656	225.78
9	France	3.601	81	79.77131369	1.228686305	1.509670037	12.967201	291.681
10	Germany	3.588	80	79.75782528	0.2421747162	0.05864859315	12.873744	287.04
11	Greece	2.727	79.5	78.86447745	0.6355225492	0.4038889106	7.436529	216.7965
12	Hungary	1.388	73.3	77.47517112	-4.175171123	17.4320539	1.926544	101.7404
13	Iceland	3.319	81.2	79.47871893	1.721281066	2.962808508	11.015761	269.5028
14	Ireland	3.424	79.7	79.58766379	0.1123362082	0.01261942367	11.723776	272.8928
15	Italy	2.686	81.4	78.82193708	2.578062922	6.646408431	7.214596	218.6404
16	Japan	2.581	82.6	78.71299222	3.88700778	15.10882948	6.661561	213.1906
17	Korea	1.688	79.4	77.78644214	1.613557855	2.603568952	2.849344	134.0272
18	Luxembourg	4.162	79.4	80.35339051	-0.9533905059	0.9089534567	17.322244	330.4628
19	Mexico	0.823	75	76.88894403	-1.888944031	3.568109554	0.677329	61.725
20	Netherlands	3.837	80.2	80.01618023	0.1838197679	0.03378970708	14.722569	307.7274
21	N.Zealand	2.454	80.2	78.58122082	1.618779179	2.620446031	6.022116	196.8108
22	Norway	4.763	80.6	80.97697012	-0.3769701199	0.1421064713	22.686169	383.8978
23	Poland	1.035	75.4	77.10890889	-1.708908887	2.920369584	1.071225	78.039
24	Portugal	2.15	79.1	78.26579952	0.8342004815	0.6958904433	4.6225	170.065
25	Slovakia	1.555	74.3	77.64844532	-3.348445325	11.21208609	2.418025	115.5365
26	Spain	2.671	81	78.80637353	2.193626473	4.811997104	7.134241	216.351
27	Sweden	3.323	81	79.48286921	1.517130786	2.301685821	11.042329	269.163
28	Switzerland	4.417	81.9	80.61797087	1.282029125	1.643598679	19.509889	361.7523
29	Turkey	0.618	73.4	76.67624217	-3.276242166	10.73376273	0.381924	45.3612
30	UK	2.992	79.5	79.13943352	0.3605664798	0.1300081864	8.952064	237.864
31	USA	7.29	78.1	83.59890969	-5.498909695	30.23800783	53.1441	569.349
32	Summa				0	125.0824413	320.944068	7151.8229
33	Medelvärde	2.989333333	79.13666667					
34								
35	Minsta-kvadratskattningar							
36	a	76.03502386						
37	b	1.037570073						
38								

gif-time!

- På webbsidan ligger en animerad gif som visar hur olika regressionlinjer ger olika SSE. gif

Regression i R

```
library(SUdatasets) # läser in datamaterialen
library(regkurs)    # läser in funktionen regsummary
fit = lm(lifespan ~ spending, data = healthbudget) # skattar regression
regsummary(fit)     # skriver ut sammanfattning av regressionsresultat
```

```
Analysis of variance - ANOVA
-----------------------------------------------
      df      SS      MS      F    Pr(>F)
Regr   1  56.907 56.9072 12.739 0.0013164
Error 28 125.082  4.4672
Total 29 181.990

Measures of model fit
-----------------------------------------------
Root MSE       R2   R2-adj
 2.11358  0.31269  0.28815

Parameter estimates
-----------------------------------------------
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)  76.0350    0.95084 79.9663 1.3416e-34
spending      1.0376    0.29071  3.5691 1.3164e-03
```

Residualvarians

- **Residualvariansen** - hur bra regressionslinjen passar data:

$$s_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

- Kom ihåg: stickprovsvariansen delar med $n-1$ eftersom vi måste beräkna $\bar{y}$ först:

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

- Residualvariansen delar med $n-2$ eftersom vi måste beräkna både a och b först. **Väntevärdesriktig**. Se F3.

- **Residualstandardavvikelsen** (residualspridningen):

$$s_e = \sqrt{s_e^2}$$

- Hälsobudgetdata

$$s_e^2 = \frac{125.0824413}{30-2} \approx 4.467 \qquad s_e = \sqrt{4.467} \approx 2.11 \text{ år}$$

Regression i R

```
Analysis of variance - ANOVA
-----------------------------------------------
      df      SS      MS      F    Pr(>F)
Regr   1  56.907 56.9072 12.739 0.0013164
Error 28 125.082  4.4672
Total 29 181.990

Measures of model fit
-----------------------------------------------
Root MSE       R2   R2-adj
 2.11358  0.31269  0.28815

Parameter estimates
-----------------------------------------------
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)  76.0350    0.95084 79.9663 1.3416e-34
spending      1.0376    0.29071  3.5691 1.3164e-03
```

Korrelation

- **Korrelationskoefficienten** mäter **graden av linjärt samband**

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}}$$

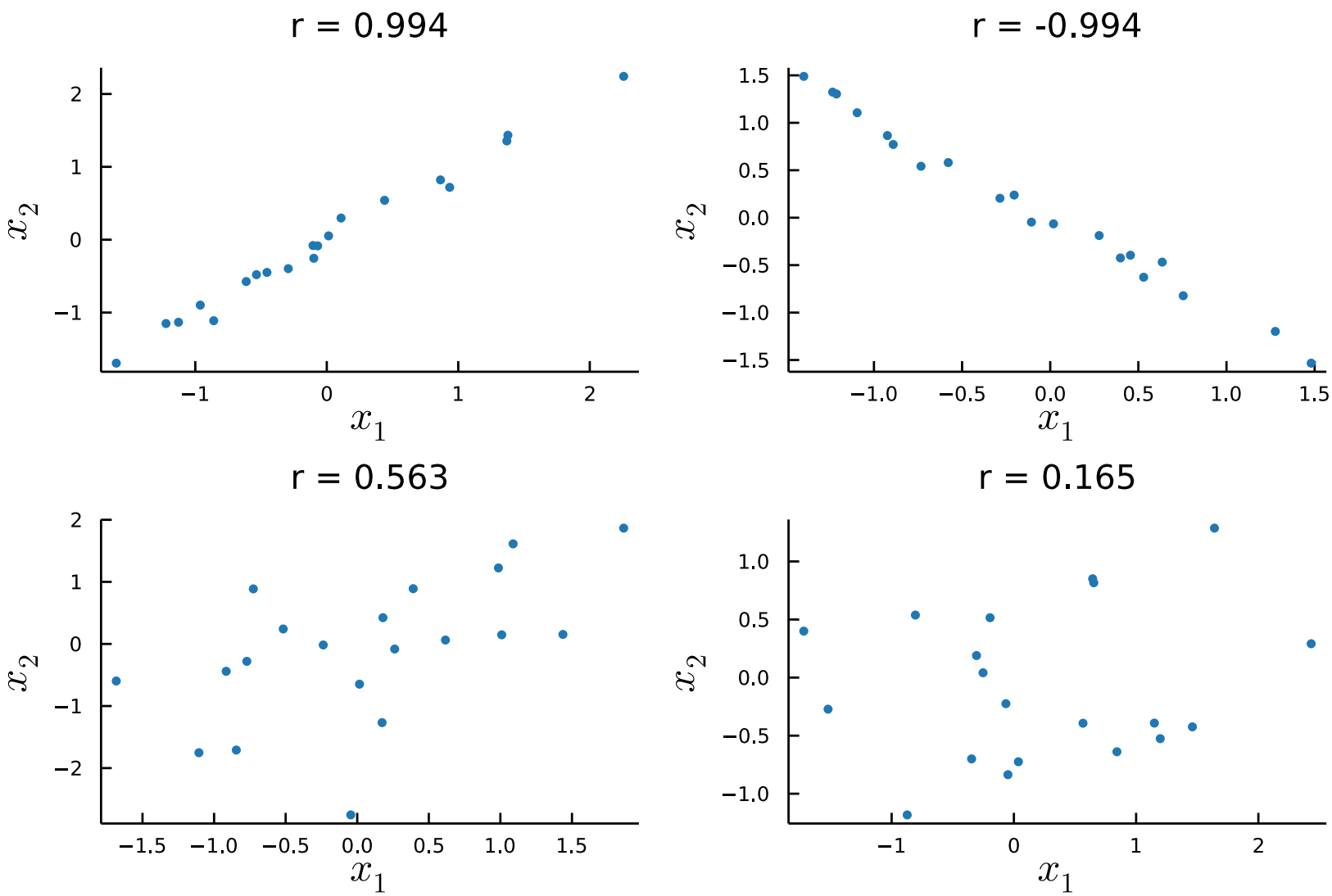
- Alternativ formel för handräkning

$$r = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{\left(n\sum x_i^2 - \left(\sum x_i\right)^2\right)\left(n\sum y_i^2 - \left(\sum y_i\right)^2\right)}}$$

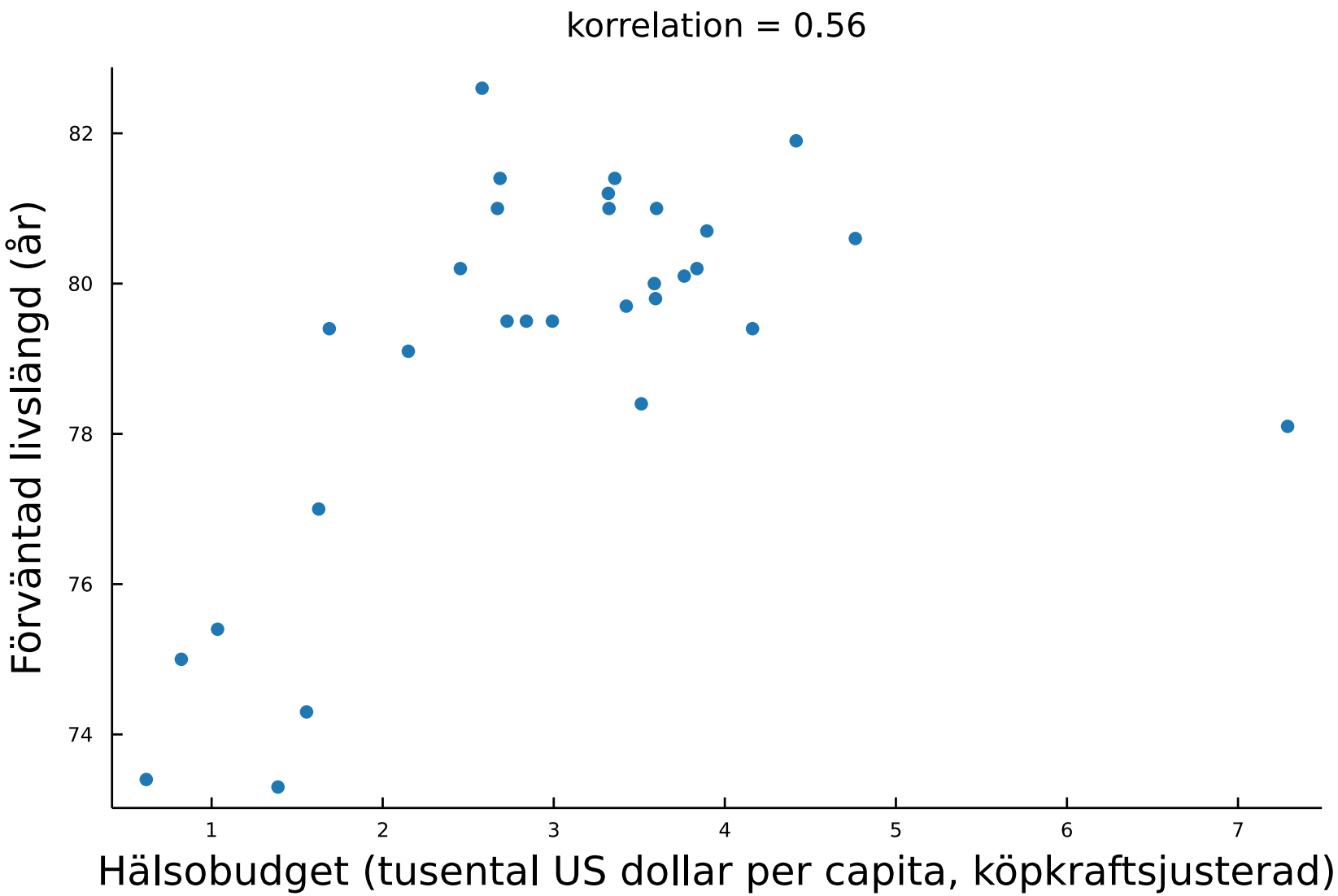
- Korrelationskoefficienten är ett normerat mått:

$$-1 \leq r \leq 1$$

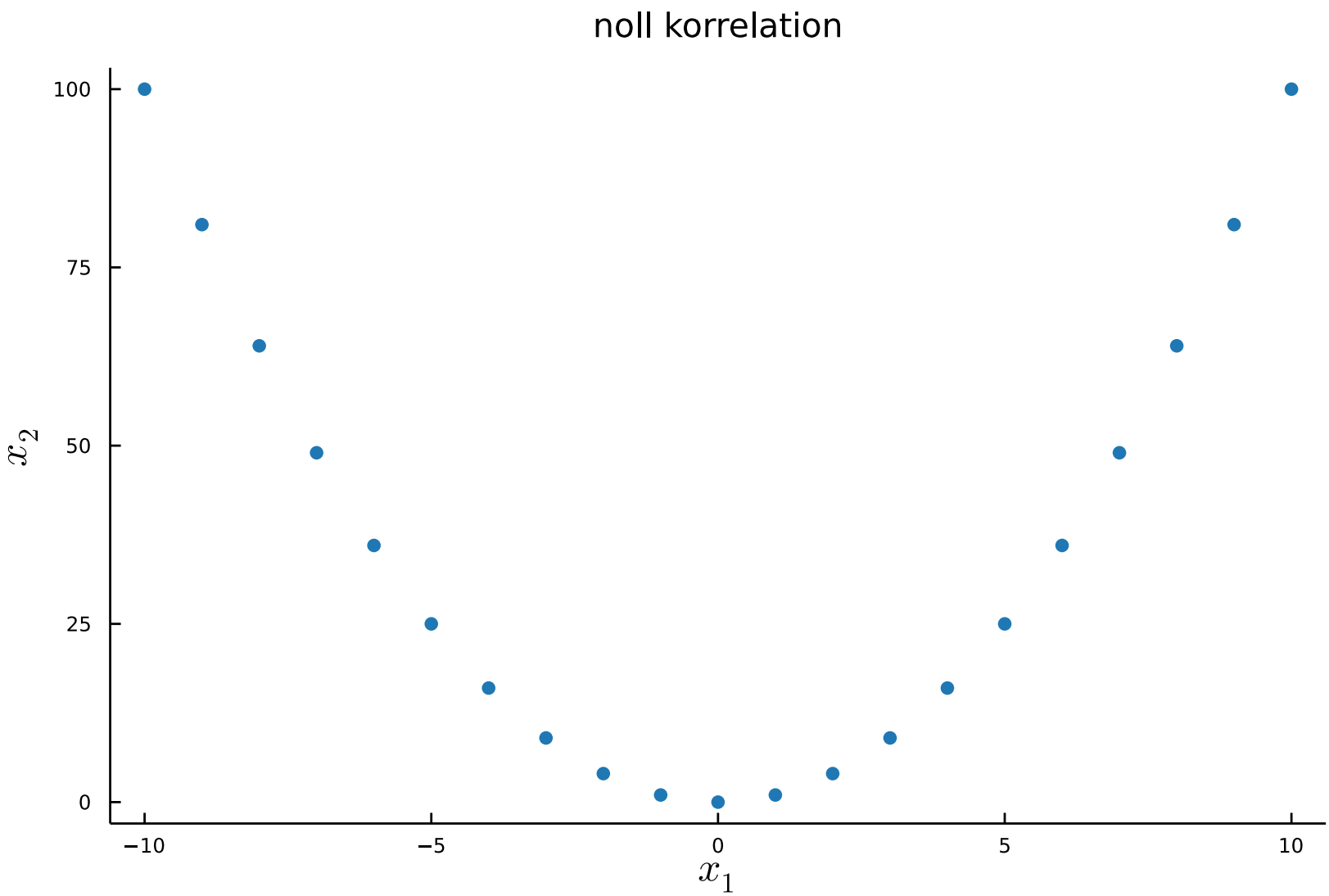
Korrelation



Korrelation hälsobudget vs livslängd



Korrelation mäter linjärt samband



Regression är korrelation, inte kausalitet

- Regression handlar om **korrelation**. **Samvariation**.
- Korrelation kan användas för **prediktion**.
- **Kausala samband** (orsak $\rightarrow$ verkan):
  - Studietimmar $\rightarrow$ Tentaresultat.
  - Smärtstillande $\rightarrow$ Smärtlindring.
  - Marknadsföring $\rightarrow$ Försäljning.

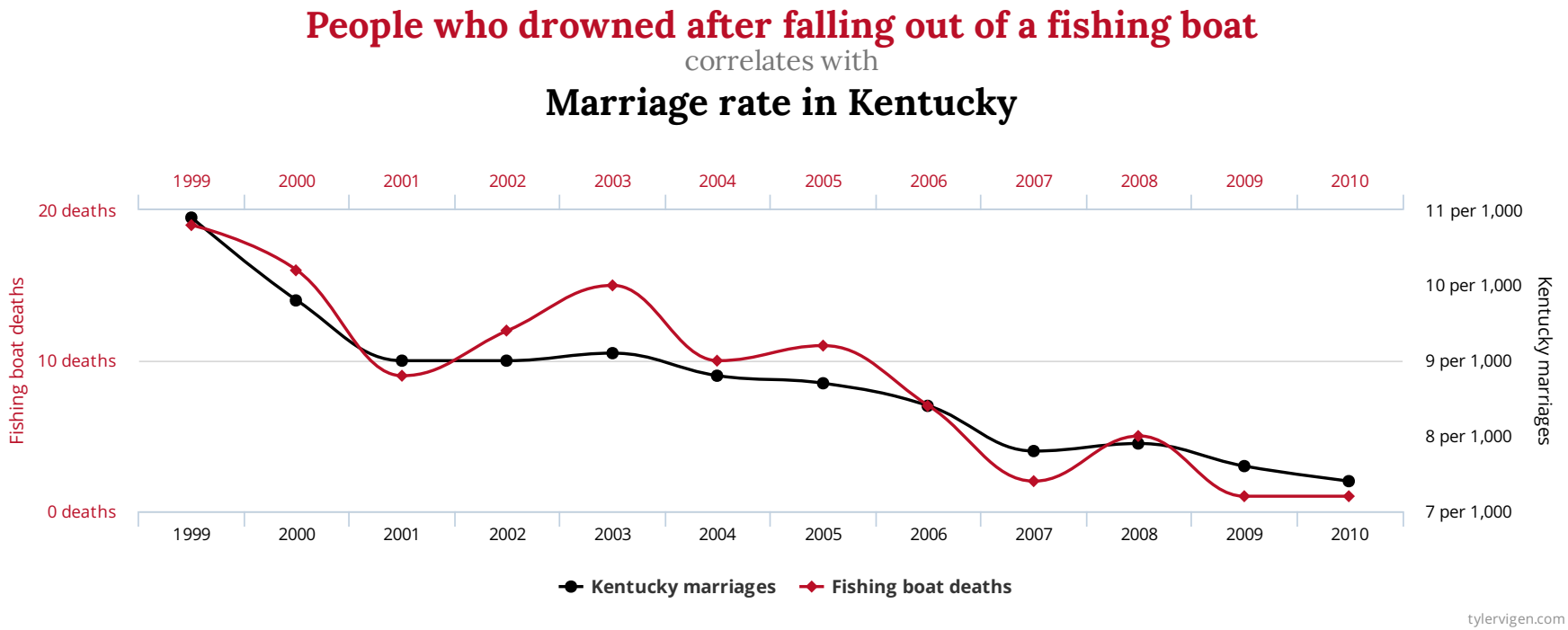
  Eller kan det också vara tvärtom? 🤔

David Hume
Filosof

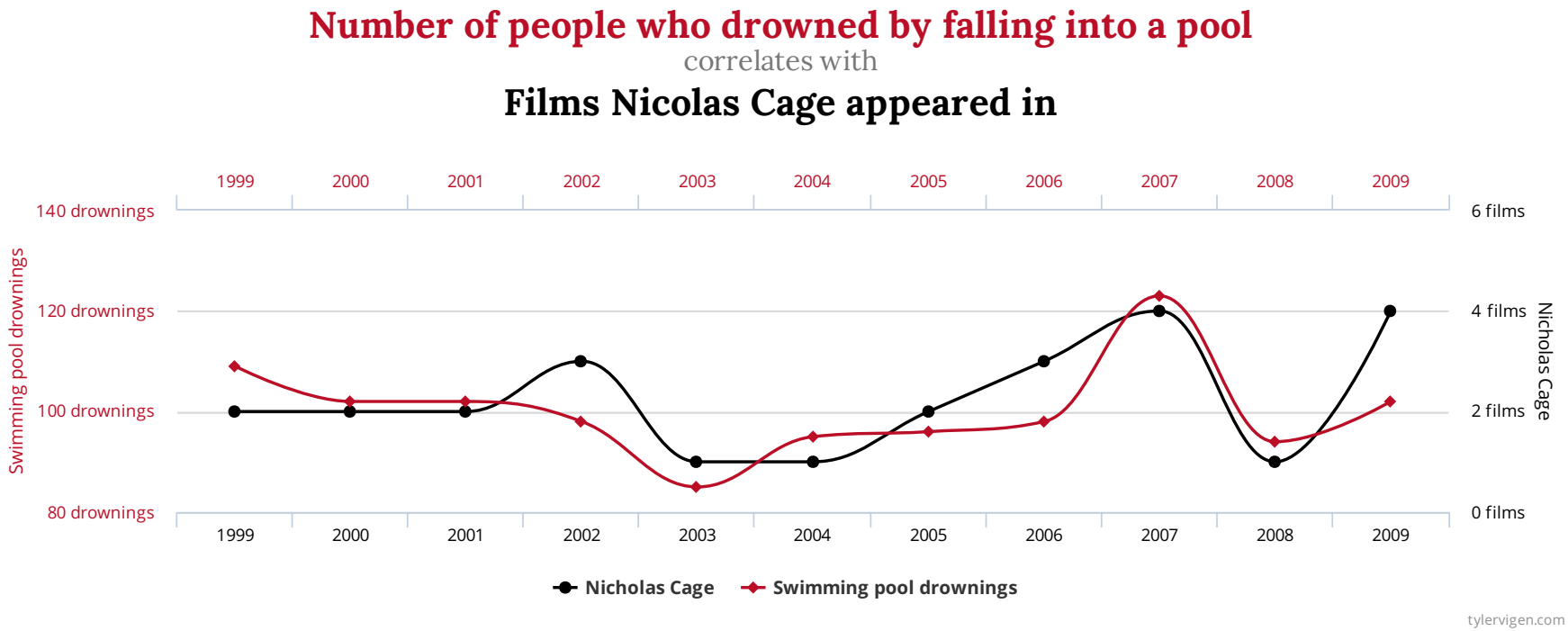
Donald Rubin
Statistiker

Judea Pearl
Datavetare

Korrelation innebär inte kausalitet $\hat{\rho} = 0.952$



Korrelation innebär inte kausalitet $\hat{\rho} = 0.666$



Variansanalys (Analysis of Variance - ANOVA)

- **ANOVA-uppdelningen:**

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR}}$$

Total variation i y = Oförklarad variation i y + Förklarad variation

$$\text{SST} = \text{SSE} + \text{SSR}$$

- Hälsobudgetdata:
  - $\text{SSE} = 125.082$ (Excelark ovan)
  - $\text{SST} = 181.990$ (kan beräknas med liknande Excelark)
  - $\text{SSR} = \text{SST} - \text{SSE} = 56.908$.

ANOVA i R

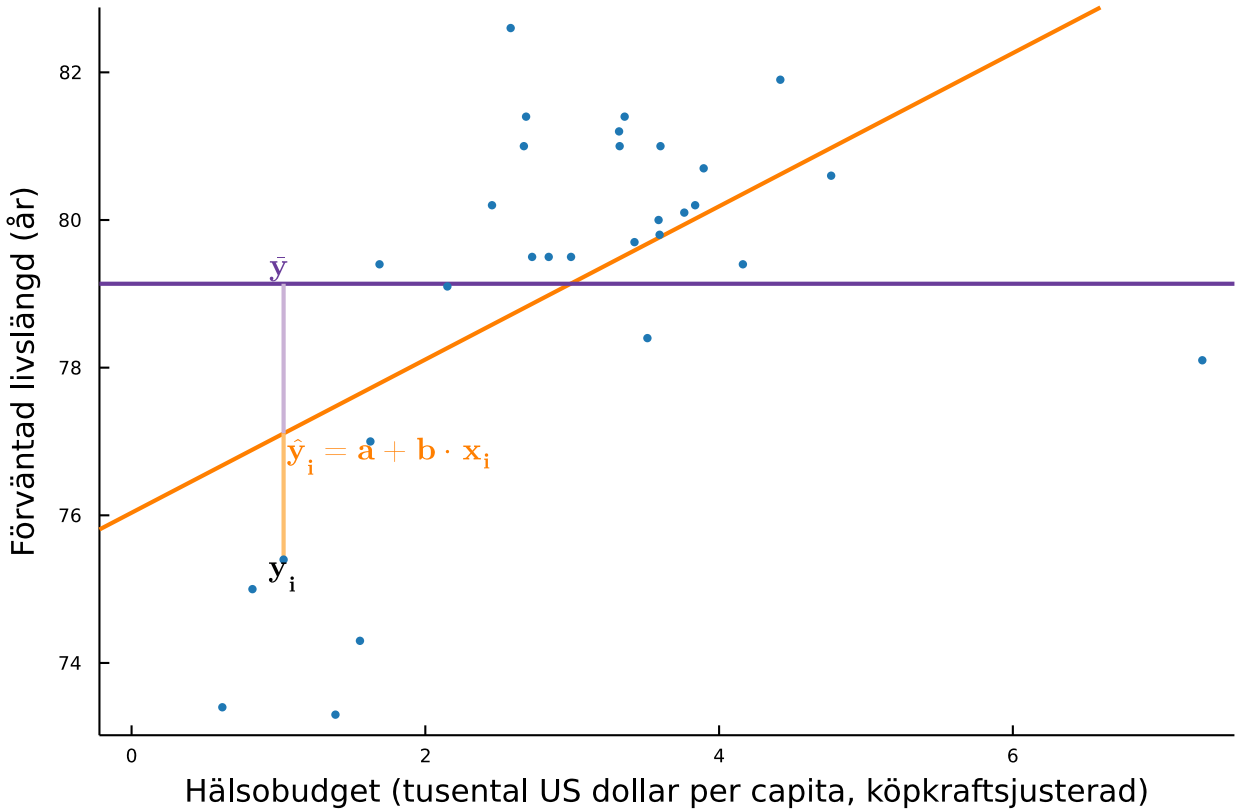
```
Analysis of variance - ANOVA
-------------------------------------------------
      df      SS      MS      F    Pr(>F)
Regr   1  56.907 56.9072 12.739 0.0013164
Error 28 125.082  4.4672
Total 29 181.990

Measures of model fit
-------------------------------------------------
Root MSE       R2   R2-adj
 2.11358  0.31269  0.28815

Parameter estimates
-------------------------------------------------
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)  76.0350    0.95084 79.9663 1.3416e-34
spending      1.0376    0.29071  3.5691 1.3164e-03
```

Variansanalys (ANOVA)

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SSR}}$$



Andel förklarad variation - R^2

- ANOVA:

$$\text{SST} = \text{SSE} + \text{SSR}$$

- Andel förklarad variation (**determinationskoefficienten**)

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

- För regression med **en** förklarande variabel gäller att

$$R^2 = r^2 \quad (r \text{ är korrelationskoefficienten})$$

- Hälsobudgetdata:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{56.908}{181.990} \approx 0.313.$$

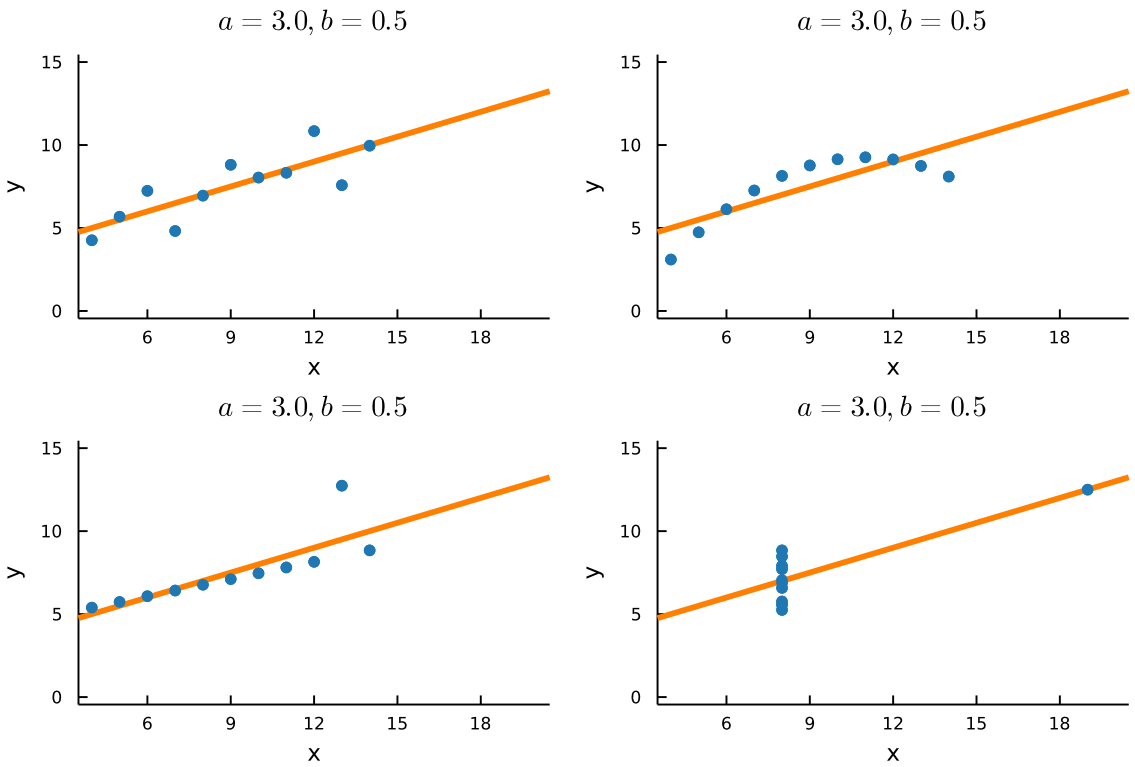
R^2 i R

```
Analysis of variance - ANOVA
------------------------------------------------
      df      SS      MS      F    Pr(>F)
Regr   1  56.907 56.9072 12.739 0.0013164
Error 28 125.082  4.4672
Total 29 181.990

Measures of model fit
------------------------------------------------
Root MSE      R2   R2-adj
 2.11358 0.31269 0.28815

Parameter estimates
------------------------------------------------
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)  76.0350    0.95084 79.9663 1.3416e-34
spending      1.0376    0.29071  3.5691 1.3164e-03
```

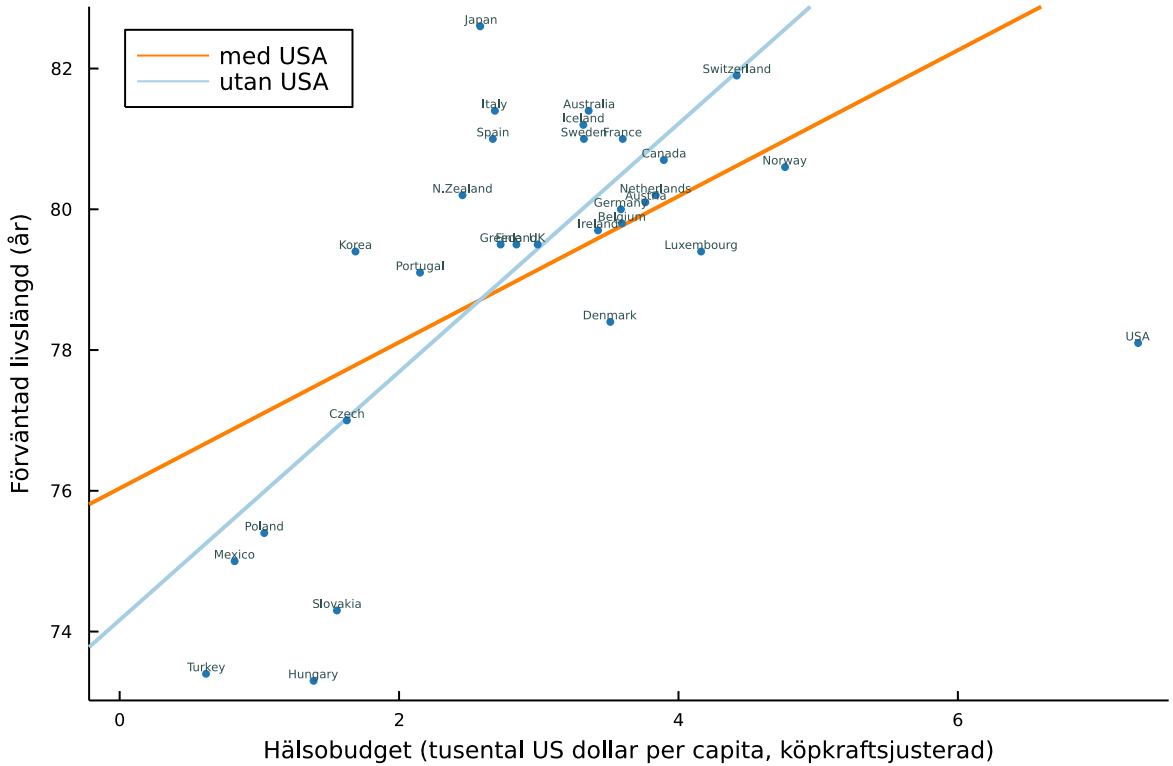
Samma regression på väldigt olika data 🤯



- Samma linjära regression trots väldigt olika samband.
- Se upp för:
  - **icke-linjära samband**
  - **outliers** (både i x och y)
  - **observationer med stor påverkan** på anpassningen.

SUdatasets anscombe

Inflytelserika observationer



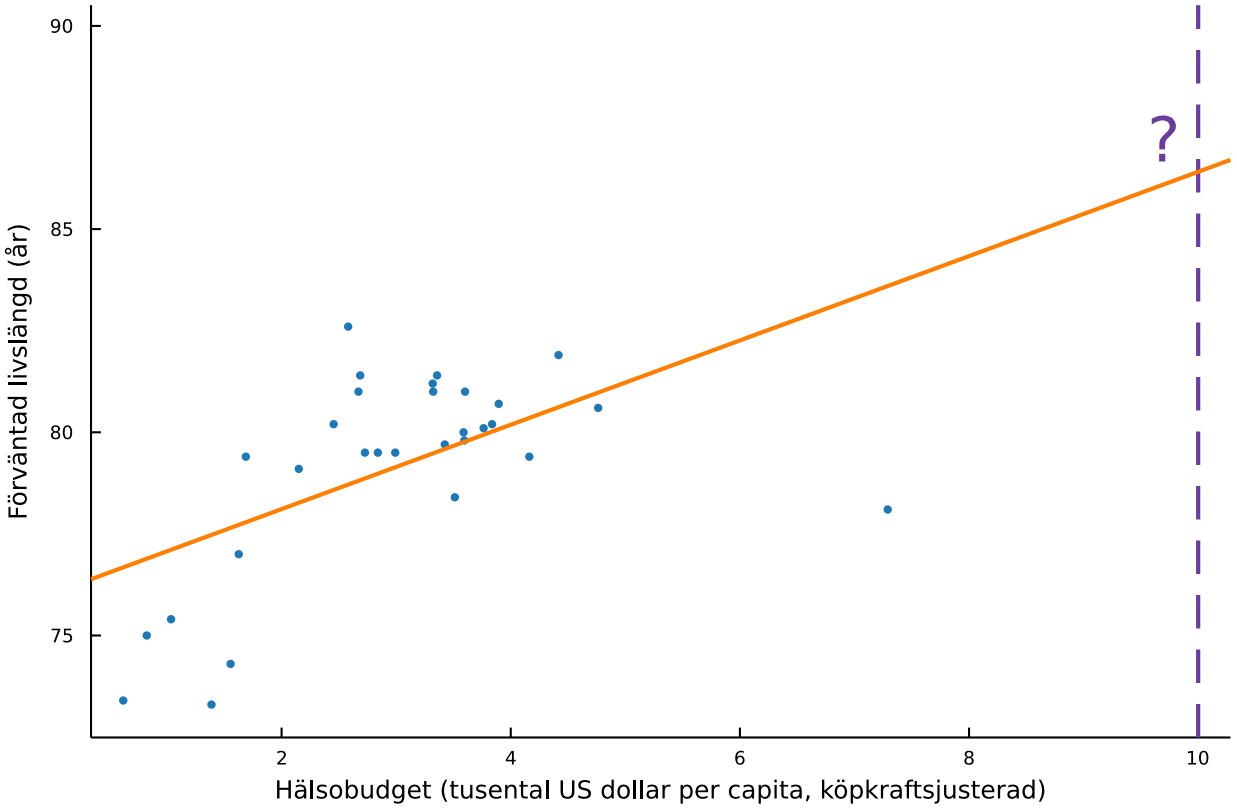
- Med USA

$$\text{livslängd} = 76.035 + 1.038 \cdot \text{hälsobudget}$$

- Utan USA

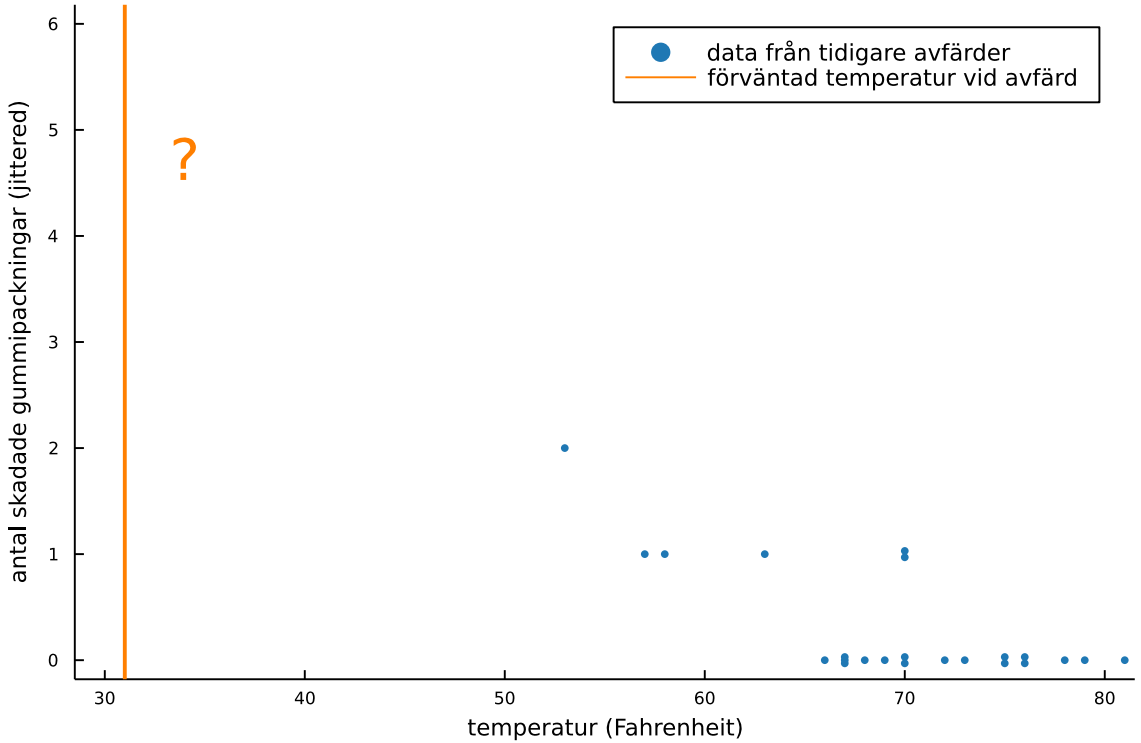
$$\text{livslängd} = 74.164 + 1.763 \cdot \text{hälsobudget}$$

Extrapolering



Extrapolering

- Rymdfärjan Challenger exploderade strax efter start.
- Gummi-packningar (O-rings) hade skadats av kylan.



SUdatasets orings