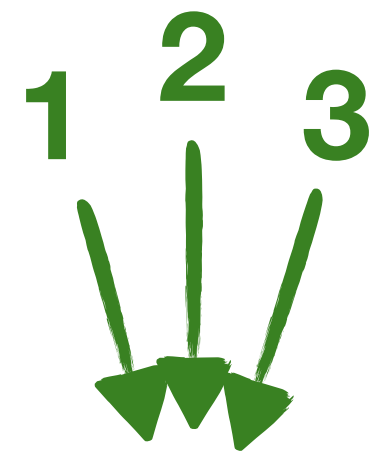


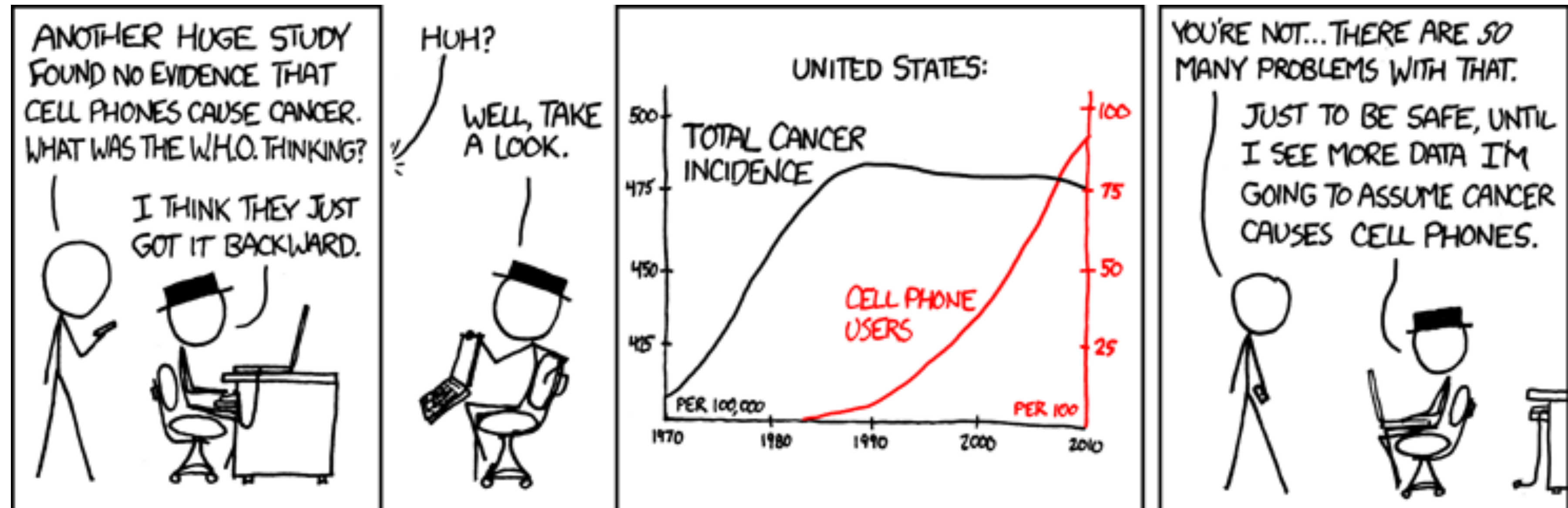


ISTT100y

Uke 9: Enkel lineær regresjon (del 3 av statistisk inferens)

Thea Bjørnland
Institutt for matematiske fag
NTNU

Husk: Gi innspill til referansegruppa om dere har ris/ros/bekymringer/ønsker



https://imgs.xkcd.com/comics/cell_phones.png

Minste kvadratsums rette linje $y = a + bx$

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

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

Uke 1

Hendelser

Stokastiske
variabler

Normalfordeling

Estimering

Konfidensintervall

Hypotesetesting

9 Enkel lineær regresjon

La $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$ være n uavhengige par der x -ene er kjente tall, og Y -ene er stokastiske variabler slik at

$$Y_i | X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma) \text{ for } i = 1, \dots, n$$

En annen måte å formulere modellen på er:

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad e_i \sim N(0, \sigma) \\ \text{for } i = 1, \dots, n$$

Uke 9

Minste kvadratsums estimatorer for β_0 og β_1 :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \text{og}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Estimert regresjonslinje:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Estimator for varians σ^2 er:

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ = \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

og estimator for standardavviket σ , er $S = \sqrt{S^2}$.

Agenda

Eksempel 1: Twist

Eksempel 2: Radius og kulevolum

Eksempel 1:

Chocolate toffee
 $x_1 = 10g$

Eclairs
 $x_2 = 7g$

Daim
 $x_3 = 6g$

Banan
 $x_4 = 8g$

Lakris
 $x_5 = 6g$

Fapp
 $x_6 = 7g$

Fapp
 $x_7 = 6g$

Cocos
 $x_8 = 6g$

Eclairs
 $x_9 = 6g$

Choc. toff
 $x_{10} = 9g$

Banan
 $x_{11} = 7g$

Lakris
 $x_{12} = 5g$

Fransk
nougat
 $x_{13} = 5g$

F. noug
 $x_{14} = 6g$

Caramel
 $x_{15} = 8g$

Nougat
wisp
 $x_{16} = 7g$

Daim
 $x_{17} = 5g$

Golden
toffee
 $x_{18} = 7g$

Eclairs
 $x_{19} = 7g$

Nougat
wisp
 $x_{20} = 7g$



Eksempel 1:

Chocolate toffee
 $x_1 = 10g$

Eclairs
 $x_2 = 7g$

Daim
 $x_3 = 6g$

Banan
 $x_4 = 8g$

Lakris
 $x_5 = 6g$

Fapp
 $x_6 = 7g$

Fapp
 $x_7 = 6g$

Cocos
 $x_8 = 6g$

Eclairs
 $x_9 = 6g$

Choc. toff
 $x_{10} = 9g$

Banan
 $x_{11} = 7g$

Lakris
 $x_{12} = 5g$

Fransk
nougat
 $x_{13} = 5g$

F. noug
 $x_{14} = 6g$

Caramel
 $x_{15} = 8g$

Nougat
visp
 $x_{16} = 7g$

Daim
 $x_{17} = 5g$

Golden
toffee
 $x_{18} = 7g$

Eclairs
 $x_{19} = 7g$

Nougat
visp
 $x_{20} = 7g$

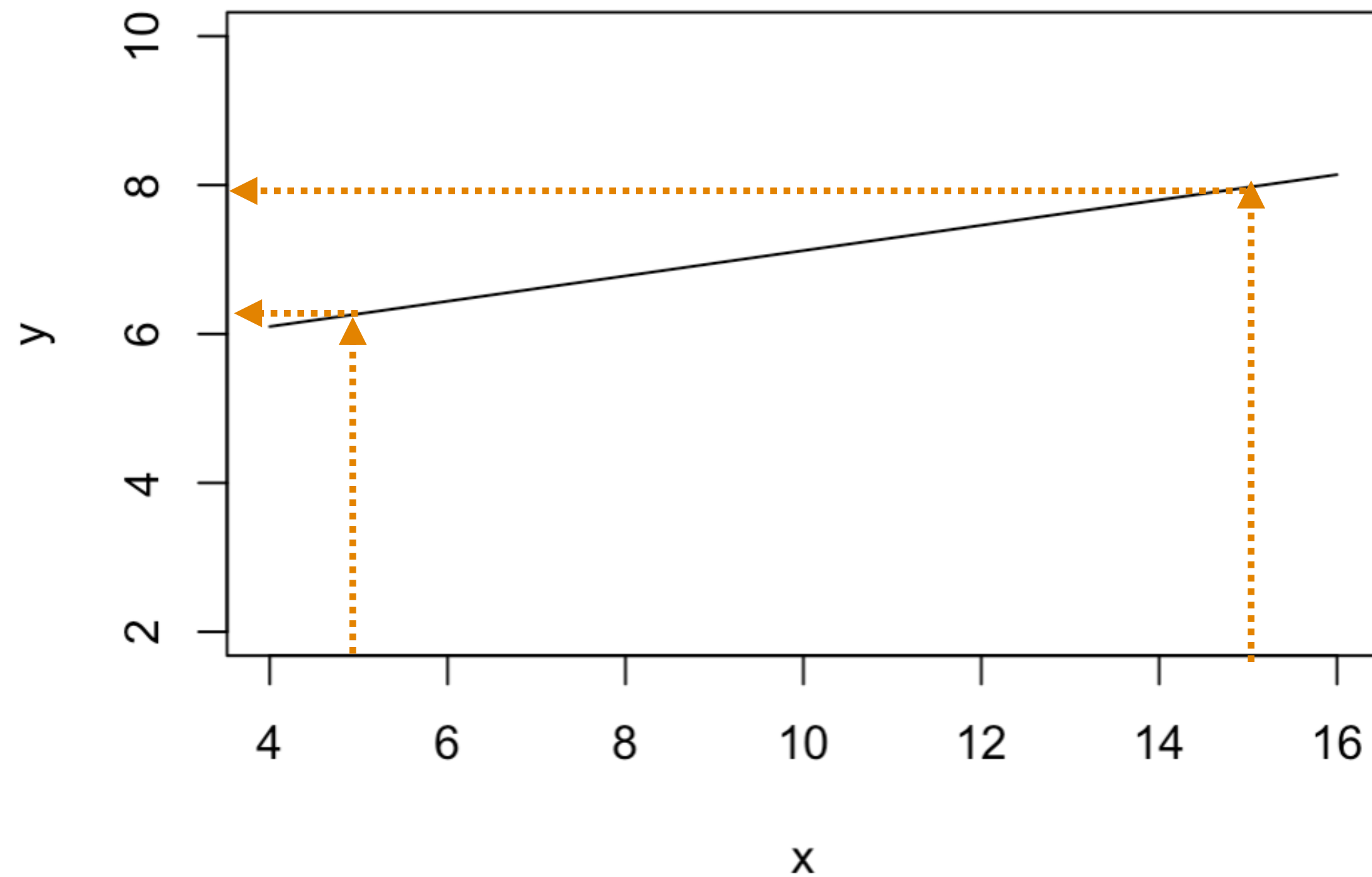


Flere bokstaver i navnet = høyere vekt??

Eksempel 1:

```
regresjon = smf.ols('vekt~bokstaver', data=df).fit()  
regresjon.params
```

Intercept	5.417307
bokstaver	0.167634



Eksempel 1:

Minste kvadratsums estimatorer for β_0 og β_1 :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \text{og}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Estimert regresjonslinje:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Estimator for varians σ^2 er:

$$\begin{aligned} S^2 &= \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \\ &= \frac{1}{n-2} \sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \end{aligned}$$

og estimator for standardavviket σ , er $S = \sqrt{S^2}$.

Konfidensintervall for β_1 :

$$\left[\hat{\beta}_1 - t_{\alpha/2, n-2} \cdot \text{SE}(\hat{\beta}_1), \quad \hat{\beta}_1 + t_{\alpha/2, n-2} \cdot \text{SE}(\hat{\beta}_1) \right], \text{ der}$$

$$\text{SE}(\hat{\beta}_1) = \frac{S}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Testobservator for $H_0 : \beta_1 = 0$ mot $H_0 : \beta_1 \neq 0$:

$$T = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} \stackrel{H_0}{\sim} t_{n-2}$$

Eksempel 1:

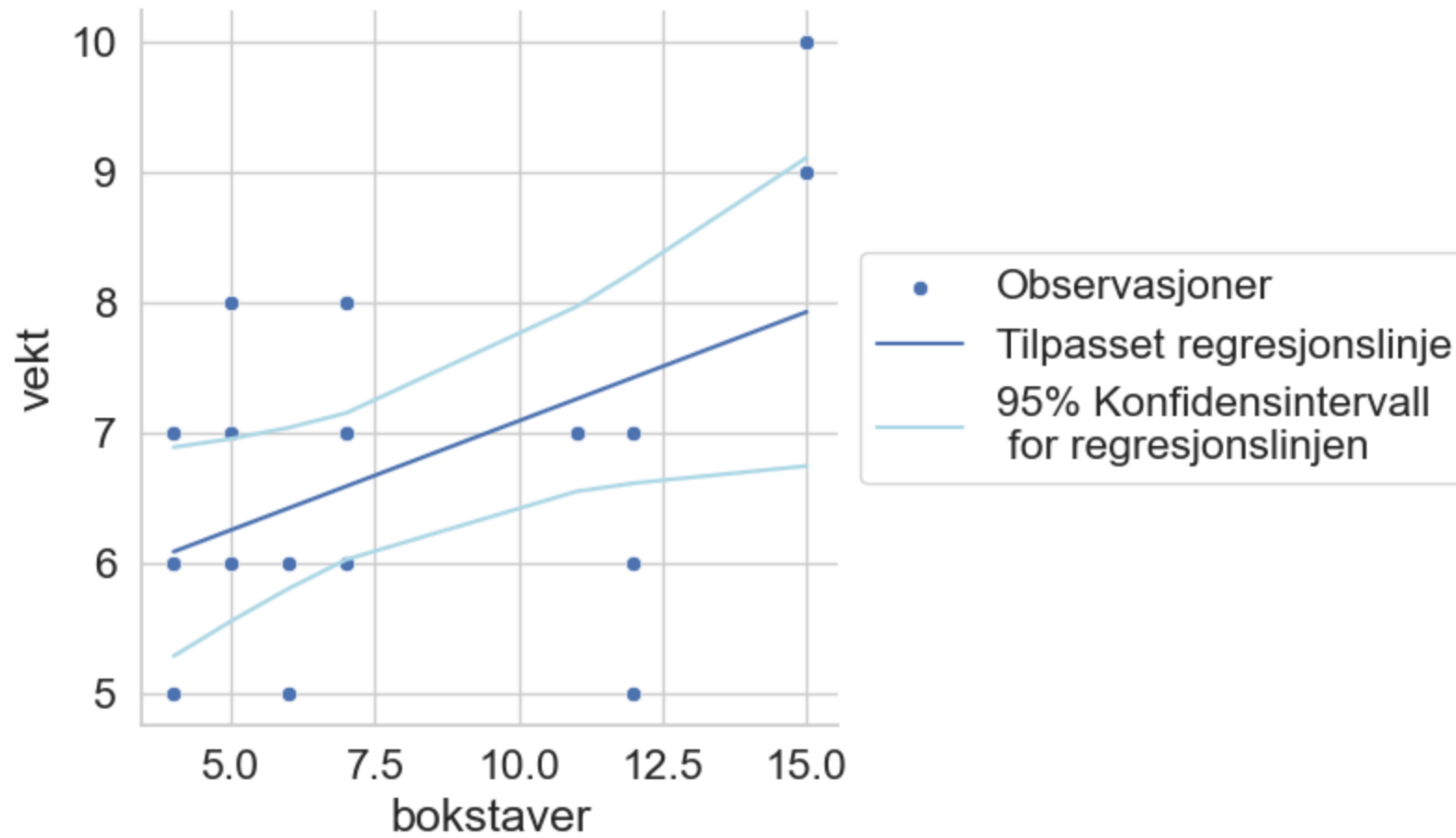
```
regresjon.summary()
```

OLS Regression Results						
Dep. Variable:		vekt		R-squared:		0.236
Model:		OLS		Adj. R-squared:		0.194
Method:		Least Squares		F-statistic:		5.569
Date:		Mon, 14 Oct 2024		Prob (F-statistic):		0.0298
Time:		19:30:21		Log-Likelihood:		-30.305
No. Observations:		20		AIC:		64.61
Df Residuals:		18		BIC:		66.60
Df Model:		1				
Covariance Type:		nonrobust				
	coef	std err	t	P> t	[0.025	0.975]
Intercept	5.4173	0.622	8.716	0.000	4.112	6.723
bokstaver	0.1676	0.071	2.360	0.030	0.018	0.317

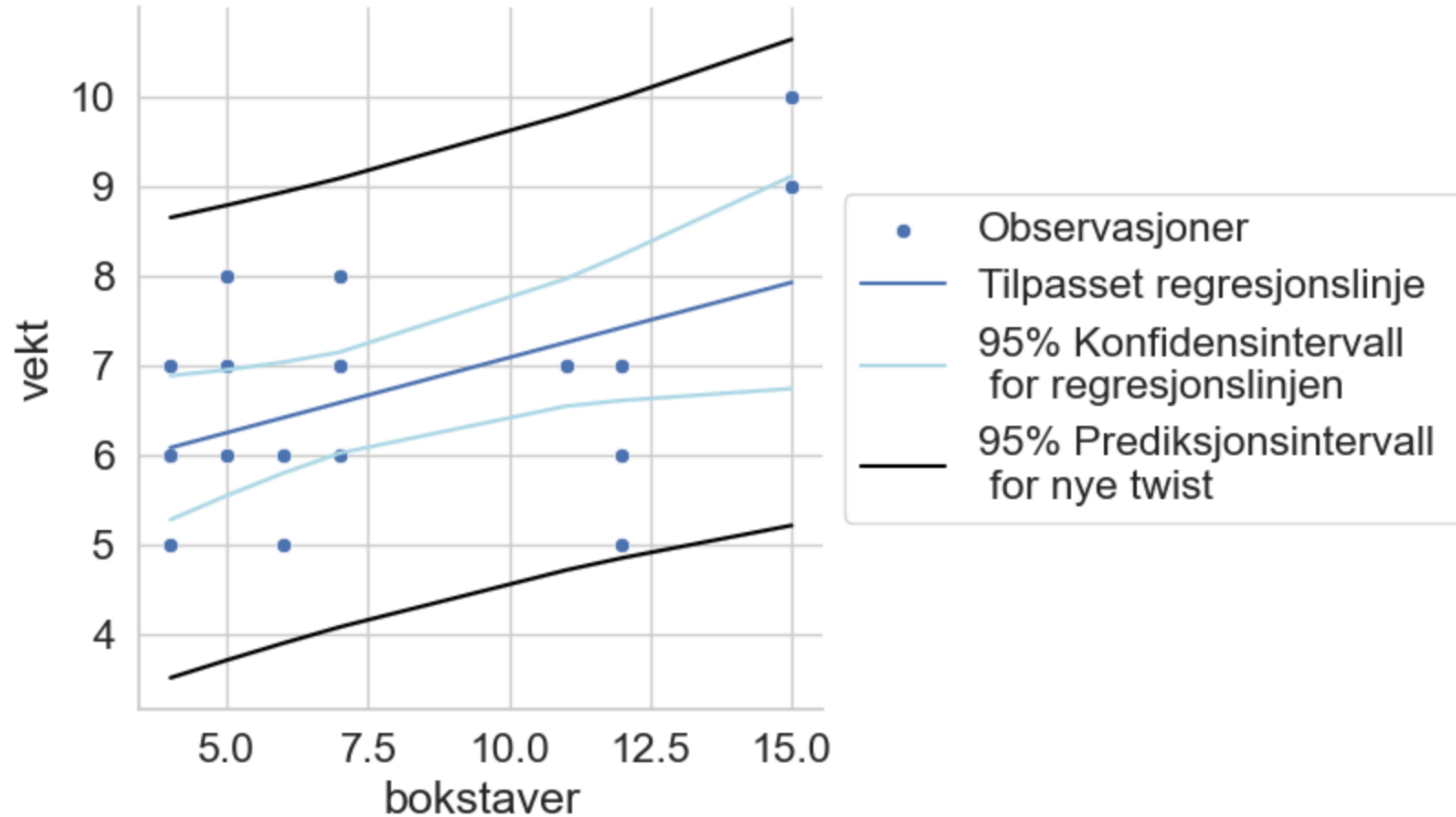
Finn:

- Estimerer på β_0 og β_1
- Standardfeil i estimatene
- Resultatet av en t-test for β_1
- 95% konfidensintervall for β_1

Eksempel 1:

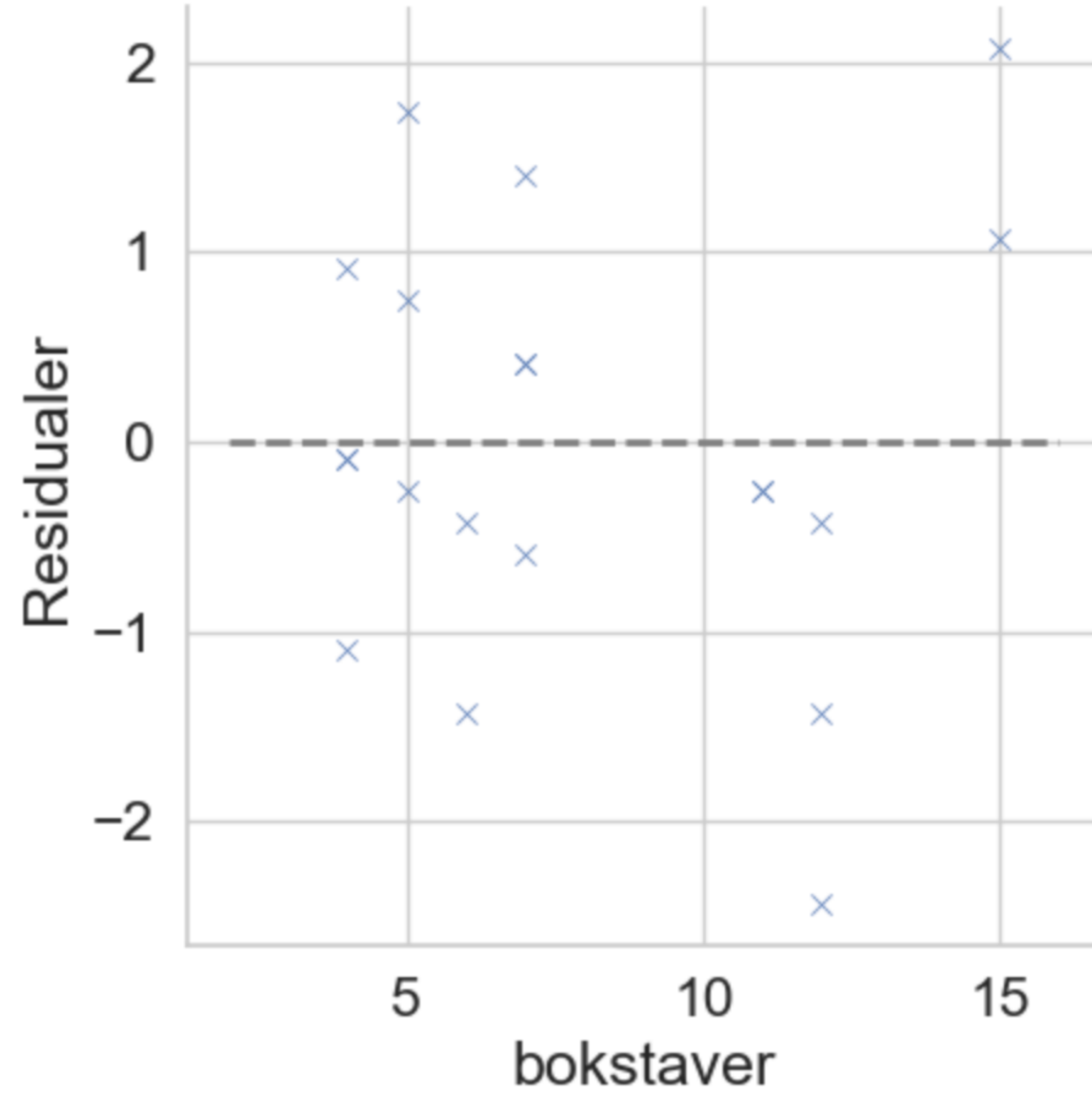


Eksempel 1:

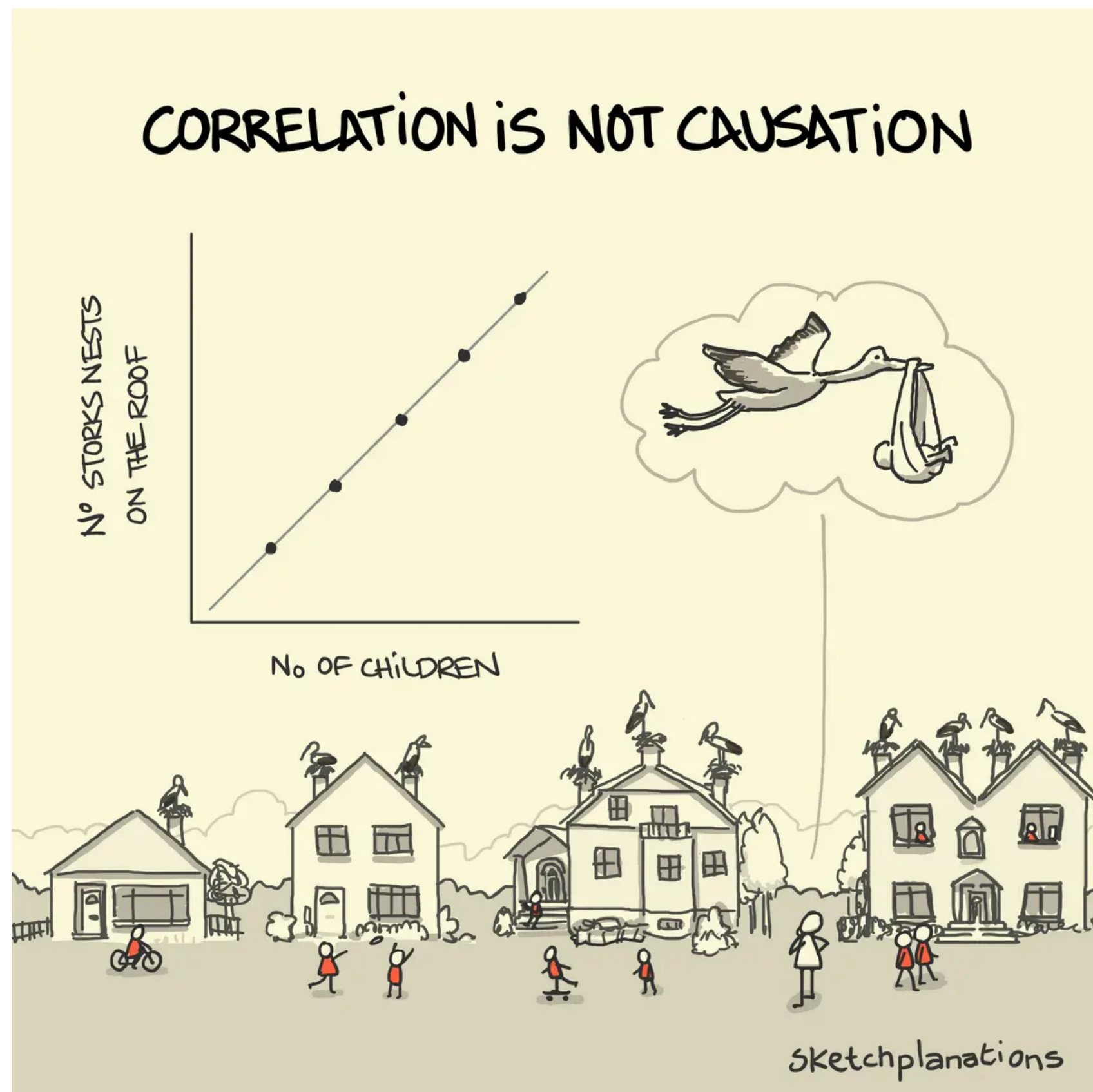


Hva tenker dere om antagelsen om lik variasjon rundt linja for alle verdier av x?

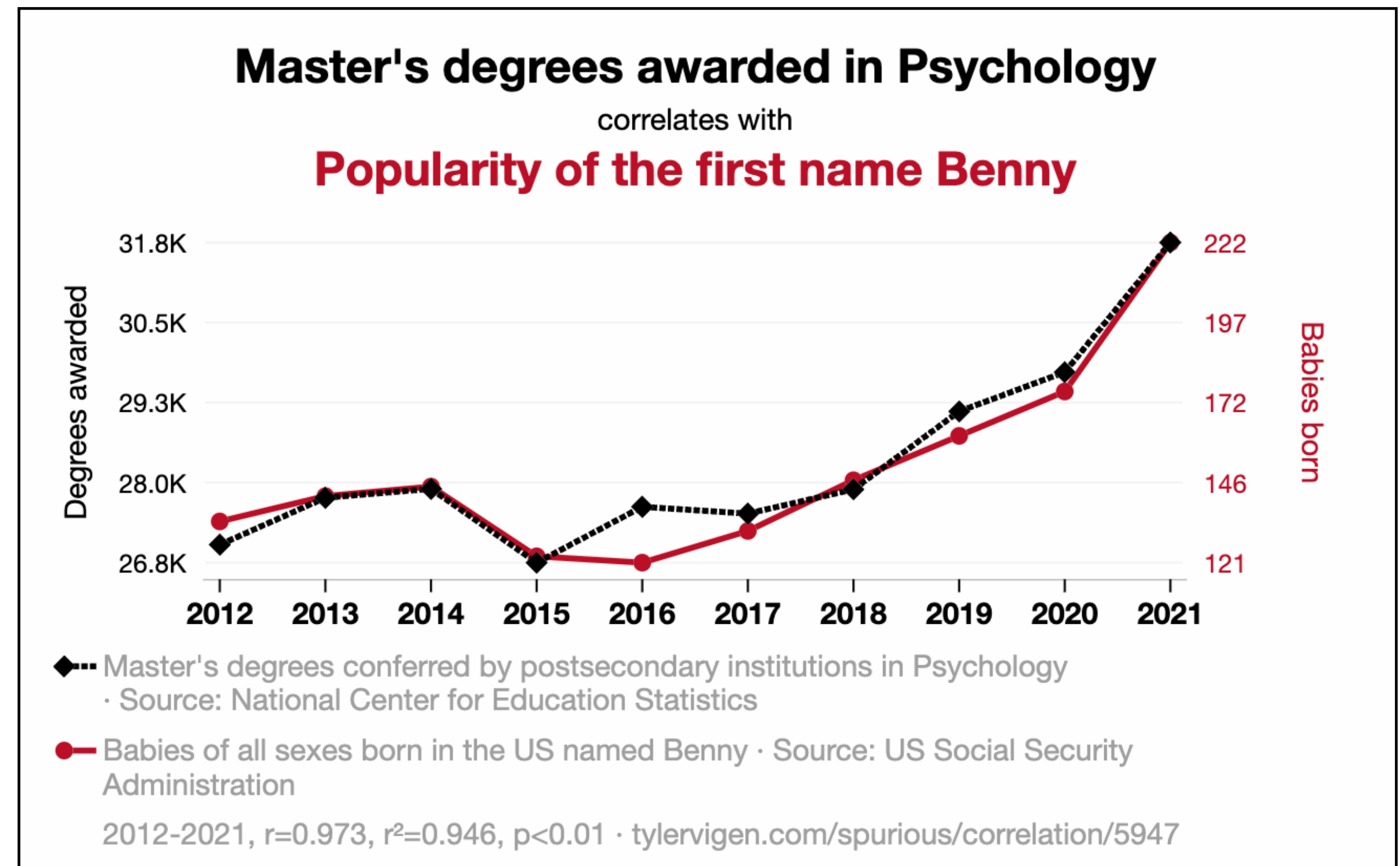
Eksempel 1:



Samvariasjon eller kausalitet?



<https://sketchplanations.com/correlation-is-not-causation>



<https://www.tylervigen.com/spurious-correlations>

Samvariasjon eller kausalitet?

Eksempel 1: Twist *Samvariasjon!*

Eksempel 2: Radius og kulevolum *Kausal sammenheng*

Merk: Regresjonsmodeller 'leter' etter sammenhenger i data

*Kausalitet er noe vi kan undersøke med spesielle studiedesign
(f.eks. "randomised controlled trials" i medisin)*

der en regresjonsmodell kan inngå for å kvantifisere en (lineær) sammenheng

Eksempel 2: π



```
# Målinger
```

```
# Målt radius (fra minste til største)
```

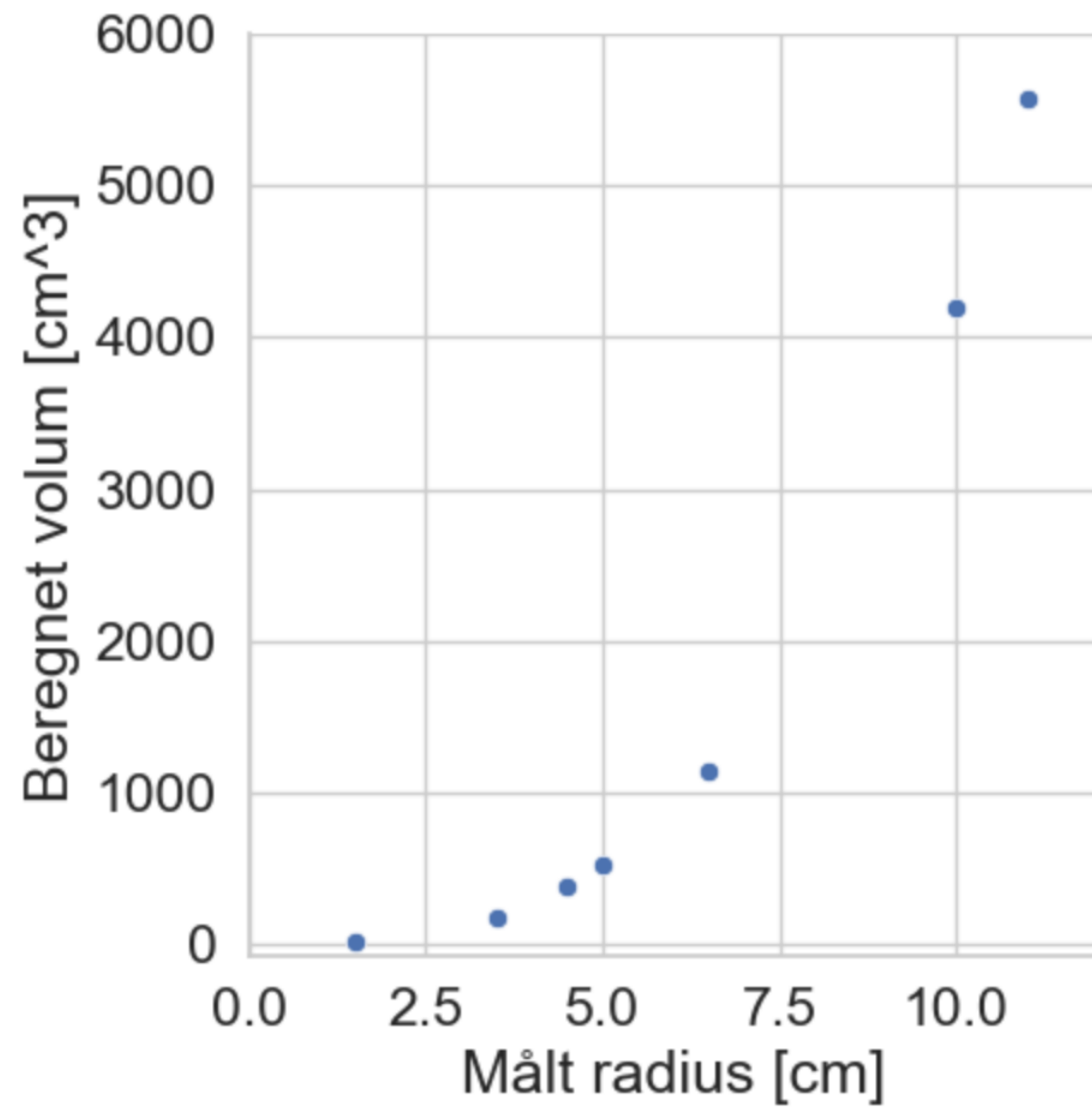
```
r = [1.5 , 3.5, 4.5, 5.0, 6.5, 10.0, 11.0] # cm
```

```
# Målt volum
```

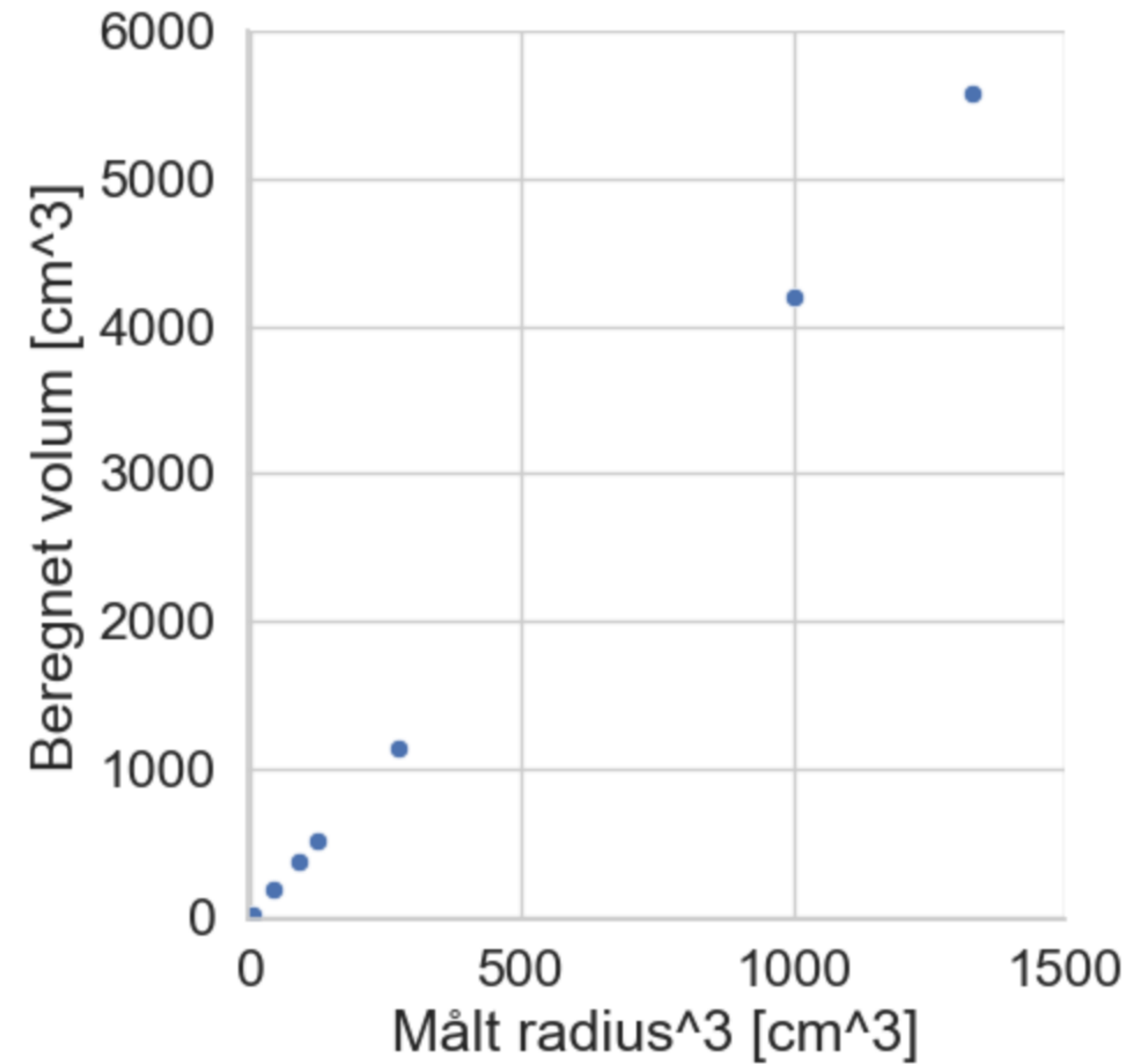
```
v = [13.8, 181.6, 380.1, 521.7, 1149.6, 4195.4, 5574.3 ] # cm^3
```

Eksempel 2: π

Volum mot radius



Volum mot radius³



Eksempel 2: π

OLS Regression Results						
Dep. Variable:		v		R-squared:		1.000
Model:		OLS		Adj. R-squared:		1.000
Method:		Least Squares		F-statistic:		3.189e+06
Date:		Mon, 14 Oct 2024		Prob (F-statistic):		1.05e-15
Time:		19:46:14		Log-Likelihood:		-16.579
No. Observations:		7		AIC:		37.16
Df Residuals:		5		BIC:		37.05
Df Model:		1				
Covariance Type:		nonrobust				
	coef	std err	t	P> t	[0.025	0.975]
Intercept	-0.4217	1.504	-0.280	0.790	-4.287	3.443
r3	4.1909	0.002	1785.663	0.000	4.185	4.197

Finn:

- Estimert regresjonslinje
- Er skjæringspunktet β_0 signifikant ulikt 0? (signifikansnivå 0.05)
- Hva blir estimatet vårt for π ?

Husk oppstart av
prosjektmoduler
neste uke!