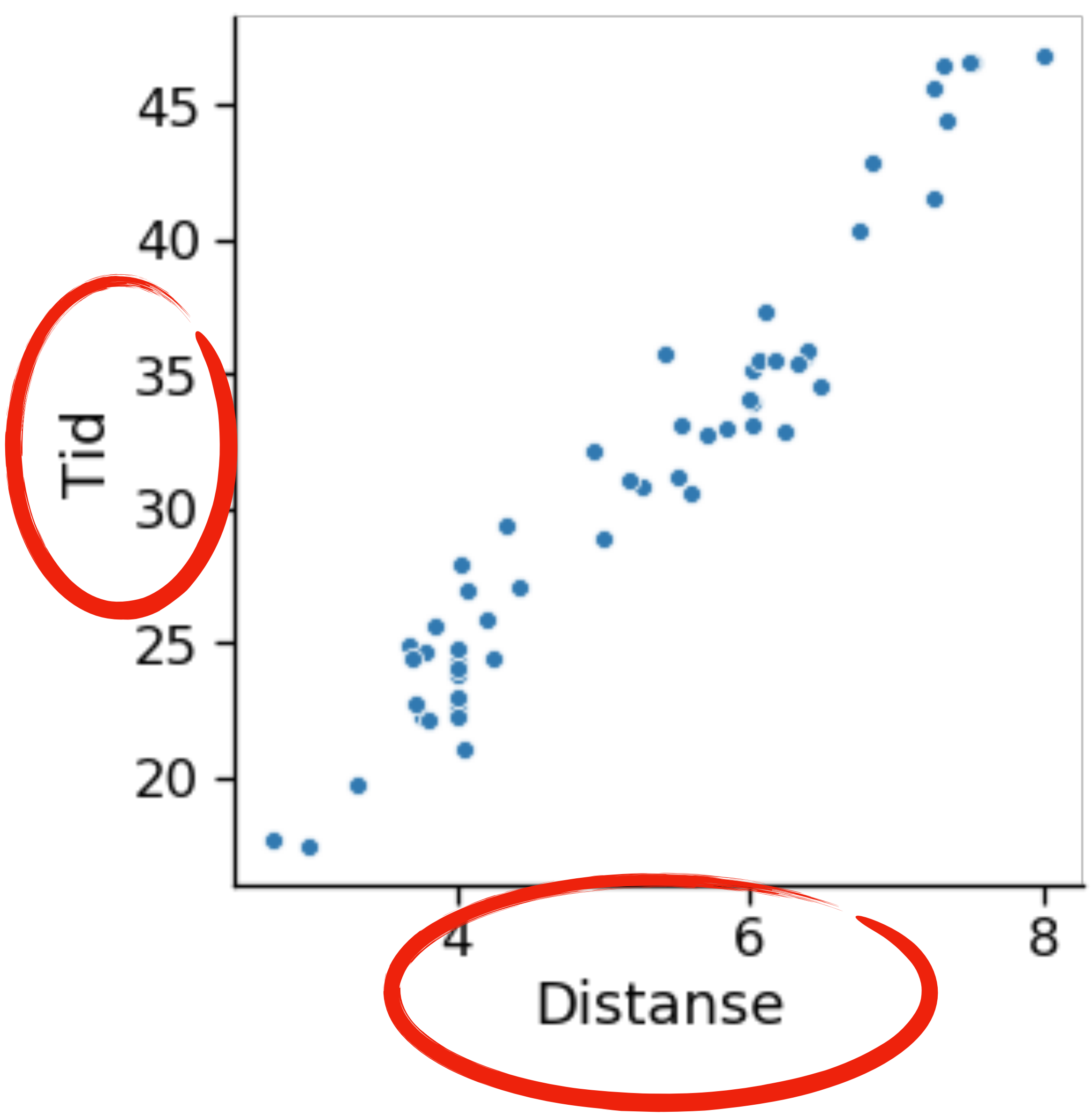
Lineær regresjon
Del 3

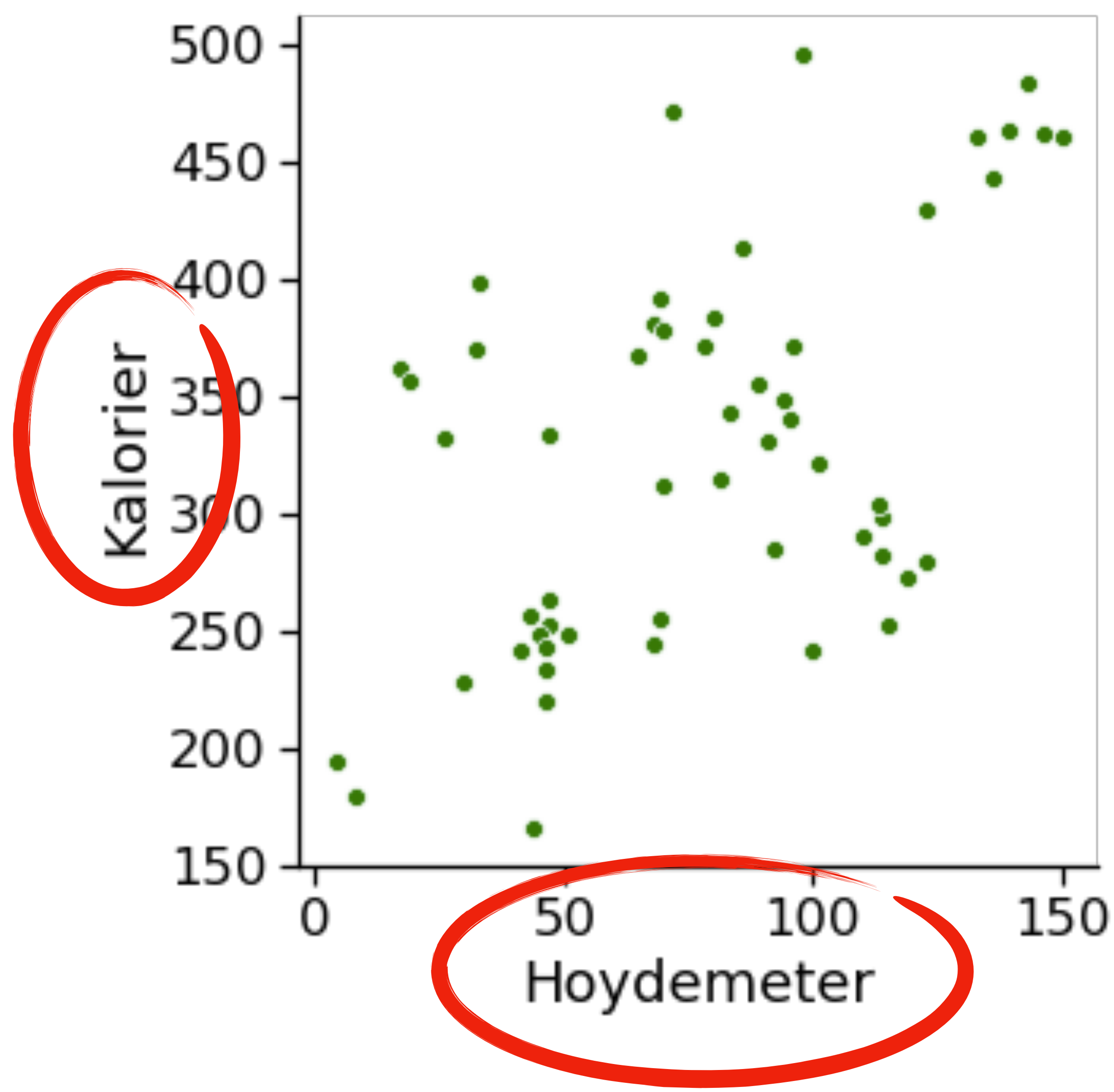
Thea Bjørnland
Institutt for matematiske fag
NTNU

Eksempel: Treningsdata

Data: 55 løpeturer lastet ned fra connect.garmin.com   Juni - Oktober 2020

Parvise observasjoner $(x_1, y_1), \quad (x_2, y_2), \quad \ldots, \quad (x_{55}, y_{55})$

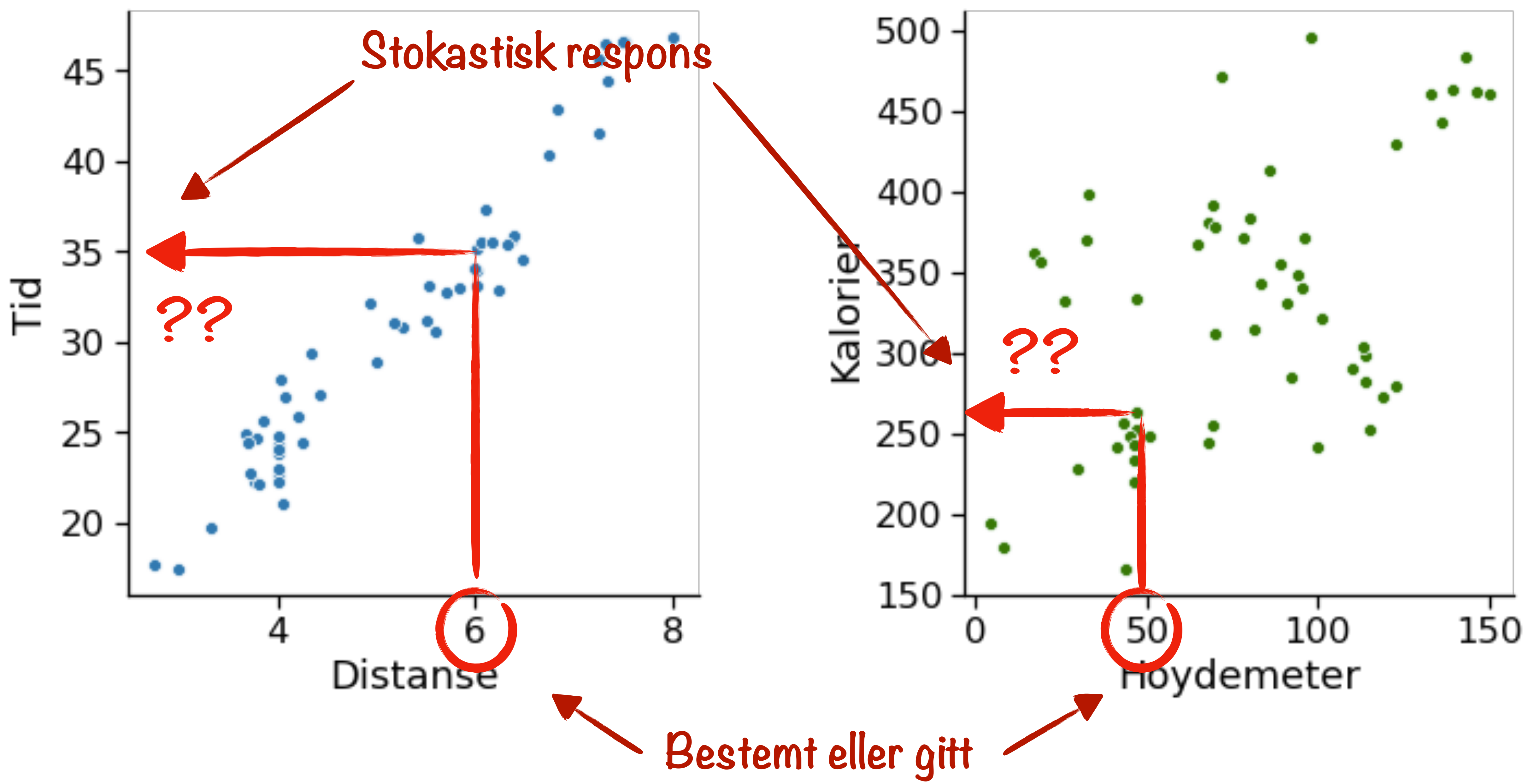




Eksempel: Treningsdata

Data: 55 løpeturer lastet ned fra connect.garmin.com Juni - Oktober 2020

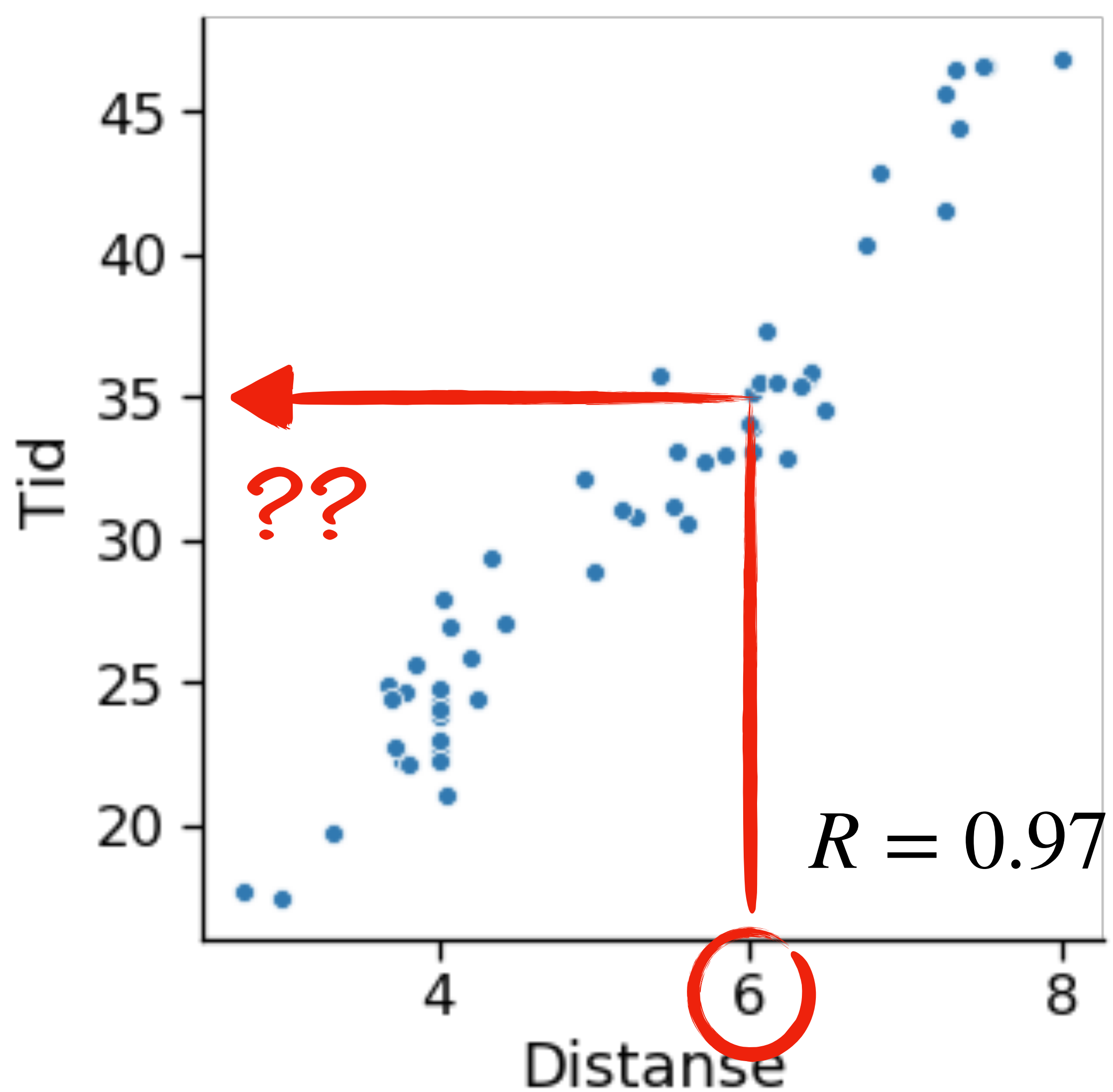
Parvise observasjoner $(x_1, y_1), \quad (x_2, y_2), \quad \ldots, \quad (x_{55}, y_{55})$

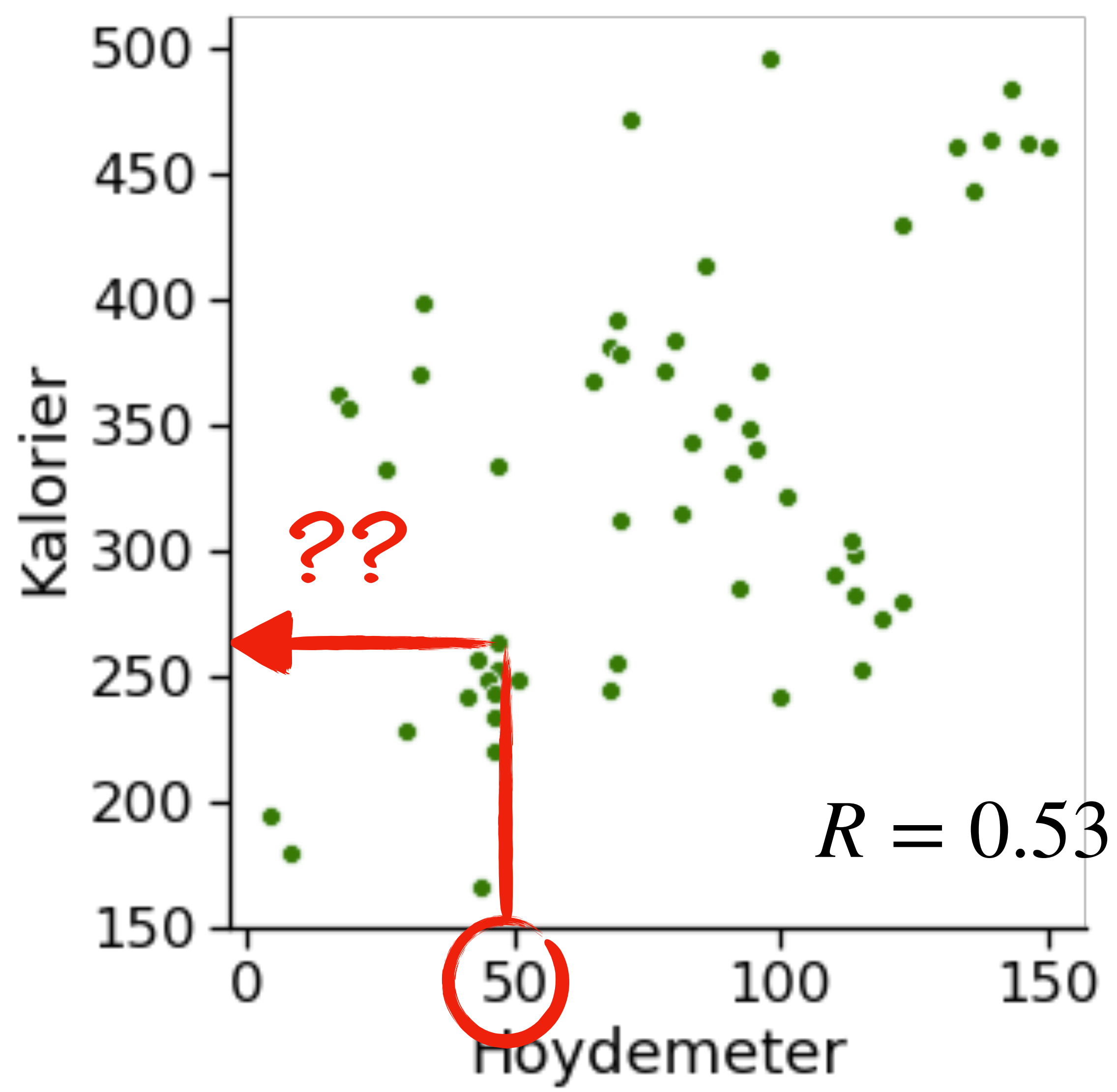


Eksempel: Treningsdata

Data: 55 løpeturer lastet ned fra connect.garmin.com   Juni - Oktober 2020

Parvise observasjoner $(x_1, y_1), \quad (x_2, y_2), \quad \ldots, \quad (x_{55}, y_{55})$





Lineær regresjon: Modellantagelser

Mål: $Y \mid X = x$

Antagelse 1: Lineær sammenheng $E(Y \mid X = x) = \beta_0 + \beta_1 x$ Regresjonslinja

Antagelse 2: Y, betinget på X=x, er normalfordelt med standardavvik σ uansett x

$$Y \mid X = x \sim N(\beta_0 + \beta_1 x, \sigma)$$

3 parametere

Ser vi egentlig at Y endrer seg lineært i x?

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

Hvor god er modellen?

Hvor godt kan kovariaten x forklare / predikere responsen Y?

Estimatorer for parameterne

Tilfeldig utvalg: $(x_1, Y_1), \quad (x_2, Y_2), \quad \ldots, \quad (x_n, Y_n)$ Anta: uavhengige variabler!

$$Y_i \,|\, X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma) \quad i = 1, \ldots, n$$

Estimatorer: $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Forventningsrette... $E(\hat{\beta}_1) = \beta_1 \qquad E(\hat{\beta}_0) = \beta_0$ ... og normalfordelte

$$\hat{\beta}_1 \sim N(\beta_1, SE(\hat{\beta}_1)) \qquad SE(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Estimatorer for parameterne

Tilfeldig utvalg: $(x_1, Y_1), \quad (x_2, Y_2), \quad \ldots, \quad (x_n, Y_n)$ Anta: uavhengige variabler!

$$Y_i \,|\, X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma) \qquad i = 1, \ldots, n$$

Estimatorer: $$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Forventningsrette... $E(\hat{\beta}_1) = \beta_1 \qquad E(\hat{\beta}_0) = \beta_0$ ... og normalfordelte

$$\hat{\beta}_1 \sim N(\beta_1, SE(\hat{\beta}_1)) \qquad SE(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

Konfidensintervall:

$$[\hat{\beta}_1 - z_{\alpha/2} SE(\hat{\beta}_1), \hat{\beta}_1 + z_{\alpha/2} SE(\hat{\beta}_1)]$$

Hypotesetest: $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$

$$Z = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Forkast dersom:

$$|z| > z_{\alpha/2}$$

Estimatorer for parameterne

Tilfeldig utvalg: $(x_1, Y_1), \quad (x_2, Y_2), \quad \ldots, \quad (x_n, Y_n)$ Anta: uavhengige variabler!

$$Y_i \mid X_i = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma) \quad i = 1, \ldots, n$$

Estimatorer: $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Forventningsrette... $E(\hat{\beta}_1) = \beta_1 \qquad E(\hat{\beta}_0) = \beta_0$ ... og normalfordelte

$$\hat{\beta}_1 \sim N(\beta_1, SE(\hat{\beta}_1)) \qquad SE(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

ukjent??

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

Konfidensintervall:

$$[\hat{\beta}_1 - \cancel{z_{\alpha/2}} SE(\hat{\beta}_1), \hat{\beta}_1 + \cancel{z_{\alpha/2}} SE(\hat{\beta}_1)]$$

$$t_{\alpha/2, n-2}$$

Hypotesetest: $H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$

$$\cancel{Z} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

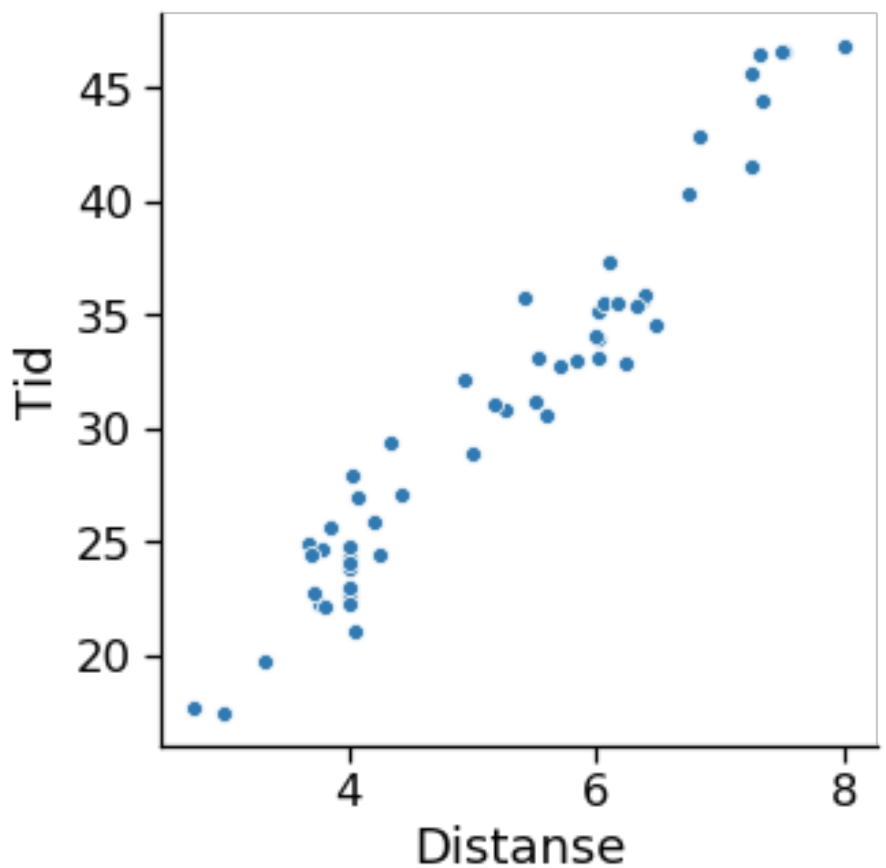
$$T$$

Forkast dersom:

$$\cancel{|z| > z_{\alpha/2}}$$

$$|t| > t_{\alpha/2, n-2}$$

Eksempel: Treningsdata




```
X = df[["Distanse"]]
X = sms.add_constant(X)
y = df["Tid"]

modell = sms.OLS(y, X).fit()

modell.summary()
```

R-squared:	0.940

$\hat{\beta}_0$

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6003	1.057	1.513	0.136	-0.521	3.721
Distanse	5.6525	0.196	28.782	0.000	5.259	6.046

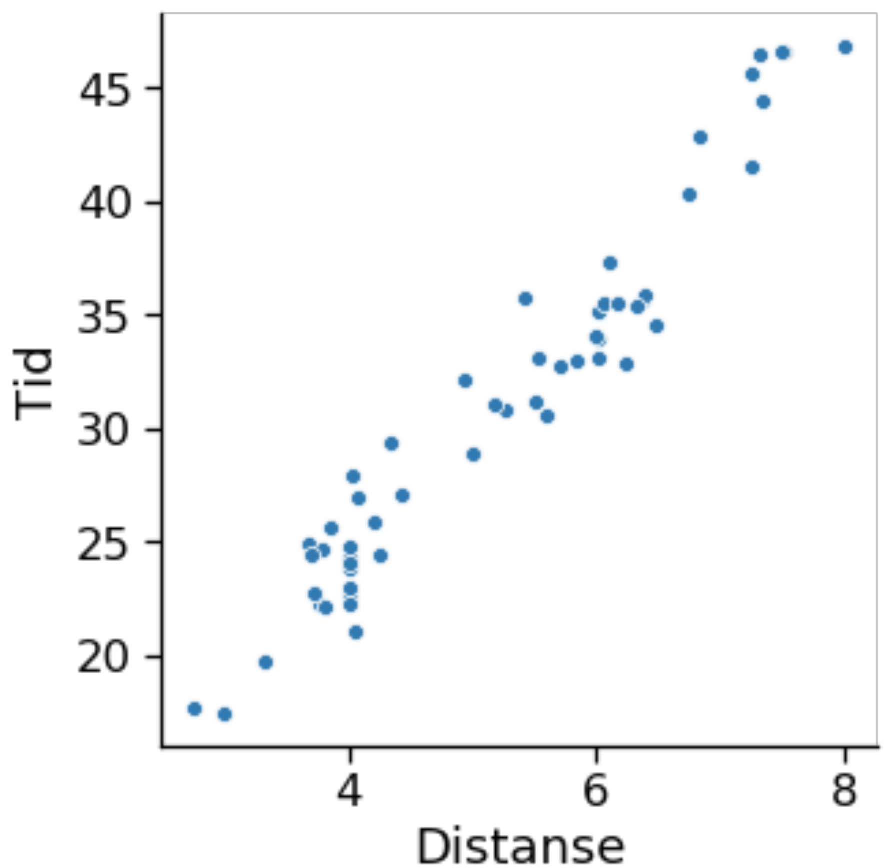
95% KI for β_1

$\hat{\beta}_1$ $SE(\hat{\beta}_1)$

Forkast dersom $|t| > t_{0.025,53} \approx z_{0.025} = 1.96$

Forkast. Linja er skrå

Eksempel: Treningsdata




```python
X = df[["Distanse"]]
X = sms.add_constant(X)
y = df["Tid"]

modell = sms.OLS(y, X).fit()

modell.summary()
```

R-squared:	0.940

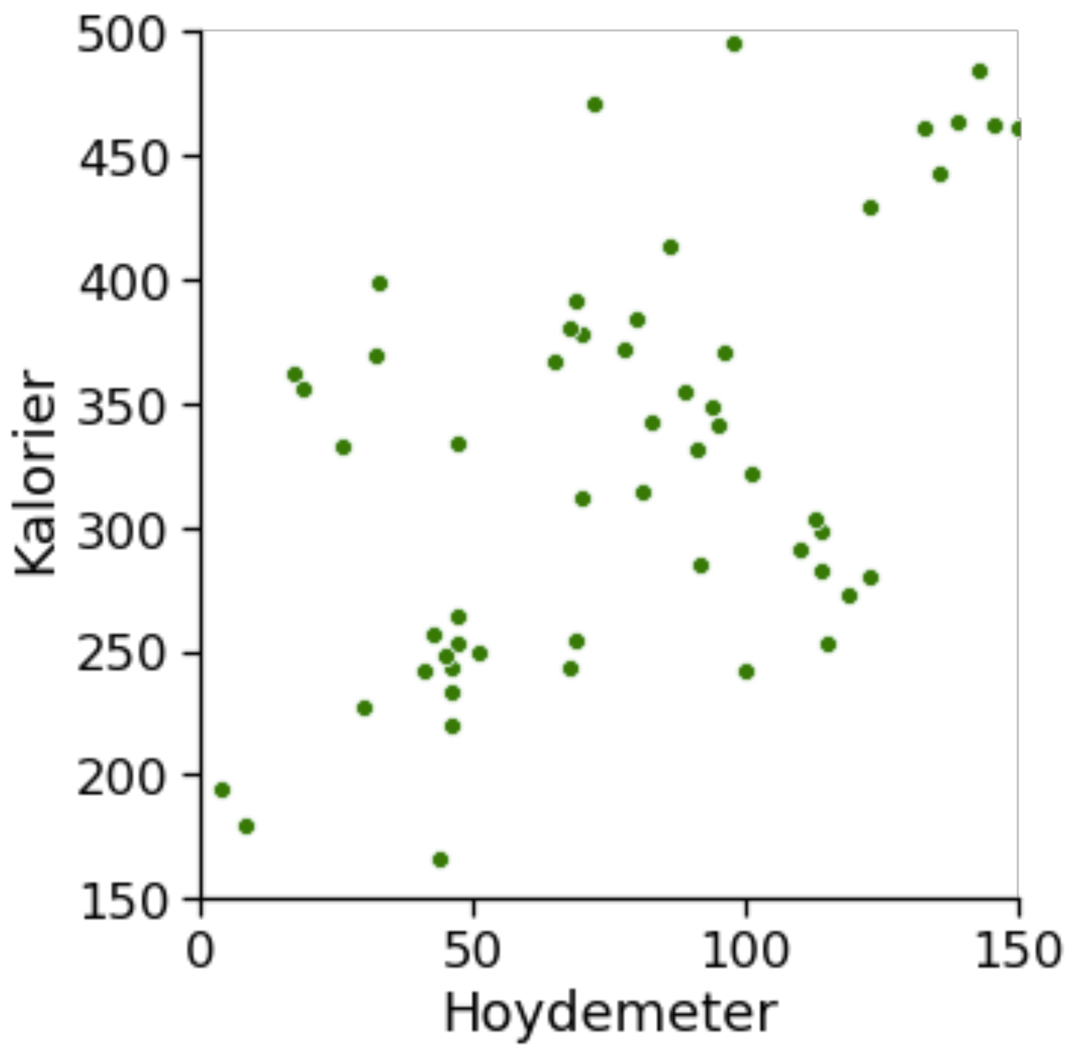
"Andelen varians i Y som forklares av den lineære regresjonsmodellen"

$$R = \frac{\hat{Cov}(X, Y)}{s_x \cdot s_y} \approx 0.97$$

$$R^2 = \frac{SST - SSE}{SST} \qquad SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

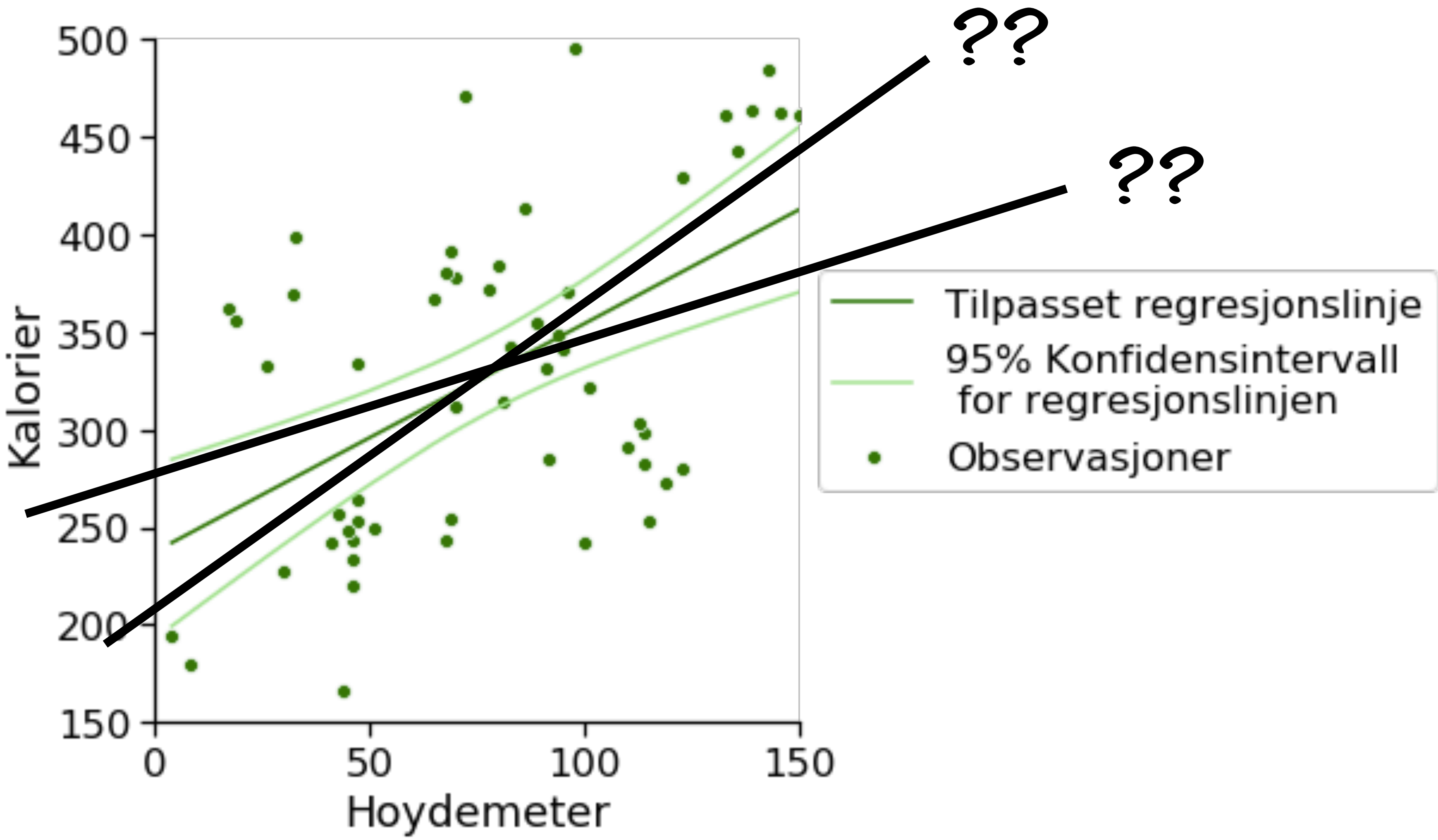
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Eksempel: Treningsdata



R-squared:	0.279

Eksempel: Treningsdata



Eksempel: Treningsdata