Lineær regresjon
Del 2

Thea Bjørnland
Institutt for matematiske fag
NTNU

Eksempel: Treningsdata

X : distanse løpt (km)

Y : tid brukt (minutter)

Bestemt:
$x = 5$ km

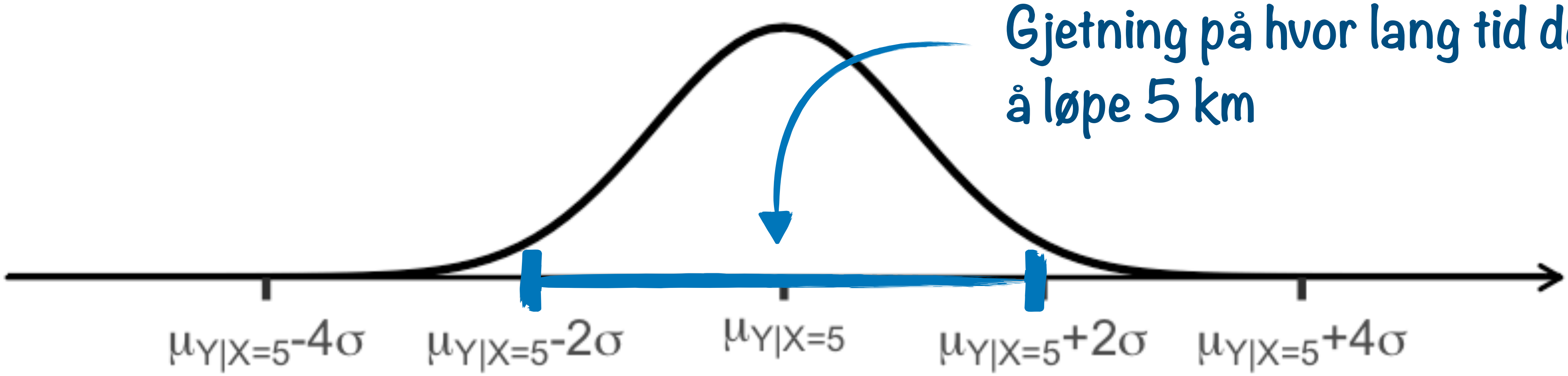
Hvis jeg skal løpe 5 km i morgen,
hvor langt tid bør jeg regne med å bruke?

$Y \mid X = x$ $\qquad E(Y \mid X = 5) = \,?$

Estimere parametere
i en normalfordeling

Antagelse $\quad Y \mid X = 5 \sim N(\mu_{Y|X=5}, \sigma)$

Gjetning på hvor lang tid det vil ta
å løpe 5 km



ca 95%

Eksempel: Treningsdata

X : distanse løpt (km)

Y : tid brukt (minutter)

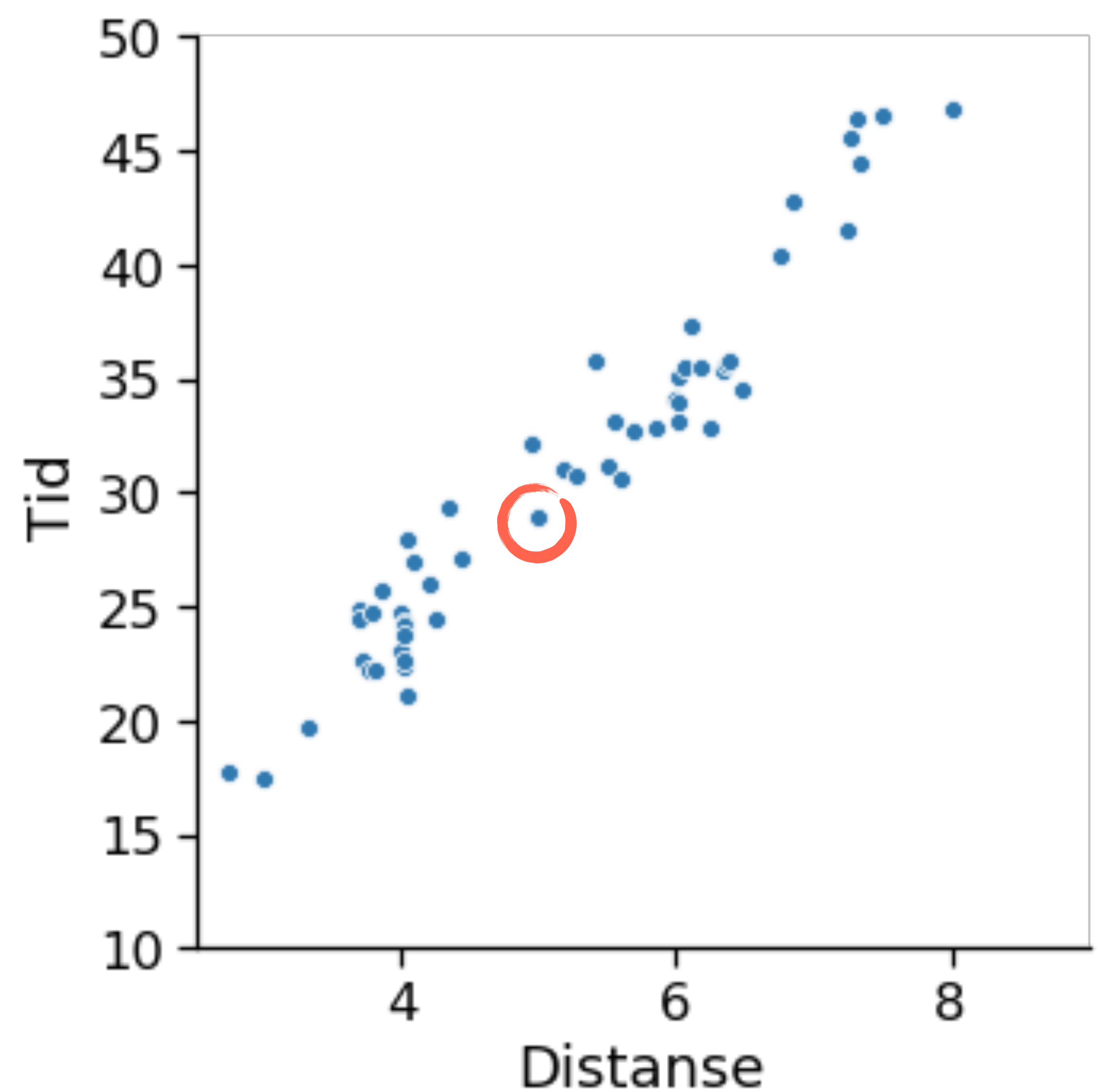
Bestemt:
$x = 5$ km

Hvis jeg skal løpe 5 km i morgen,
hvor langt tid bør jeg regne med å bruke?

$Y \,|\, X = x$ $\qquad E(Y \,|\, X = 5) = \,?$

Bruke hele datasettet?

Antagelse $\quad Y \,|\, X = 5 \;\sim N(\mu_{Y|X=5}, \sigma)$



Lineær regresjon: Modellantagelser

X : distanse løpt (km) Y : tid brukt (minutter) **Mål:** $Y \,|\, X = x$

Respons

Kovariat

Forklaringsvariabel

Lineær regresjon: Modellantagelser

X : distanse løpt (km) Y : tid brukt (minutter) **Mål:** $Y \mid X = x$

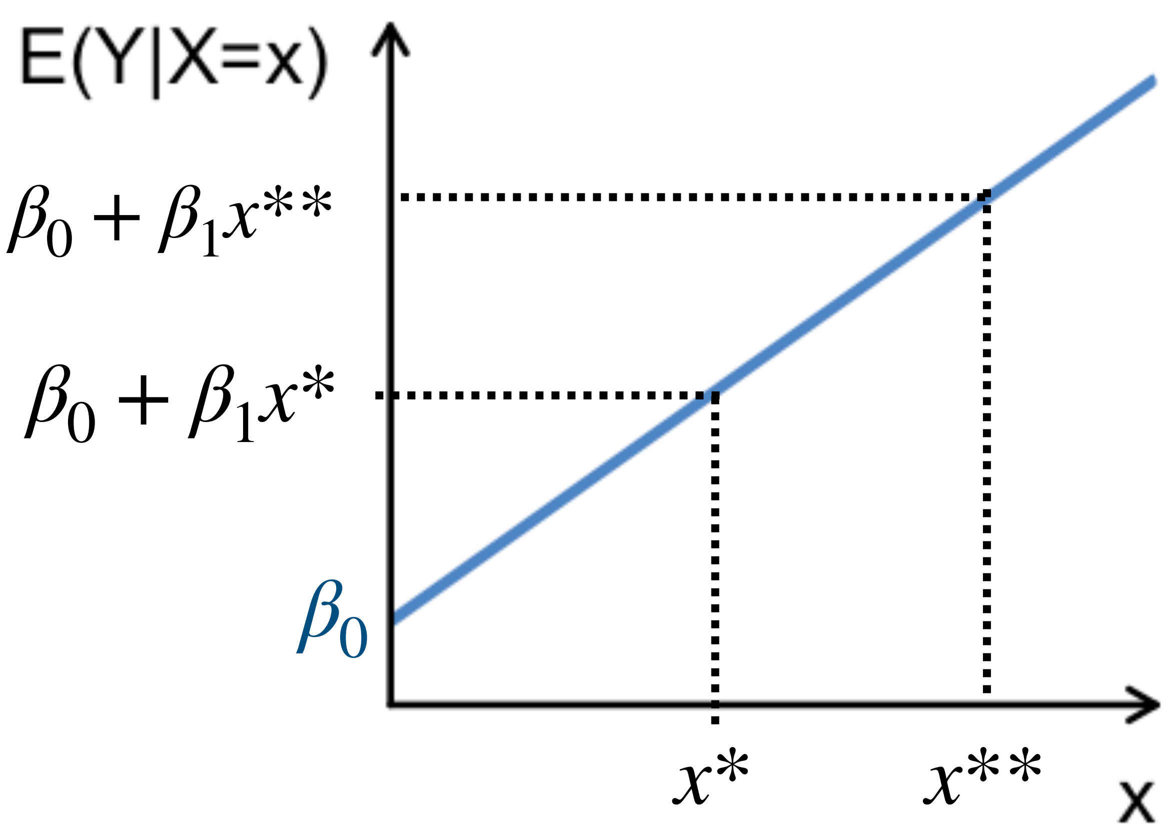
Antagelse 1: Lineær sammenheng $\rho(X, Y) \neq 0$ (Plott observasjoner av x og y)

Forventningsverdien til Y, betinget på X=x, er lineær i x

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

Stigningstallet til linja

Kryssningspunkt y-aksen



Lineær regresjon: Modellantagelser

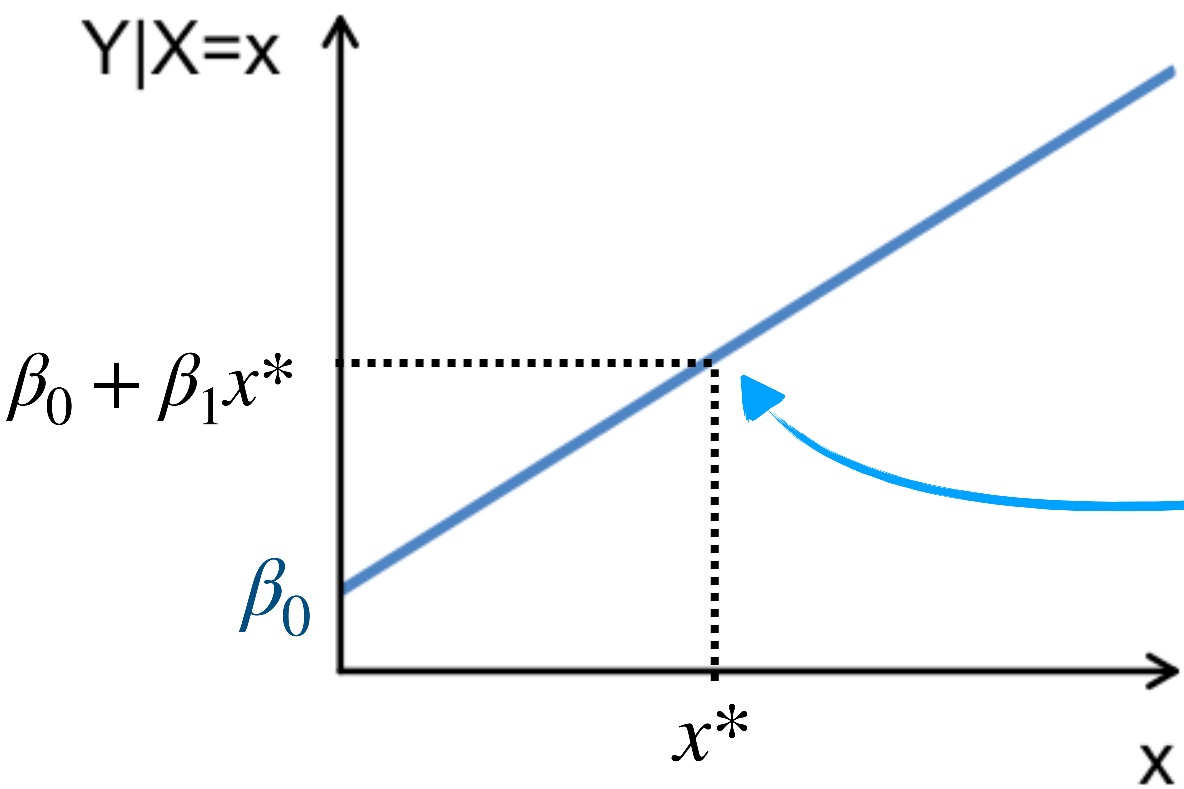
X : distanse løpt (km) Y : tid brukt (minutter) **Mål:** $Y \mid X = x$

Antagelse 1: Lineær sammenheng $\rho(X, Y) \neq 0$ (Plott observasjoner av x og y)

Forventningsverdien til Y, betinget på X=x, er lineær i x

$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

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



NB! Observasjoner av Y, når X = x*, vil spre seg i en viss avstand rundt linja

Lineær regresjon: Modellantagelser

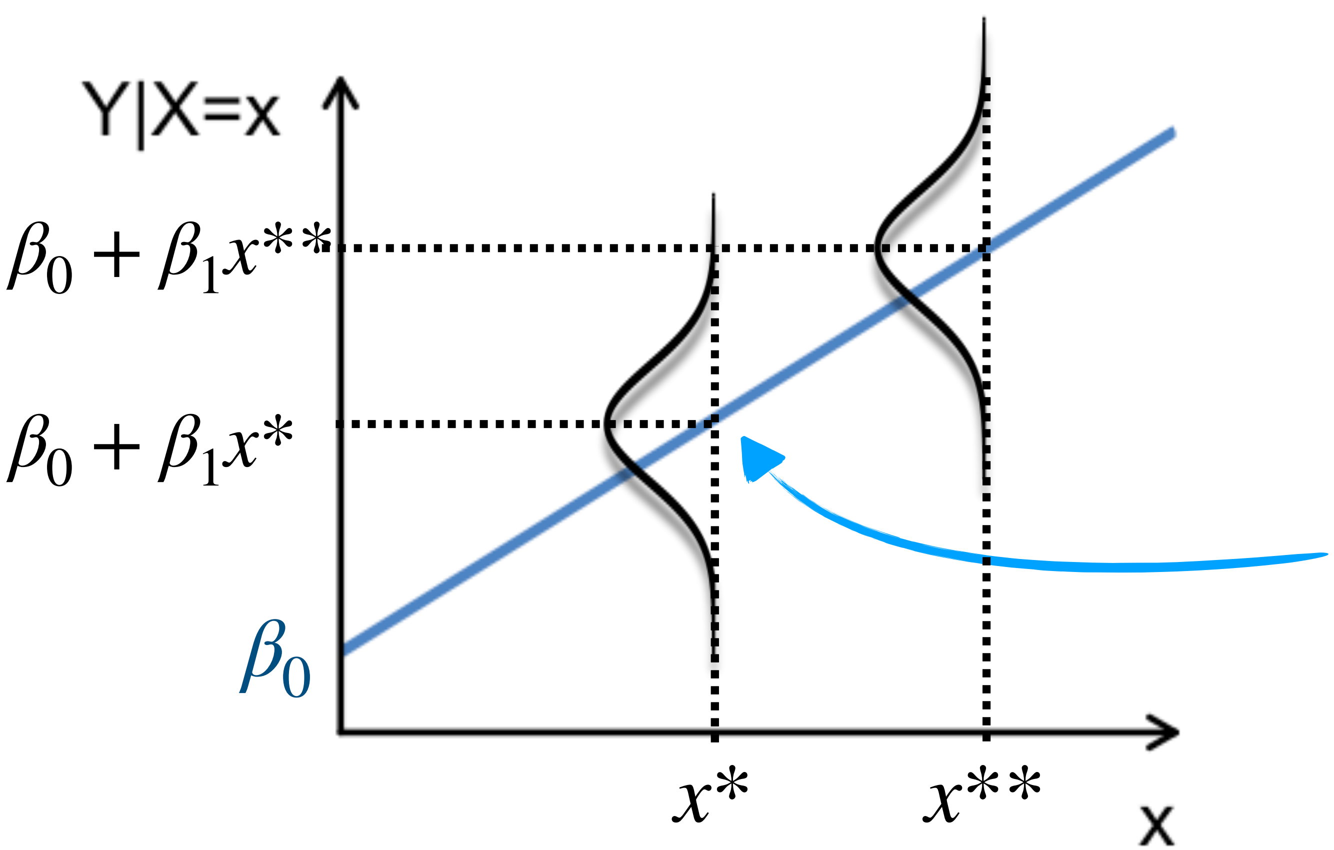
X : distanse løpt (km) Y : tid brukt (minutter) **Mål:** $Y \mid X = x$

Antagelse 1: Lineær sammenheng $\rho(X, Y) \neq 0$ (Plott observasjoner av x og y)

Forventningsverdien til Y, betinget på X=x, er lineær i x

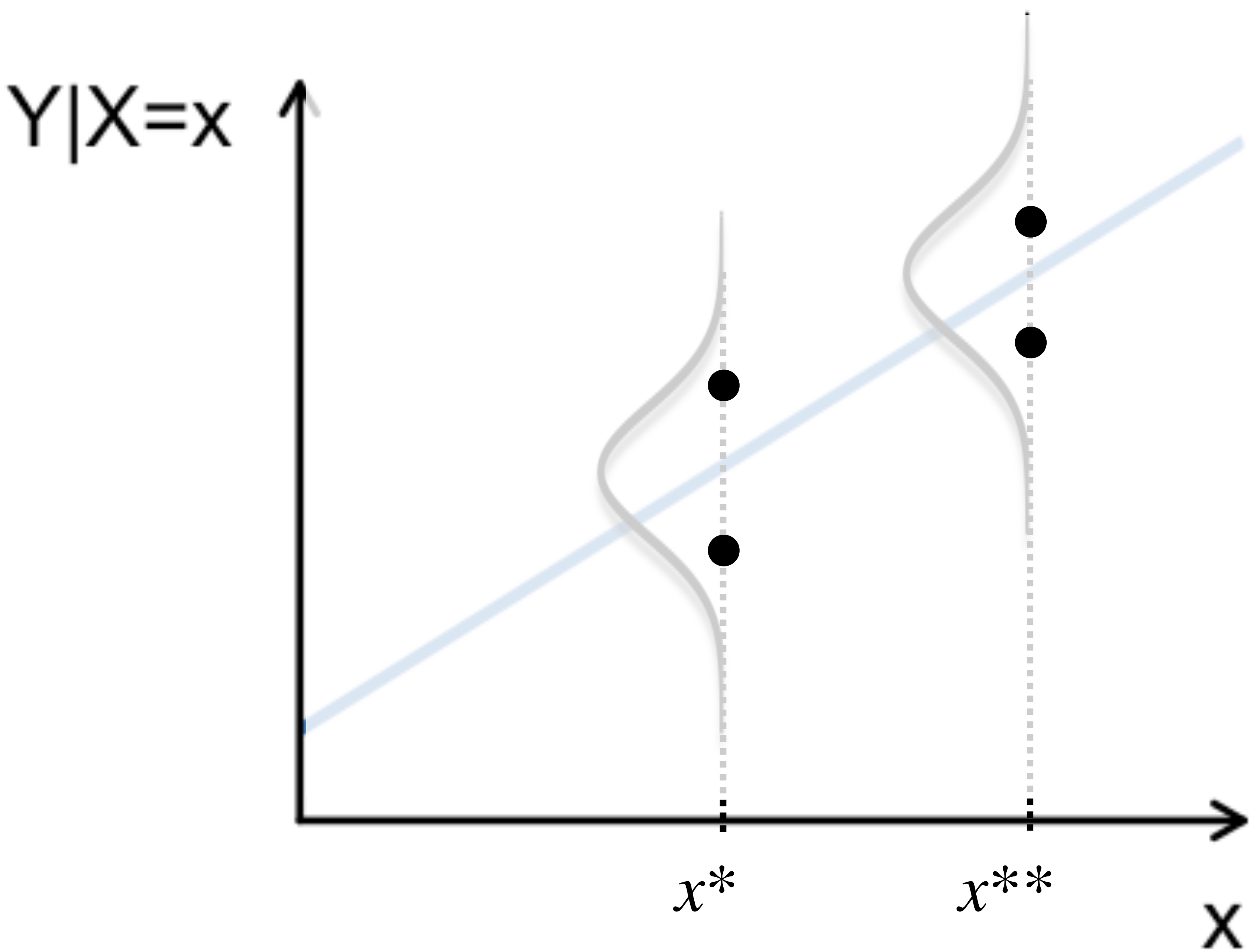
$$E(Y \mid X = x) = \beta_0 + \beta_1 x$$

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

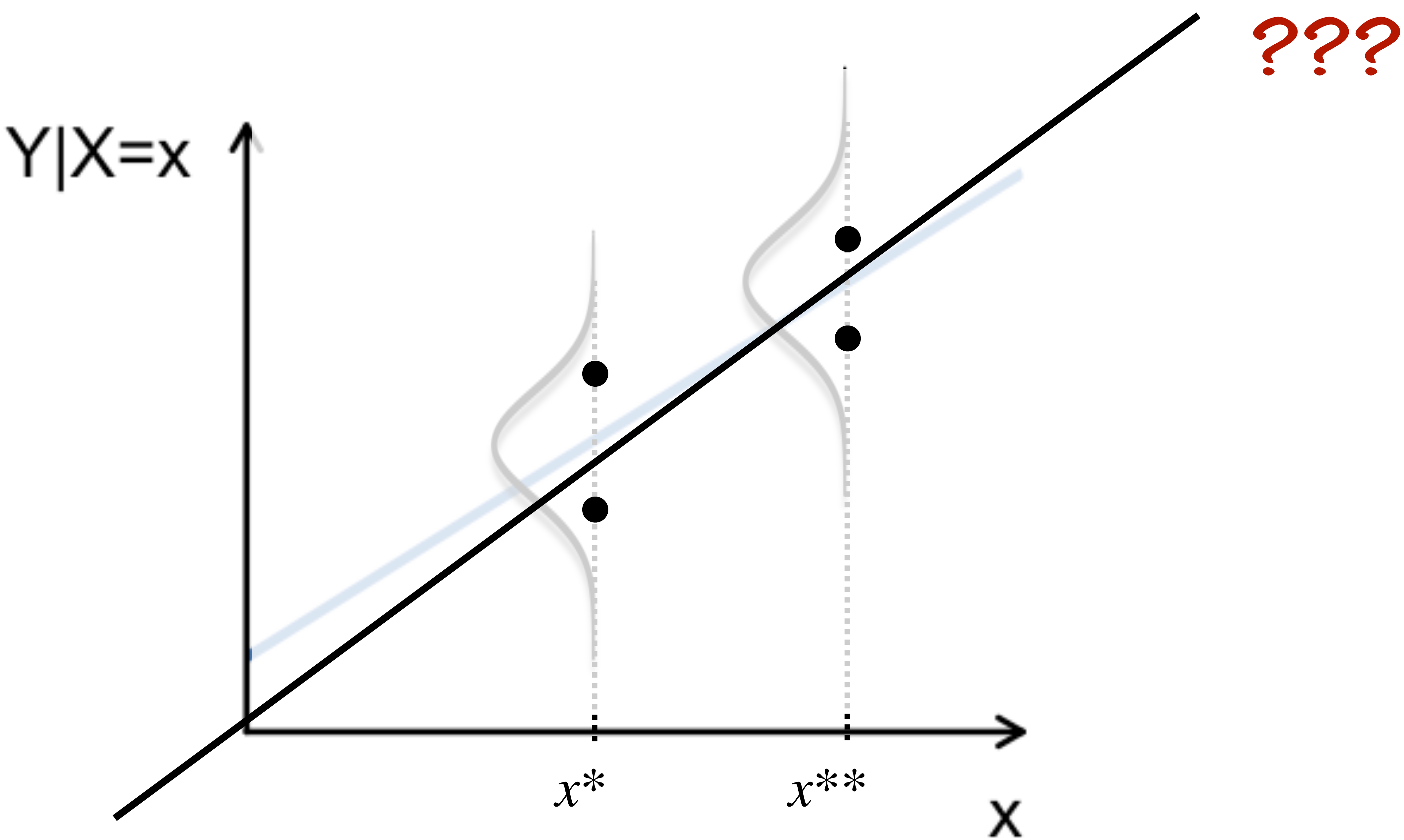


NB! Observasjoner av Y, når X = x*,
vil spre seg i en viss avstand rundt linja

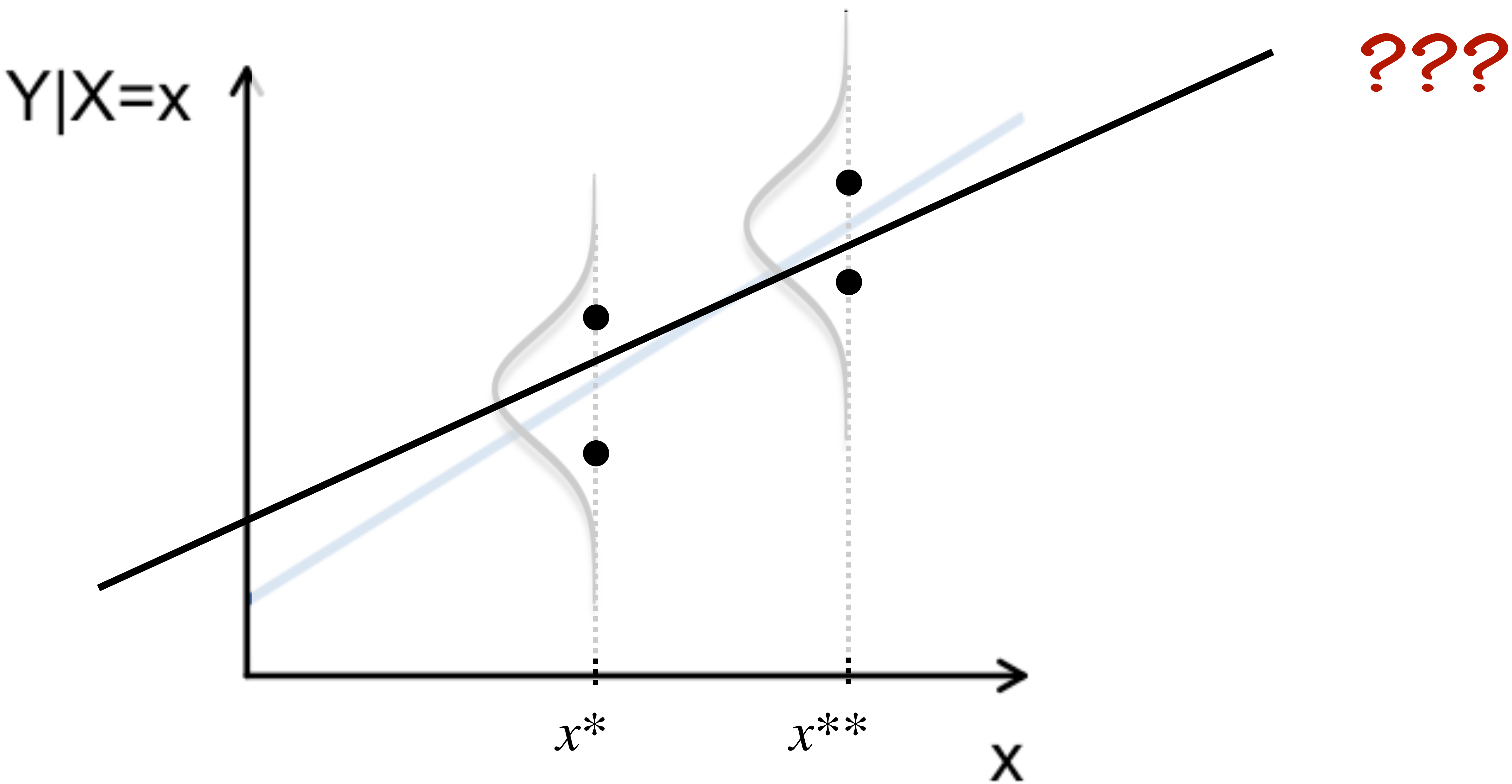
Lineær regresjon: Hvordan ser dataene ut?



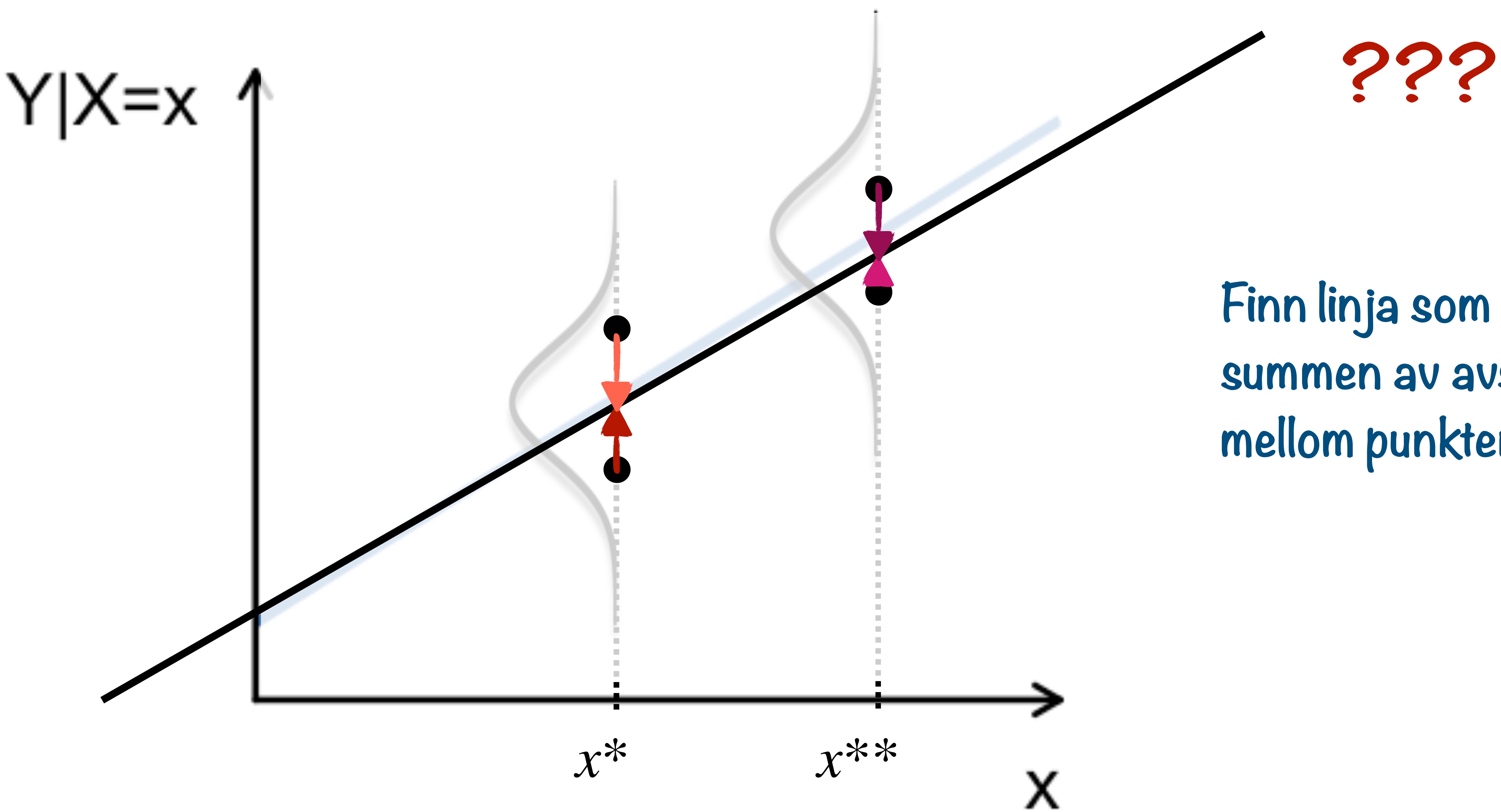
Lineær regresjon: Hvordan estimerer vi linja?



Lineær regresjon: Hvordan estimerer vi linja?

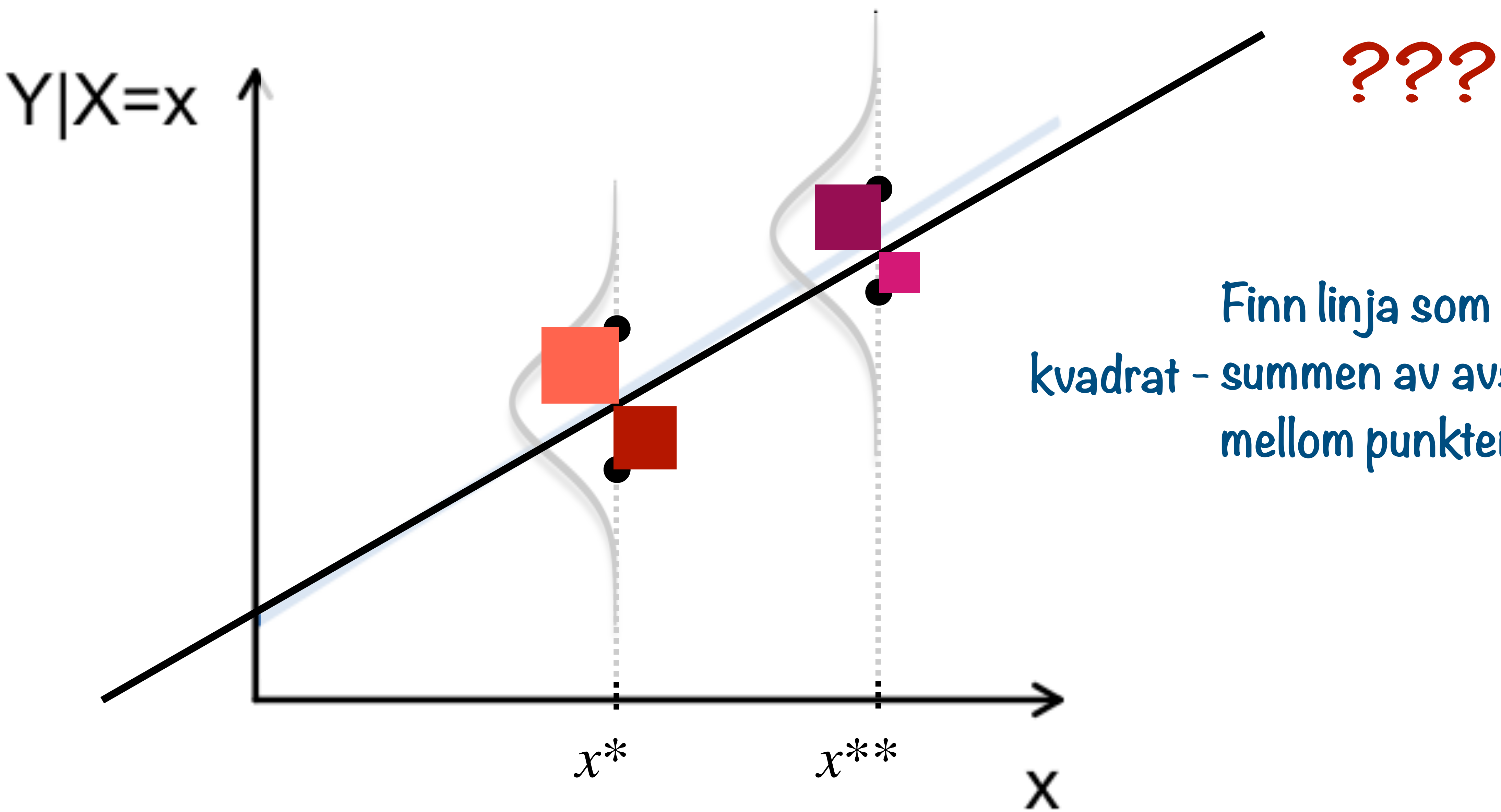


Lineær regresjon: Hvordan estimerer vi linja?



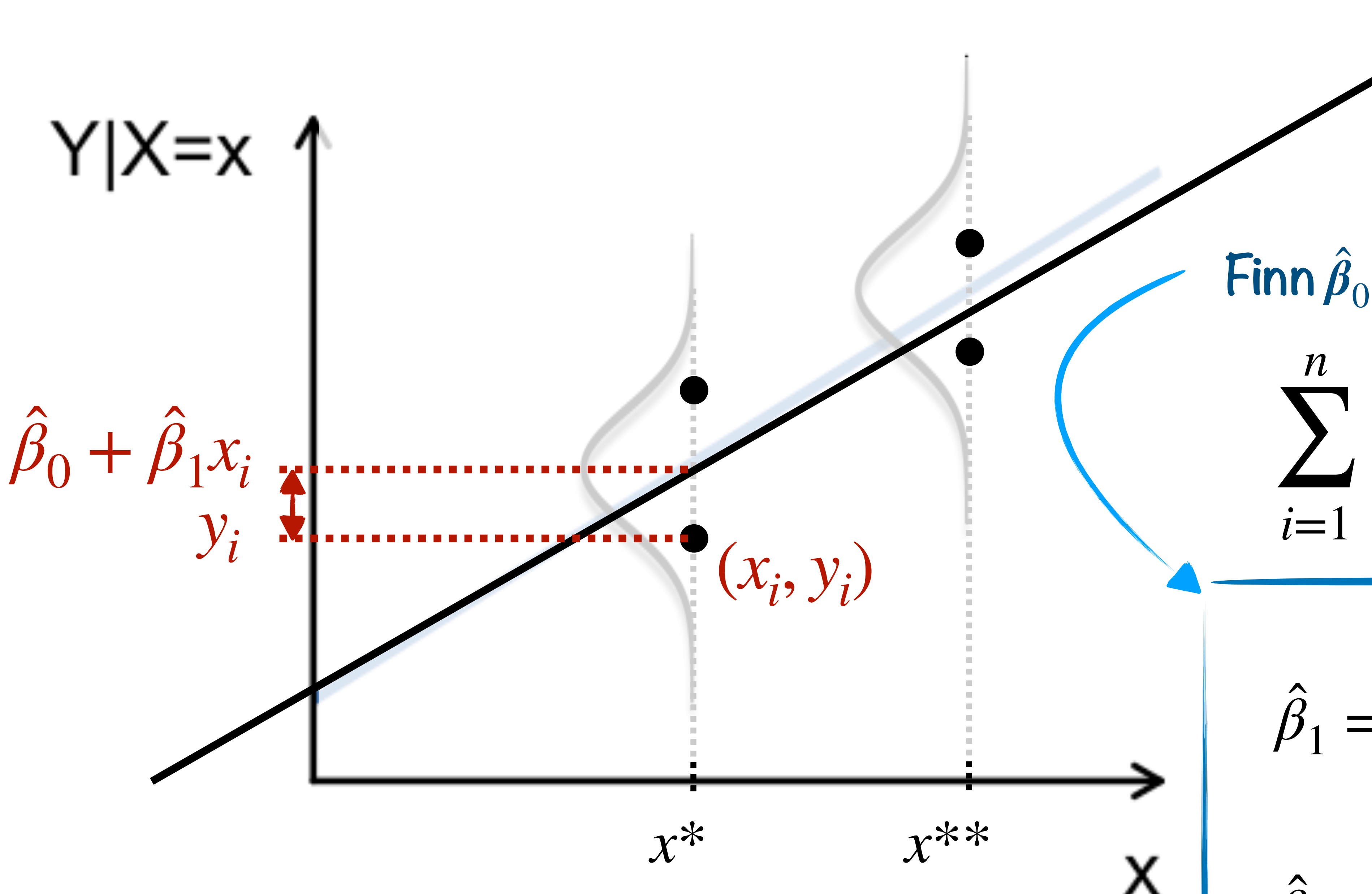
Finn linja som minimerer summen av avstandene mellom punktene og linja?

Lineær regresjon: Hvordan estimerer vi linja?



Finn linja som minimerer
kvadrat - summen av avstandene
mellom punktene og linja

Lineær regresjon: Hvordan estimerer vi linja?



Finn $\hat{\beta}_0$ og $\hat{\beta}_1$ som minimerer:

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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

Estimator for regresjonslinja

Tilfeldig utvalg: $(x_1, Y_1), \quad (x_2, Y_2), \quad \dots, \quad (x_n, Y_n)$

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

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

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

Estimator for linja: $\hat{\beta}_0 + \hat{\beta}_1 x \qquad E(\hat{\beta}_0 + \hat{\beta}_1 x) = \beta_0 + \beta_1 x$

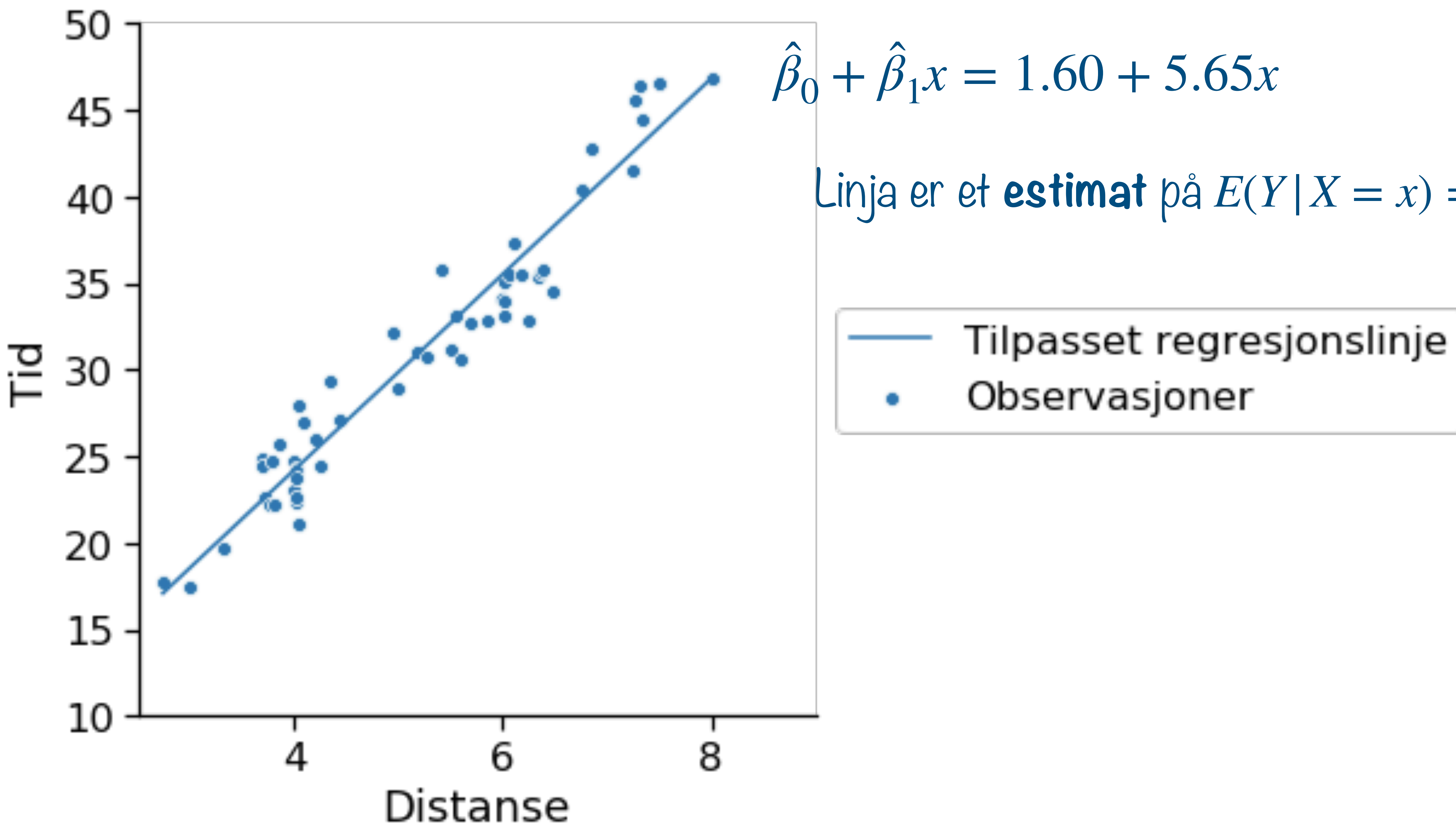
Eksempel: Treningsdata

X : distanse (km) $\quad Y$: tid (minutter)

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

$$\hat{\beta}_0 + \hat{\beta}_1 x = 1.60 + 5.65x$$

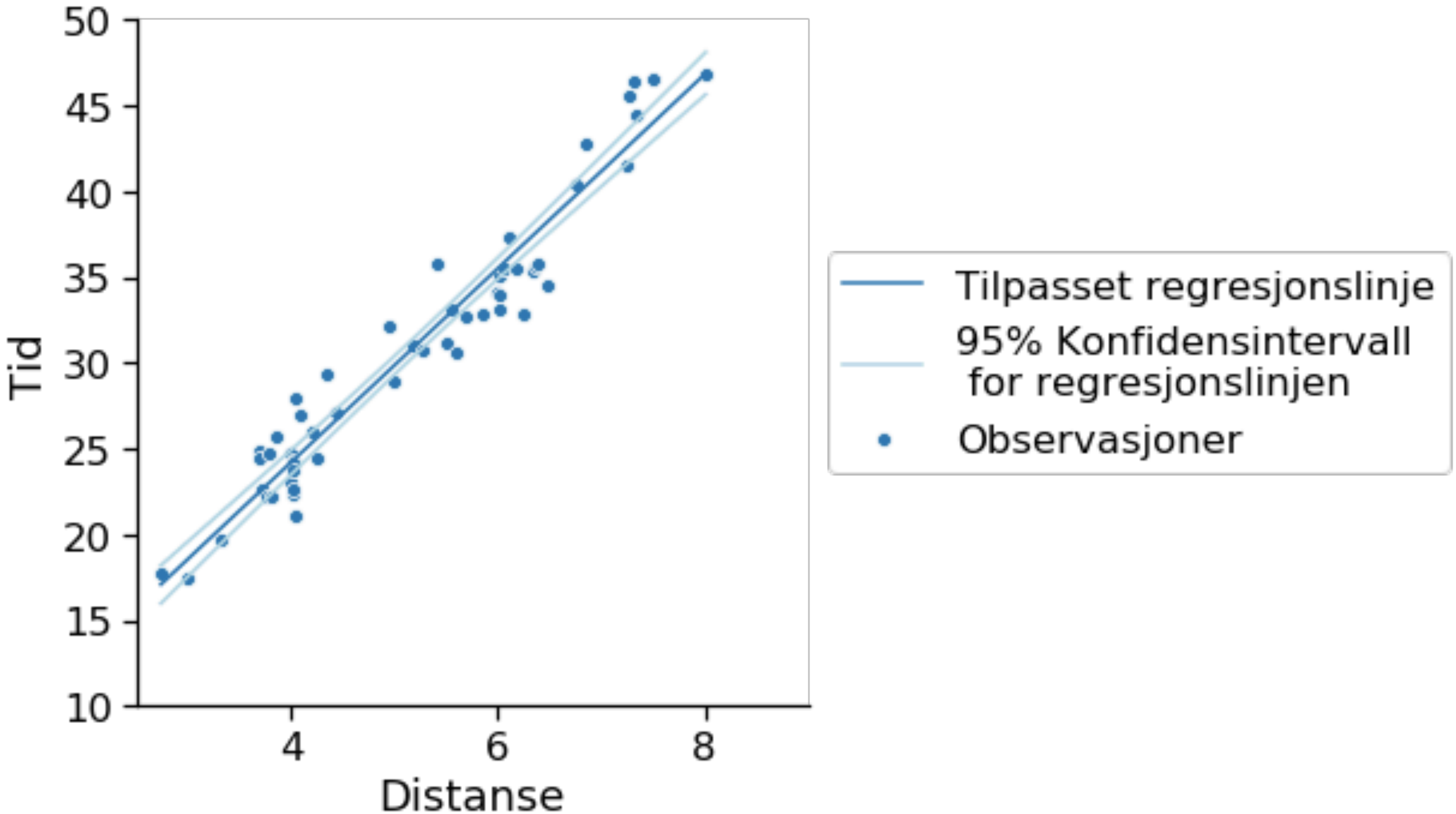
Linja er et **estimat** på $E(Y \mid X = x) = \beta_0 + \beta_1 x$



Eksempel: Treningsdata

X : distanse (km) Y : tid (minutter)

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



Eksempel: Treningsdata

X : distanse (km) Y : tid (minutter)

$$(x_1, y_1), \quad (x_2, y_2), \quad \dots, \quad (x_{55}, y_{55})$$