Statistikk

Jo Eidsvik

Matematiske fag, NTNU

Regresjon

Data kommer i form av kjente kovariater (eller forklaringsvariable) og målinger eller responsvariable.

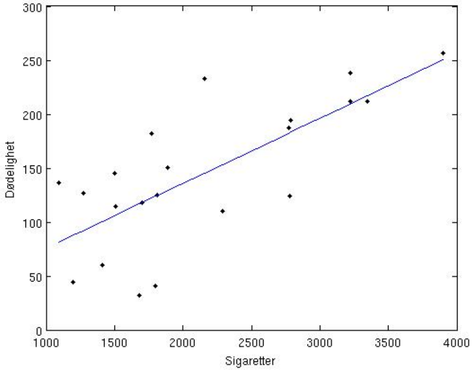
- Forklaringsvariable: $x_1, \dots, x_n$
- Responsvariable: $Y_1, \dots, Y_n$.

Modell for linear regresjon

For $i = 1, \ldots, n$:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \text{uavhengige}$$

Dødelighet av hjerte og karsykdommer (Y) mot sigaretter (x)



Mål er parameterestimering og prediksjon

1. **Parameterestimering:** Finn $\hat{\beta}_0$ og $\hat{\beta}_1$ utfra data: responser og kovariater.
2. **Prediksjon:** Finn $E(Y_0)$ og $\mathrm{Var}(Y_0)$ der Y_0 er en ny måling med kovariat x_0.

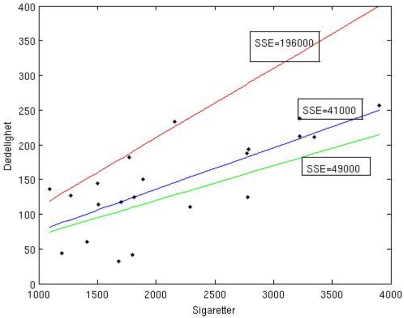
Lag konfidensintervall, gjennomfør tester.

Parameterestimering

Minste kvadratsums metode finner linja som minimerer kvadratiske avvik til data.

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

Maximum likelihood estimation (MLE) gir tilsvarende målfunksjon.



Sannsynlighetsestimering

$$L = L(\beta_0, \beta_1) = \prod_{i=1}^{n} f(y_i; \beta_0, \beta_1),$$

$$l = \log L = \text{const} - \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}$$

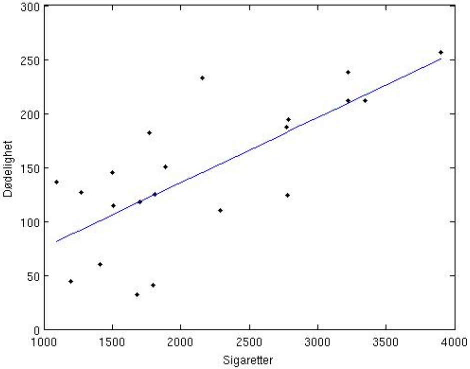
Uttrykket med kvadratiske avvik er samme som i Minste kvadratsums metode.

Estimatorer

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

.

$\hat{\beta}_1 = 0.06$ Er det signifikant? **JA! - Konfidensintervall for β_1 dekker ikke 0**



Fordeling til $\hat{\beta}_1$ (tilsvarende for $\hat{\beta}_0$)

$$E(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x}) E(Y_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1$$

.

$$Var(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 Var(Y_i)}{[\sum_{i=1}^n (x_i - \bar{x})^2]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

.

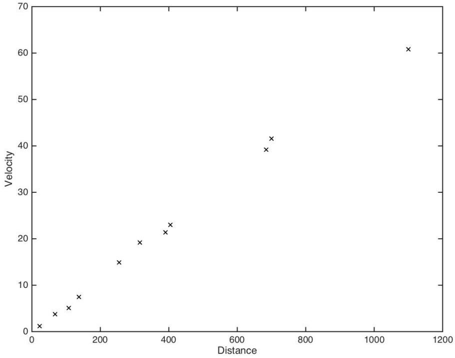
$$\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim N(0,1)$$

$$\frac{\hat{\beta}_1 - \beta_1}{s / \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \sim t_{n-2}$$

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

Hubble eksempel

x_i =Avstand til galakse i, Y_i =Hastighet til galakse i, $i = 1, \ldots, 11$.
Modell $Y_i = \beta x_i + \epsilon_i$, $i = 1, \ldots, 11$, $\epsilon_i \sim N(0, \sigma^2)$



Hubble eksempel - tilpasning

x_i =Avstand til galakse i, Y_i =Hastighet til galakse i, $i = 1, \ldots, 11$.

Estimator

$$\hat{\beta} = \frac{\sum_{i=1}^{11} x_i Y_i}{\sum_{i=1}^{11} x_i^2}$$