STK1110 Høsten 2026

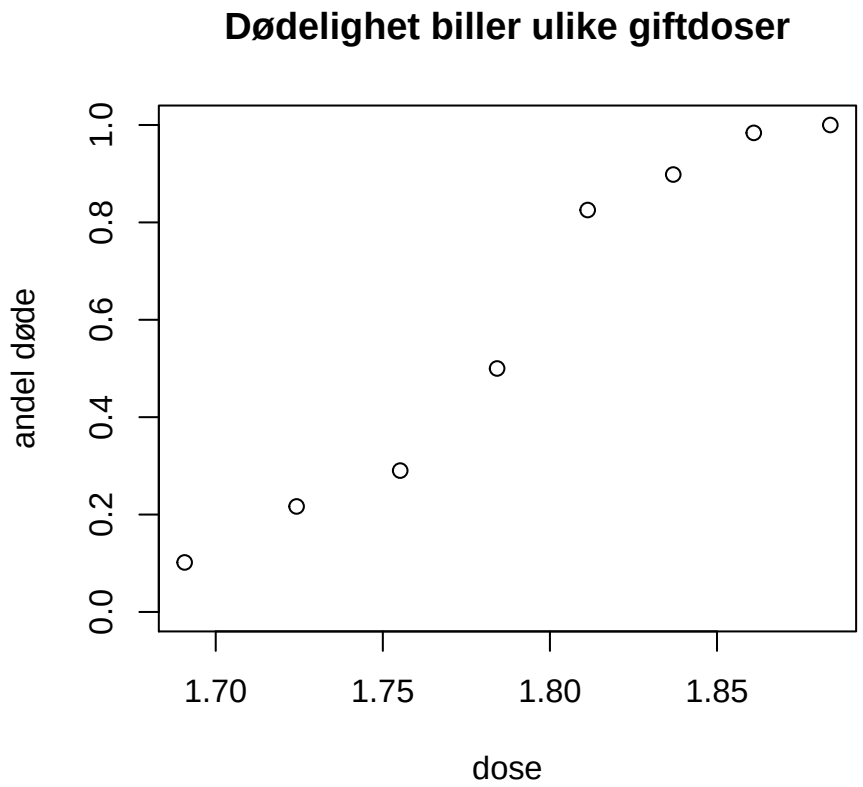
Logistisk regresjon

Tilsvarer Avsnitt 12.10

Eksempel: Dødelighet biller ved ulike giftdoser

Dødelighet ved hver av 8 ulike giftdoser ble testet for ca. 60 biller.

Dose	antall	døde
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60



Ønsker å studere hvordan sannsynligheten for at biller dør når dosen endres.

Eksempel: Challenger-ulykken (Eks. 12.34 i boka)



Challenger-ulykken

Temperatur (i $^\circ F$)	Feil
53	Ja
57	Ja
58	Ja
63	Ja
66	Nei
67	Nei
67	Nei
68	Nei
69	Nei
70	Nei
70	Nei
70	Ja
70	Ja
72	Nei
73	Nei
75	Nei
75	Ja
76	Nei
76	Nei
78	Nei
79	Nei
81	Nei

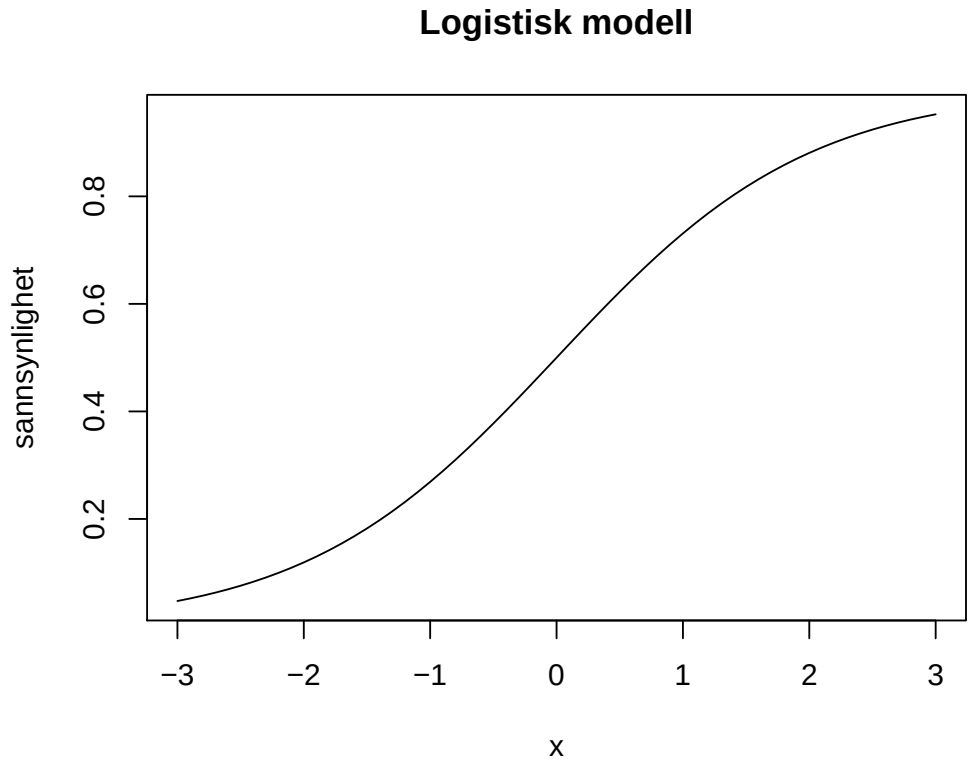
Logistisk regresjon

- Anta at $(x_i, y_i), i = 1, \ldots, n$, der y_i-ene er observerte verdier av stokastiske variable $Y_1, \ldots, Y_n$, mens x_i-ene er gitt.
- $Y_i \in \{0, 1\}$, dvs binær respons
- Regresjonssetting, ønsker å "forklare" Y_i med x_i
- Anta Y_i er uavhengige (Bernoulli-fordelt) med sannsynligheter

$$p(x_i) = \mathrm{P}(Y_i = 1 | x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}.$$

- Dette kalles den logistiske regresjonsmodellen.
- Parameterne β_0 og β_1 estimeres ved hjelp av maksimum likelihood-estimering.

Illustrasjon:



Modellen $p(x) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$ gir en "sigmoid" kurve med x.

- Har $0 < p(x) < 1$ (sannsynlighet)
- Hvis $\beta_1 > 0$ er $p(x)$ voksende i x
- Hvis $\beta_1 < 0$ er $p(x)$ avtagende i x
- Kurven er symmetrisk om $(-\frac{\beta_0}{\beta_1}, \frac{1}{2})$

Fortolkning av $\exp(\beta_1)$ som OR=Odds-ratio

Vi definere odds som $\text{Odds}(x) = \frac{p(x)}{1-p(x)}$.

Siden $1 - p(x) = \frac{1}{1+\exp(\beta_0+\beta_1 x)}$ får vi da

$$\text{Odds}(x) = \frac{p(x)}{1-p(x)} = \exp(\beta_0 + \beta_1 x)$$

Dette gir at forholdet mellom $\text{Odds}(x+1)$ og $\text{Odds}(x)$ gis ved

$$OR = \frac{\text{Odds}(x+1)}{\text{Odds}(x)} = \frac{\exp(\beta_0 + \beta_1(x+1))}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1)$$

En økning av x til $x+1$ gir altså en endring i odds lik odds-ratioen $OR = \exp(\beta_1)$.

Når $p(x)$ er liten har vi at $\text{Odds}(x) = \frac{p(x)}{1-p(x)} \approx p(x)$ og dermed at odds-ratioen er tilnærmet lik

$$\text{Relativ risk } = RR = \frac{p(x+1)}{p(x)}$$

Estimering av parametere

- Likelihood-funksjonen er gitt ved

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i},$$

hvilket gir log-likelihooden

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(\beta_0 y_i + \beta_1 x_i y_i - \log\left(1 + e^{\beta_0 + \beta_1 x_i}\right)\right).$$

- Maksimum likelihood-estimatene $\hat{\beta}_0$ og $\hat{\beta}_1$ får vi ved å maksimere $\ell(\beta_0, \beta_1)$ med hensyn på β_0 og β_1 (numerisk).

Eksempel

Eks. 12.34 fra boka.
Får $e^{\hat{\beta}_1} = e^{-0.2322} \approx 0.79$, odds minker med 21%

Egenskaper ved $\hat{\beta}_0$ og $\hat{\beta}_1$

- La $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$, der $\hat{\beta}_0$ og $\hat{\beta}_1$ er maksimum likelihood-estimatorene for β_0 og β_1.
- Når n er stor, er da $\hat{\boldsymbol{\beta}}$ tilnærmet binormalt fordelt med forventning $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ og kovariansmatrise $\boldsymbol{\Sigma}$, som vi kan finne ved hjelp av informasjonsmatrisen $\mathbf{I}(\boldsymbol{\beta})$.
- Vi har

$$\mathbf{I}(\boldsymbol{\beta}) = \begin{pmatrix} I_{00}(\boldsymbol{\beta}) & I_{01}(\boldsymbol{\beta}) \\ I_{10}(\boldsymbol{\beta}) & I_{11}(\boldsymbol{\beta}) \end{pmatrix}$$

med

$$\begin{aligned}
I_{00}(\boldsymbol{\beta}) &= -\mathrm{E}\left(\frac{\partial^2}{\partial \beta_0^2}\ell(\boldsymbol{\beta})\right) = \sum_{i=1}^n \frac{e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} \\
I_{01}(\boldsymbol{\beta}) &= -\mathrm{E}\left(\frac{\partial^2}{\partial \beta_0 \partial \beta_1}\ell(\boldsymbol{\beta})\right) = \sum_{i=1}^n x_i \frac{e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2} = I_{10}(\boldsymbol{\beta}) \\
I_{11}(\boldsymbol{\beta}) &= -\mathrm{E}\left(\frac{\partial^2}{\partial \beta_1^2}\ell(\boldsymbol{\beta})\right) = \sum_{i=1}^n x_i^2 \frac{e^{\beta_0+\beta_1 x_i}}{\left(1+e^{\beta_0+\beta_1 x_i}\right)^2}.
\end{aligned}$$

Egenskaper ved $\hat{\beta}_0$ og $\hat{\beta}_1$ (forts.)

- Da har vi

$$\boldsymbol{\Sigma} = \mathbf{I}(\boldsymbol{\beta})^{-1} = \begin{pmatrix} I^{00}(\boldsymbol{\beta}) & I^{01}(\boldsymbol{\beta}) \\ I^{10}(\boldsymbol{\beta}) & I^{11}(\boldsymbol{\beta}) \end{pmatrix},$$

  som vi kan estimere med

$$\hat{\boldsymbol{\Sigma}} = \mathbf{I}(\hat{\boldsymbol{\beta}})^{-1} = \begin{pmatrix} I^{00}(\hat{\boldsymbol{\beta}}) & I^{01}(\hat{\boldsymbol{\beta}}) \\ I^{10}(\hat{\boldsymbol{\beta}}) & I^{11}(\hat{\boldsymbol{\beta}}) \end{pmatrix}.$$

- Da finnes standardfeilen til $\hat{\beta}_j$ ved $S_{\hat{\beta}_j} = \sqrt{I^{jj}(\hat{\boldsymbol{\beta}})}, j = 0, 1.$

Eksempel

Eks. 12.34 i boka.

Hypotesetest vedrørende β_1

- Vi vil teste $H_0 : \beta_1 = 0$ mot $H_a : \beta_1 \neq 0$.
- Testobservator $Z = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$, der $Z \overset{tiln.}{\sim} N(0,1)$ under H_0.
- Alternativ: likelihood ratio-testen.
- Likelihood ratio er gitt ved

$$LR = \frac{L(\beta_0^*, 0)}{L(\hat{\beta}_0, \hat{\beta}_1)},$$

  der β_0^* er ML-estimatoren for β_0 under H_0.
- La $p = \frac{e^{\beta_0}}{1+e^{\beta_0}}$, som er $p(x_i)$ under H_0, og la p^* være ML-estimatet for p under H_0.

Hypotesetest vedrørende β_1 (forrts.)

- Da er $p^* = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$ og $\beta_0^* = \log\left(\frac{p^*}{1-p^*}\right) = \log\left(\frac{\bar{y}}{1-\bar{y}}\right)$.
- Når n er stor, er $-2\log(LR) \overset{tiln.}{\sim} \chi_1^2$ under H_0.
- Vi får dermed en test med signifikansnivå α dersom vi forkaster H_0 hvis $-2\log(LR) \geq \chi_{\alpha,1}^2$.

Eksempel

Eks. 12.34 i boka.

Multippel logistisk regresjon

- Anta nå k forklaringsvariable $(x_1, ..., x_k) = \boldsymbol{x}$
- Modell:

$$p(\boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1 + e^{\beta_0+\beta_1 x+\cdots+\beta_k x_k}}$$

- Log-odds:

$$\text{Log-Odds} = \log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x + \cdots + \beta_k x_k$$

  Lineær struktur i $x_1, ..., x_k$
- Estimering og testing: Tilsvarende som for $k = 1$

Eksempel

Eks. 12.34 i boka.

Tolkning av parametre i multippel logistisk regresjon

- Modell:

$$p(\boldsymbol{x}) = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}}$$

- Odds:

$$\text{Odds} = \frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})} = e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k}$$

- Hvis x_1 øker en en enhet, og $x_2, \ldots, x_k$ holdes konstant, så øker odds med e^{β_1}
- Log-Odds:

$$\text{Log-Odds} = \log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x$$

Log-odds øker med β_1.