Maximum likelihood (ML)

- Likelihood og sannsynlighet
- ML-prinsippet og ML-estimering

Maximum likelihood (ML)-estimering

Introduksjon til prinsippet gjennom et eksempel

Mange melder seg på arrangementer på Facebook, men alle påmeldte møter ikke opp. *Anta at påmeldte møter opp (eller ikke) uavhengig av hverandre.*

La parameteren p = P(møter opp| påmeldt), og definer for hver påmeldt person i den tilfeldige variabelen Y_i med verdi 1 hvis personen møter, 0 ellers.

Observerer et tilfeldig utvalg av 10 påmeldte, resultatet blir
1, 0, 0, 1, 1, 0, 0, 0, 0, 0

Sannsynligheten for akkurat dette resultatet –
før vi visste hvor mange som skulle komme- var

$P(Y_1=1 \text{ og } Y_2=0 \text{ og } \ldots \text{ og } Y_{10}=0; p) = p(1-p)\ldots..(1-p) = p^3(1-p)^7$

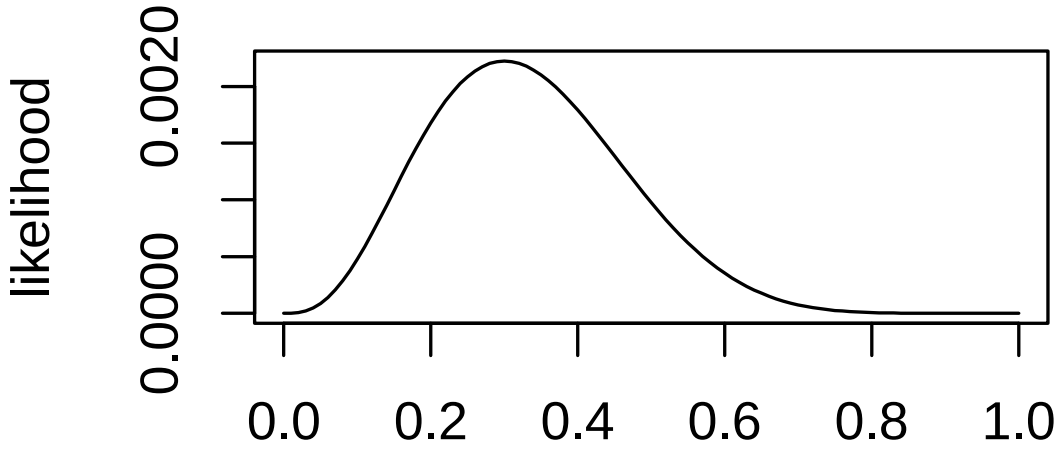
Hvilken verdi for parameteren **p** ville gitt **høyest mulig sannsynlighet** for utfallet vi har observert**?**

Maximum likelihood-estimatet for en parameter er den verdien som ville gitt det vi har observert høyest mulig «sannsynlighet»

egentlig: størst mulig likelihood for dataene

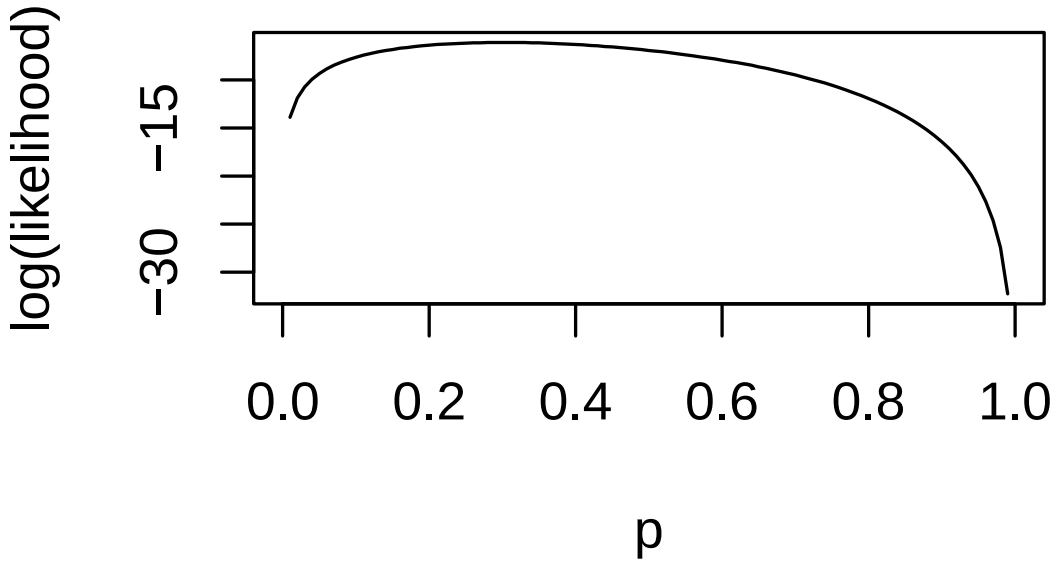
Likelihood L og log-likelihood logL vil maksimeres av samme verdi for p.

Det er ofte enklere å beregne maksimum for log-likelihood.



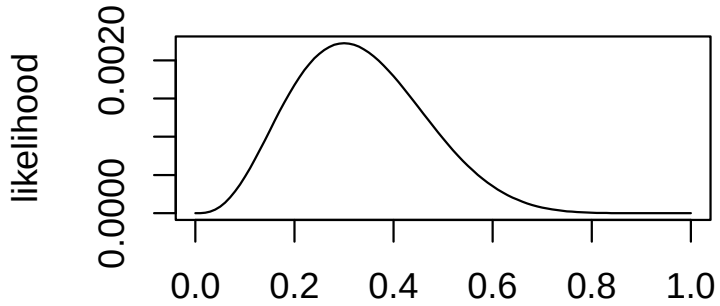
Likelihood L:

Dette er et plott av $p^3(1-p)^7$ som funksjon av p



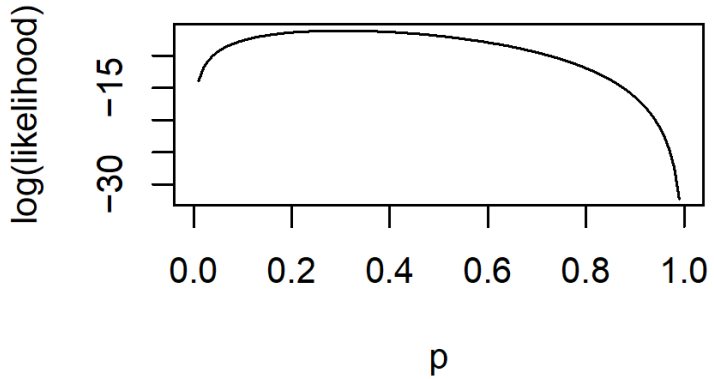
Log-likelihood logL:

Dette er et plott av $\log(p^3(1-p)^{7})$ som funksjon av p



Likelihood for oppmøte-dataene er $L(y_i, y_2, \ldots, y_{10}; p) = p^3(1-p)^7$

Loglikelihooden er $\log L(y_i, y_2, \ldots, y_{10}; p) = 3 \log p + 7\log(1-p)$



$d \log L(y_i, y_2, \ldots, y_{10}; p) / dp = 3/p - 7/(1-p)$

Vi finner topp-punktet ved å finne p-verdien der den deriverte av loglikelihooden er lik 0:

p = 3/10 maksimerer loglikelihood, og er ML-parameterestimatet

Likelihood måler hvor godt dataene støtter en viss verdi for en parameter

- Likelihooden er simultan sannsynlighets-tetthet for dataene, betraktet som en funksjon av parameteren eller parameterne

- Estimatorer funnet ved maximum likelihood (ML)-metoden har spesielt gode statistiske egenskaper, spesielt når n er stor.

Sannsynlighetsmaksimeringsprinsippet

eng: Principle of maximum likelihood (ML)

- $y_1, \ldots, y_n \overset{uif}{\sim} f(y;\theta)$
- $L(\theta;\mathbf{y}) = f(y_1, \ldots, y_n;\theta) = \prod_i f(y_i;\theta)$
- $\log L(\theta;\mathbf{y}) = \sum_i \log f(y_i;\theta)$
- $\hat{\theta} = \text{argmax}_\theta L(\theta;\mathbf{y})$
  - Konsistent, asymptotisk effisient
  - Analyttiske løsninger for lineære/Gaussiske modeller
    Ett-/to- utvalgs modeller, variansanalyse, lineær regresjon
  - Generelt: Numerisk optimering

Likelihood

Loglikelihood

Les : ”best”

Når utvalgsstørrelsen n er stor, er ML-estimatoren tilnærma forventningsrett, tilnærma normalfordelt, *og* vi kan beregne standardfeilen

- For stor n:

$$\hat{\theta} \approx N(\theta, I(\hat{\theta})^{-1}) \approx N(\theta, J(\hat{\theta}; \mathbf{y})^{-1})$$

$$J(\theta; \mathbf{y}) = -\frac{\partial^2}{\partial\theta\partial\theta^T} \log L(\theta; \mathbf{y})$$

Observert informasjon

$$I(\theta) = E[J(\theta; \mathbf{y})] \quad \text{Alltid pos. (semi)definit}$$

Fisher-informasjonen (matrise)

Maximum likelihood (ML) for enkel lineær regresjon

ML = sannsynlighetsmaksimering

- I enkel lineær regresjon setter vi opp en modell for forventningsverdien μ_i til responsvariabelen y som funksjon av populasjonsparameterne β_0 og β_1 og forklaringsvariabelen x.

- Når vi antar at individuell variasjon ε_i om forventningsverdien μ_i er normalfordelt, får vi en sannsynlighetsfordeling for en tilfeldig y fra underpopulasjonen definert ved en angitt x-verdi.

- Denne sannsynlighetsfordelinga blir bestemt av verdien på forklaringsvariabelen x, og populasjonsparameterne $\beta_{0,}$ β_1 og σ.

Maximum likelihood (ML) for enkel lineær regresjon

ML = sannsynlighetsmaksimering

Oppsettet for enkel, lineær regresjon:

- $y_i \sim \mathrm{N}(\mu_i, \sigma^2)$ der $\mu_i = \beta_0 + \beta_1 x_i$
- y_i-ene er uavhengige

For enkelthets skyld antar vi at σ^2 er kjent

Sannsynlighetsfordelinga til y_i er da en normalfordeling:

$$f(y_i, \mu_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\}$$

For en lineær regresjonsmodell, støtter maximum likelihood-prinsippet oss i valget om å bruke minste kvadraters metode

Likelihood er den simultane tettheten for alle n observasjonene

$$L = \prod_{i=1}^{n} f(y_i, \mu_i) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2\}$$

betraktet som en funksjon av populasjonsparameterne β_0 og β_1.

Vi estimerer de ukjente parameterverdiene for β_0 og β_1 ved å velge de verdiene som tildeler de observerte y_i-verdiene så høy likelihood som mulig (høyest mulig «sannsynlighet»)

Igjen, å maksimere likelihood L er det samme som å maksimere

$$\log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2,$$

som i dette tilfellet er det samme som å minimere $\sum_{i=1}^{n}(y_i - \mu_i)^2$

I logistisk regresjon får likelihooden samme grunn-formel som i oppmøteeksempelet, $\mathbf{L} = \prod_{i=1}^{n} p_i^{\,y_i} (1 - p_i)^{1-y_i}$

fordi responsvariabelen y indikerer om noe *skjedde* eller *ikke*.

Men logistisk regresjon er også regresjon, med forklaringsvariabel x og ukjente populasjonsparametere β_0 og β_1 .

Maximum likelihood (ML) for logistisk regresjon

ML = sannsynlighetsmaksimering

Vi skal senere (kap 14) introdusere dette mer grundig, men:

I logistisk regresjon setter vi opp en modell for suksess-sannsynligheten p for hver y som funksjon av populasjonsparameterne β_0 og β_1 og forklaringsvariabelen x :

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

For logistisk regresjon finnes det ikke *formler* for parameter-estimatene b_0 og b_1, men programvare (eks R) finner tallverdier med numerisk optimering av L $= \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$

Disse estimatene kalles maximum likelihood estimater (MLE).

7 Maksimum likelihood metoden

Ikke pensum, men for de som er interessert

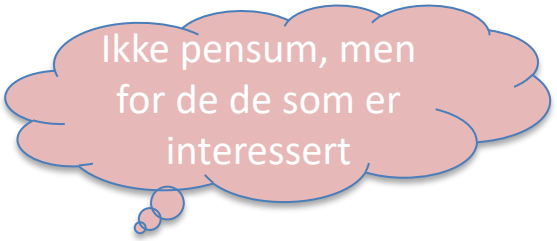
Anta at $X_1, X_2, \ldots, X_n$ har simultan punktsannsynlighet/sannsynlighetstetthet $f(x_1, x_2, \ldots, x_n \,|\, \boldsymbol{\theta})$, der $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ er en parametervektor (skalar hvis $p = 1$). Vi antar at $f(x_1, x_2, \ldots, x_n \,|\, \boldsymbol{\theta})$ tilfredsstiller visse deriverbarhetsbetingelser.

(*a*) Gitt observerte verdier $X_i = x_i$; $i = 1, \ldots, n$; er likelihood-funksjonen $\text{lik}(\boldsymbol{\theta}) = f(x_1, x_2, \ldots, x_n \,|\, \boldsymbol{\theta})$ og loglikelihood-funksjonen $l(\boldsymbol{\theta}) = \log(\text{lik}(\boldsymbol{\theta}))$.

(*b*) Maksimum likelihood *estimatet* er den verdien av $\boldsymbol{\theta}$ som maksimerer $\text{lik}(\boldsymbol{\theta})$ eller ekvivalent maksimerer $l(\boldsymbol{\theta})$. Hvis vi erstatter de observerte x_i-ene med de stokastiske X_i-ene, får vi maksimum likelihood *estimatoren*.

(*c*) Maksimum likelihood estimatet $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$ er en løsning av ligningene $s_j(\boldsymbol{\theta}) = 0$; $j = 1, \ldots, p$; der $s_j(\boldsymbol{\theta}) = (\partial/\partial\theta_j) l(\boldsymbol{\theta})$ er score-funksjonene. Vektoren av scorefunksjoner er $\boldsymbol{s}(\boldsymbol{\theta}) = (s_1(\boldsymbol{\theta}), \ldots, s_p(\boldsymbol{\theta}))^T$.

(*d*) Den observerte informasjonsmatrisen $\bar{\boldsymbol{J}}(\boldsymbol{\theta})$ er $p \times p$ matrisen med element (i, j) gitt ved $\bar{J}_{ij}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\boldsymbol{\theta})$.
Den forventede informasjonsmatrisen (eller Fishers informasjonsmatrise) $\bar{\boldsymbol{I}}(\boldsymbol{\theta})$ er $p \times p$ matrisen med element (i, j) gitt ved $\bar{I}_{ij}(\boldsymbol{\theta}) = \text{E}[\bar{J}_{ij}(\boldsymbol{\theta})]$.
For uavhengige og identisk fordelte observasjoner har vi at $\bar{\boldsymbol{I}}(\boldsymbol{\theta}) = n\boldsymbol{I}(\boldsymbol{\theta})$ der $\boldsymbol{I}(\boldsymbol{\theta})$ er forventet informasjon til en observasjon.



(e) Når ligningene i punkt (c) ikke har en eksplisitt løsning, kan vi finne maksimum likelihood estimatet ved å bruke Newton-Raphsons metode:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + \bar{\boldsymbol{J}}^{-1}(\boldsymbol{\theta}^{(s)})\boldsymbol{s}(\boldsymbol{\theta}^{(s)})$$

, ved å bruke Fishers scoringsalgoritme:

$$\boldsymbol{\theta}^{(s+1)} = \boldsymbol{\theta}^{(s)} + \bar{\boldsymbol{I}}^{-1}(\boldsymbol{\theta}^{(s)})\boldsymbol{s}(\boldsymbol{\theta}^{(s)}),$$

eller ved passende modifikasjoner av disse.

(f) Når vi har "tilstrekkelig mye" data, er $\hat{\theta}_i$ tilnærmet normalfordelt med forventning θ_i og med varians lik det i-te diagonalelementet til $\bar{\boldsymbol{I}}^{-1}(\boldsymbol{\theta})$. Kovariansen mellom $\hat{\theta}_i$ og $\hat{\theta}_j$ er tilnærmet lik element (i, j) i $\bar{\boldsymbol{I}}^{-1}(\boldsymbol{\theta})$. Vi kan estimere varianser/kovarianser ved å sette inn $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ i $\bar{\boldsymbol{I}}^{-1}(\boldsymbol{\theta})$ eller i $\bar{\boldsymbol{J}}^{-1}(\boldsymbol{\theta})$.