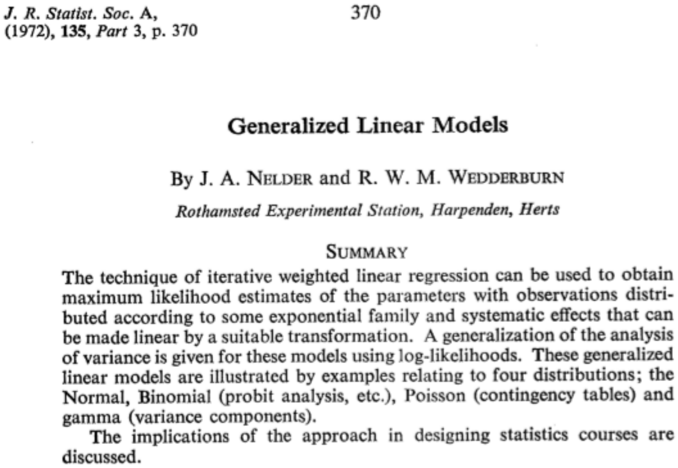
Lecture 1. Introduction, preparation.

Course Designer: Jesper Rydén, SLU, Uppsala, Sweden

Course Lecturer: Raazesh Sainudiin, UU, Uppsala, Sweden

Generalized Linear Models 1MS369 • Autumn 2018

Nelder and Wedderburn (1972)



J. R. Statist. Soc. A,
(1972), **135**, *Part* 3, p. 370

Generalized Linear Models

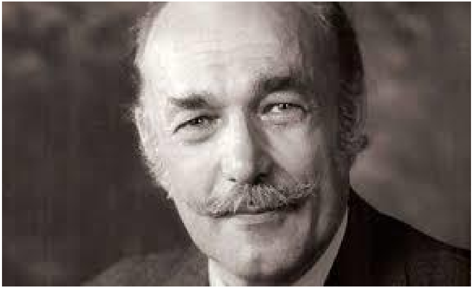
By J. A. Nelder and R. W. M. Wedderburn

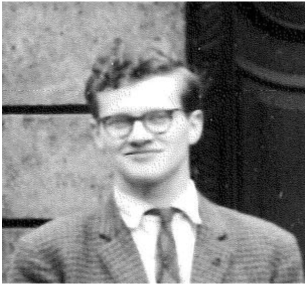
Rothamsted Experimental Station, Harpenden, Herts

Summary

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.





John Ashworth Nelder (1924-2010)
Robert William Maclagan Wedderburn (1947-1975)

A conversation with John Nelder

Statistical Science, Vol. 18, No. 1 (2003). Stephen Senn.

SS: I must confess to having some confusion when I was a young statistician between general linear models and generalized linear models. Do you regret the terminology?

A conversation with John Nelder

Statistical Science, Vol. 18, No. 1 (2003). Stephen Senn.

SS: I must confess to having some confusion when I was a young statistician between general linear models and generalized linear models. Do you regret the terminology?

JN: I think I probably do. I suspect we should have found some more fancy name for it that would have stuck and not been confused with the general linear model, although general and generalized are not quite the same. I can see why it might have been better to have thought of something else.

Apropos the 1972 paper:

JN: I should add, perhaps, that a rather eminent statistical journal to which it was submitted first, turned it down flat without any opportunity to resubmit.

Apropos the 1972 paper:

JN: I should add, perhaps, that a rather eminent statistical journal to which it was submitted first, turned it down flat without any opportunity to resubmit.

SS: What were the reasons for turning it down? Can you remember?

Apropos the 1972 paper:

JN: I should add, perhaps, that a rather eminent statistical journal to which it was submitted first, turned it down flat without any opportunity to resubmit.

SS: What were the reasons for turning it down? Can you remember?

JN: Not enough mathematics, I suspect.

Apropos the 1972 paper:

JN: I should add, perhaps, that a rather eminent statistical journal to which it was submitted first, turned it down flat without any opportunity to resubmit.

SS: What were the reasons for turning it down? Can you remember?

JN: Not enough mathematics, I suspect.

SS: Wedderburn died very young. Did you have plans for further collaboration?

Apropos the 1972 paper:

JN: I should add, perhaps, that a rather eminent statistical journal to which it was submitted first, turned it down flat without any opportunity to resubmit.

SS: What were the reasons for turning it down? Can you remember?

JN: Not enough mathematics, I suspect.

SS: Wedderburn died very young. Did you have plans for further collaboration?

JN: I'm sure we would have gone on. He was full of ideas, and I would have liked to develop many things. He died of anaphylatic shock from an insect bite on a canal holiday, aged 28. It was very sad.

Numerical algorithm

Algorithms and software was crucial.

GLIM: **G**eneralized **L**inear **I**nteractive **M**odelling

A statistical software program for fitting generalized linear models (GLMs). It was developed by the Royal Statistical Society's Working Party on Statistical Computing (later renamed the GLIM Working Party), chaired initially by John Nelder.

First release: 1974. Last major release: GLIM4, in 1993.
GLIM was distributed by the Numerical Algorithms Group (NAG).

(GLIM is no longer actively developed or distributed.)

Components of a linear model

Essentially three components:

- **Random component.** Specifies the response variable Y and its probability distribution. Observations: $\mathbf{y} = (y_1, \ldots, y_n)^\mathsf{T}$, outcomes of $\mathbf{Y}$.
- **Linear predictor.** A parameter vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\mathsf{T}$ and a $n \times p$ design matrix $\mathbf{X}$; linear predictor: $\mathbf{X}\boldsymbol{\beta}$. (Also called systematic component.)
- **Link function.** A function g applied to each component of $\mathrm{E}[\mathbf{Y}]$, relating it to the linear predictor:

$$g(\mathrm{E}[\mathbf{Y}]) = \mathbf{X}\beta$$

Linear models and transformations

Nelder and Wedderburn (1972) introduced the class of GLMs and the algorithm for fitting.

However, many models in the class were in practice by then.

Transforming the response and then performing usual regression models could be an option. Common suggestions, examples:

Distribution of Y	V$[Y]$ in terms of mean μ	Transformation
Poisson	μ	$\sqrt{Y}$
Binomial	$\mu(1-\mu)/n$	$\sin^{-1}(\sqrt{Y})$

Warning

However, a transformation should be able to fix assumptions of constant variance as well as normality.

In fact, some authors today propose *not* to transform, e.g.

OHara, R., & Kotze, D. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, **1** (2), 118-122.

Example: Transformation of Poisson

Let $X \sim \text{Po}(\mu)$. It can be shown, by using the delta method, that

(a) If $Y = X^{2/3}$, then the skewness $\text{E}[(Y - \text{E}[Y])^3]/\sigma_Y^3 \approx 0$ (where $\sigma_Y^2 = \text{V}[Y]$).

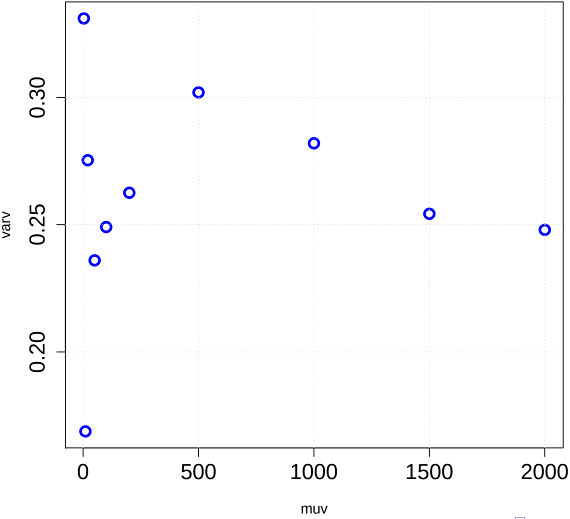
(b) If $Y = X^{1/2}$, then its variance $\sigma_Y^2 \approx 1/4$ is stable.

(c) If $Y = \ln(X + \frac{1}{2})$, then $\text{E}[Y] \approx \ln \mu$ (log-linearity).

None of the above transformations at the same time leads to skewness zero, homoskedasticity and additive systematic effects.

Example: Simulation study, transformation of Poisson

Investigation of the approximation $\mathrm{V}[Y] \approx 1/4$, influence on the mean parameter μ. Sample size for each μ: 200 observations.



Epidemicological model

In the early stages of a disease epidemic, the rate at which new cases occur can often increase exponentially over time.

Let μ_i be the expected number of new cases on day t_i. Model:

$$\mu_i = \gamma \mathrm{e}^{\delta t_i}$$

where γ and δ are unknown parameters.

Turn into GLM form:

$$\ln(\mu_i) = \ln \gamma + \delta t_i = \beta_0 + \beta_1 t_i.$$

Comments:

- Response: a count, maybe try a Poisson distribution
- Link function: Log link
- Linear predictor

Example from Wood (2006)

Rate of capture of prey animals

The rate of capture of prey items. y_i, by a hunting animal, tends to increase with increasing density of prey, x_i.

Eventually, the rate levels off, when the predator is catching as much as it can cope with.

Model:

$$\mu_i = \frac{\alpha x_i}{h + x_i}$$

where α and h are unknown parameters.

α: the maximum capture rate.

h: the prey density at which the capture rate is half the maximum rate.

Use a *reciprocal link*, the right-hand side can be made linear in the parameters:

$$\frac{1}{\mu_i} = \frac{1}{\alpha} + \frac{h}{\alpha}\frac{1}{x_i} = \beta_0 + \frac{1}{x_i}\beta_1.$$

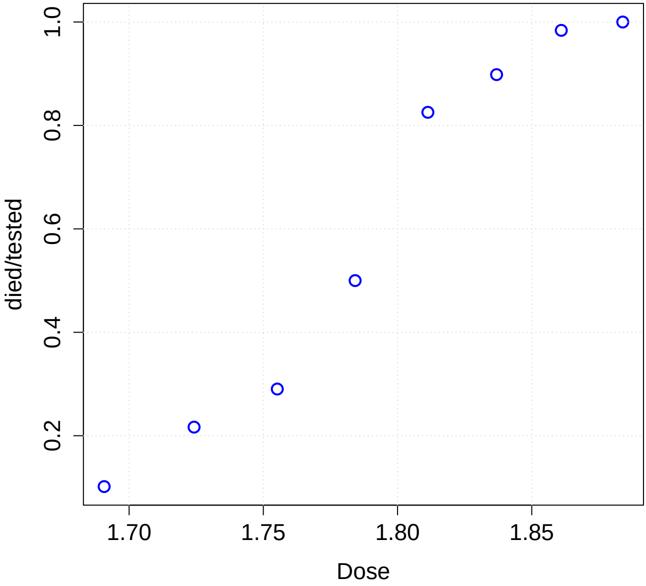
Example from Wood (2006)

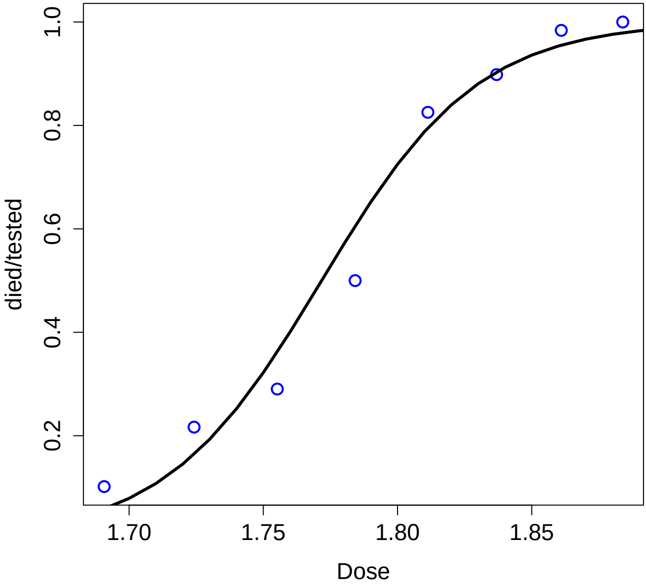
Example. Beetle mortality data

Insects were exposed to gaseous carbon disulphide ($\log \mathrm{CS}_2$ mgl^{-1} for a period of 5 hours (Bliss, 1935). The dose, the number of insects and the number of killed insects were registred for 8 experiments:

Dose x_i	Number tested n_i	Number killed z_i
`Dose`	`tested`	`died`
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

Proportion killed insects as a function of dose?


died/tested
0.2
0.4
0.6
0.8
1.0
1.70
1.75
1.80
1.85
Dose


died/tested
0.2
0.4
0.6
0.8
1.0
1.70
1.75
1.80
1.85
Dose

R code

```
# Reading data
library(Flury); data(dead.beetles); attach(dead.beetles)

# Fitting model
m1 = glm(cbind(died,tested-died)˜Dose,
  family=binomial(link=logit))
x <- seq(1.6,1.9,by=0.01)
y <- ilogit(m1$coef[1]+m1$coef[2]*x)

# Plotting
plot(Dose,died/tested)
lines(x,y,type=’l’)
```

Score function

DEFINITION. Consider $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\mathsf{T} \in \Theta^k$ and assume that Θ^k is an open subspace of $\mathbb{R}^k$ and that the log-likelihood is continuously differentiable. The function

$$\ell'_\theta(\boldsymbol{\theta}; \mathbf{y}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathbf{y}) = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \ell(\boldsymbol{\theta}; \mathbf{y}) \\ \vdots \\ \frac{\partial}{\partial \theta_k} \ell(\boldsymbol{\theta}; \mathbf{y}) \end{pmatrix}$$

is called the **score function**, often denoted $S(\boldsymbol{\theta}, \mathbf{y})$.

REMARK. The score function depends on $\mathbf{y}$. Hence, the score function $S(\boldsymbol{\theta}, \mathbf{Y})$ is a random variable.

Theorem: score functions

THEOREM. Under normal regularity conditions,

$$\mathsf{E}_{\boldsymbol{\theta}_0}\left[\frac{\partial}{\partial\boldsymbol{\theta}}S(\boldsymbol{\theta};\mathbf{Y})\right]=\mathbf{0}$$

where the subscript $\boldsymbol{\theta}_0$ means that the expectation is carried out with respect to the density $f(\boldsymbol{\theta}_0;\mathbf{y})$.

REMARK. This result, combined with properties of exponential families, is a key issue in the foundations of GLM.

Observed information

DEFINITION. The matrix

$$\mathbf{j}(\boldsymbol{\theta}; \mathbf{y}) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\mathsf{T}}\ell(\boldsymbol{\theta}; \mathbf{y})$$

with the elements

$$\mathbf{j}(\boldsymbol{\theta}; \mathbf{y})_{i,j} = -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\boldsymbol{\theta}; \mathbf{y})$$

is called the **observed information** corresponding to the observation $\mathbf{y}$ and the parameter $\boldsymbol{\theta}$.

REMARK. The observed information is equal to the *Hessian* (with opposite sign) of the log-likelihood function evaluated at $\boldsymbol{\theta}$.

Expected information

DEFINITION. The expectation of the observed information

$$\mathbf{i}(\boldsymbol{\theta}) = \mathsf{E}[\mathbf{j}(\boldsymbol{\theta}; \mathbf{Y})]$$

where the expectation is determined under the distribution corresponding to $\boldsymbol{\theta}$, is called the **expected information** or the **information matrix** corresponding to the parameter $\boldsymbol{\theta}$.

It is also known as **Fisher's information matrix**.

Invariance property of likelihood estimators

THEOREM. Assume that $\widehat{\boldsymbol{\theta}}$ is an ML estimator for $\boldsymbol{\theta}$ and let $\boldsymbol{\psi} = \boldsymbol{\psi}(\boldsymbol{\theta})$ denote a one-to-one mapping of $\Omega \subset \mathbb{R}^k$ onto $\Psi \subset \mathbb{R}^k$.

Then the estimator $\boldsymbol{\psi}(\widehat{\boldsymbol{\theta}})$ is an ML estimator for the parameter $\boldsymbol{\psi}(\boldsymbol{\theta})$.

Distribution of the ML estimate

THEOREM. Assume that $\widehat{\boldsymbol{\theta}}$ is consistent. Then, under some regularity conditions,

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \to \mathsf{N}(0, \mathbf{i}(\boldsymbol{\theta})^{-1}).$$

Distribution of the ML estimate

THEOREM. Assume that $\widehat{\boldsymbol{\theta}}$ is consistent. Then, under some regularity conditions,

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \to \mathsf{N}(0, \mathbf{i}(\boldsymbol{\theta})^{-1}).$$

REMARK. In practice, we use the observed Fisher information:

$$\widehat{\boldsymbol{\theta}} \sim \mathsf{N}(\boldsymbol{\theta},\ \mathbf{j}^{-1}(\widehat{\boldsymbol{\theta}})).$$

The standard error of $\widehat{\theta}_i$ is given by the ith diagonal element of $\mathbf{j}^{-1}(\widehat{\boldsymbol{\theta}})$. An estimate of the *dispersion matrix* (variance–covariance matrix) is thus $\mathbf{j}^{-1}(\widehat{\boldsymbol{\theta}})$.

Example: Invariance

Consider the model

$$y_i \sim \mathrm{Po}(\mathrm{e}^{\theta_1 x_i + \theta_2}), \quad i = 1, \ldots, 10.$$

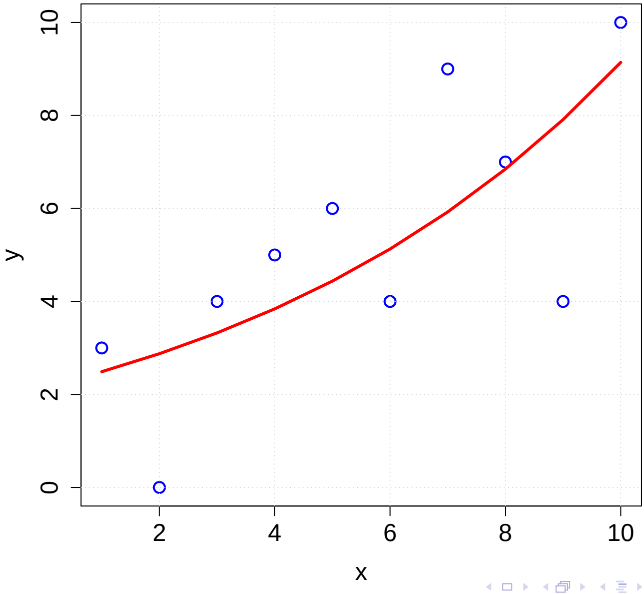
Numerical example in R:

```
# Data
x = c(1:10); y = c(3, 0, 4, 5, 6, 4, 9, 7, 4, 10);
# Define loglikelihood function
logl <- function(th){
  -sum(dpois(y,exp(th[1]*x + th[2]),log=TRUE))
}
# Optimisation
fitml = optim(par=c(1,mean(y)),fn=logl,hessian=TRUE)
```

Estimates of parameters θ_1 and θ_2 and related standard errors? Confidence intervals for the parameters? Testing that $\theta_1 = 0$?

R and Blackboard

Example: Fitted model



Likelihood ratio

Consider the hypothesis

$$H_0 : \ \boldsymbol{\theta} \in \Omega_0$$

against the alternative

$$H_1 : \ \boldsymbol{\theta} \in \Omega \backslash \Omega_0$$

where $\dim(\Omega_0) = m$ and $\dim(\Omega) = k$.

For given observations $\boldsymbol{y} = y_1, y_2, \ldots, y_n$, the likelihood ratio is defined as

$$\lambda(\boldsymbol{y}) = \frac{\sup_{\boldsymbol{\theta} \in \Omega_0} L(\boldsymbol{\theta}; \boldsymbol{y})}{\sup_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\theta}; \boldsymbol{y})}$$

Wilk's likelihood ratio test

THEOREM. For $\lambda(\boldsymbol{y})$, under the null hypothesis H_0

$$-2\log\lambda(\boldsymbol{Y}) \to \chi^2(k-m)$$

(convergence in distribution)

The null model and full model

Null model:
$\Omega_{\text{null}} = \mathbb{R}, \ (\dim(\Omega_{\text{null}} = 1):$ a model with only one parameter.

Full model:
$\Omega_{\text{full}} = \mathbb{R}^n, \ (\dim(\Omega_{\text{null}} = n):$ a model whose dimension is equal to the number of observations. The model fits each observation perfectly.

The Deviance: first encounter

Introduce

$$L_0 = \sup_{\boldsymbol{\theta} \in \Omega_0} L(\boldsymbol{\theta}; \boldsymbol{y}) \quad \text{and} \quad L = \sup_{\boldsymbol{\theta} \in \Omega_{\text{full}}} L(\boldsymbol{\theta}; \boldsymbol{y})$$

It follows that

$$-2 \log \lambda(\boldsymbol{Y}) = -2 \log(L_0 - \log L = 2(\log L - \log L_0)$$

The statistic

$$D = -2 \log \lambda(\boldsymbol{Y}) = 2(\log L - \log L_0)$$

is called the *deviance* by Nelder and Wedderburn (1972).

Example: Likelihood-ratio test

Likelihood-ratio test for NH: $\theta_1 = 0$ in the model $Y_i \sim \mathrm{Po}(\exp(\theta_1 x + \theta_2))$.

```
lA <- function(th) {
      -sum(dpois(y, exp(th[1] * x + th[2]), log = TRUE)) }

fitA <- optim(par = c(0, 0), fn = lA, hessian = TRUE)

lB <- function(th) { -sum(dpois(y, exp(0 * x + th[1]), log = TRUE)) }

fitB <- optim(par = c(0), fn = lB, hessian = TRUE)

D <- 2 * (fitB$value - fitA$value)

1 - pchisq(D, 1)
```

R and Blackboard

Multivariate normal distribution

Consider $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^{\mathsf{T}}$ and assume that

$$\mathbf{Y} \sim \mathrm{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma}).$$

Multivariate normal distribution

Consider $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\mathsf{T}$ and assume that

$$\mathbf{Y} \sim \mathrm{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma}).$$

DEFINITION. The deviance in this context is

$$D(\mathbf{y}; \boldsymbol{\mu}) = (\mathbf{y} - \boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$$

Interpretation: A quadratic norm of the vector $(\mathbf{y} - \boldsymbol{\mu})$ corresponding to the inner product defined by $\boldsymbol{\Sigma}^{-1}$.

Multivariate normal distribution

Consider $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^{\mathsf{T}}$ and assume that

$$\mathbf{Y} \sim \mathrm{N}_n(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma}).$$

DEFINITION. The deviance in this context is

$$D(\mathbf{y}; \boldsymbol{\mu}) = (\mathbf{y} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})$$

Interpretation: A quadratic norm of the vector $(\mathbf{y} - \boldsymbol{\mu})$ corresponding to the inner product defined by $\boldsymbol{\Sigma}^{-1}$.

The normal density can be written

$$f(\mathbf{y}; \boldsymbol{\mu}, \sigma^2) = \frac{1}{(\sqrt{2\pi})^n \sigma^n \sqrt{\det \boldsymbol{\Sigma}}} \exp\left[-\frac{1}{2\sigma^2} D(\mathbf{y}; \boldsymbol{\mu})\right].$$

Likelihood function, score function, information matrices:

Blackboard

Profile likelihood

DEFINITION. Consider a vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with the likelihood function $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; y)$.

Suppose that $\hat{\boldsymbol{\theta}}_{2.1}$ is the MLE of $\boldsymbol{\theta}_2$ for a given value of $\boldsymbol{\theta}_1$. Then the **profile likelihood** for $\boldsymbol{\theta}_1$ is $L(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_{2.1}; y)$.

Profile likelihood

DEFINITION. Consider a vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with the likelihood function $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; y)$.

Suppose that $\hat{\boldsymbol{\theta}}_{2.1}$ is the MLE of $\boldsymbol{\theta}_2$ for a given value of $\boldsymbol{\theta}_1$. Then the **profile likelihood** for $\boldsymbol{\theta}_1$ is $L(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_{2.1}; y)$.

REMARK 1. Note that the profile likelihood is not a genuine likelihood function. For instance, the properties of score functions generally do not apply.

Profile likelihood

DEFINITION. Consider a vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with the likelihood function $L(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2; y)$.

Suppose that $\hat{\boldsymbol{\theta}}_{2.1}$ is the MLE of $\boldsymbol{\theta}_2$ for a given value of $\boldsymbol{\theta}_1$. Then the **profile likelihood** for $\boldsymbol{\theta}_1$ is $L(\boldsymbol{\theta}_1, \hat{\boldsymbol{\theta}}_{2.1}; y)$.

REMARK 1. Note that the profile likelihood is not a genuine likelihood function. For instance, the properties of score functions generally do not apply.

REMARK 2. The profile likelihood is one of several modifications of the likelihood, collectively known as pseudo-likelihoods. Other examples: marginal likelihood, conditional likelihood, partial likelihood.

Example on blackboard (Ex. 2.10)