Introduction to General and Generalized Linear Models

The Likelihood Principle - part II

Henrik Madsen
Poul Thyregod

Informatics and Mathematical Modelling
Technical University of Denmark
DK-2800 Kgs. Lyngby

Feb 2012

Last week

- Likelihood function $L(\theta) = P_\theta(Y = y)$
- Log likelihood function $\ell(\theta) = \log(L(\theta))$
- Score function $\ell'(\theta)$
- Maximum likelihood estimate $\widehat{\theta} = \underset{\theta \in \Theta}{\text{argmax}}\ \ell(\theta)$
- Observed information matrix $-\ell''(\widehat{\theta})$

This lecture

- The maximum likelihood estimate (MLE)
- Distribution of the ML estimator
- Quadratic approximation of the log-likelihood
- Model selection
- Dealing with nuisance parameters

The Maximum Likelihood Estimate (MLE)

Definition (Maximum Likelihood Estimate (MLE))

Given the observation $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ the *Maximum Likelihood Estimate (MLE)* is a function $\widehat{\boldsymbol{\theta}}(\boldsymbol{y})$ such that

$$L(\widehat{\boldsymbol{\theta}}; \boldsymbol{y}) = \sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \boldsymbol{y})$$

The function $\widehat{\boldsymbol{\theta}}(\boldsymbol{Y}) = \underset{\boldsymbol{\theta} \in \Theta}{\text{argmax}} L(\boldsymbol{\theta}; \boldsymbol{Y})$ over the sample space of observations $\boldsymbol{Y}$ is called an *ML estimator*.

In practice it is convenient to work with the log-likelihood function $l(\boldsymbol{\theta}; \boldsymbol{y})$.

The Maximum Likelihood Estimate (MLE)

The *score function* can be used to obtain the estimate, since the MLE can be found as the solution to

$$l'_\theta(\boldsymbol{\theta}; \boldsymbol{y}) = 0$$

which are called the *estimation equations for the ML-estimator*, or, just the ML equations.

- It is common practice, especially when plotting, to normalize the likelihood function to have unit maximum and the log-likelihood to have zero maximum.

Invariance property

Theorem (Invariance property)

Assume that $\widehat{\boldsymbol{\theta}}$ is a maximum likelihood 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 a maximum likelihood estimator for the parameter $\boldsymbol{\psi}(\boldsymbol{\theta})$.

Distribution of the ML estimator

Theorem (Distribution of the ML estimator)

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

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \rightarrow \mathrm{N}(0, \boldsymbol{i}(\boldsymbol{\theta})^{-1})$$

where $\boldsymbol{i}(\boldsymbol{\theta})$ is the expected information or the information matrix.

- The results can be used for inference under very general conditions. As the price for the generality, the results are only asymptotically valid.
- Asymptotically the variance of the estimator is seen to be equal to the Cramer-Rao lower bound for any unbiased estimator.
- The practical significance of this result is that the MLE makes efficient use of the available data for large data sets.

Distribution of the ML estimator

In practice, we would use

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

where $\boldsymbol{j}(\widehat{\boldsymbol{\theta}})$ is the observed (Fisher) information.

Remember how we get these in practice.

This means that asymptotically

i) $\mathrm{E}[\widehat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$

ii) $D[\widehat{\boldsymbol{\theta}}] = \boldsymbol{j}^{-1}(\widehat{\boldsymbol{\theta}})$

Distribution of the ML estimator

- The standard error of $\widehat{\theta}_i$ is given by

$$\widehat{\sigma}_{\hat{\theta}_i} = \sqrt{\mathrm{Var}_{ii}[\widehat{\theta}]}$$

  where $\mathrm{Var}_{ii}[\widehat{\theta}]$ is the i'th diagonal term of $\boldsymbol{j}^{-1}(\widehat{\boldsymbol{\theta}})$
- Hence we have that an estimate of the dispersion (variance-covariance matrix) of the estimator is

$$D[\widehat{\boldsymbol{\theta}}] = \boldsymbol{j}^{-1}(\widehat{\boldsymbol{\theta}})$$

- An estimate of the uncertainty of the individual parameter estimates is obtained by decomposing the dispersion matrix as follows:

$$D[\widehat{\boldsymbol{\theta}}] = \widehat{\boldsymbol{\sigma}}_{\hat{\boldsymbol{\theta}}} \boldsymbol{R} \widehat{\boldsymbol{\sigma}}_{\hat{\boldsymbol{\theta}}}$$

  into $\widehat{\boldsymbol{\sigma}}_{\hat{\boldsymbol{\theta}}}$, which is a diagonal matrix of the standard deviations of the individual parameter estimates, and $\boldsymbol{R}$, which is the corresponding correlation matrix. The value R_{ij} is thus the estimated correlation between $\widehat{\theta}_i$ and $\widehat{\theta}_j$.

The Wald Statistic

A test of an individual parameter

$$\mathcal{H}_0 : \theta_i = \theta_{i,0}$$

is given by the *Wald statistic*:

$$Z_i = \frac{\widehat{\theta}_i - \theta_{i,0}}{\widehat{\sigma}_{\hat{\theta}_i}}$$

which under $\mathcal{H}_0$ is approximately $\mathrm{N}(0,1)$-distributed.

Fitting Poisson regression (Ex.: Invariance Property)

- Consider the model:

$$y_i \sim \text{Pois}(e^{\theta_1 x_i + \theta_2}), \; i = 1, \ldots, 10$$

- With observations given in the following R snip:

```
> x <- 1:10
> y <- c(3, 0, 4, 5, 6, 4, 9, 7, 4, 10)
> l <- function(th) {
+     -sum(dpois(y, exp(th[1] * x + th[2]), log = TRUE))
+ }
> fit <- optim(par = c(0, 0), fn = l, hessian = TRUE)
```

Fitting object contain

```
> fit

$par
[1] 0.1395798 0.8017717

$value
[1] 21.41071

$counts
function gradient
      91       NA

$convergence
[1] 0

$message
NULL

$hessian
          [,1]      [,2]
[1,] 2665.3444 343.96835
[2,]  343.9684  51.99424
```

Fitting object contain

- Estimates of θ_1 and θ_2

```
> est <- fit$par
> est
```

```
[1] 0.1395798 0.8017717
```

- Estimates of their standard deviations

```
> std <- sqrt(diag(solve(fit$hessian)))
> std
```

```
[1] 0.05064875 0.36263335
```

- Setup confidence intervals

```
> ci <- cbind(low = est - 2 * std, high = est + 2 * std)
> ci
```

```
             low      high
[1,] 0.03828231 0.2408773
[2,] 0.07650498 1.5270384
```

Fitting object contain

- The correlation matrix R

```
> solve(fit$hessian)/(std %o% std)
```

```
             [,1]       [,2]
[1,]  1.0000000 -0.9239835
[2,] -0.9239835  1.0000000
```

- Wald test for no influence of x, or in other words $\theta_1 = 0$

```
> Z <- (est[1] - 0)/std[1]
> Z
```

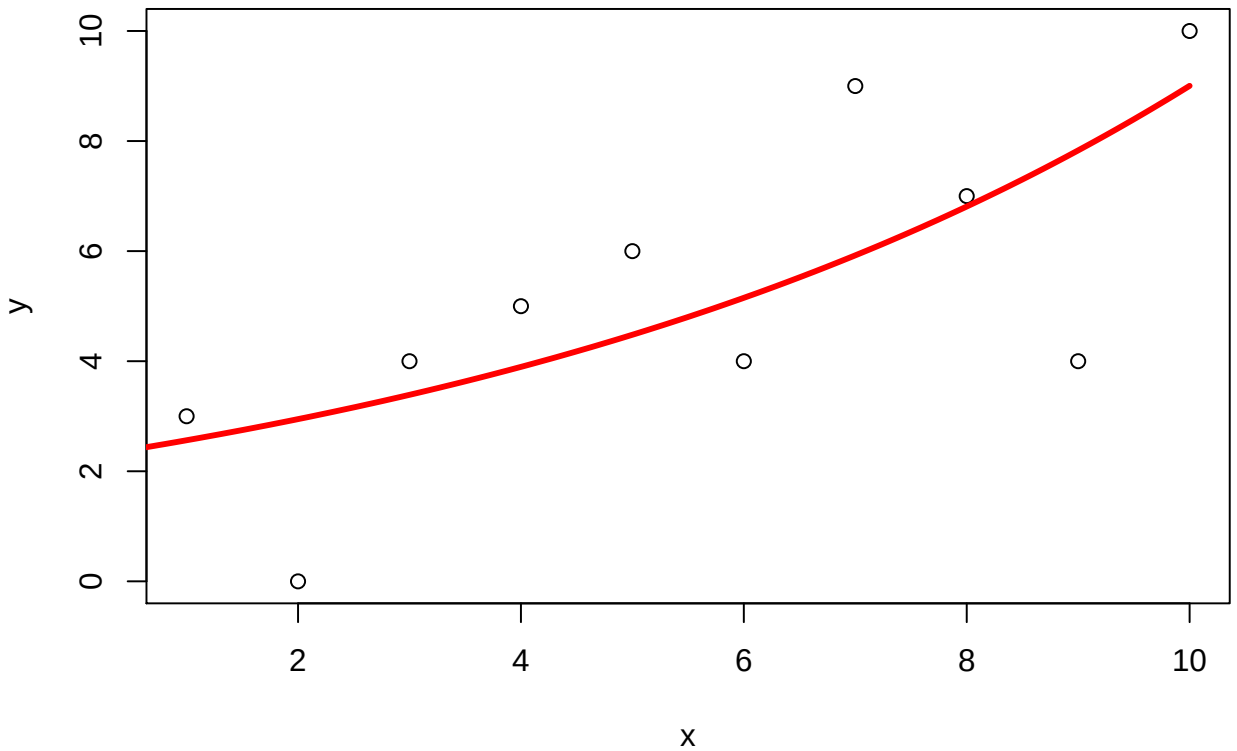
```
[1] 2.755839
```

- Greater than 2, so hypothesis can be rejected, the p-value is:

```
> 2 * (1 - pnorm(abs(Z)))
```

```
[1] 0.005854177
```

Plot it



Quadratic approximation of the log-likelihood

- A second-order Taylor expansion around $\widehat{\theta}$ provides us with a quadratic approximation of the normalized log-likelihood around the MLE.
- A second-order Taylors expansion around $\widehat{\theta}$ we get

$$l(\theta) \approx l(\widehat{\theta}) + l'(\widehat{\theta})(\theta - \widehat{\theta}) - \frac{1}{2} j(\widehat{\theta})(\theta - \widehat{\theta})^2$$

and then

$$\log \frac{L(\theta)}{L(\widehat{\theta})} \approx -\frac{1}{2} j(\widehat{\theta})(\theta - \widehat{\theta})^2$$

- In the case of normality the approximation is exact which means that a quadratic approximation of the log-likelihood corresponds to normal approximation of the $\widehat{\theta}(\boldsymbol{Y})$ estimator.

Example: Quadratic approximation of the log-likelihood

Consider again the thumbtack example.
The log-likelihood function is:

$$l(\theta) = y \log \theta + (n - y) \log(1 - \theta) + const$$

The score function is:

$$l'(\theta) = \frac{y}{\theta} - \frac{n - y}{1 - \theta},$$

and the observed information:

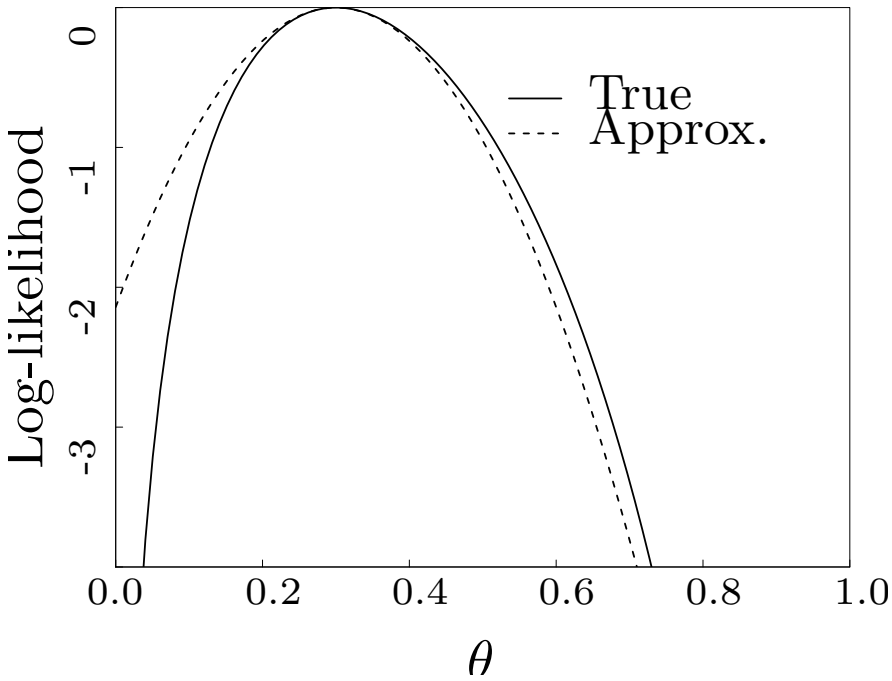
$$j(\theta) = \frac{y}{\theta^2} + \frac{n - y}{(1 - \theta)^2}.$$

For $n = 10,\ y = 3$ and $\widehat{\theta} = 0.3$ we obtain

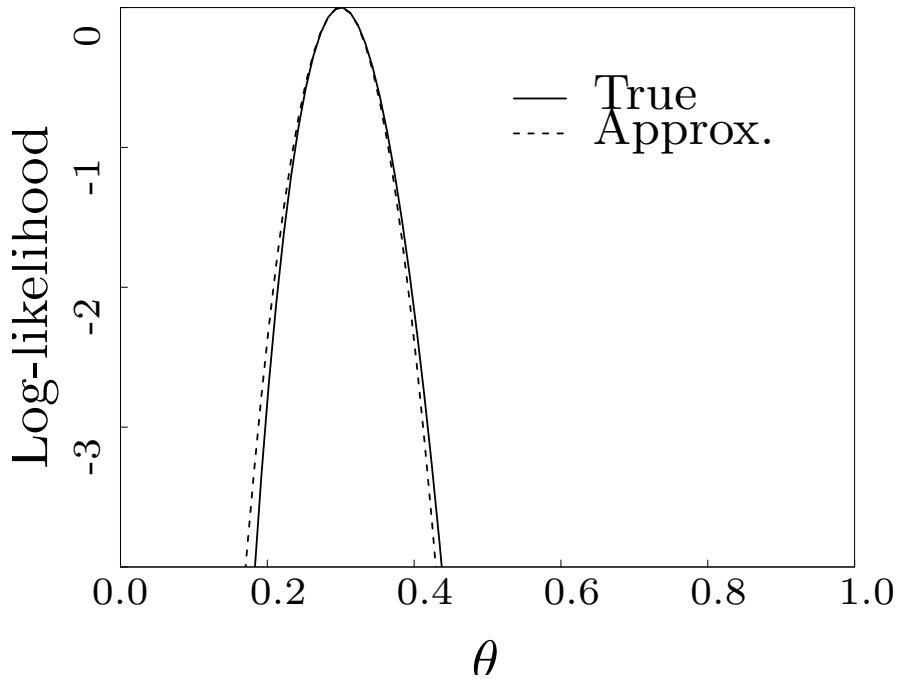
$$j(\widehat{\theta}) = 47.6$$

The quadratic approximation is poor in this case. By increasing the sample size to $n = 100$, but still with $\widehat{\theta} = 0.3$ the approximation is much better.

Example: Quadratic approximation of the log-likelihood



(a) $n = 10, y = 3$



(b) $n = 100, y = 30$

Figure: Quadratic approximation of the log-likelihood function.

Likelihood ratio tests

- Method for testing hypotheses using the likelihood function.
- Assume we have a sufficient model (A) with $\boldsymbol{\theta} \in \Omega_A$
- Want to test if we can reduce to a sub-model (B) $\mathcal{H}_0 : \boldsymbol{\theta} \in \Omega_B$
- We call B a sub-model of A if $\Omega_B \subset \Omega_A$ (for instance when setting a free parameter in A to a fixed value in B)
- The purpose is to analyze if the observations provide sufficient evidence to reject the model reduction, otherwise we accept the hypothesis $\mathcal{H}_0$
- The test statistic D (called deviance) is calculated as:

$$D = 2(\ell_A(\widehat{\boldsymbol{\theta}_A}, \boldsymbol{Y}) - \ell_B(\widehat{\boldsymbol{\theta}_B}, \boldsymbol{Y}))$$

- $D \geq 0$, so "small values" means model B is as good as A
- "Large values" means model B gives worse data description than A

Likelihood ratio tests

- Want to know the distribution of D when assuming $\mathcal{H}_0$ (model B).
- It is sometimes possible to calculate the exact distribution. This is for instance the case for the General Linear Model for Gaussian data.
- In most cases, however, we must use following important result regarding the asymptotic behavior.

Theorem (Wilk's Likelihood Ratio test)

The random variable $D = 2(\ell_A(\widehat{\boldsymbol{\theta}_A}, \boldsymbol{Y}) - \ell_B(\widehat{\boldsymbol{\theta}_B}, \boldsymbol{Y}))$ *converges in law to a* χ^2 *random variable with* $f = (\dim(\Omega_A) - \dim(\Omega_B))$ *degrees of freedom, i.e.,*

$$D \to \chi^2(f)$$

under $\mathcal{H}_0$.

Likelihood ratio tests

- With Wilk's in place we have the (approximate) distribution of D if we assume $\mathcal{H}_0$ to be true
- Further we have calculated the observed D
- The evidence against $\mathcal{H}_0$ is measured by the *p-value*.
- The p-value is the probability under $\mathcal{H}_0$ of observing a value of D equal to or more extreme as the actually observed test statistic.
- Hence, a small p-value (say ≤ 0.05) leads to a strong evidence against $\mathcal{H}_0$, and $\mathcal{H}_0$ is then said to be *rejected*. Likewise, we retain $\mathcal{H}_0$ unless there is a strong evidence against this hypothesis.
- Rejecting $\mathcal{H}_0$ given $\mathcal{H}_0$ is true is called a *Type I error*, while retaining $\mathcal{H}_0$ when the truth is actually $\mathcal{H}_1$ is called a *Type II error*.

Likelihood ratio tests

With the notation from the book

Definition (Likelihood ratio)

Consider the hypothesis $\mathcal{H}_0 : \boldsymbol{\theta} \in \Omega_0$ against the alternative $\mathcal{H}_1 : \boldsymbol{\theta} \in \Omega \setminus \Omega_0$ ($\Omega_0 \subseteq \Omega$), where $\dim(\Omega_0) = r$ and $\dim(\Omega) = k$.
For given observations $y_1, y_2, ..., y_n$ the *likelihood ratio* is defined as

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

- If λ is small, then the data are seen to be more plausible under the alternative hypothesis than under the null hypothesis.
- Hence the hypothesis ($\mathcal{H}_0$) is rejected for small values of λ.

Null model and full model

The null model

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

The full model

$\Omega_{\text{full}} = \mathbb{R}^n$ $(\dim(\Omega_{\text{full}}) = n)$, is a model where the dimension is equal to the number of observations and hence the model fits each observation perfectly.

The deviance

The deviance

Let us introduce $L_0 = \sup_{\boldsymbol{\theta} \in \Omega_0} L(\boldsymbol{\theta}; \boldsymbol{y})$ and $L = \sup_{\boldsymbol{\theta} \in \Omega_{\text{full}}} L(\boldsymbol{\theta}; \boldsymbol{y})$ then we notice that

$$\begin{aligned} -2 \log \lambda(\boldsymbol{Y}) &= -2(\log L_0 - \log L) \\ &= 2(\log L - \log L_0). \end{aligned}$$

The statistic $\mathrm{D} = -2 \log \lambda(\boldsymbol{Y}) = 2(\log L - \log L_0)$ is called the *deviance*.

Example likelihood ratio test

- Likelihood ratio test for $\theta_1 = 0$ in the model $y_i \sim \text{Pois}(e^{\theta_1 x_i + \theta_2})$

```
> 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)
> D
[1] 7.96622
> 1 - pchisq(D, 1)
[1] 0.004765838
```

Hypothesis chains

Consider a *chain* of hypotheses specified by a sequence of parameter spaces

$$\mathbb{R} \subseteq \Omega_M \ldots \subset \Omega_2 \subset \Omega_1 \subset \mathbb{R}^n.$$

For each parameter space Ω_i we define a hypothesis

$$\mathcal{H}_i : \boldsymbol{\theta} \in \Omega_i$$

with $\dim(\Omega_i) < \dim(\Omega_{i-1})$.

Partial likelihood ratio test

Definition (Partial likelihood ratio test)

Assume that the hypothesis $\mathcal{H}_i$ allows the sub hypothesis $\mathcal{H}_{i+1} \subset \mathcal{H}_i$. The *partial likelihood ratio test* for $\mathcal{H}_{i+1}$ under $\mathcal{H}_i$ is the likelihood ratio test for the hypothesis $\mathcal{H}_{i+1}$ under the assumption that the hypothesis $\mathcal{H}_i$ holds. The likelihood ratio test statistic for this partial test is

$$\lambda_{\mathcal{H}_{i+1}|\mathcal{H}_i}(y) = \frac{\sup_{\boldsymbol{\theta}\in\Omega_{i+1}} L(\boldsymbol{\theta};\boldsymbol{y})}{\sup_{\boldsymbol{\theta}\in\Omega_i} L(\boldsymbol{\theta};\boldsymbol{y})}$$

When $\mathcal{H}_{i+1}$ holds, the distribution of $\lambda_{\mathcal{H}_{i+1}|\mathcal{H}_i}(Y)$ approaches a $\chi^2(f)$ distribution with $f = \dim(\Omega_i) - \dim(\Omega_{i+1})$.

Partial tests

Theorem (Partitioning into a sequence of partial tests)

Consider a chain of hypotheses.

Now, assume that $\mathcal{H}_1$ holds, and consider the minimal hypotheses $\mathcal{H}_M : \boldsymbol{\theta} \in \Omega_M$ with the alternative $\mathcal{H}_1 : \boldsymbol{\theta} \in \Omega_1 \setminus \Omega_M$. The likelihood ratio test statistic $\lambda_{\mathcal{H}_M|\mathcal{H}_1}(y)$ for this hypothesis may be factorized into a chain of partial likelihood ratio test statistics $\lambda_{\mathcal{H}_{i+1}|\mathcal{H}_i}(y)$ for $\mathcal{H}_{i+1}$ given $\mathcal{H}_i$, $i = 1, \ldots, M$.

- The partial tests "corrects" for the effect of the parameters that are in the model at that particular stage
- When interpreting the test statistic corresponding to a particular stage in the hierarchy of models, one often says that there is "controlled for", or "corrected for" the effect of the parameters that are in the model at that stage.

Two factor experiment

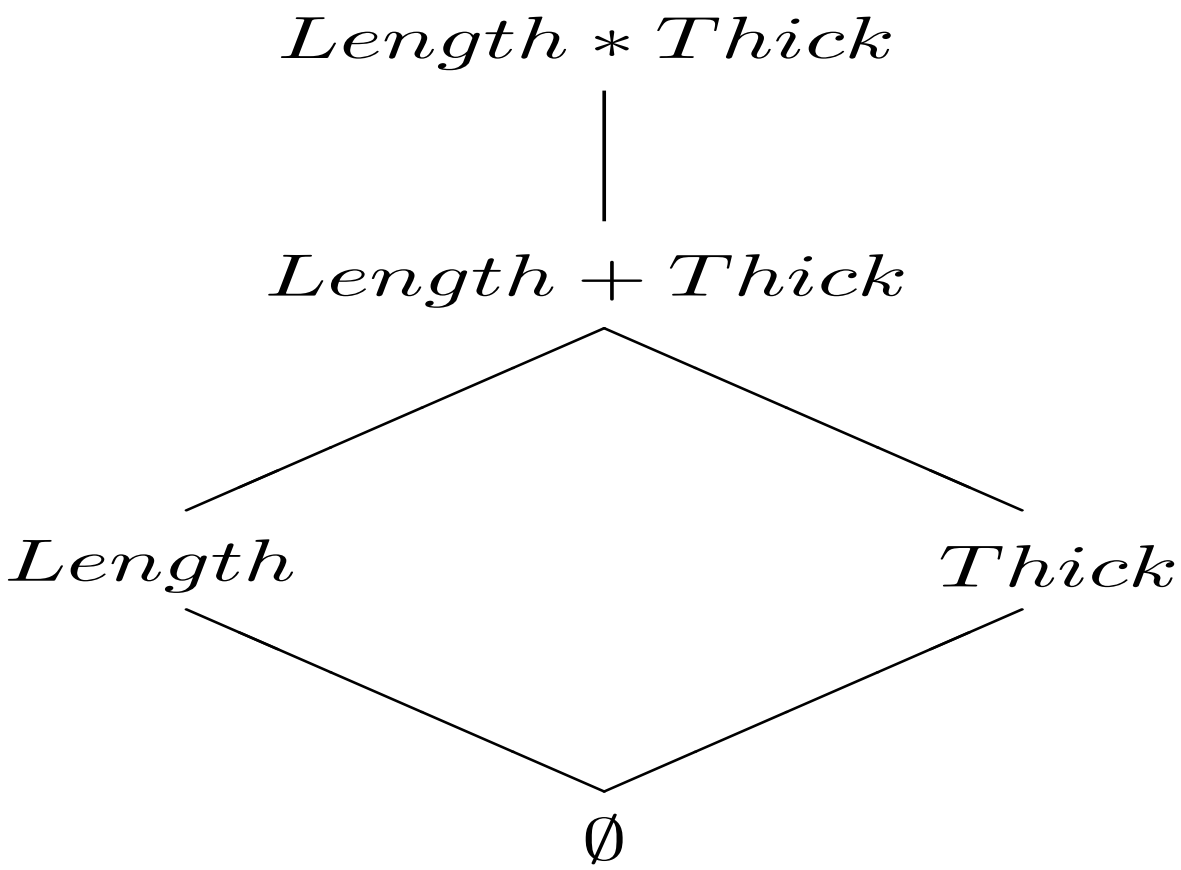


Figure: Inclusion diagram corresponding to a two-factor experiment. The notation is the same as used by the software R, i.e. Length∗Thick denotes a two-factor model with interaction between the two factors

Three factor experiment

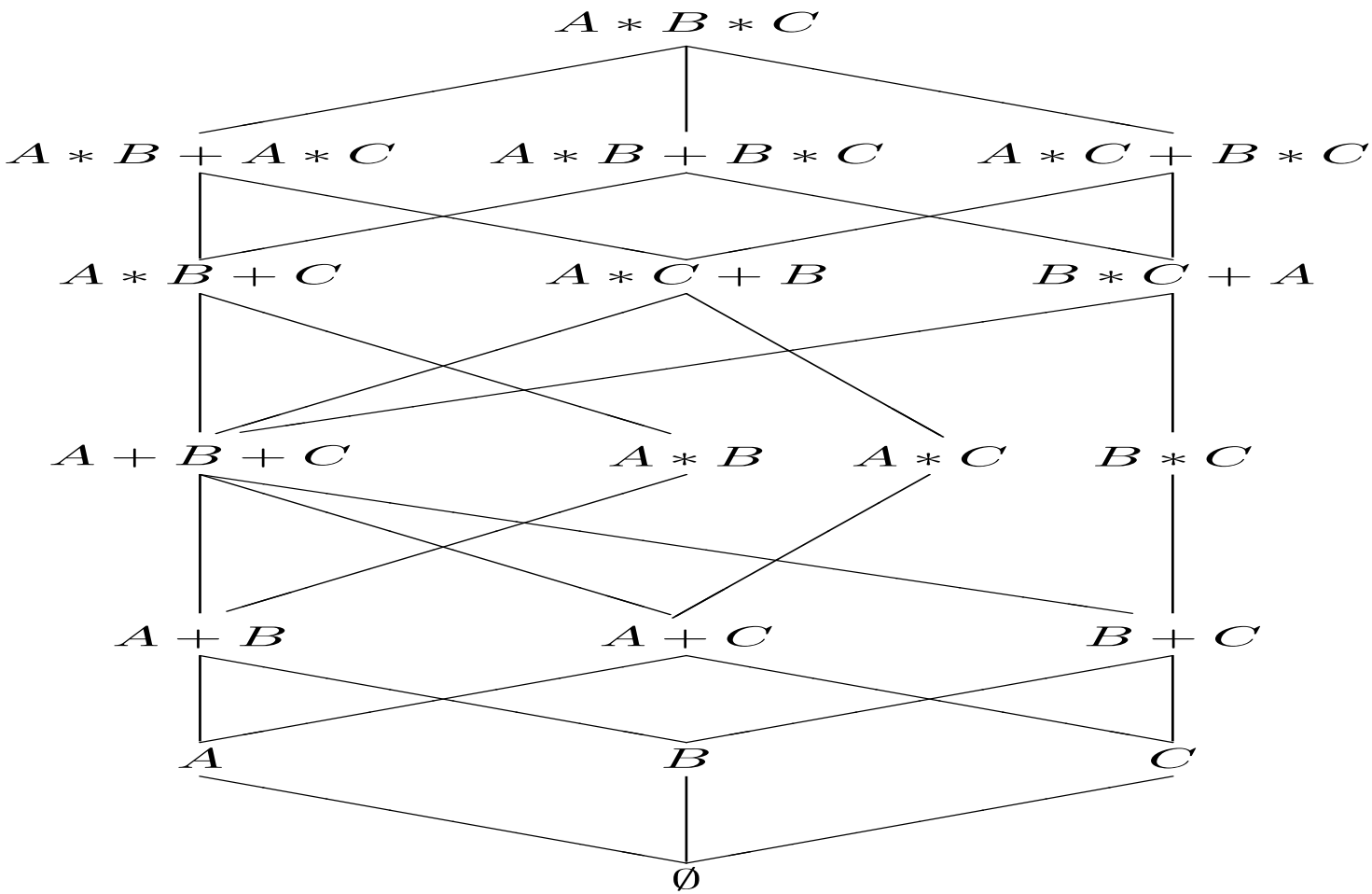


Figure: Inclusion diagram corresponding to a three-factor experiment.

Strategies for variable selection in hypothesis chains

Typically, one of the following principles for model selection is used:

a) Forward selection: Start with a null model, add at each step the variable that would give the lowest p-value of the variable not yet included in the model.

b) Backward selection: Start with a model containing all variables, variables are step by step deleted from the model. At each step, the variable with the largest p-value is deleted.

c) Stepwise selection: This is a modification of the forward selection principle. Variables are added to the model step by step. In each step, the procedure also examines whether variables already in the model can be deleted.

d) Best subset selection: For $k = 1, 2, ...$ up to a user-specified limit, the procedure identifies a specified number of best models containing k variables.

Variable selection in hypothesis chains

In-sample methods for model selection

The model complexity is evaluated using the same observations as those used for estimating the parameters of the model.

- The *training data* is used for evaluating the performance of the model.
- Any extra parameter will lead to a reduction of the loss function.
- In the in-sample case statistical tests are used to access the significance of extra parameters, and when the improvement is small in some sense the parameters are considered to be non-significant.
- The classical statistical approach.

Variable selection in hypothesis chains

Statistical learning or *data mining*

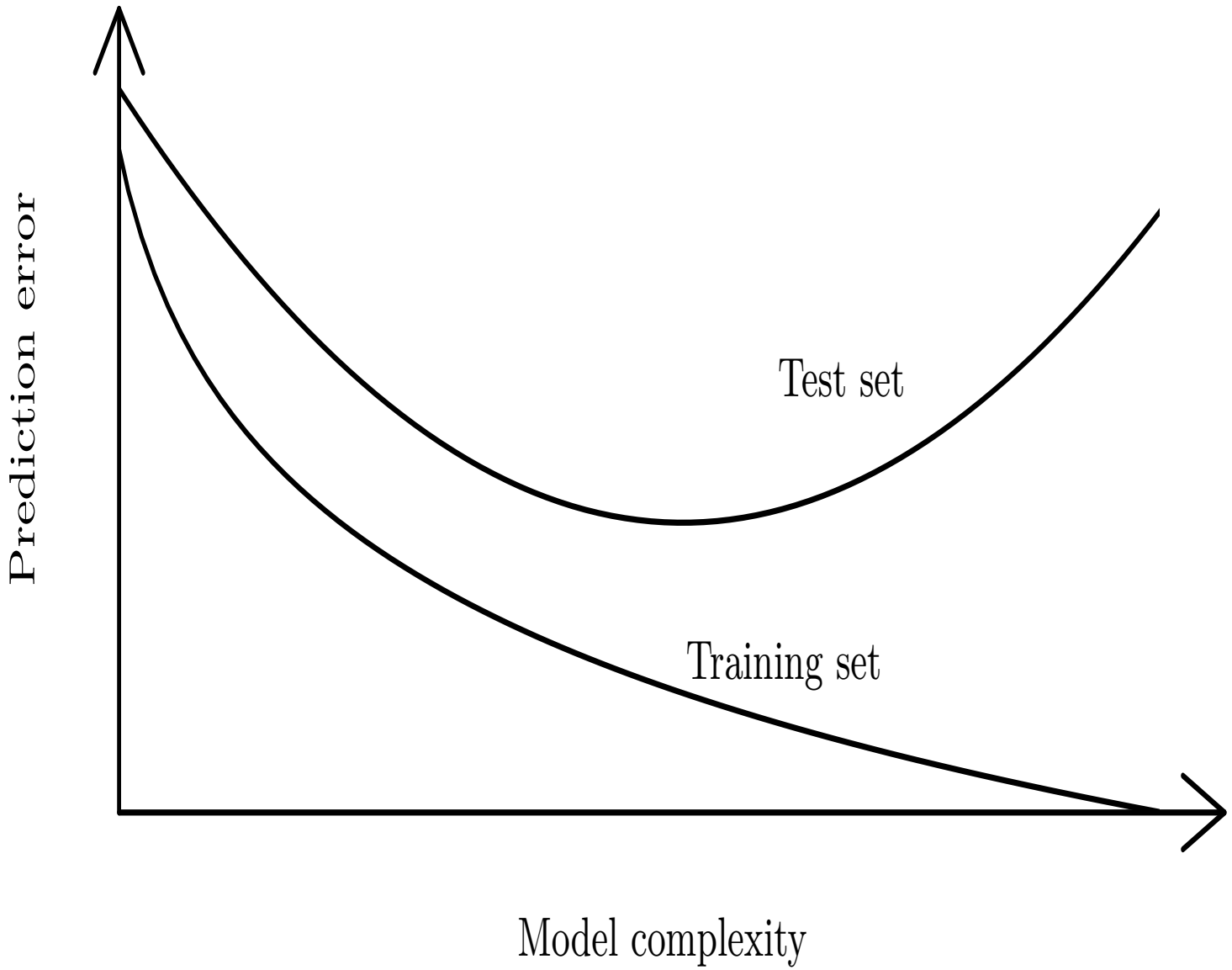
We have a data-rich situation such that only a part of the data is needed for model estimation and the rest can be used to test its performance.

- Seeking the *generalized performance* of a model which is defined as the expected performance on an independent set of observations.
- The expected performance can be evaluated as the expected value of the generalized loss function.
- The expected prediction error on an independent set of observations is called the *test error* or *generalization error*.

Variable selection in hypothesis chains

- In a data-rich situation, the performance can be evaluated by splitting the total set of observations in three parts:
  - *training set*: used for estimating the parameters
  - *validation set*: used for *out-of-sample* model selection
  - *test set*: used for assessing the generalized performance, i.e. the performance on new data
- A typical split of data is 50 pct for training and 25 pct for both validation and testing.

Variable selection in hypothesis chains



Nuisance parameters

- In many cases, the likelihood function is a function of many parameter but our interest focuses on the estimation on one or a subset of the parameters, with the others being considered as *nuisance parameters*
- Methods are needed to summarize the likelihood on a subset of the parameters by eliminating the nuisance parameters.
- Accounting for the extra uncertainty due to unknown nuisance parameters is an essential consideration, especially in small-sample cases.

Profile likelihood

Definition (Profile likelihood)

Assume that the statistical model for the observations $Y_1, Y_2, \ldots, Y_n$ is given by the family of joint densities, $\theta = (\tau, \zeta)$ and τ denoting the parameter of interest. Then the *profile likelihood function* for τ is the function

$$L_P(\tau; \boldsymbol{y}) = \sup_{\zeta} L((\tau, \zeta); \boldsymbol{y})$$

where the maximization is performed at a fixed value of τ.

Profile likelihood based confidence intervals

Consider again $\boldsymbol{\theta} = (\boldsymbol{\tau}, \boldsymbol{\zeta})$. It is seen that

$$\left\{ \boldsymbol{\tau}; \frac{L_P(\boldsymbol{\tau}; \boldsymbol{y})}{L(\widehat{\boldsymbol{\theta}}, \boldsymbol{y})} > \exp(-\frac{1}{2}\chi^2_{1-\alpha}(p)) \right\} \tag{1}$$

defines a set of values of $\boldsymbol{\tau}$ (Notice: p is the dimension of $\boldsymbol{\tau}$) that constitutes a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\tau}$.

Important special case:
In the case $p = 1$ (a single parameter) we find the lower bound of a 95% Confidence Interval (CI) as the smallest value of τ for which $\log L_P(\tau) > \log L(\widehat{\boldsymbol{\theta}}) - 1.92$ ($2 \times 1.92 = 3.84$ is equal to the 95%-percentile of the $\chi^2(1)$ distribution). The upper bound is found in a similar way.

Example: Profile likelihood confidence intervals

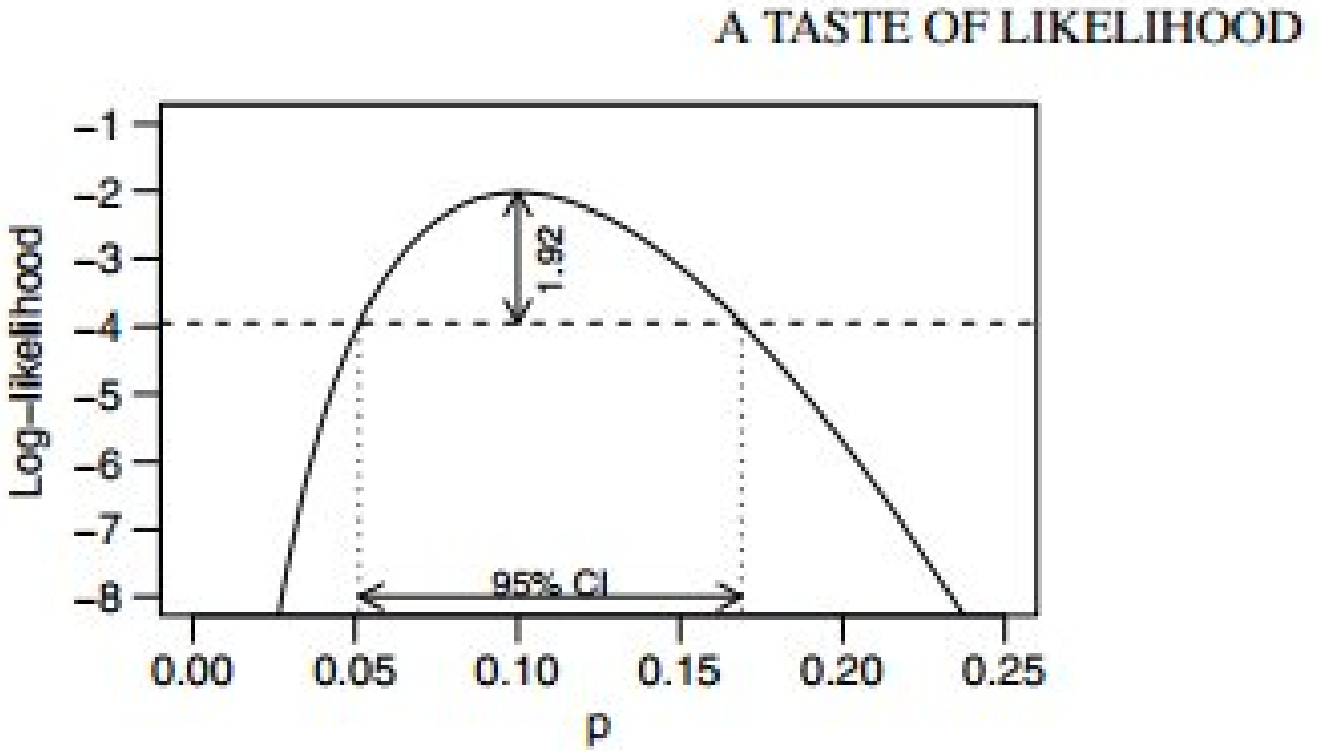


Figure: Binomial log-likelihood for 10 successes from 100 trials, and 95 pct. likelihood confidence interval.

Likelihood 95 pct. CI is $(0.051, 0.169)$. (**asymmetric**)

Question: What is the normal (**symmetric**) Wald CI?

Marginal likelihood

Definition (Marginal likelihood)

Assume that the statistical model for the observations $Y_1, Y_2, \ldots, Y_n$ is given by the family of joint densities, $\theta = (\tau, \zeta)$ and τ denoting the parameter of interest. Let (U, V) be a sufficient statistic for (τ, ζ) for which the factorization

$$f_{U,V}(u, v; (\tau, \zeta)) = f_U(u; \tau) f_{V|U=u}(v; u, \tau, \zeta)$$

holds. Provided that the likehood factor which corresponds to $f_{V|U=u}(\cdot)$ can be neglected, inference about τ can be based on the marginal model for U with density $f_U(u; \tau)$. The corresponding likelihood function

$$L_M(\tau; u) = f_U(u; \tau)$$

is called the *marginal likelihood function* based on U.

Now you know it all

- Likelihood function $L(\theta) = P_\theta(Y = y)$
- Log likelihood function $\ell(\theta) = \log(L(\theta))$
- Score function $\ell'(\theta)$
- Maximum likelihood estimate $\widehat{\theta} = \underset{\theta \in \Theta}{\text{argmax}}\ \ell(\theta)$
- Observed information matrix $-\ell''(\widehat{\theta})$
- Distribution of the ML estimator $\widehat{\theta} \sim \mathrm{N}(\theta, (-\ell''(\widehat{\theta}))^{-1})$
- Likelihood ratio test $2(\ell_A(\widehat{\theta_A}, Y) - \ell_B(\widehat{\theta_B}, Y)) \sim \chi^2_{\dim(A)-\dim(B)}$
- Dealing with nuisance parameters