Chapter 4
Likelihood-based inference

4.1 Likelihoods

Data $\mathbf{x} = \{x_1, \ldots, x_n\}$, joint distribution of $\mathbf{x}$ depends on unknown θ.
Likelihood is density (or probability if x_i is discrete) of the data x conditional on the parameter θ, i.e.

$$f(\mathbf{x}|\theta).$$

Function of θ *for fixed* $\mathbf{x}$, so denote the likelihood function by $L(\theta; \mathbf{x})$:

$$L(\theta; \mathbf{x}) = f(\mathbf{x}|\theta).$$

If $x_1, \ldots, x_n$ are independent, then
$f(\mathbf{x}|\theta) = f(x_1|\theta) \times f(x_2|\theta) \times \ldots \times f(x_n|\theta)$, and so

$$L(\theta; \mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta).$$

Used for point and interval estimation, and hypothesis testing.

Score statistics, Fisher information and the Cramer-Rao minimum variance bound

The score statistic is defined to be $\frac{\partial}{\partial\theta} l(\theta; \mathbf{x}) = \frac{\partial}{\partial\theta} \log f(\mathbf{x}|\theta)$.
$\mathbf{X}$: unobserved value of $\mathbf{x}$. Define the *random variable*

$$\frac{\partial}{\partial\theta} l(\theta; \mathbf{X}) = \frac{\partial}{\partial\theta} \log f(\mathbf{X}|\theta).$$

Transformation of a r.v. $\mathbf{X}$, where transformation is derivative, w.r.t. θ, of the log of the density of $\mathbf{X}$.

N.B. We treat $l(\theta; \mathbf{X})$ as a function of the random data $\mathbf{X}$, *evaluated at the true value of* θ, rather than a function of the parameter θ for fixed data $\mathbf{x}$.

$$
\begin{aligned}
\frac{\partial}{\partial\theta} l(\theta;\mathbf{X}) &= \frac{\partial}{\partial\theta}\log L(\theta;\mathbf{X}) \\
&= \left\{\frac{\partial}{\partial\theta} L(\theta;\mathbf{X})\right\} \times \frac{1}{L(\theta;\mathbf{X})} = \left\{\frac{\partial}{\partial\theta} f(\mathbf{X}|\theta)\right\} \times \frac{1}{f(\mathbf{X}|\theta)}.
\end{aligned}
$$

$$
\begin{aligned}
E\left\{\frac{\partial}{\partial\theta} l(\theta;\mathbf{X})\right\} &= \int \left\{\frac{\partial}{\partial\theta} l(\theta;\mathbf{x})\right\} f(\mathbf{x}|\theta)d\mathbf{x} \\
&= \int \left\{\frac{\partial}{\partial\theta} f(\mathbf{x}|\theta)\right\} \times \frac{1}{f(\mathbf{x}|\theta)} f(\mathbf{x}|\theta)d\mathbf{x} \\
&= \frac{\partial}{\partial\theta}\int f(\mathbf{x}|\theta)d\mathbf{x} \\
&= \frac{\partial}{\partial\theta} 1 = 0.
\end{aligned}
$$

Expected value of the derivative of the log-likelihood at the true value of θ is 0.

Consider example of $X \sim exp(rate = \theta)$. Then $f(x|\theta) = \theta e^{-\theta x}$

$l(\theta; X) = \log\theta - \theta X$ and

$$\text{Score statistic} = \frac{\partial}{\partial\theta} l(\theta; \mathbf{X}) = \frac{1}{\theta} - X,$$

so $\mathbb{E}(\text{score})$

$$\begin{aligned} E\left\{\frac{\partial}{\partial\theta} l(\theta; \mathbf{X})\right\} &= \int \left(\frac{1}{\theta} - x\right)\theta\exp(-\theta x)dx \\ &= \frac{1}{\theta}\int \theta\exp(-\theta x)dx - \int x\theta\exp(-\theta x)dx \\ &= \frac{1}{\theta} - \frac{1}{\theta} = 0. \end{aligned}$$

wrt the value of θ

However, the expected value of the derivative of the log-likelihood evaluated at the *wrong* value of θ, say θ^*, is not 0. For example,

Score (θ^*)
$$\frac{\partial}{\partial\theta} l(\theta; \mathbf{X})\bigg|_{\theta=\theta^*} = \frac{1}{\theta^*} - X,$$

with

$$\begin{aligned} E_{\theta}\left\{ \frac{\partial}{\partial\theta} l(\theta; \mathbf{X})\bigg|_{\theta=\theta^*} \right\} &= \int \overbrace{\left(\frac{1}{\theta^*} - x\right)}^{\text{score}(\theta^*)} \underbrace{\theta \exp(-\theta x)}_{f(x|\theta)} dx \\ &= \frac{1}{\theta^*} - \frac{1}{\theta}, \end{aligned}$$

which is non-zero for $\theta^* \neq \theta$.

In general $\mathbb{E}[\text{score}(\theta)] = \begin{cases} 0 & \text{if } \theta \text{ is true } \theta \\ \neq 0 & \text{if } \theta \neq \text{true } \theta \end{cases}$

To derive an expression for the variance of $\frac{\partial}{\partial\theta} l(\theta; \mathbf{X})$, we note that

$$\mathrm{Var}\left(\frac{\partial}{\partial\theta} l(\theta)\right) = \mathbb{E}\left(\left(\frac{\partial\, l(\theta)}{\partial\theta}\right)^2\right) - \cancel{\left(\mathbb{E}\frac{\partial l}{\partial\theta}\right)^2} \nearrow 0$$

$$\begin{aligned}
& 0 = \int \frac{\partial}{\partial\theta} l(\theta; \mathbf{x}) f(\mathbf{x}|\theta) d\mathbf{x} \\
\Rightarrow\quad & 0 = \frac{\partial}{\partial\theta} \int \frac{\partial}{\partial\theta} l(\theta; \mathbf{x}) f(\mathbf{x}|\theta) d\mathbf{x} \\
\Rightarrow\quad & 0 = \int \left\{ \frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{x}) \right\} f(\mathbf{x}|\theta) d\mathbf{x} + \int \left\{ \frac{\partial}{\partial\theta} l(\theta; \mathbf{x}) \right\} \left\{ \frac{\partial}{\partial\theta} f(\mathbf{x}|\theta) \right\} d\mathbf{x} \\
\Rightarrow\quad & 0 = \int \left\{ \frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{x}) \right\} f(\mathbf{x}|\theta) d\mathbf{x} \\
& \quad + \int \left\{ \frac{\partial}{\partial\theta} l(\theta; \mathbf{x}) \right\} \left\{ \frac{\partial}{\partial\theta} l(\theta; \mathbf{x}) \right\} f(\mathbf{x}|\theta) d\mathbf{x} \\
\Rightarrow\quad & \underline{E\left[\left\{ \frac{\partial}{\partial\theta} l(\theta; \mathbf{X}) \right\}^2 \right] = -E\left\{ \frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{X}) \right\}}. = \mathrm{Var}(\text{score}(\theta))
\end{aligned}$$

$$E\left[\left\{\frac{\partial}{\partial\theta}l(\theta;\mathbf{X})\right\}^2\right] = -E\left\{\frac{\partial^2}{\partial\theta^2}l(\theta;\mathbf{X})\right\}$$

Since $E\{\frac{\partial}{\partial\theta}l(\theta;\mathbf{X})\} = 0$, we have

$$Var\left\{\frac{\partial}{\partial\theta}l(\theta;\mathbf{X})\right\} = -E\left\{\frac{\partial^2}{\partial\theta^2}l(\theta;\mathbf{X})\right\}.$$

The term $-E\left\{\frac{\partial^2}{\partial\theta^2}l(\theta;\mathbf{X})\right\}$ is known as the **Fisher information** about parameter θ in data $\underline{x}$. which we will denote by $\mathcal{I}_E(\theta)$:

$$\mathcal{I}_E(\theta) \equiv -E\left\{\frac{\partial^2}{\partial\theta^2}l(\theta;\mathbf{X})\right\}.$$

Fisher information: measure of amount of information a sample size of n contains about θ. For independent observations $X_1, \ldots, X_n$,

$$f(\underline{x}|\theta) = \prod f(x_i|\theta)$$

$$l(\theta; \mathbf{X}) = \sum_{i=1}^{n} \log f(X_i|\theta),$$

$$\mathcal{I}_E(\theta) = -nE\left\{\frac{\partial^2}{\partial\theta^2} l(\theta; X_i)\right\},$$

hence Fisher information is proportional to sample size.

• Example. Consider $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ with σ^2 known. Then

$$\begin{aligned} -E\left\{\frac{\partial^2}{\partial\theta^2} l(\theta; \mathbf{X})\right\} &= -E\left\{\frac{\partial^2}{\partial\theta^2}\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \theta)^2\right\} \\ &= \frac{n}{\sigma^2}, \end{aligned}$$

Fisher information is n/σ^2. As σ^2 decreases, the observations more likely to be close to θ, so data more informative about θ.

$T(X)$ unbiased estimator of θ

Fisher information can be used to give a bound on the variance of an estimator.
Let $T(\mathbf{X})$ be an unbiased estimator, with $X_1, \ldots, X_n$ independent. Then it is possible to prove that

$$Var(T) \geq \frac{1}{\mathcal{I}_E(\theta)}.$$

This is known as the **Cramer-Rao minimum variance bound**.

.

Asymptotic normality

$n = \#$ data points

$\hat{\theta} = \arg\max_{\theta} \ell(\theta)$

$\hat{\theta}$ solves $\text{score}(\theta) = 0$

For large n, the distribution of the m.l.e $\hat{\theta}$ is approximately normal, with

$$\hat{\theta} \sim N\{\theta, \mathcal{I}_E(\theta)^{-1}\}.$$

Thus for large n, the m.l.e. $\hat{\theta}$ is *approximately* unbiased, and achieves the Cramer-Rao minimum variance bound.

$\text{Score}(\theta) \in \mathbb{R}^d \qquad \mathcal{I}_E(\theta) \in \mathbb{R}^{d\times d}$

In the multivariate case with $\underline{\theta} = (\theta_1, \ldots, \theta_d)$ we have

$$\text{Score}(\theta) = \begin{pmatrix} \frac{\partial l}{\partial \theta_1} \\ \vdots \\ \frac{\partial l}{\partial \theta_d} \end{pmatrix} \qquad \mathcal{I}_E(\theta) = \begin{pmatrix} e_{1,1}(\theta) & \cdots & e_{1,d}(\theta) \\ \vdots & & \vdots \\ e_{d,1}(\theta) & \cdots & e_{d,d}(\theta) \end{pmatrix}, = -\mathbb{E}\left(\frac{\partial^2 l}{\partial \underline{\theta}^2}\right)$$

with

$$e_{i,j}(\theta) = E\left\{-\frac{\partial^2}{\partial\theta_i\,\partial\theta_j} l(\theta)\right\}.$$

So for large n, the distribution of the m.l.e of θ is approximately multivariate normal:

$$\underline{\hat{\theta}} \sim N_d(\underline{\theta}, \underbrace{\mathcal{I}_E(\theta)^{-1}}_{\text{covariance matrix of } \hat{\theta}}),$$

Example: normally distributed data

Consider $X_1, \ldots, X_n$ with $X_i \sim N(\theta_1, \theta_2)$, with both θ_1 and θ_2 unknown. We write $\theta = (\theta_1, \theta_2)^T$.

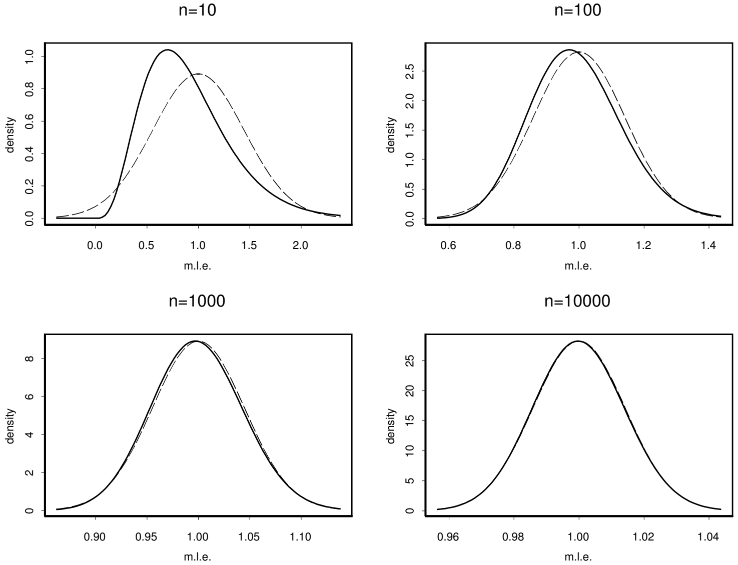
$$l(\theta; \mathbf{x}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\theta_2 - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2,$$

$$\begin{aligned} \hat{\theta}_1 &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\theta}_2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{aligned}$$

$$\mathcal{I}_E(\theta) = \begin{pmatrix} \frac{n}{\theta_2} & 0 \\ 0 & \frac{n}{2\theta_2^2} \end{pmatrix}.$$

For large n, the approximate distribution of $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)^T$ is

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix} \sim N \left\{ \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \begin{pmatrix} \frac{\theta_2}{n} & 0 \\ 0 & \frac{2\theta_2^2}{n} \end{pmatrix} \right\}$$


n=10
n=100
n=1000
n=10000
density
m.l.e.
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
2.5
0.6
0.8
1.2
1.4
0.90
0.95
1.00
1.05
1.10
0.96
0.98
1.02
1.04
0
2
4
6
8
5
10
15
20
25

Confidence intervals based on asymptotic normality

$\hat{\underline{\theta}} \sim N_d(\underline{\theta}, \mathcal{I}_E^{-1})$ marginal distⁿ of $\hat{\theta}_j$ is $N(\theta_j, \gamma_{jj})$

Suppose we want a $100(1-\alpha)\%$ confidence interval for any particular element of θ, say θ_j. For suitably large n, we have

$$\hat{\theta}_j \sim N(\theta_j, \gamma_{j,j}),$$

$\mathcal{I}_E = \mathbb{E}_\theta\left(-\frac{\partial^2 \ell}{\partial \theta^2}\right)$

wrt true value of θ.

where $\gamma_{j,j}$ is the $\{j, j\}$ element of $\mathcal{I}_E(\theta)^{-1}$. This then gives us an approximate interval as

$$(\hat{\theta}_j - z_{1-\frac{\alpha}{2}}\sqrt{\gamma_{j,j}}, \hat{\theta}_j + z_{1-\frac{\alpha}{2}}\sqrt{\gamma_{j,j}}),$$

$1-\frac{\alpha}{2}$ quantile of $N(0,1)$ random variable

e.g. 95% interval

$\hat{\theta}_j \pm 1.96\sqrt{\gamma_{jj}}$

$$\mathcal{I}_E = \mathbb{E}(\mathcal{I}_O)$$

θ unknown, so approximate $\mathcal{I}_E(\theta)$ by observed information matrix

$$\mathcal{I}_O(\theta) = \begin{pmatrix} -\frac{\partial^2}{\partial\theta_1^2} l(\theta) & \cdots & -\frac{\partial^2}{\partial\theta_1 \partial\theta_d} l(\theta) \\ \vdots & & \vdots \\ -\frac{\partial^2}{\partial\theta_d \partial\theta_1} (\theta) & \cdots & -\frac{\partial^2}{\partial\theta_d^2} l(\theta) \end{pmatrix},$$

evaluated at $\theta = \hat{\theta}$.

For large n we hope $\mathcal{I}_O \approx \mathcal{I}_E$

We use $\mathcal{I}_O$ as an estimator of $\mathcal{I}_E$

Note $\tilde{\gamma}_{ij} \neq \frac{1}{\left(\frac{\partial^2 l}{\partial\theta_i \partial\theta_j}\right)}$ – You cant use element wise inverses.

$\tilde{\gamma}_{i,j}$: the i,jth element of the inverse of $\mathcal{I}_O(\theta)$, we use

$$(\hat{\theta}_j - z_{1-\frac{\alpha}{2}}\sqrt{\tilde{\gamma}_{j,j}}, \hat{\theta}_j + z_{1-\frac{\alpha}{2}}\sqrt{\tilde{\gamma}_{j,j}}),$$

as an approximate confidence interval. Since we know that $\hat{\theta} \to \theta$ as $n \to \infty$, with probability 1, we would expect $\mathcal{I}_O(\theta)$ to be similar to $\mathcal{I}_E(\theta)$ for large sample sizes.

4.2 Profile Likelihood

Eg $X_i \sim N(\mu, \sigma^2)$ We may only want to learn about μ. σ^2 may be a 'nuisance' parameter

- RV X, density function f, parameters $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_d\}$
- Given $\mathbf{x} = (x_1, \ldots, x_n)$, only want inferences about *subset* of $\boldsymbol{\theta}$.
- Partition $\boldsymbol{\theta}$ into $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ with $\boldsymbol{\theta}_1$ the parameters of direct interest.
- $\boldsymbol{\theta}_2$, the parameters not of direct interest are known as **nuisance parameters**.

- Example: $X \sim N(\mu, \sigma^2)$ with both μ and σ^2 unknown, though we may only be interested in the mean parameter μ.
- Can use asymptotic distribution of m.l.e. to derive confidence intervals for individual parameters.
- Will now consider an alternative form of likelihood function which in some cases can produce more accurate confidence intervals.

Partitioning $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, **profile** log-likelihood function for $\boldsymbol{\theta}_1$ is

fnc^n of $\underline{\theta}_1$

$$l_p(\boldsymbol{\theta}_1; \mathbf{x}) = \max_{\boldsymbol{\theta}_2} l(\boldsymbol{\theta}). \tag{1}$$

To get the profile log-likelihood function for θ_1:

1. Treat $\boldsymbol{\theta}_1$ as a constant in $l(\boldsymbol{\theta}; \mathbf{x})$.
2. Find the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_2$ in terms of the data $\mathbf{x}$ *and* $\boldsymbol{\theta}_1$.
3. Plug in this expression for $\hat{\boldsymbol{\theta}}_2$ into the full log-likelihood $l(\boldsymbol{\theta}; \mathbf{x})$ to get the profile log-likelihood $l_p(\boldsymbol{\theta}_1; \mathbf{x})$.

- Writing $\boldsymbol{\theta} = (\theta_i, \boldsymbol{\theta}_{-i})$, plotting $l_p(\theta_i)$ gives us profile of log-likelihood surface viewed from θ_i axis. - informs us about how easy it is to learn θ_i.
- If $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ maximises $l(\boldsymbol{\theta})$, then $\hat{\boldsymbol{\theta}}_1$ maximises $l_p(\boldsymbol{\theta}_1)$ and $\hat{\boldsymbol{\theta}}_2$ maximises $l_p(\boldsymbol{\theta}_2)$.
- Useful exploratory tool; allows you to plot a likelihood $l_p(\theta_i)$ for a single parameter θ_i.
- Can be used to derive more accurate confidence intervals. than relying on asymptotic normality.

Example 1 $\theta = (\mu, \sigma^2) \in \mathbb{R}^2$

$$f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ i.i.d.

$L(\theta; \underline{x}) = \prod f(x_i|\theta)$

$$l(\mu, \sigma^2; \mathbf{x}) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2. \tag{2}$$

Find profile likelihood for μ: $\frac{\partial l}{\partial \sigma^2} = 0$ & solve

Fixing μ, the MLE of σ^2 is $\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$.Substituting this back into the full log-likelihood $l(\mu, \sigma^2; \mathbf{x})$, we get

$$l_p(\mu; \mathbf{x}) = -\frac{n}{2} \log\{\frac{1}{n} \sum (x_i - \mu)^2\} - \frac{n}{2}. \tag{3}$$

Find profile likelihood for σ^2:

Fixing σ^2, the MLE of μ is $\bar{x}$.The profile log-likelihood for σ^2 is

$$l_p(\sigma^2; \mathbf{x}) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{4}$$

We can plot both of these to get an idea of the value of the parameters & how much info there is about them.

Inference using the deviance function

- Can construct CI for $\boldsymbol{\theta}$ based on asymptotic normality of MLE. Alternative approach: use **deviance function**.
- For arbitrary $\boldsymbol{\theta}^*$,

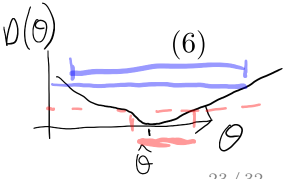
log like at MLE   log like at θ^*

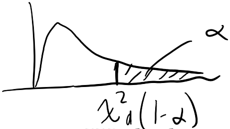
$$D(\boldsymbol{\theta}^*) = 2\{l(\hat{\boldsymbol{\theta}}; \mathbf{x}) - l(\boldsymbol{\theta}^*; \mathbf{x})\}. \geq 0 \qquad (5)$$

  $\hat{\boldsymbol{\theta}}$ maximises log-likelihood, so $D(\boldsymbol{\theta}^*) \geq 0$.
- If $D(\boldsymbol{\theta}^*)$ is small, then $l(\boldsymbol{\theta}^*)$ must be close to $l(\hat{\boldsymbol{\theta}})$, which suggests that $\boldsymbol{\theta}^*$ is a plausible estimate for the true unknown value of $\boldsymbol{\theta}$.
- A confidence interval (or region if $\boldsymbol{\theta}$ is a vector) could then be of the form

$$C = \{\boldsymbol{\theta}^* : D(\boldsymbol{\theta}^*) \leq c\}, \qquad (6)$$

  for some suitable value of c.





Wilk's theorem

- With data $x_1, \ldots, x_n$, for sufficiently large n, it can be shown that at the true value of $\boldsymbol{\theta}$, $D(\boldsymbol{\theta}) \sim \chi^2_d$, where d is the dimensionality of $\boldsymbol{\theta}$.
- An approximate $(1-\alpha)$ confidence region for $\boldsymbol{\theta}$ is then given by

$$C_\alpha = \{\boldsymbol{\theta}^* : D(\boldsymbol{\theta}^*) \leq c_\alpha\}, \tag{7}$$

  with c_α the $(1-\alpha)$ percentage point of the χ^2_d distribution.
- Usually more accurate than asymptotic normality approximation, may require greater computational effort.

Profile likelihood and the deviance function

- $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, with $\boldsymbol{\theta}_1$ a k-dimensional subset of $\boldsymbol{\theta}$. **Profile deviance**: NB $\hat{\theta}_1$ maxs $l_p(\theta_1)$

$$D_p(\boldsymbol{\theta}_1^*) = 2\{l(\hat{\boldsymbol{\theta}}; \mathbf{x}) - l_p(\boldsymbol{\theta}_1^*; \mathbf{x})\}, \geq 0 \tag{8}$$

$= 0$ at $\theta_1^* = \hat{\theta}_1$

with $\hat{\boldsymbol{\theta}}$ the maximum likelihood estimator of $\boldsymbol{\theta}$.

- Based on a sample of size n, with n sufficiently large,

profile deviance for θ_1

$$D_p(\boldsymbol{\theta}_1) \sim \chi^2_k. \tag{9}$$

$k = \dim(\theta_1)$

- Can obtain a confidence interval for any element θ_i as

$$C_\alpha = \{\theta_i^* : D_p(\theta_i^*) \leq c_\alpha\}, \tag{10}$$

again, with c_α the $(1-\alpha)$ percentage point of the χ^2_1 distribution.

- This will often be more accurate than the interval

$$\hat{\theta}_i \pm z_{\frac{\alpha}{2}} \sqrt{\psi_{i,i}} \tag{11}$$

stated earlier.

Example: leukaemia data

- Leukaemia patients given drug, 6-mercaptopurine (6-MP), and the number of days t_i until freedom from symptoms is recorded:

  $6^*, 6, 6, 6, 7, 9^*, 10^*, 10, 11^*, 13, 16, 17^*, 19^*, 20^*, 22, 23, 25^*, 32^*, 32^*,$

  A * denotes an observation censored at that time.
- Weibull model:

$$f_T(t) = \alpha\beta(\beta t)^{\alpha-1} \exp\{-(\beta t)^\alpha\} \tag{12}$$

  for $t > 0$. $\alpha = 1$ gives exponential distribution.
- For censored data

$$P(T > t) = \exp\{-(\beta t)^\alpha\}. \tag{13}$$

  d: no. of uncensored observations, $\sum_u \log t_i$: sum of all logs of the uncensored observations.

Find $\hat{\beta}$ which maxs $l(\alpha, \beta)$

$\frac{dl}{d\beta} = \frac{\alpha d}{\beta} - \alpha \beta^{\alpha-1} \sum t_i^\alpha$. Set $\frac{dl}{d\beta} = 0$ & solve for β.

$$l(\alpha, \beta; \mathbf{x}) = d \log \alpha + \alpha d \log \beta + (\alpha - 1) \sum_u \log t_i - \beta^\alpha \sum_{i=1}^n t_i^\alpha. \quad (14)$$

Treat α as fixed, and find MLE of β as function of data and α.

$$\hat{\beta} = \left(\frac{d}{\sum_{i=1}^n t_i^\alpha} \right)^{\frac{1}{\alpha}}. \quad (15)$$

The profile log-likelihood of α is then given by

$$\begin{aligned} l_p(\alpha) &= l(\alpha, \hat{\beta}) \\ &= d \log \alpha + \alpha d \log \left(\frac{d}{\sum_{i=1}^n t_i^\alpha} \right)^{\frac{1}{\alpha}} + (\alpha - 1) \sum_u \log t_i - d \end{aligned}$$

- Finding the full MLE $(\hat{\alpha}, \hat{\beta})$ cannot be done analytically, so numerical methods have to be used. - Use 'optim' in R
- To construct the confidence interval, onlyneed $\hat{\alpha}$ that maximises $l_p(\hat{\alpha})$, as $l_p(\hat{\alpha}) = l(\hat{\alpha}, \hat{\beta})$.
- For a 95% confidence interval, the 95th percentage point of the χ_1^2 distribution is 3.841.The confidence interval is then given by $\chi^2_1(0.95) = 3.841$

$$
\begin{aligned}
C_{0.05} &= \{\alpha^* : D_p(\alpha^*) \leq 3.841\} && (16)\\
&= [\alpha^* : 2\{l_p(\hat{\alpha}) - l_p(\alpha^*)\} \leq 3.841] && (17)\\
&= \{\alpha^* : l_p(\alpha^*) > l_p(\hat{\alpha}) - 3.841/2\}. && (18)
\end{aligned}
$$

- Numerically, we estimate the MLE $\hat{\alpha}$ to be 1.35, with $l_p(\hat{\beta}) = -41.66$.
- From the graph, we can then read off the 95% confidence interval for α as (0.73,2.2).
- This contains the value 1, so the simpler exponential distribution is plausible for this dataset.

Example: machine component failure

- Level of corrosion w in a machine component recorded and component tested until a failure is observed, at time t.
- Denote each observation by (w_i, t_i), where w_i is the level of corrosion, and t_i is the failure time.
- Possible model: $T \sim Exponential(\lambda)$ distribution, with λ a function of the corrosion level w:

$$\lambda = \alpha w^{\beta}. \tag{19}$$

  w treated as fixed, i.e. model distribution of the failure time conditional on the corrosion.
- $\beta = 0$ implies same expected time to failure, α^{-1} for all components, regardless of the corrosion level w.

The density of a single observation (w, t) is given by

$$f_T(t) = \alpha w^{\beta} \exp\{-\alpha w^{\beta} t\}. \tag{20}$$

$$l(\alpha, \beta; \mathbf{x}) = n \log \alpha + \beta \sum_{i=1}^{n} \log w_i - \alpha \sum_{i=1}^{n} w_i^{\beta} t_i. \tag{21}$$

We can derive an expression for the profile log-likelihood of β:
Treating β as fixed, we obtain the MLE of α as

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^{n} w_i^{\beta} t_i}. \tag{22}$$

We then substitute this expression for α in the full log-likelihood $l(\alpha, \beta)$ to get the profile log-likelihood for β:

$$l_p(\beta; \mathbf{x}) = n \log \left(\frac{n}{\sum_{i=1}^{n} w_i^{\beta} t_i} \right) + \beta \sum_{i=1}^{n} \log w_i - n. \tag{23}$$

- Numerically, estimate $\hat{\beta} = 0.473$, with $l_p(\hat{\beta}; \mathbf{x}) = -20.01$.
- From graph, read off 95% confidence interval for β as (0.11,0.95).
- Doesn't contain zero, and so there is clear evidence that $\beta \neq 0$
- For comparison, compute confidence interval for β using normal approximation.
- Observed information matrix is given by

$$\begin{pmatrix} -\frac{\partial^2 l}{\partial \alpha^2} & -\frac{\partial^2 l}{\partial \alpha \partial \beta} \\ -\frac{\partial^2 l}{\partial \alpha \partial \beta} & -\frac{\partial^2 l}{\partial \beta^2} \end{pmatrix} = \begin{pmatrix} n\alpha^{-2} & \sum w_i^{\beta} t_i \log w_i \\ \sum w_i^{\beta} t_i \log w_i & \alpha \sum w_i^{\beta} t_i (\log w_i)^2 \end{pmatrix} \tag{24}$$

- Obtain $\hat{\alpha}$ by substituting $\beta = 0.473$ into formula, gives $\hat{\alpha} = 1.099$.
- Substitute $\alpha = 1.099$, $\beta = 0.473$ into observed information matrix, invert to get

$$V = \begin{pmatrix} 0.0534 & -0.0241 \\ -0.0241 & 0.0442 \end{pmatrix}. \tag{25}$$

- CI for β using asymptotic normality is

$$\hat{\beta} \pm 1.96 \times 0.0442^{0.5}, \tag{26}$$

which gives (0.0611,0.8849).