1.6 Inference

Inference

The two main tools in statistics to make conclusions are

- **confidence intervals:** the width of a confidence interval provides a measure of the precision of the point estimates
- **hypothesis testing:** it compares how well two related models fit the data. The logic can be summarised as follows:
    - specify a model M_0 corresponding to H_0 and a more general model M_1 corresponding to H_1
    - fit model M_0 and compute measure of goodness of fit, G_0; repeat for M_1 to obtain G_1
    - calculate the improvement in fit
    - test the null hypothesis $G_0 = G_1$
    - if $G_0 = G_1$ is not rejected, then H_0 is not rejected and M_0 is the preferred model

For confidence intervals and hypothesis testing, sampling distributions are required.

- for **normally distributed r.v.**, the sampling distribution can be determined exactly
- for **other distributions** we need to rely on large-sample asymptotic results based on the CLT

The basic idea is that, under appropriate conditions (such as i.i.d), the statistic of interest S is approximately

$$\frac{S - \mathbb{E}(S)}{\sqrt{\mathbb{V}\mathrm{ar}(S)}} \dot{\sim} \mathcal{N}(0,1) \tag{1.6.1}$$

or equivalently

$$\frac{[S - \mathbb{E}(S)]^2}{\mathbb{V}\mathrm{ar}(S)} \dot{\sim} \chi^2_1 \tag{1.6.2}$$

and, in case, of p-multivariate statistics

$$[\mathbf{S} - \mathbb{E}(\mathbf{S})]^\top \mathbf{V}^{-1} [\mathbf{S} - \mathbb{E}(\mathbf{S})] \dot{\sim} \chi^2_p \tag{1.6.3}$$

Sampling distribution for score statistics

Suppose $\mathrm{Y}_1, \ldots, \mathrm{Y}_N$ are independent random variables from a distribution which belongs to the exponential family, with parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

The **score statistics** are such that

$$E[\mathrm{U}_j] = 0, \quad \text{for all } j = 1, \ldots, p. \tag{1.6.4}$$

The **variance-covariance matrix of the score statistics** is the information matrix $\mathcal{I}$, with elements

$$\mathcal{I}_{jk} = \mathbb{E}[\mathrm{U}_j \mathrm{U}_k] \tag{1.6.5}$$

If $p = 1$, the score statistic has the asymptotic sampling distribution

$$\frac{\mathrm{U}}{\sqrt{\mathcal{I}}} \dot{\sim} \mathcal{N}(0, 1),$$

where the $\dot{\sim}$ symbol means "approximately distributed as". Equivalently we can write

$$\frac{\mathrm{U}^2}{\mathcal{I}} \dot{\sim} \chi^2_1$$

If $p > 1$,

$$\mathbf{u} \dot{\sim} MVN(\mathbf{0}, \mathcal{I})$$

or, equivalently,

$$\mathbf{u}^\top \mathcal{I}^{-1} \mathbf{u} \dot{\sim} \chi^2_p$$

Example: Binomial distribution

If $\mathrm{Y} \sim Bin(n, p)$, the log-likelihood function is

$$\ell(p; y) = y \log p + (n - y) \log(1 - p) + \log \binom{n}{y}$$

and the score statistic is

$$\mathrm{U} = \frac{d\ell}{dp} = \frac{\mathrm{Y}}{p} - \frac{n - \mathrm{Y}}{1 - p} = \frac{\mathrm{Y} - np}{p(1 - p)}$$

Since $\mathbb{E}(\mathrm{Y}) = np$, $\mathbb{E}(\mathrm{U}) = 0$ as expected.

Since $\mathbb{V}\mathrm{ar}(\mathrm{Y}) = np(1 - p)$, the information is

$$\mathcal{I} = \mathbb{V}\mathrm{ar}(\mathrm{U}) = \frac{1}{p^2(1 - p)^2}\mathbb{V}\mathrm{ar}(\mathrm{Y}) = \frac{n}{p(1 - p)}$$

and, hence,

$$\frac{\mathrm{U}}{\sqrt{\mathcal{I}}} = \frac{\mathrm{Y} - np}{\sqrt{np(1 - p)}} \dot{\sim} \mathcal{N}(0, 1)$$

This is known as the *normal approximation to the binomial distribution*.

Taylor approximation

To obtain the asymptotic sampling distributions for various statistics, it is useful to use **Taylor approximations** for generic functions f in a neighbourhood of t

$$f(x) = f(t) + (x-t)\left[\frac{df}{dx}\right]_{x=t} + \frac{1}{2}(x-t)^2\left[\frac{d^2 f}{dx^2}\right]_{x=t} + \ldots$$

For a **log-likelihood**, the first *three* terms are

$$\ell(\theta) = \ell(\tilde{\theta}) + (\theta - \tilde{\theta})\mathrm{U}(\tilde{\theta}) - \frac{1}{2}(\theta - \tilde{\theta})^2\mathcal{I}(\tilde{\theta}) \tag{1.6.6}$$

where $\tilde{\theta}$ is an estimate of θ and $\mathrm{U}(\tilde{\theta})$ is the score function evaluated at $\theta = \tilde{\theta}$. Note that $U' = \frac{d^2\ell}{d\theta^2}$ is approximated by its expected value $E(U') = -\mathcal{I}$.

For a p-dimensional vector $\boldsymbol{\theta}$

$$\ell(\boldsymbol{\theta}) = \ell(\tilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\top\mathbf{u}(\tilde{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})^\top\mathcal{I}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \tag{1.6.7}$$

Similarly, for the **score function** of a one-dimensional parameter θ, the first *two* terms of the Taylor approximation are

$$\mathrm{U}(\theta) \approx \mathrm{U}(\tilde{\theta}) + (\theta - \tilde{\theta})\mathrm{U}'(\tilde{\theta}) = \mathrm{U}(\tilde{\theta}) - (\theta - \tilde{\theta})\mathcal{I}(\tilde{\theta}) \tag{1.6.8}$$

and the score function of a p-dimensional parameter $\boldsymbol{\theta}$ becomes

$$\mathbf{u}(\boldsymbol{\theta}) \approx \mathbf{u}(\tilde{\boldsymbol{\theta}}) - \mathcal{I}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}) \tag{1.6.9}$$

Sampling distribution of the MLE

Let's define the MLE as $\hat{\boldsymbol{\theta}}$.

By definition, the MLE is the estimator which maximises $\ell(\hat{\boldsymbol{\theta}})$, i.e. $\mathbf{u}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$

$$\begin{aligned} \mathbf{u}(\boldsymbol{\theta}) &\approx -\mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \\ -\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{u}(\boldsymbol{\theta}) &\approx (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \end{aligned}$$

Properties:

- **consistency:** since $\mathbb{E}(\mathbf{u}) = 0$

$$\begin{aligned} \mathbb{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = 0 \longrightarrow \mathbb{E}(\hat{\boldsymbol{\theta}}) &= \boldsymbol{\theta} \\ \mathbb{E}(\mathbf{u}) &= \mathbb{E}(\mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})) \\ \mathbf{0} &= \mathcal{I}(\hat{\boldsymbol{\theta}})\mathbb{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ \mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{0} &= \mathbb{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \end{aligned}$$

- **variance-covariance matrix**

$$\begin{aligned} \mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\right] &= \mathbb{E}[\mathcal{I}^{-1}\mathbf{u}\mathbf{u}^\top\mathcal{I}^{-1}] \\ &= \mathcal{I}^{-1}\mathbb{E}[\mathbf{u}\mathbf{u}^\top]\mathcal{I}^{-1} = \mathcal{I}^{-1} \end{aligned}$$

- **asymptotic sampling distribution**

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \dot{\sim} \chi^2(p)$$

Remarks

- $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is also known as **Wald statistics**
- for one-dimensional parameter, you can write $\hat{\theta} \dot{\sim} \mathcal{N}(\theta, \mathcal{I}^{-1})$
- if the response variable is normally distributed, the results are exact; for other GLM, the results are asymptotic