Statistical Models: background theory

Simon Wood
Mathematical Sciences, University of Bath, U.K.

Statistical models

- Statistical model

  A mathematical cartoon of how some data, $\mathbf{y}$*,*
  might have been generated

- The model depends on some *unknowns*, $\boldsymbol{\theta}$, usually parameters.
- Key features of a statistical model. Given $\boldsymbol{\theta}$
  1. the model can be used to simulate data that are like $\mathbf{y}$.
  2. *in principle* the model determines $f_\theta(\mathbf{y})$, the pdf of $\mathbf{y}$.

Statistical Inference

- Learn about unknown $\boldsymbol{\theta}$ from observed data $\mathbf{y}$.
- 4 main questions.
  1. What value of $\boldsymbol{\theta}$ is most consistent with $\mathbf{y}$?
  2. What range of values of $\boldsymbol{\theta}$ are consistent with $\mathbf{y}$?
  3. Is some specified value of $\boldsymbol{\theta}$, or restriction on $\boldsymbol{\theta}$, consistent with $\mathbf{y}$?
  4. Are *any* values of the $\boldsymbol{\theta}$ consistent with $\mathbf{y}$?
- Answers to these questions are provided by
  1. Point estimation.
  2. Interval estimation.
  3. Hypothesis testing (more generally model selection).
  4. Model checking.

2 approaches to inference

- There are two main approaches to inference. We will need both.
- Maximum likelihood estimation.
    - $\boldsymbol{\theta}$ are treated as fixed states of nature, about which we want to learn.
    - Use the notion that $\boldsymbol{\theta}$ values are 'likely' if they make $\mathbf{y}$ appear 'probable'.
- Bayesian inference.
    - The unknowns, $\boldsymbol{\theta}$, are treated as random variables.
    - Our knowledge of $\boldsymbol{\theta}$, described by a pdf, is updated using $\mathbf{y}$.

Likelihood

- The log pdf of $\mathbf{y}$ evaluated at the observed $\mathbf{y}$, considered as a function of $\boldsymbol{\theta}$, is the *log likelihood function* $l(\boldsymbol{\theta})$.
- i.e. $l(\boldsymbol{\theta}) = \log f_{\theta}(\mathbf{y})$ where $\mathbf{y}$ is the actual observed data.
- Values of $\boldsymbol{\theta}$ have relatively high log likelihood if they make the observed data appear relatively probable.
- Parameter values that are plausible given the data should have relatively high log likelihood.
- Notice that $l(\boldsymbol{\theta})$ is defined using the marginal distribution of the *observed* data $\mathbf{y}$, only.

Maximum Likelihood Estimation

- The maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

- $\hat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ ‘most consistent’ with the data.
- In general $\hat{\boldsymbol{\theta}}$ is found by numerical optimization.

Interval estimation

- How would $\hat{\theta}$ vary under repeated sampling of the data, $\mathbf{y}$?
- Treating $\mathbf{y}$ as random and considering the *estimator* $\hat{\theta}$, then as $n = \dim(\mathbf{y}) \to \infty$

$$\hat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}_{\text{true}}, \hat{\mathcal{I}}^{-1}\right) \quad \text{where} \quad \hat{\mathcal{I}} = -\frac{\partial^2 l}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}}$$

- Mild regularity conditions apply! The expected information can be substituted.
- Confidence intervals for the elements of $\boldsymbol{\theta}$ can be obtained directly from this result.

Hypothesis testing

- Consider testing $\mathrm{H}_0 : r(\boldsymbol{\theta}) = \mathbf{0}$ for p dimensional function r.
- Define

$$\hat{\boldsymbol{\theta}}_0 = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) \text{ subject to } r(\boldsymbol{\theta}) = \mathbf{0}$$

- Under repeated re-sampling of $\mathbf{y}$, then in the limit $n \to \infty$

$$2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_0)\} \sim \chi^2_p$$

if H_0 *is true*. Otherwise $2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_0)\} > \chi^2_p$.

- A test based on this result is known as a *generalized likelihood ratio test* (GLRT).
- The test can be used to compare nested models.
- Note that the GLRT result breaks down if H_0 restricts $\boldsymbol{\theta}$ to an edge of the feasible parameter space.

Model comparison by AIC

- The log likelihood ratio used in the GLRT measures the discrepancy between two models.
- Ideally we would like to select the model which has the minimum discrepancy from the truth.
- Let $f_t(\mathbf{y})$ be the true pdf of $\mathbf{y}$. The Kullback-Leibler distance is the expected log likelihood ratio of model and truth

$$K(f_{\hat{\theta}}, f_t) = \int \{\log f_t(\mathbf{y}) - \log f_{\hat{\theta}}(\mathbf{y})\} f_t(\mathbf{y}) d\mathbf{y}$$

- Selecting the model that minimizes an estimate of K, amounts to selecting the model that minimizes

$$\text{AIC} = -2l(\hat{\boldsymbol{\theta}}) + 2\text{dim}(\boldsymbol{\theta}).$$

Random effects

- In many models $\mathbf{y}$'s distribution depends on *unobserved* random variables, $\mathbf{z}$, and only $f_\theta(\mathbf{y}, \mathbf{z})$ is straightforward.
- Variables like $\mathbf{z}$ are known as *random effects* (unless they are simply 'missing data' from the observation of $\mathbf{y}$).
- To obtain a likelihood we need

$$f_\theta(\mathbf{y}) = \int f_\theta(\mathbf{y}, \mathbf{z}) d\mathbf{z}$$

  ...which is often intractable.
- Common solutions...
  1. If $\mathbb{E}_{z|y} \log f_\theta(\mathbf{z}, \mathbf{y})$ is tractable, then the *EM algorithm* allows $l(\boldsymbol{\theta}) = \log f_\theta(\mathbf{y})$ to be maximized *without evaluating* $\log f_\theta(\mathbf{y})$.
  2. Alternatively, the integral can be approximated.

Laplace approximation

- Let $\hat{\mathbf{z}}$ denote the maximizer of $f_\theta(\mathbf{y}, \mathbf{z})$ for a given $\mathbf{y}$.
- Let
$$\nabla_z^2 \log f_\theta = \left. \frac{\partial^2 \log f_\theta(\mathbf{y}, \mathbf{z})}{\partial \mathbf{z} \partial \mathbf{z}^{\mathrm{T}}} \right|_{\hat{\mathbf{z}}}$$
- Then by Taylor's theorem

$$\begin{aligned}
\log f_\theta(\mathbf{y}, \mathbf{z}) &\simeq \log f_\theta(\mathbf{y}, \hat{\mathbf{z}}) + (\mathbf{z} - \hat{\mathbf{z}})^{\mathrm{T}} \nabla_z^2 \log f_\theta (\mathbf{z} - \hat{\mathbf{z}})/2 \\
\Rightarrow f_\theta(\mathbf{y}, \mathbf{z}) &\simeq f_\theta(\mathbf{y}, \hat{\mathbf{z}}) e^{-\frac{1}{2}(\mathbf{z}-\hat{\mathbf{z}})^{\mathrm{T}}(-\nabla_z^2 \log f_\theta)(\mathbf{z}-\hat{\mathbf{z}})} \\
\Rightarrow f_\theta(\mathbf{y}) &\simeq f_\theta(\mathbf{y}, \hat{\mathbf{z}}) \frac{(2\pi)^{\dim(\mathbf{z})/2}}{\sqrt{|-\nabla_z^2 \log f_\theta|}}
\end{aligned}$$

since a MVN pdf integrates to 1.

Model checking

- Does the model fit *at all*?
- If it does not, then all the preceding theory is useless.
- All model checking amounts to looking for evidence that the observed data do not come from the pdf specified by the model.
- i.e. we look for evidence that

$$\mathbf{y} \nsim f_{\hat{\theta}}(\mathbf{y}).$$

- Formal goodness of fit testing is sometimes useful, but won't indicate *how* a model fails.
- Graphical checks are often helpful, as they can help to pin-point the way in which a model fails.

Bayesian inference

- If your target of inference is a random variable, then you are being Bayesian.
- We must specify a prior distribution $\theta \sim f(\theta)$ as part of modelling process.
- The prior is updated using the observed $\mathbf{y}$ via Bayes rule.
- Bayes rule is a re-arrangement of $f(\theta, \mathbf{y}) = f(\mathbf{y}, \theta)$

$$\begin{aligned} f(\theta|\mathbf{y})f(\mathbf{y}) &= f(\mathbf{y}|\theta)f(\theta) \\ \Rightarrow f(\theta|\mathbf{y}) &= f(\mathbf{y}|\theta)f(\theta)/f(\mathbf{y}) \end{aligned}$$

- $f(\mathbf{y})$ is usually intractable, but it is a constant, so ...
    1. Sometimes the form of $f(\theta|\mathbf{y})$ can be recognised from $f(\mathbf{y}|\theta)f(\theta)$.
    2. It is possible to simulate from $f(\theta|\mathbf{y})$ without knowing $f(\mathbf{y})$.

The MLE Bayesian connection

- Suppose we use improper uniform priors $f(\theta) = \text{constant}$.
- Then $f(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)$. i.e. the *posterior distribution*, $f(\theta|\mathbf{y})$ is directly proportional to the likelihood, $f(\mathbf{y}|\theta)$.
- So the most probable value of θ according to the posterior will be the MLE, $\hat{\theta}$.
- Actually, as the sample size $n \to \infty$ the likelihood dominates *any* prior that is non-zero over all the parameter space. Hence the posterior modes $\to \hat{\theta}$.
- Furthermore $f(\theta|\mathbf{y}) \to k \exp\{-(\theta - \hat{\theta})^{\mathrm{T}}\mathcal{I}(\theta - \hat{\theta})/2\}$ as $n \to \infty$ for any regular posterior about which $\mathbf{y}$ is informative, by Taylor's theorem.
- i.e. in the large sample limit $\theta|\mathbf{y} \sim N(\hat{\theta}, \mathcal{I}^{-1})$.

Linear predictor regression models

- In this course we will consider only statistical models in which we want to model observations of a *response variable*, y, using some *predictor variables* that accompany each observation.
- We will consider only the case in which $E(y_i)$ is completely determined by a single variable η_i, which depends flexibly on the predictor variables, but only *linearly* on the model parameters and any random effects.
- η_i is known as a *linear predictor*.
- We will further assume that given η_i the y_i are independent.
- Inference with these models uses the preceding theory, but numerical estimation, model specification and checking are greatly facilitated by the special structure.