Model Selection

Day 1 – concepts and theory

Caveats - warnings

- Personal views – pragmatic but of course theory is important to understand why some approaches might work better than others
- Very complex issues – no simple answer that can be used in all cases. Depends on the objectives, the data, previous knowledge, …
- Tools – theory and simulations are important to evaluate their properties, not if they are «true»

What is a statistical model and what it is used for (Cox 1990)

- Substantive models
- Empirical models
- Randomization theory
- Indirect models

- Exploratory-Description
- Confirmatory-Inference
- Causal-Estimation-Association-Prediction

Criteria for Models

- Link with underlying knowledge
- Link with previous (or future) published work
- Pointer towards a process that might have generated the data
- Parameters in the primary aspects of the model should have specific interpretations
- Secondary aspects should give adequate description of the random variation
- Model should capture the main features of interest
- Model should be consistent with the data

(Cox & Wermuth 1996, p 18:19)

Criteria for Models

- Principle of parsimony or Ockham's razor

  (Lazar 2010 for extensive info on Ockham +)

  «**entities or assumptions should not be multiplied unnecessarily»**
- Good theories are those that explain all the known facts in a fashion as uncomplicated as possible
- simpler models should be preferred until the data justify more complex models

Criteria for Models

- Chamberlin + Platt: Multiple working hypotheses

1) Devising alternative hypotheses;

2) Devising a crucial experiment (or several of them), with alternative possible outcomes, each of which will, as nearly as possible, exclude one or more of the hypotheses;

3) Carrying out the experiment so as to get a clean result;

1′) Recycling the procedure, making subhypotheses or sequential hypotheses to refine the possibilities that remain; and so on.

Strong Inference

Certain systematic methods of scientific thinking may produce much more rapid progress than others.

John R. Platt

How do we measure Statistical Evidence

- P-values
- Likelihood
- AIC
- Bayes factor and BIC
- DIC

P-values

- Test statistic; X_{obs}=Data, $T(X_{obs})=T_{obs}$
- H_0: Model for X, generating a distribution for T
- Large values are unexpected under H0
- P-val = $\text{Prob}(T(X) \geq T_{obs} \mid H_0)$

  Significance level

P-values

- P-val = Prob($T(X) \geq T_{obs}$ | H_0)
- Indirect Evidence against H_0
- NOT   Prob(H_0 | T_{obs})
- Two differences
  - Conditional probabilities, Bayes theorem
  - $T(X) = T_{obs}$ vs $T(X) \geq T_{obs}$

Likelihood

- Statistical Model describing how data can be generated, as a function of parameters θ: $f(y|\theta)$
- Linear regression model: (x_i, y_i)

  $\theta = (\beta_0, \beta_1, \sigma)$

  $y_i \sim \text{Norm}(\beta_0 + \beta_1 x_i, \sigma)$

- $f(y_i|\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}}$
- Independence: $f((y_1,\ldots,y_n)|\theta) = \prod_{i=1}^{n} f(yi|\theta)$

Likelihood

- lik(θ) proportional to f(y|θ), as a function of θ
- Use log(lik(θ)) because it is simpler and most theoretical results refer to this function
- Linear regression model

$$\text{Log}(f(y_i|\theta)) = -\log(2\pi) - \log(\sigma) - (y_i - (\beta_0 + \beta_1 x_i))^2 / 2\sigma^2$$

$$\text{Log}(\text{lik}(\theta = (\beta_0, \beta_1, \sigma))) = -n \log(\sigma) - \frac{\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2)}{2\sigma^2}$$

$$= -n \log(\sigma) - \frac{SSE}{2\sigma^2}$$

Likelihood

- Estimation of parameters: MLE
- For a linear regression model
  - minimize SCE to estimate β_0 and β_1
  - $\frac{\partial LL}{\partial \sigma} = - \frac{n}{\sigma} + \frac{SSE}{\sigma^3} \Rightarrow \sigma^2 = \frac{SSE}{n}$
- Illustration using R (liknorm.R)

Likelihood

- Likelihood Principle

Two models A and B

Likelihoods lik(data | A) and lik(data | B)

Evidence given by $\frac{\text{lik(data | A)}}{\text{lik(data | B)}}$

- Models with different parameter values
  - Likelihood ratio test
  - Different models – MLE for each model

Probability, Frequency, Belief, Likelihood

- Probability theory (mathematics) does not care about the meaning of probability (axioms-Kolmogorov)
- Probability comes always with two flavours: long-term frequency and belief; one is «objective» (can be measured), the other is «subjective»
- They can be mixed in equations but one should be careful about their meanings
- P-values are frequencies (frequentist statistics), prob(H | data) is a belief (Bayesian statistics)
- They can be mixed (long-term frequencies of Bayesian statistics)

AIC

- Akaike's Information Criterion
  - Hirotogu Akaike (1927-2009)
- Linear models with increasing number of predictor variables: SCE↓ as p↑
- A «simplistic» application of the likelihood principle would lead to choosing the most complex model…

AIC

- Akaike realized that loglik($\hat{\theta}$), with $\hat{\theta}$ the MLE, is a biased estimate of E[log(f(X|$\hat{\theta}$))], where the expectation is taken wrt to X and $\hat{\theta}$
- The theory behind the derivation is rather complicated, and there have been some disagreements

AIC and KL distance

- Assume a true generating density g
- KL(g,f(θ)) = $\int g(y)\, log \frac{g(y)}{f(y,\theta)} dy$
  - Distance between f(θ) and the «truth»
  - MLE $\hat{\theta}$ aims at providing the best parametric approximation inside the class f(θ) to g

AIC and KL distance

- KL(g,f($\hat{\theta}$)) = $\int g(y)\ log \frac{g(y)}{f(y,\hat{\theta})} dy$

  = $\int g(y)\ log\ g(y)dy$ - $\int g(y)\ log\ f\left(y,\hat{\theta}\right)dy$
- Q_n = E_g [$\int g(y)\ log\ f\left(y,\hat{\theta}\right)dy$]
- Naive estimate: $\hat{Q}_n = \frac{1}{n}\sum_{i=1}^{n} \log(f(y_i, \hat{\theta})$ = $\frac{1}{n} l_n(\hat{\theta})$
- E($\hat{Q}_n - Q_n$) = p*/n
- AIC(M) = - 2$l_n(\hat{\theta})$ + 2 length(θ)
- -2 for «historical» reasons, - 2$l_n(\hat{\theta})$ = deviance
- Sometines defined without -2, with 2, divided by n…

AIC and AIC_c

- AIC is an unbiased first-order estimate
- Asymptotically unbiased, biased for small samples
- For linear models Y=Xβ+ε, with dim(X)=(n,p):
- $AIC_C = -2l_n\left(\hat{\theta} = \left(\hat{\beta}, \hat{\sigma}\right)\right) + 2\ \frac{n(p+1)}{n-p-2}$
  - $\hat{\sigma}$ is the MLE of σ, known to be a biased estimate
- $AIC_{C2} = -2l_n\left(\hat{\theta} = \left(\hat{\beta}, \widehat{\sigma^*}\right)\right) + 2(p+1)$

  $\widehat{\sigma^*}$ being the unbiased estimate SSE/(n-p-2)

  Claeskens and Hjort: not obvious why one is better…
- No theory to justify the same correction for other models (eg generalized linear models)

AIC and Evidence

- AIC : relative likelihood and weights
  - scale AIC values relative to minimum value: ΔAIC
  - Relative Likelihood Model i: $\exp\left(-\frac{1}{2}\Delta AIC_i\right)$
- AIC weights = $\dfrac{\exp(-\frac{1}{2}\Delta AIC_i)}{\sum_{model\ 1}^{model\ n}\exp(-\frac{1}{2}\Delta AIC_j)}$
- Unclear what it is when e.g. B&A define prob($Model_i$|Data)

Bayes factors

- Two hypotheses H_1 and H_2
- Prior probabilities (beliefs) $p(H_1)$ and $p(H_2)$
- Likelihood $p(data|H_1)$ and $p(data|H_2)$
- Posterior probabilities

$$p(H_1|data) = p(data|H_1)\ p(H_1) / p(data)$$

$$p(H_2|data) = p(data|H_2)\ p(H_2) / p(data)$$

- Ratio of posterior probabilities

$$\frac{p(H_1|data)}{p(H_2|data)} = \frac{p(H_1)}{p(H_2)}\frac{p(data|H_1)}{p(data|H_2)}$$

- Bayes factor = $p(data|H_1) / p(data|H_2)$

Bayes factors

- If the two hypotheses involve parameters

H_1: β and H_2: θ

BF=$\frac{\int p_1(y|\beta)\pi_1(\beta)d\beta}{\int p_2(y|\theta)\pi_2(\theta)d\theta}$

Where $\pi_1(\beta)$ and $\pi_2(\theta)$ are prior distributions of the parameters

- Parameters are integrated out (LRT: use MLEs)
- Can be calculated numerically (examples in R)
- Can be sensitive to the choice of priors

Bayes Factors and BIC

- BIC as an approximation to Bayes Factors
- BIC = 2 loglik($\hat{\theta}$) - log(n) length($\hat{\theta}$)
- $\exp\left(-\frac{1}{2}\Delta BIC_i\right)$
- BIC weights

DIC

- Developed in the context of MCMC simulations for Bayesian modelling
- Deviance D(y, θ) = -2log(f(y, θ))
- Prior distribution π(θ); posterior π(θ|data)
- DIC = D(y, $\bar{\theta}$) + 2p_D
  - $\bar{\theta}$ is the posterior mean
  - p_D is the effective number of parameters
  - $p_D = \overline{D(Y, \theta)} - D(Y, \bar{\theta})$

Measuring the goodness of fit of a statistical model

- Assumptions of statistical models
- Systematic component (main structure)
- Stochastic component
  - Independence
  - Variance function
  - Distribution

Goodness of Fit

- Not all assumptions are equally important
- Linear Models
  1) Independence
  2) Constant Variance
  3) Normal Distribution
- Mixed Models

  Constant variance of residuals and random effects