STAT 740: Testing & Model Selection

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 740: Statistical Computing

Testing & model choice, likelihood-based

A common way to test a larger model vs. a nested sub-model is through hypothesis testing. Three ways: Wald, score, and LRT tests. Large-sample versions lead to χ^2 test statistics.

Bootstrapped p-values do not rely on asymptotics (coming up).

Non-nested model selection is carried out through information criteria: AIC, BIC, etc.

Other methods: LASSO, LAD, elastic net, cross-validation, best subsets.

Testing & model choice, Bayesian

Bayes factors compare models in a coherent way. Savage-Dickey ratio allows computation of Bayes factors for point null hypotheses. Can also do "usual" hypothesis test based on credible interval (or region).

Non-nested model selection is carried out through information criteria: DIC, WIC, etc.

Other methods: Bayesian LASSO, LPML (leave-one-out cross validated predictive density), model averaging, reversible jump (Green, 1995), pseudo-priors (Carlin and Chib, 1995), stochastic search variable selection (SSVS).

Wald, likelihood ratio, and score tests

These are three ways to perform large sample hypothesis tests based on the model likelihood $L(\boldsymbol{\theta}|\mathbf{x})$.

Wald test

Let $\mathbf{M}$ be a $m \times k$ matrix. Many hypotheses can be written $H_0 : \mathbf{M}\boldsymbol{\theta} = \mathbf{b}$ where $\mathbf{b}$ is a known $m \times 1$ vector.

For example, let $k = 3$ so $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$. The test of $H_0 : \theta_2 = 0$ is written in matrix terms with $\mathbf{M} = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$ and $b = 0$. The hypothesis $H_0 : \theta_1 = \theta_2 = \theta_3$ has $\mathbf{M} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$ and $\mathbf{b} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Wald test, cont.

The large sample result for MLEs is

$$\hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N_k(\boldsymbol{\theta}, [-\nabla^2 \log L(\hat{\boldsymbol{\theta}}|\mathbf{x})]^{-1}).$$

So then

$$\mathbf{M}\hat{\boldsymbol{\theta}} \overset{\bullet}{\sim} N_m(\mathbf{M}\boldsymbol{\theta}, \mathbf{M}[-\nabla^2 \log L(\hat{\boldsymbol{\theta}}|\mathbf{x})]^{-1}\mathbf{M}').$$

If $H_0 : \mathbf{M}\boldsymbol{\theta} = \mathbf{b}$ is true then

$$\mathbf{M}\hat{\boldsymbol{\theta}} - \mathbf{b} \overset{\bullet}{\sim} N_m(\mathbf{0}, \mathbf{M}[-\nabla^2 \log L(\hat{\boldsymbol{\theta}}|\mathbf{x})]^{-1}\mathbf{M}').$$

So

$$W = (\mathbf{M}\hat{\boldsymbol{\theta}} - \mathbf{b})'[\mathbf{M}[-\nabla^2 \log L(\hat{\boldsymbol{\theta}}|\mathbf{x})]^{-1}\mathbf{M}']^{-1}(\mathbf{M}\hat{\boldsymbol{\theta}} - \mathbf{b}) \overset{\bullet}{\sim} \chi^2_m.$$

W is called the Wald statistic and large values of W indicate $\mathbf{M}\boldsymbol{\theta}$ is far away from $\mathbf{b}$, i.e. that H_0 is false. The p-value for $H_0 : \mathbf{M}\boldsymbol{\theta} = \mathbf{b}$ is given by p-value $= P(\chi^2_m > W)$.

The simplest, most-used Wald test is the familiar test that a regression effect is equal to zero, common to multiple, logistic, Poisson, and ordinal regression models.

Nonlinear tests

For nonlinear function $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^m$,

$$\mathbf{g}(\boldsymbol{\theta}) = \begin{bmatrix} g_1(\boldsymbol{\theta}) \\ \vdots \\ g_m(\boldsymbol{\theta}) \end{bmatrix},$$

we have

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) \overset{\bullet}{\sim} N_m(\mathbf{g}(\boldsymbol{\theta}), [\nabla \mathbf{g}(\hat{\boldsymbol{\theta}})]'[-\nabla^2 \log L(\hat{\boldsymbol{\theta}}|\mathbf{x})]^{-1}[\nabla \mathbf{g}(\hat{\boldsymbol{\theta}})]).$$

Let $\mathbf{g}_0 \in \mathbb{R}^m$ be a fixed, known vector.

An approximate Wald test of $H_0 : \mathbf{g}(\boldsymbol{\theta}) = \mathbf{g}_0$ has test statistic

$$W = (\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}_0)'\{[\nabla \mathbf{g}(\hat{\boldsymbol{\theta}})]'[-\nabla^2 \log L(\hat{\boldsymbol{\theta}}|\mathbf{x})]^{-1}[\nabla \mathbf{g}(\hat{\boldsymbol{\theta}})]\}^{-1}(\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}_0).$$

In large samples $W \overset{\bullet}{\sim} \chi^2_m$. **Example:** V.A. data.

Score test

In general, the $\text{cov}(\hat{\boldsymbol{\theta}})$ is a function of the unknown $\boldsymbol{\theta}$. The Wald test replaces $\boldsymbol{\theta}$ by its MLE $\hat{\boldsymbol{\theta}}$ yielding $[-\nabla^2 \log L(\hat{\boldsymbol{\theta}}|\mathbf{x})]^{-1}$. The score test replaces $\boldsymbol{\theta}$ by the the MLE $\hat{\boldsymbol{\theta}}_0$ obtained under the constraint imposed by H_0

$$\hat{\boldsymbol{\theta}}_0 = \text{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}: \mathbf{M}\boldsymbol{\theta} = \mathbf{b}} L(\boldsymbol{\theta}|\mathbf{x}).$$

Let $[-\nabla^2 \log L(\boldsymbol{\theta}|\mathbf{x})]^{-1}$ be the asymptotic covariance for *unconstrained* MLE.

The resulting test statistic

$$S = [\tfrac{\partial}{\partial \boldsymbol{\theta}} \log L(\hat{\boldsymbol{\theta}}_0|\mathbf{x})]'[-\nabla^2 \log L(\hat{\boldsymbol{\theta}}_0|\mathbf{x})]^{-1}[\tfrac{\partial}{\partial \boldsymbol{\theta}} \log L(\hat{\boldsymbol{\theta}}_0|\mathbf{x})] \overset{\bullet}{\sim} \chi^2_m.$$

Sometimes it is easier to fit the reduced model rather than the full model; the score test allows testing whether new parameters are necessary from a fit of a smaller model.

Likelihood ratio tests

The likelihood ratio test is easily constructed and carried out for nested models. The full model has parameter vector $\boldsymbol{\theta}$ and the reduced model obtains when $H_0 : \mathbf{M}\boldsymbol{\theta} = \mathbf{b}$ holds. A common example is when $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and we wish to test $H_0 : \boldsymbol{\theta}_1 = \mathbf{0}$ (e.g. a subset of regression effects are zero). Let $\hat{\boldsymbol{\theta}}$ be the MLE under the full model

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}|\mathbf{x}),$$

and $\hat{\boldsymbol{\theta}}_0$ be the MLE under the constraint imposed by H_0

$$\hat{\boldsymbol{\theta}}_0 = \text{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}: \mathbf{M}\boldsymbol{\theta} = \mathbf{b}} L(\boldsymbol{\theta}|\mathbf{x}).$$

LRT, cont.

If $H_0 : \mathbf{M}\boldsymbol{\theta} = \mathbf{b}$ is true,

$$L = -2[\log L(\hat{\boldsymbol{\theta}}_0|\mathbf{x}) - \log L(\hat{\boldsymbol{\theta}}|\mathbf{x})] \overset{\bullet}{\sim} \chi^2_m.$$

The statistic L is the likelihood ratio test statistic for the hypothesis $H_0 : \mathbf{M}\boldsymbol{\theta} = \mathbf{b}$. The smallest L can be is zero when $\hat{\boldsymbol{\theta}}_0 = \hat{\boldsymbol{\theta}}$. The more different $\hat{\boldsymbol{\theta}}$ is from $\hat{\boldsymbol{\theta}}_0$, the larger L is and the more evidence there is that H_0 is false. The p-value for testing H_0 is given by $p - \text{value} = P(\chi^2_m > L)$.

To test whether additional parameters are necessary, LRT tests are carried out by fitting two models: a "full" model with all effects and a "reduced" model. In this case the dimension m of $\mathbf{M}$ is the difference in the numbers of parameters in the two models.

LRT, cont.

For example, say we are fitting the standard regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

where $e_i \overset{iid}{\sim} N(0, \sigma^2)$. Then $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2)$ and we want to test $\boldsymbol{\theta}_1 = (\beta_2, \beta_3) = (0, 0)$, that the 2^{nd} and 3^{rd} predictors aren't needed. This test can be written using matrices as

$$H_0 : \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

LRT, cont.

The likelihood ratio test fits the full model above and computes $L_f = \log L_f(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\sigma})$.

Then the reduced model $Y_i = \beta_0 + \beta_1 x_{i1} + e_i$ is fit and $L_r = \log L_r(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma})$ computed.

The test statistic is $L = -2(L_r - L_f)$; a p-value is computed as $P(\chi^2_2 > L)$. If the p-value is less than, say, $\alpha = 0.05$ we reject $H_0 : \beta_2 = \beta_3 = 0$.

Of course we wouldn't use this approximate LRT test here! We have outlined an approximate test, but there is well-developed theory that instead uses a different test statistic with an exact F-distribution.

Comments

Note that:

- The Wald test requires maximizing the unrestricted likelihood.
- The score test requires maximizing the restricted likelihood (under a nested submodel).
- The Likelihood ratio test requires both of these.

So the likelihood ratio test uses more information and both Wald and Score tests can be viewed as approximations to the LRT.

However, Wald tests of the form $H_0 : \mathbf{M}\boldsymbol{\theta} = \mathbf{b}$ are easy to get in both SAS and R. In large samples the tests are equivalent.

Confidence intervals

A plausible range of values for a parameter β_j (from ϵ) is given by a confidence interval (CI). Recall that a CI has a certain fixed probability of containing the unknown β_j before data are collected. After data are collected, nothing is random any more, and instead of "probability" we refer to "confidence."

A common way of obtaining confidence intervals is by *inverting* hypothesis tests of $H_0 : \beta_k = b$. Without delving into why this works, a $(1 - \alpha)100\%$ CI is given by those b such that the p-value for testing $H_0 : \beta_k = b$ is larger than α.

CIs, cont.

For Wald tests of $H_0 : \beta_k = b$, the test statistic is $W = (\hat{\beta}_k - b)/\text{se}(\hat{\beta}_k)$. This statistic is approximately $N(0, 1)$ when $H_0 : \beta_k = b$ is true and the p-value is larger than $1 - \alpha$ only when $|W| < z_{\alpha/2}$ where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of a $N(0, 1)$ random variable. This yields the well known CI

$$(\hat{\beta}_k - z_{\alpha/2}\text{se}(\hat{\beta}_k),\ \hat{\beta}_k + z_{\alpha/2}\text{se}(\hat{\beta}_k)).$$

The likelihood ratio CI operates in the same way, but the log-likelihood must be computed for all values of b.

Example: V.A. data.

Non-nested models

Wald, score, and likelihood ratio tests all work for nested models. The constraint $\mathbf{M}\boldsymbol{\theta} = \mathbf{b}$ implies a model that is nested in the larger model without the constraint.

For non-nested models, model selection criteria are commonly employed. All information criteria have a portion reflecting model fit (that increases with the number of parameters) and a portion that penalizes for complexity. Smaller information criteria $\Rightarrow$ better model.

AIC for model selection

The AIC (Akaike, 1974) is a widely accepted statistic for choosing among models. The AIC is asymptotically justified as attempting to minimize the estimated Kullback-Liebler distance between the true probability model and several candidate models. As the true model is often more complex than our simple statistical models, the AIC will tend to pick larger, more complex models as more data are collected and more is known about the true data generating mechanism.

The AIC is

$$\text{AIC} = 2k - 2\log L(\hat{\boldsymbol{\theta}}|\mathbf{x}),$$

where k is the number of parameters in $\boldsymbol{\theta}$.

Marginal & conditional AIC

Random effects models can be dealt with in one of two ways.

If the random effects $\mathbf{u}$ can be integrated out, the usual likelihood is obtained

$$L(\boldsymbol{\theta}|\mathbf{x}) = \int_{\mathbf{u}\in\mathbb{R}^p} L(\boldsymbol{\theta}|\mathbf{x},\mathbf{u})p(\mathbf{u}|\boldsymbol{\theta})d\mathbf{u}.$$

Then the (marginal) AIC is computed as usual. Here the model "focus" is on population effects $\boldsymbol{\theta}$ only. This is the default AIC provided in `lmer` and `glmer` in the `lme4` package.

Example: (board) Oneway ANOVA $y_{ij} = \alpha_i + \epsilon_{ij}$ where α_i are (a) fixed effects, (b) random effects.

Conditional AIC

Otherwise, the "conditional AIC" can be used which focuses on both the fixed effects and the random effects. The conditional AIC has a somewhat complicated definition, but its use is automated for fitting GLMMs using `glmer` or `lmer` in the `lme4` package with the `cAIC4` package. Read Saefken, Kneib, van Waveren, and Greven (2014, *Electronic J. of Statistics*) for more details.

In many models the conditional AIC is $2\hat{k} - 2\log L(\hat{\theta}|\mathbf{x})$, where $\hat{k}$ is the "effective number of parameters," often a function of the trace of a "hat matrix," motivated by "effective degrees of freedom" from the smoothing literature.

BIC

The BIC (Schwarz, 1978) will pick the *correct* model as the sample size increases (it is consistent) *as long as the correct model is among those under consideration*. Since we do not know whether the true model is among those we are considering, I tend to use AIC and possibly err on the side of a more complex model, but one that better predicts the actual data that we saw.

The BIC is

$$\text{BIC} = k \log(n) - 2 \log L(\hat{\theta}|\mathbf{x}).$$

This penalizes for adding predictors moreso than AIC when $n \geq 8$, and so tends to give simpler models.

One troublesome aspect of the BIC is the sample size n. It is unclear what to use for n when data are missing, censored, or where data are highly dependent.

Bayesian model selection: DIC

The deviance information criterion (DIC, Spiegelhalter et al., 2002) is a Bayesian version of conditional AIC. The model deviance is defined as $S - 2\log L(\hat{\boldsymbol{\theta}}|\mathbf{x})$ where S is $2 \times$ log-likelihood under a "saturated model" and $\hat{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$. Typically S is left off for model selection.

The version of DIC used by JAGS is $\text{DIC} = 2\hat{k} - 2\log L(\bar{\boldsymbol{\theta}}|\mathbf{x})$ where $\bar{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}|\mathbf{x}}\{\boldsymbol{\theta}\}$ and $\hat{k} = \frac{1}{2} var_{\boldsymbol{\theta}|\mathbf{x}}\{-2\log L(\boldsymbol{\theta}|\mathbf{x})\}$ are the "effective number of parameters."

If there are random effects then $\text{DIC} = 2\hat{k} - 2\log L(\bar{\boldsymbol{\theta}}|\mathbf{x}, \bar{\mathbf{u}})$ where $\hat{k} = \frac{1}{2} var_{\boldsymbol{\theta},\mathbf{u}|\mathbf{x}}\{-2\log L(\boldsymbol{\theta}|\mathbf{x}, \mathbf{u})\}$.

SAS' DIC uses the original $\hat{k} = -2E_{\boldsymbol{\theta}|\mathbf{x}}\{\log L(\boldsymbol{\theta}|\mathbf{x})\} + 2\log L(\bar{\boldsymbol{\theta}}|\mathbf{x})$.
Example: Ache hunting with and without (a) quadratic terms, (b) random effects.

Bayesian model selection: Bayes factors

Consider two models with likelihoods $L_1(\boldsymbol{\theta}_1|\mathbf{x})$ and $L_2(\boldsymbol{\theta}_2|\mathbf{x})$ and priors $\pi_1(\boldsymbol{\theta}_1)$ and $\pi_2(\boldsymbol{\theta}_2)$. The marginal joint distribution of the data $\mathbf{x}$ for model j is

$$f_j(\mathbf{x}) = \int_{\boldsymbol{\theta}_j \in \boldsymbol{\Theta}_j} \underbrace{L_j(\boldsymbol{\theta}_j|\mathbf{x})\pi_j(\boldsymbol{\theta}_j)}_{f(\boldsymbol{\theta}_j,\mathbf{x})} d\boldsymbol{\theta}_j.$$

The Bayes factor for comparing the two models is

$$BF_{12} = \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}.$$

Note: $BF_{13} = BF_{12}BF_{23} \Rightarrow$ BF's coherently rank models.

Bayes factors can give evidence towards null! Chib has several papers (first: Chib, 1995) on how to compute Bayes factors from MCMC output. Lots of newer approaches in the last 10 years!

Bayes factors

Kass & Raftery (1995, JASA) modified Jeffreys' original scale to

BF_{12}	Evidence strength
1 to 3	not worth more than a bare mention
3 to 20	positive
20 to 150	strong
over 150	very strong

Using p-values to bound Bayes factors. Sellke, Bayarri, and Berger (2001) show $BF \geq \frac{-1}{e\ p \log(p)}$. A BF over 150 ($p \leq 0.0003$) is considered "very strong" evidence against H_0 by Kass and Raftery (1995) and "decisive" by Jeffreys (1961).

To simply achieve "strong" evidence, one needs $p < 0.003$. There is a movement among statisticians to make the "new scientific standard" to be $\alpha = 0.005$ instead of 0.05 based on arguments such as this.

Savage-Dickey ratio for point nulls

The setup is a bit more general in Verdinelli and Wasserman (1995), but here's a common situation.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ and consider testing $H_0 : \boldsymbol{\theta}_2 = \mathbf{b}$ vs. $H_a : \boldsymbol{\theta}_2 \neq \mathbf{b}$. Let $\pi_1(\boldsymbol{\theta}_1)$ be the prior for $\boldsymbol{\theta}_1$ under H_0 and $\pi_{12}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ be the prior under H_a. If $\pi_1(\boldsymbol{\theta}_1) = \pi_{12}(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2 = \mathbf{b})$ the Bayes factor for comparing the nested model H_0 to the alternative H_a is

$$BF = \frac{\pi_{\boldsymbol{\theta}_2|\mathbf{x}}(\mathbf{b})}{\pi_2(\mathbf{b})},$$

where $\pi_2(\boldsymbol{\theta}_2) = \int_{\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1} \pi_{12}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) d\boldsymbol{\theta}_1$. Note: $\pi_1(\boldsymbol{\theta}_1) = \pi_{12}(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2 = \mathbf{b})$ automatically holds under independent priors $\pi_{12}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1) = \pi_1(\boldsymbol{\theta}_1)\pi_2(\boldsymbol{\theta}_2)$.

Savage-Dickey ratio for point nulls

If $[\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{x}]$ has a closed-form density $\pi_{21}(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \mathbf{x})$ and an MCMC sample is available

$$\pi_{\boldsymbol{\theta}_2|\mathbf{x}}(\mathbf{b}) \approx \frac{1}{M} \sum_{m=1}^{M} \pi_{21}(\mathbf{b}|\boldsymbol{\theta}_1^m, \mathbf{x}).$$

Otherwise, if the posterior is approximately Gaussian one can compute $\boldsymbol{\mu}_2 = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\theta}_2^m$ and
$\boldsymbol{\Sigma}_2 = \frac{1}{M} \sum_{m=1}^{M} (\boldsymbol{\theta}_2^m - \boldsymbol{\mu}_2)(\boldsymbol{\theta}_2^m - \boldsymbol{\mu}_2)'$ and use

$$\pi_{\boldsymbol{\theta}_2|\mathbf{x}}(\mathbf{b}) \approx \phi_{k_2}(\mathbf{b}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2).$$

Note: like Walkd tests, only requires fitting the full model!

Comparing two means via Savage-Dickey

$$y_{11}, \ldots, y_{1n_1} | \boldsymbol{\theta} \overset{iid}{\sim} N(\mu_1, \tau_1),$$

$$y_{21}, \ldots, y_{2n_2} | \boldsymbol{\theta} \overset{iid}{\sim} N(\mu_1 + \delta, \tau_2),$$

where $\boldsymbol{\theta} = (\mu_1, \delta, \tau_1, \tau_2)'$. Let $\mu_1 \sim N(m_1, v_1)$ indep. $\delta \sim N(0, v_d)$.

Approximate BF for testing $H_0 : \delta = 0$ vs. $H_a : \delta \neq 0$ is

$$BF = \frac{\phi(0|\bar{\delta}, s_\delta^2)}{\phi(0|0, v_d)},$$

where $\bar{\delta} = \frac{1}{M}\sum_{m=1}^{M} \delta^m$ and $s_\delta^2 = \frac{1}{M}\sum_{m=1}^{M}(\delta^m - \bar{\delta})^2$. Can also find $[\delta|\mu_1, \tau_1, \tau_2, \mathbf{y}_1, \mathbf{y}_2]$ and average over MCMC iterates for numerator; more accurate.

Blood flow, exercise, & high altitude

Myocardial blood flow (MBF) was measured for two groups of subjects after five minutes of bicycle exercise. The normoxia ("normal oxygen") group was provided normal air to breathe whereas the hypoxia group was provided with a gas mixture with reduced oxygen, to simulate high altitude.

Comment on Savage-Dickey

Note that we simply need to specify a full model. If we *assume* the induced prior $\pi_1(\boldsymbol{\theta}_1) = \pi_{12}(\boldsymbol{\theta}_1 | \boldsymbol{\theta}_2 = \mathbf{b})$ under H_0 the Savage-Dickey ratio works.

Technically don't even have to think about the null model, but probably should make sure induced prior makes sense in practice!

Very useful for testing, e.g. $H_0 : \beta_2 = 0$ in a regression model. Alternative needs to be two-sided though.

Note: flat priors *ruin Bayes factors*! A flat prior pleads complete ignorance, when, in almost all situations, something is known about the model parameters. We know height, age, IQ, etc. are positive and have a natural upper bound. We know correlations are between -1 and 1. Known as "Lindley's paradox" – look at Wikipedia entry. **Example:** Ache hunting.

Some more recent Bayesian model selection criteria

- Various fixes to traditional Bayes factors to make them robust to prior assumptions: posterior Bayes factors (Aitkin 1991), intrinsic Bayes factors (Berger and Pericchi 1996), & fractional Bayes factor (O'Hagan 1995). Useful for simpler classes of models; not broadly applicable and not in widespread use.
- LPML for data $\mathbf{y} = (y_1, \ldots, y_n)'$ that are independent given $(\boldsymbol{\theta}, \mathbf{u})$ also used for Bayesian model comparison:

$$\text{LPML} = \sum_{i=1}^{n} \log f(y_i | \mathbf{y}_{-i}).$$

Plagued by simple but unstable estimates over the last 20+ years.

Some more recent Bayesian model selection criteria

- $\exp(\text{LPML}_1 - \text{LPML}_2)$ is "pseudo Bayes factor comparing models 1 to 2. Not sensitive to prior assumptions and does not suffer Lindley's paradox.
- WAIC ('widely-applicable information criteria', developed in Watanabe, 2010) attempts to estimate the same quantity as DIC but uses a much more stable estimate of the effective number of parameters. WAIC and LPML are *asymptotically equivalent* estimators of out-of-sample prediction error. DIC can be viewed as an approximation to WAIC.

Some very recent Bayesian model selection criteria

- Both a stable LPML (a somewhat different version termed 'LOO') and WAIC are available in the `loo` R package; simply need to monitor the log-likelihood (not posterior!) contribution $L_i = \log f(y_i|\boldsymbol{\theta}, \mathbf{u})$ of each datum (assuming independence given $\boldsymbol{\theta}$ and/or $\mathbf{u}$); see Vehtari, Gelman, & Gabry (2017, *Statistics and Computing*).
- DIC is automatically given in JAGS; WAIC and LPML can are computed by the `loo` package from standard JAGS or STAN output in two different ways (one simpler to code but requires more storage, $M \times n$).
- DIC is also given by SAS, although a more primitive (and more problematic) version than used by JAGS.
- Ideal situation: use LPML (or LOO). DIC, LOO, and WAIC all favor models with smaller values.

Why pick one model? BMA!

Model averaging assigns weights to each model; prediction averages over models! Let

$$w_j = P(M = j|\mathbf{x}) = \frac{f(\mathbf{x}|M = j)P(M = j)}{\sum_{k=1}^{J} f(\mathbf{x}|M = k)P(M = k)},$$

i.e.

$$w_j = P(M = j|\mathbf{x}) = \frac{f_j(\mathbf{x})P(M = j)}{\sum_{k=1}^{J} f_k(\mathbf{x})P(M = k)},$$

Noting that BIC_j consistently picks the true model, and $f_j(\mathbf{x})$ is difficult to compute, the approximation

$$w_j = \frac{\exp(-\frac{1}{2}BIC_j)}{\sum_{k=1}^{J} \exp(-\frac{1}{2}BIC_k)}$$

has been suggested, assuming $P(M = j) = \frac{1}{J}$.

Consistency only happens when one of the J models is *true*. Aikake actually suggested

$$w_j = \frac{\exp(-\frac{1}{2}AIC_j)}{\sum_{k=1}^{J}\exp(-\frac{1}{2}AIC_k)}.$$

Assuming independence data given $\boldsymbol{\theta}_j$ in each model, the predictive density for a new observation is simply

$$f(x|\mathbf{x}) = \sum_{j=1}^{J} w_j E_{\boldsymbol{\theta}|\mathbf{x}}\{f(x|\boldsymbol{\theta})\}.$$

Sest subsets via SSVS

In regression models, we often want to know which predictors are important. Stochastic search variable selection places "spike and slab" priors on each coefficient; the simplest version is due to Kuo and Mallick (1998).

Let the ith linear predictor (for logistic regression, Poisson regression, beta regression, normal-errors regression, etc.) be

$$\eta_i = \beta_0 + \beta_1\gamma_1 x_{i1} + \cdots + \beta_p\gamma_p x_{ip}.$$

Assume some sort of prior $\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta})$, e.g. a g-prior, and

$$\gamma_1, \ldots, \gamma_p \overset{iid}{\sim} \text{Bern}(q).$$

Easily carried out in JAGS or STAN or SAS! See O'Hara and Sillanpää (2009, *Bayesian Analysis*).

Posterior inference is summarized by a table that looks like, e.g. when $p = 3$

Model			Prob.
x_1	x_2	x_3	p_{123}
x_1	x_2		p_{12}
x_1		x_3	p_{13}
	x_2	x_3	p_{23}
x_1			p_1
	x_2		p_2
		x_3	p_3
intercept only			p_0

Note: basic idea extends to zero-inflated models.