Hypothesis Testing and Bayes factor

Two hypotheses: Null hypothesis $H_0$ and Alternative hypothesis $H_a$.

After determining an appropriate test statistic $T(Y)$, one then compute the observed significance or p-value, of the test as,

$$p \text{-value} = P\{T(Y) \text{more extreme than } T(Y_{obs} | \theta, H_0)\}$$

where extremeness is in the direction of the alternative hypothesis.

If the p-value is less than some prespecified Type I error rate, $H_0$ is rejected; otherwise it is not.

1. It can only be applied when two hypotheses in question are nested, one within the other. That is $H_0$ must constitute a simplification of $H_a$ (say, by setting one of the model parameters in $H_a$ equal to constant-usually 0).

   But many practical hypothesis testing problems involve a choice between two (or more) models that are not nested (choosing between quadratic and exponential growth model).

2. Tests of this type can only offer evidence against the null hypothesis. A small p-value indicates that the larger alternative model has significantly more explanatory power.

   However, a large p-value does not suggest that the two models are equivalent, but only that we lack evidence that they are not.

   So we say fail to reject rather than accept

P-value is long-term probability (in a hypothetical repetition of the same experiment) of obtaining data at least as unusual as what was actually observed.

But it is not the probability that $H_0$ is true.

In general case suppose want to compare two models $M_1$ and $M_2$.

In Bayesian way calculate $p(M_i|y)$: probability of the model conditioned on the data. or the posterior probability of the model.

Compare the posterior probabilities.

For model $M_i$, parameters $\theta_i$, prior for the parameters: $p(\theta_i)$ and prior for the model: $p_i(M_i); i = 1, \cdots k$.

Marginal density $p(y|M_i) = \int p(y|\theta_i, M_i)p_i(\theta_i)d\theta_i$.

Now using Bayes Theorem:

$$p(M_i|y) = \frac{p(y|M_i)p(M_i)}{\sum_{j=1}^{k} p(y|M_j)p(M_j)}$$
Bayes factor $B$ is defined on the basis of ratio of the posterior odds of $M$ to the prior odds of $M$ as

$$M = \frac{p(M_1|y)/p(M_2|y)}{p(M_1)/p(M_2)} = \frac{p(y|M_1)}{p(y|M_2)} \quad (1)$$

If $p(M_1) = p(M_2) = 0.5$ then $B = p(M_1|y)/p(M_2|y)$.

When both models share the same parameterization ($\theta_1 = \theta_2 = \theta$) and both hypotheses are simple $M_1 : \theta = \theta_1$ and $M_2 : \theta = \theta_2$.

Then $p_i(\theta_i)$ consists of a point mass at $\theta_i$ and so

$$B = \frac{p(y|\theta_1)}{p(y|\theta_2)} \text{ which is the likelihood ratio between two models.}$$

Hence in the simple-versus-simple setting, the Bayes factor is precisely the odds in favor of $M_1$ over $M_2$ given solely by data.

**Occam’s razor**

The Bayesian framework contains a natural penalty against against over complex models sometimes called Occam’s razor: a term refers to the principle of parsimony as advocated by William Occam 1320:

“Multiplicity is not to be preferred when it is unnecessary”.

That is, a simpler theory is to be favored to a more complex one, other things being equal.

Bayes factor contains this natural penalty within it.

**Other Criterion based model choice**

These are mainly penalized likelihood ratio models for $M_1$ and $M_2$ with parameters $p_1$ and $p_2$

Define $W = -2\log[\sup_{M_1} p(y|\theta_1)/\sup_{M_2} p(y|\theta_2)]$

Akaike Information Criterion (AIC): $\Delta \text{AIC} = W - 2(p_2 - p_1)$.

$\Delta$ denotes the change from Model 1 to Model 2 and the second term acts as a penalty term which corrects for differences in size between the model.

Practitioners noticed that it keep too many terms in the model.

Schwartz criterion (Bayesian information criterion) (BIC):

$\Delta \text{BIC} = W - (p_2 - p_1)\log n$
\[ \exp\left(-\frac{1}{2}\Delta BIC\right) \] provides a rough approximation to the Bayes factor which is independent of the priors on \( \theta \).

**Difficulties with Bayes factor**

(i) Computation: Need to evaluate
\[ p(y|M_i) = \int \cdots \int p(y|\theta_i, M_i)p(\theta_i)d\theta_i \]
which could be a high dimension integral.

(ii) If \( p_i(\theta_i) \) is improper then
\[ p(y|M_i) = \int \cdots \int p(y|\theta_i, M_i)p(\theta_i)d\theta_i \]
necessarily is as well, and so \( B \) is not well defined.

To repair deficiency (ii), one approach is to divide the data into two parts, \( y = (y_1, y_2) \) and use the first portion to compute model-specific posterior densities \( p(\theta_i|y_1), i = 1, 2. \)

These two posteriors are then used as the priors with the remaining data \( y_2 \) producing the partial Bayes factor
\[ B(y_2|y_1) = \frac{p(y_2|y_1, M_1)}{p(y_2|y_1, M_2)} \]

Questions: (i) How many data points to allocate to the training sample \( y_1 \)
(ii) How to select them?

Berger and Pericchi (1996) propose using an arithmetic or geometric mean of the partial Bayes factor obtained using all possible minimal training samples (where \( n_1 \) is taken to be the smallest sample size leading to a proper posterior \( p(\theta_i|y_1). \)

**Intrinsic Bayes Factor**

**Model Mixing (averaging)** Rather than selecting a single model can mix or average over models incorporating model uncertainty.

Suppose \( (M_1, M_2, \cdots, M_m) \) index our collection of candidate models, and \( \Delta \) is a quantity of interest, assumed to be well defined for every model.

Given a set of prior model probabilities \( \{p(M_1, \cdots, p(M_m))\} \)
the posterior distribution of \( \Delta \) is given by
\[ \pi(\Delta|y) = \sum_{j=1}^{m} \pi(\Delta|M_j, y)\pi(M_j|y) \]
where \( p(\Delta|M_i, y) \) is the posterior for \( |\Delta| \) under the \( i \)th model, and \( p(M_i|y) \) is the posterior probability of the model computable as
\[ \pi(M_i|y) = \frac{p(y|M_i)p_0(M_i)}{\sum_{j=1}^{m}p(y|M_j)p_0(M_j)}. \]

Comments: (i) Averaging over all the models result in a better prediction than any one of the model individually.
(ii) It considers model uncertainty.

(iii) Implementation is difficult due to large number of potential models ($m$ is extremely large.

Say for 20-predictor regression $m = 2^{20}$.

To limit the class of models to those which meet some minimum level of posterior support, say the class

$$A = \{M_i : \frac{\max_j p(M_j|y)}{p(M_i|y)} \leq c\}$$

say for $c = 20$ suggested by Madigan and Raftery (1994).

So eliminate set of models that have less posterior supports.

Alternative would to use search algorithms over the entire model space using MCMC methods and locate the high-probability models.

The problem could be more complicated if the parameter $\theta$ can’t be integrated out analytically.

Then we need to search over model-parameter spaces which is more complicated.